

NEJLT

**Northern
European
Journal**

of

Language Technology



www.nejlt.org

Volume 9, December 2023
ISSN 2000-1553

NEJLT Editorial Team 2023

Leon Derczynski, IT University of Copenhagen, Editor-in-Chief

Isabelle Augenstein, University of Copenhagen

Nikolaos Aletras, University of Sheffield

Francesco Barbieri, Snap

Jasmijn Bastings, Google

Rachel Bawden, INRIA, Paris

Yonatan Belinkov, Technion

Emily M. Bender, University of Washington

Christos Christodoulopoulos, Amazon

Manuel R. Ciosici, USC Information Sciences Institute

Miryam de Lhoneux, KU Leuven

Lucia Donatelli, Saarland University

Yanai Elazar, University of Washington

Angela Fan, Meta

Yang Feng, Chinese Academy of Sciences

Mark Fishel, University of Tartu

Antske Fokkens, Vrije Universiteit Amsterdam

Hila Gonen, Meta / University of Washington

Yufang Hou, IBM

Zhijing Jin, Max Planck Institute & ETH Zurich

Xiang Lorraine Li, Allen Institute for Artificial Intelligence

Sasha Luccioni, Hugging Face

Benjamin Marie, 4i

Nafise Sadat Moosavi, The University of Sheffield

Debora Nozza, Bocconi University

Ellie Pavlick, Brown University

Verena Rieser, Deepmind

Kay Rottmann, Amazon Alexa AI

Vered Shwartz, University of British Columbia

Song Linfeng, Tencent

Dhanasekar Sundararaman, Microsoft

Jörg Tiedemann, University of Helsinki

Emiel van Miltenburg, Tilburg University

Adina Williams, Meta

Steve Wilson, Oakland University

Resource papers as registered reports: a proposal

Emiel van Miltenburg, Tilburg University, The Netherlands

c.w.j.vanmiltenburg@tilburguniversity.edu

Letter abstract This is a proposal for publishing resource papers as registered reports in the Northern European Journal of Language Technology. The idea is that authors write a data collection plan with a full data statement, to the extent that it can be written before data collection starts. Once the proposal is approved, publication of the final resource paper is guaranteed, as long as the data collection plan is followed (modulo reasonable changes due to unforeseen circumstances). This proposal changes the reviewing process from an antagonistic to a collaborative enterprise, and hopefully encourages NLP resources to develop and publish more high-quality datasets. The key advantage of this proposal is that it helps to promote *responsible resource development* (through constructive peer review) and to avoid *research waste*.

1 Introduction

A common sentiment in NLP is that the creation of corpora and benchmarks is under-appreciated (Rogers, 2020; Sambasivan et al., 2021), even though resources are one of the driving factors of progress in our field. Moreover, the measurement of progress critically depends on having solid benchmarks. If authors are weary of producing new resources, we all suffer the consequences. How can we avoid this?

1.1 Barriers to resource production

Generally speaking, there seem to be two barriers to resource production: funding and appreciation. Building resources requires time and money, and researchers may only be willing to invest time in a project if it could lead to a publication in a respectable venue.

To make resource-building an attractive proposition, we somehow need to convince potential resource authors that their time will be well-spent. One way to do this is to provide a guarantee that their paper will be published. Of course, we would need to have some form of quality control, to make sure that the final resource will be useful to our community. Luckily, such a process already exists in the form of registered reports.

1.2 Registered reports

Registered reports are papers that are reviewed in two phases (Chambers, 2019; Henderson and Chambers, 2022). First, authors submit a research proposal, with a clear motivation and outline of the methodology. (Sim-

ilar to a preregistration, see van Miltenburg et al. 2021.) This proposal is reviewed until authors and reviewers agree on the research plan. This agreement means that the paper is accepted in principle. Once the approval is in, authors carry out their study and report their results as specified in the proposal. Then they submit their final paper for the second review phase. In this phase, reviewers check whether the authors followed their proposed methodology. Any changes should be indicated by the authors, with a clear motivation for why those changes were made. Reviewers may not criticize the methodology anymore, but can only comment on the quality of the reporting. Once this is approved, the paper is published.

1.3 Earlier discussion in NLP

Van Miltenburg et al. (2021) proposed preregistration and registered reports as potentially helpful innovations in NLP. They suggested that virtually all paper types in NLP are amenable to preregistration.¹ In response, Søgaard et al. (2023) argued that there are also some downsides to preregistration that may outweigh the benefits.^{2,3} Nevertheless, they also see

¹The argument is mostly based on what Lakens (2019) calls the *positive externalities* of preregistration. He argues that the core value of preregistration is “to allow others to transparently evaluate the capacity of a test to falsify a prediction, or the severity of a test.” This idea is often not applicable in NLP, but many benefits remain. See Sarafoglou et al. 2022 for a survey among researchers to determine the benefits of preregistration.

²E.g., preregistration may increase administrative workload, as also pointed out by Sarafoglou et al. (2022); Hostler (2023).

³A full discussion of the authors’ arguments goes beyond the scope of this letter, especially since the authors agree preregistra-

enough value in the idea of registered reports to make a counter-proposal: “limit preregistration to research for which our risk tolerance is low” (p.90).

Søgaard et al. (2023) roughly define risk as *the cost of being wrong*, which in NLP often means that we lose compute and human hours. They argue that this cost is often acceptable, especially in comparison with the human tragedy that may result from clinical trials, so we do not need to burden ourselves with the overhead that risk minimization strategies (such as registered reports) typically bring. On reflection, it does seem true that the risk in NLP is often lower than in the medical field, but the cost of being wrong can still be significant.

Grainger et al. (2020) coin the term *research waste*, and highlight different ways in which we may produce such waste. If you take the wrong approach, you lose researcher and GPU time, and waste the efforts of the volunteers, crowd workers, or consultants involved in your research. Registered reports can be used to prevent this situation. At the same time, they also enable us to carry out ethics review where it is most relevant: in the preparation stage. This immediately solves the problem of after-the-fact ethics reviewing, where we may spot issues, but authors may no longer be able to resolve them.⁴

Contribution. This letter proposes registered reports to support the creation of resources (through peer feedback early on in the process) and to avoid research waste. The proposal is painted in broad brush strokes to emphasise the big picture. If this proposal is successful, we can work out the finer details.⁵

2 Process outline

The general writing process for registered reports has been described elsewhere (e.g., Henderson and Chambers 2022; Kiyonaga and Scimeca 2019). What would the process look like for resource papers? Here is a brief sketch of what this process could look like if NEJLT would accept registered reports.

2.1 Review phase 1

The first review phase is all about your plans. This means that authors will have to write about:

1. **The purpose of the resource.** Why do you want to collect the data? What secondary purposes could the resource also be used for? These

tion/registered reports can be beneficial for our field—we should just work out the proper conditions and guidelines.

⁴Lakens (2023) makes a similar argument, but his solution is to make institutional review boards also review research methodology, as part of their ethics approval procedure.

⁵The Center for Open Science provides a useful set of resources to get started with registered reports: <https://www.cos.io/initiatives/registered-reports>

questions serve as a guide to inform your answers to the other questions. After having listed the different use cases that you (don't) want to support, you can carry out a requirements analysis to see what is needed (split up into essential or nice-to-have) to actually carry out the relevant task. For benchmark datasets it is important to have a clear definition of the skills that you want to assess or the dependent variables that you aim to operationalise. (See Schlangen 2021; Shimorina and Belz 2022 for inspiration.)

2. **The composition of the resource.** What properties should your resource have, and how do you plan to ensure that the resource will indeed have those properties? Additionally: at what level of granularity should you collect different kinds of information?⁶ It is a good idea to prepare a draft data statement (Bender and Friedman, 2018) for your resource.⁷
3. **The development process.** How will you go about developing the resource? How will you ensure that the requirements are met? (Also taking practical and technological limitations into account.) If your project requires a large amount of computing power, what strategies are you using to minimise your carbon emissions? (See Lucioni et al. (2020) for recommendations.)
4. **Ethical considerations.** How are the rights and well-being of participants/crowd-workers, data subjects and other direct/indirect stakeholders taken into account, both during and after the development of the resource? Jamieson et al. (2023) provide questions and considerations to make the resource development process more reflexive. As Henderson and Chambers (2022) note, it is important to consider when to submit a proposal to your local institutional review board (IRB) for ethics approval. For most NLP studies it seems reasonable to first apply to your local IRB before submitting the proposal to a journal. This would strengthen your proposal, and any important changes that are requested during the review process could be approved via an amendment to the original IRB application.
5. **Data stewardship.** How will the data be stored, and what measures will be put in place to maintain the resource and take care of any issues that arise from the publication or use of the resource? For discussion, see for example: Peng et al. 2021; Jernite et al. 2022. As with ethics review, it is rea-

⁶Here one might also consider *k-anonymity* for participants/crowd-workers/data subjects (Sweeney, 2002), i.e. ensuring that each property or combination of properties is shared by at least *k* individuals.

⁷A step-by-step guide for writing data statements is available at this URL: <https://techpolicylab.uw.edu/data-statements/>

sonable to contact your local data steward about the measures you should take to responsibly collect and share data. (In some cases, you may be required to carry out a [Data Protection Impact Assessment](#).) At some universities, the IRB process already incorporates a form on data management to protect any data subjects.

This is more or less equivalent to writing an introduction, theoretical framework, methodology, and ethical considerations section.

When to submit a proposal?

What is the right time to submit a research proposal? This is an open question, as we know that the annotation process is often cyclical, with multiple rounds of revision before an appropriate model and a set of guidelines has been developed (see [Pustejovsky et al. 2017](#), for example). However, most project parameters are likely to be known after a small-scale pilot study. (By keeping the pilot small, we are still minimising research waste.) Even if the exact model and annotation scheme are not fully fixed yet, the methodology and feasibility of the study are clear. At that point, research proposals may be submitted for review.⁸

Reviewing

Reviewing the proposal is similar to how it is currently done at NEJLT: you submit the paper to the journal, and an editor assigns reviewers to your proposal. The reviews themselves should be constructive, focusing mainly on the methodological and ethical issues:

1. Does this resource address a current need in NLP research?
2. Is the proposed dataset representative of the intended genre or domain?
3. Is the methodology appropriate, valid, and described in sufficient detail?
4. Will the data be responsibly collected and maintained?

What sets registered reports apart from regular submissions is that reviewers can actively contribute to the methodology; they can propose changes to improve the quality of the dataset to be more considerate of any stakeholders, or to make it more broadly us-

⁸A related and common question is: what happens if authors want to change the design of their study, after their research proposal has been accepted? The answer depends on the nature of the changes. Small modifications should be noted and motivated in the final report. Larger modifications may need to be reviewed, or at least flagged to the editor. The Center for Open Science notes in their [Frequently Asked Questions](#) that it is also possible to carry out *sequential registrations* for studies where the design and hypotheses for each subsequent study in a paper is based on previous results.

able. Authors can then refine their proposal before the manuscript is provisionally accepted.

Should we publish Stage 1 protocols?

An open question here is whether the proposal should be published at this stage, or only when the final report has been accepted for publication. Publication policies differ between different journals: [The Royal Society \(ND\)](#) does *not* publish Stage 1 registered reports before the final manuscript is approved. [Nature Scientific Reports \(ND\)](#) does not publish Stage 1 registered reports either, but *does* require authors to preregister their study in a recognised repository. The preregistration can either be made public, or put under embargo until Stage 2. This matches the recommendations from [Chambers et al. \(2023\)](#), who note that “the journal can also perform the Stage 1 registration process on behalf of authors.” Finally, the publisher [Wiley \(2018\)](#) recommends that journals publish registered reports after passing Stage 1 peer review, but also allows its journals to instead require authors to preregister their study design (similar to Nature Scientific Reports).

[Wiley \(2018\)](#) notes that publishing registered protocols has the advantage of providing transparency and accountability both for journals (showing what reports are in principle accepted, publicly committing to the publication of the final result) and authors (showing *what* they are working on, *when* they developed the ideas for their final publication, and publicly committing to finish the resource). Of course researchers could also feel uncomfortable sharing their research-in-progress for all sorts of reasons, so it may be good to at least offer them the option to put their research proposal under embargo.

2.2 Review phase 2

With an in-principle acceptance in hand, authors should aim to carefully follow their original proposal. Deviations from the original plans are possible, but these should be clearly indicated in the report and well-motivated by the authors. Once the dataset has been collected and a full report has been written, the paper can undergo the final review.

Reviewing

Having already approved the methodology, reviewers now comment on the execution of the project:

1. Has the resource been compiled according to plan, with all deviations clearly marked?
2. Does the report contain all relevant details about the creation and composition of the resource?
3. Is the presentation clear and accessible?
4. Is the resource accessible and easy-to-use?

We should expect resources to be publicly available, unless there are strong arguments in favor of limited accessibility (for example: copyright issues, or privacy of the data subjects).

Should reviewing be anonymous?

An open question here is whether reviewing in the second phase should proceed anonymously, or whether it is also OK for author names to be revealed at this time. This would certainly make it easier to assess the final resource (which may be hard to anonymise), but may unduly influence the reviewing process.

2.3 Publication

Once the paper is ready for prime-time, it can be published as usual. An open question is whether the reviews should be published as well and, if so, whether the reviews should be kept anonymous or not. For transparency reasons, it would be really insightful to publish all the correspondence between authors and reviewers along side the final report. This way, we would get to see the original intentions of the authors, and how the approach was transformed during the review process. Reviewer names could be published on an opt-in basis, so that they might claim credit for the provided service. (This avoids the issue of reviewers holding back their criticism for fear of retribution if their name is published alongside their review, see e.g. [Ali and Watson 2016](#) for discussion.)

3 Eligibility

What kinds of resources should be eligible for publications through registered reports? So far this proposal has not set any strict requirements to determine what makes a resource worth publishing in a journal like NE-JLT. To some extent, we can be pragmatic about this issue: authors tend to prefer conferences for smaller contributions, and journals for larger contributions. The administrative hassle for smaller projects may just not be worth the effort of writing a registered report (in [Søgaard et al.](#)'s terms: there is less 'risk' involved), so authors of small studies are not very likely to submit a research proposal.⁹ What matters is that authors have clearly thought through their proposal, and are not just letting reviewers do their work. In the latter case, desk rejection seems appropriate. If we do need more guidelines, we can always fall back on the existing ones, that easily carry over to (and indeed overlap with some reviewing questions in) this proposal.¹⁰

⁹But if authors think that their work should be published as a registered report, there is little harm in letting them carry out a small but high-quality study.

¹⁰<https://www.nejlt.org/review/>

4 Conclusion

This letter proposed to offer potential resource authors the opportunity to publish their resources as registered reports, as an addition to the existing paper types. (The resource category would not need be removed.)

The proposal outlined here is more modest than the one put forth by [Van Miltenburg et al. \(2021\)](#), who suggest that *all* types of NLP papers (except position papers) could in theory be published as registered reports. This modesty is not for a lack of ambition; instead, this proposal is offered as a first step, to see if registered reports could actually work for NLP research. And what better way to start, than to support the creation of fundamental resources?

References

- Ali, Parveen Azam and Roger Watson. 2016. Peer review and the publication process. *Nursing Open*, 3(4):193–202.
- Bender, Emily M. and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.
- Chambers, Chris. 2019. What's next for registered reports? *Nature*, 573:187–189.
- Chambers, Chris, George Christopher Banks, Dorothy Bishop, Sara Bowman, Kate Button, Molly Crockett, Zoltan Dienes, Timothy M. Errington, Agneta Fischer, Alex O. Holcombe, Kai Jonas, Edward Miguel, Marcus Munafò, Brian A. Nosek, Brendan J. Nyhan, David Rand, Daniel J. Simons, Carien van Reekum, Andrew Sallans, Steven Rogelberg, , and David Thomas Mellor. 2023. Registered reports implementation checklist. Available through: <https://osf.io/2m4ct> Originally published on February 4, 2014. The current version was modified on May 8, 2023. Retrieved 13 July 2023.
- Grainger, Matthew J., Friederike C. Bolam, Gavin B. Stewart, and Erlend B. Nilsen. 2020. Evidence synthesis for tackling research waste. *Nature Ecology & Evolution*, 4(4):495–497.
- Henderson, Emma L. and Christopher D. Chambers. 2022. Ten simple rules for writing a registered report. *PLOS Computational Biology*, 18(10):1–9.
- Hostler, Thomas J. 2023. The Invisible Workload of Open Research. *Journal of Trial & Error*. <https://journal.trialanderror.org/pub/the-invisible-workload>.

- Jamieson, Michelle K., Gisela H. Govaart, and Madeleine Pownall. 2023. Reflexivity in quantitative research: A rationale and beginner's guide. *Social and Personality Psychology Compass*, 17(4):e12735.
- Jernite, Yacine, Huu Nguyen, Stella Biderman, Anna Rogers, Maraim Masoud, Valentin Danchev, Samson Tan, Alexandra Sasha Luccioni, Nishant Subramani, Isaac Johnson, Gerard Dupont, Jesse Dodge, Kyle Lo, Zeerak Talat, Dragomir Radev, Aaron Gokaslan, So-maieh Nikpoor, Peter Henderson, Rishi Bommasani, and Margaret Mitchell. 2022. Data governance in the age of large-scale data-driven language technology. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 2206–2222, New York, NY, USA. Association for Computing Machinery.
- Kiyonaga, Anastasia and Jason M. Scimeca. 2019. Practical considerations for navigating registered reports. *Trends in Neurosciences*, 42(9):568–572.
- Lakens, Daniël. 2019. The value of preregistration for psychological science: A conceptual analysis. *Japanese Psychological Review*, 62(3):221–230.
- Lakens, Daniël. 2023. Is my study useless? Why researchers need methodological review boards. *Nature*, 613(7942):9–9.
- Luccioni, Alexandra, Alexandre Lacoste, and Victor Schmidt. 2020. Estimating carbon emissions of artificial intelligence [opinion]. *IEEE Technology and Society Magazine*, 39(2):48–51.
- van Miltenburg, Emiel, Chris van der Lee, and Emiel Krahmer. 2021. Preregistering NLP research. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 613–623, Online. Association for Computational Linguistics.
- Nature Scientific Reports. ND. Registered reports. Retrieved from <https://www.nature.com/srep/journal-policies/registered-reports> on 13 July 2023.
- Peng, Kenneth, Arunesh Mathur, and Arvind Narayanan. 2021. Mitigating dataset harms requires stewardship: Lessons from 1000 papers. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. Curran.
- Pustejovsky, James, Harry Bunt, and Annie Zaenen. 2017. Designing annotation schemes: From theory to model. *Handbook of Linguistic Annotation*, pages 21–72.
- Rogers, Anna. 2020. Peer review in nlp: resource papers. Published on the Hacking Semantics blog on 16 April 2020, retrieved 13 July 2023 from: <https://hackingsemantics.xyz/2020/reviewing-data/>.
- Sambasivan, Nithya, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. “everyone wants to do the model work, not the data work”: Data cascades in high-stakes ai. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA. Association for Computing Machinery.
- Sarafoglou, Alexandra, Marton Kovacs, Bence Bakos, Eric-Jan Wagenmakers, and Balazs Aczel. 2022. A survey on how preregistration affects the research workflow: Better science but more work. *Royal Society Open Science*, 9(7):211997.
- Schlangen, David. 2021. Targeting the benchmark: On methodology in current natural language processing research. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 670–674, Online. Association for Computational Linguistics.
- Shimorina, Anastasia and Anya Belz. 2022. The human evaluation datasheet: A template for recording details of human evaluation experiments in NLP. In *Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 54–75, Dublin, Ireland. Association for Computational Linguistics.
- Søgaard, Anders, Daniel Herscovich, and Miryam de Lhoneux. 2023. A two-sided discussion of preregistration of NLP research. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 83–93, Dubrovnik, Croatia. Association for Computational Linguistics.
- Sweeney, Latanya. 2002. k-anonymity: A model for protecting privacy. *International journal of uncertainty, fuzziness and knowledge-based systems*, 10(05):557–570.
- The Royal Society. ND. Registered reports. Retrieved from <https://royalsocietypublishing.org/rsos/registered-reports> on 13 July 2023.
- Wiley. 2018. Registered reports policy on publishing at stage 1 and stage 2. Version 1.0 dated 10 January 2018, retrieved from <https://authorservices.wiley.com/author-resources/Journal-Authors/submission-peer-review/>

[registered-reports-policy.html](#) on 13 July
2023.

PARSEME Meets Universal Dependencies: Getting on the Same Page in Representing Multiword Expressions

Agata Savary, University of Paris-Saclay, CNRS, LISN, France agata.savary@universite-paris-saclay.fr

Sara Stymne, Uppsala University, Sweden sara.stymne@lingfil.uu.se

Verginica Barbu Mititelu, Romanian Academy Research Institute for Artificial Intelligence, Romania
vergi@racai.ro

Nathan Schneider, Georgetown University, USA nathan.schneider@georgetown.edu

Carlos Ramisch, Aix Marseille Univ, CNRS, LIS, Marseille, France carlos.ramisch@lis-lab.fr

Joakim Nivre, Uppsala University & RISE Research Institutes of Sweden, Sweden
joakim.nivre@lingfil.uu.se

Abstract Multiword expressions (MWEs) are challenging and pervasive phenomena whose idiosyncratic properties show notably at the levels of lexicon, morphology, and syntax. Thus, they should best be annotated jointly with morphosyntax. In this position paper we discuss two multilingual initiatives, Universal Dependencies and PARSEME, addressing these annotation layers in cross-lingually unified ways. We compare the annotation principles of these initiatives with respect to MWEs, and we put forward a roadmap towards their gradual unification. The expected outcomes are more consistent treebanking and higher universality in modeling idiosyncrasy.

1 Introduction

Multiword expression (MWE) is an umbrella term spanning a range of linguistic phenomena whose common property is *idiosyncrasy* or, more specifically, *idiomaticity*, which may manifest in many different respects: lexical, morphological, syntactic, semantic, pragmatic, and statistical (Baldwin and Kim, 2010).

MWEs are challenging and pervasive. For instance, in an MWE-annotated corpus of French (Candito et al., 2021), over 11% of all tokens belong to MWEs. Moreover, MWEs likely exist in any natural language. Therefore, modeling idiosyncrasy in language resources and tools is a natural quest. This position paper addresses two language annotation frameworks, Universal Dependencies and PARSEME, from the point of view of MWEs.

Universal Dependencies¹ (UD; Nivre et al., 2016, 2020; de Marneffe et al., 2021) is a framework for consistent annotation of grammar (parts of speech, morphological features, and syntactic dependencies) across many languages. It is an open community effort with over 300 contributors producing nearly 200 treebanks in over 100 languages. PARSEME² (Savary et al., 2018; Ramisch et al., 2020) is a scientific network which evolved from a homonymous COST action dedicated to parsing and MWEs. One of its major outcomes is a multilingual corpus annotated for verbal MWEs (VMWEs) in 26 languages by over 160 native annotators.

The common objective of UD and PARSEME is *universality*, i.e., the development of cross-linguistically consistent and applicable language descriptions. Such consistency leads to valuable

¹<https://universaldependencies.org/>

²<https://gitlab.com/parseme/corpora/>

insights about linguistic phenomena (including idiosyncrasy), contributes to contrastive studies, and promotes progress in NLP across many languages. Concretely, both UD and PARSEME (i) develop cross-lingually unified and continuously enhanced annotation guidelines, (ii) annotate, enhance, and release corpora on the basis of these guidelines, and (iii) use these corpora to develop NLP tools for syntactic parsing and MWE identification.

Despite their common goals, UD and PARSEME have operated relatively independently, ending up with partly divergent and competing terminologies and methods. Some of the MWE types addressed by PARSEME, such as light-verb constructions, are annotated to some extent also within UD, but typically not consistently across languages, as we will discuss in Section 3.6. We think it is desirable to keep morphosyntactic annotations separate from MWE-related annotations.³

The desire for greater convergence between UD and PARSEME practices has steadily grown as the initiatives have matured. PARSEME has relied on the UD format (cf. Sec. 3.2) and data in its latest corpus releases. In August 2021, a joint Dagstuhl Seminar on *Universals of Linguistic Idiosyncrasy in Multilingual Computational Linguistics* brought together the two initiatives (Baldwin et al., 2021).⁴ Finally, September 2022 saw the start of a new COST action entitled UniDive (*Universality, Diversity and Idiosyncrasy in Language Technology*), with UD/PARSEME unification on the agenda.

This paper aims at providing a roadmap towards this unification. We first survey the dimensions of MWE idiosyncrasy (Sec. 2) and compare the two frameworks' annotation principles that bear on MWEs (Sec. 3). Then, we offer short-, mid- and long-term proposals for adjusting the frameworks, paving the way towards eventually unifying them (Sec. 4). Sec. 5 concludes with future perspectives.

³The current status led to problems for the VMWE identifiers evaluated in the PARSEME shared tasks (Ramisch et al., 2020), which were given UD morphosyntactic annotations as input, and were expected to predict VMWE annotations. Since some MWE-related phenomena currently are annotated in the morphosyntactic layer, this type of evaluation is biased (since part of the information to be predicted is already given as input).

⁴<https://www.dagstuhl.de/seminars/seminar-calendar/seminar-details/21351>

2 Dimensions of Idiosyncrasy in MWEs

MWEs deviate from compositionality norms, as seen in the examples from the PARSEME languages below. The MWE in (1) contains a cranberry word *oścież*, i.e. a token having no status of a standalone word but only occurring in a MWE.⁵

- (1) **na oścież** (pl)
on 'oścież'
'wide (open)'

The MWE in (2) is exocentric, since it is a nominal phrase whose head is a finite-form verb.

- (2) **um deus nos acuda** (pt)
a god us.ACC help.IMP.2.SG
lit. 'a god-help-us' | 'a mess'

In (3), the verb is modified by an adjective and an infinitive, which is not a regular syntactic structure.

- (3) Elle **a beau** pleurer. (fr)
she has pretty.M cry.INF
lit. 'She has pretty to cry.' | 'She cries in vain.'

In (4), the possessive *her* must agree with the subject, otherwise the MWE is understood literally, as in (5).

- (4) She **knows her stuff**. (en)
'She is skilled.'
(5) #She knows my stuff (en)

Concrete nouns in verb-object constructions can inflect for number, but pluralizing the noun in (6) implies losing the idiomatic reading, as shown in (7).

- (6) a **întoarce foaia** (ro)
to turn sheet.DEF
lit. 'to turn the sheet' | 'to become harsher'
(7) #a întoarce foile (ro)
to turn sheet.PL.DEF
'to turn the sheets'

Given these examples, MWE idiosyncrasy can be considered along two orthogonal dimensions.

⁵Examples follow the PMWE conventions (Markantonatou et al., 2021). POS and morphological features use UD. We use the IETF BCP-47 standardized language codes in all examples.

Occurrences vs. Types Some idiomatic properties of MWEs display at the level of individual occurrences of MWEs (Savary et al., 2019). Conversely, others are visible at the level of *types*, that is, sets of surface realizations of the same MWE. For instance, the cranberry word (1), irregular agreement (2), and irregular syntax (3) can be observed in every single occurrence of these MWEs. On the other hand, compulsory agreement (4) or restricted inflection (6) can only be attested while considering several possible surface realizations of the given MWE, so as to test whether different inflection, agreement or syntactic alternations do or do not preserve the idiomatic reading.

Lichte et al. (2019) propose a different but isomorphic terminology, contrasting restrictive vs. defective idiosyncrasy. A *defective* property excludes a literal interpretation of a given MWE. This is observable precisely at the level of individual MWE occurrences, as in (1)–(3). A *restrictive* property reduces the number of possible surface realizations of a given MWE relative to the corresponding literal interpretation. This amounts to idiosyncrasy at the level of MWE types, as in (4)–(6).

Morphosyntactic vs. Semantic Idiosyncrasy

The idiosyncratic properties discussed above occur at the morphosyntactic level. However, the most salient property of MWEs is *semantic non-compositionality*: their meaning cannot be deduced from the meanings of their components and from their syntactic structure in a way deemed regular (Sag et al., 2002). Examples (1)–(6) can safely be considered as semantically non-compositional.

Distinguishing morphosyntactic from semantic idiosyncrasy is a hard nut to crack. First, the borders between morphology, syntax and semantics are fuzzy. For instance, the notions of syntactic and semantic arguments are closely related in the linguistic debate about arguments vs. adjuncts (Przepiórkowski and Patejuk, 2018). Second, idiosyncratic properties in MWEs usually cross multiple layers of linguistic description. For instance, the MWE in (3) exhibits not only unusual syntax but also restricted inflection, as in (6). Third, semantic non-compositionality is hard to test directly and reliably at the level of occurrences. Nonetheless, it can be more accurately approximated by lexical and morphosyntactic inflexibility, by testing it at the level of types (Gross, 1988; Gibbs and Nayak,

1989). This again suggests that morphosyntactic and semantic idiosyncrasies are entangled.

Kahane et al. (2017) propose considering syntactic and semantic idiosyncrasy as separate dimensions. They consider: (i) regular constructions, subsystems and irregular constructions, (ii) compositional, semi-compositional and non-compositional expressions, along the syntactic and semantic axes. Various expressions are then placed in this two-dimensional space. For instance, syntactically irregular constructions can be semantically compositional, e.g. (fr) *peser lourd* (lit. ‘to weigh heavy’) ‘to be very heavy’. While this classification is promising, it fails to provide an operational definition of semantic non-compositionality. In particular, assuming that formal semantics accurately approximates semantic compositionality, there can be no constructions with irregular syntax but compositional semantics.⁶ Still, what we retain from Kahane et al. (2017) is the premise that syntactic and semantic properties of MWEs should be annotated at different layers as much as possible. In particular, it is useful to display regular syntax in MWEs despite their semantic idiosyncrasy.

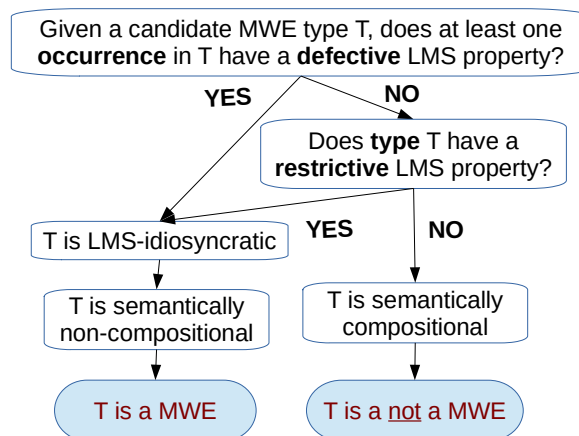


Figure 1: Implications among lexical and/or morphosyntactic (LMS) and semantic idiosyncrasy of MWE occurrences and types.

In short, we distinguish occurrence vs. type and lexical/morphosyntactic vs. semantic idiosyncrasies in MWEs, but we note that these dimensions are closely linked, as shown in Fig. 1. First, if at least one MWE occurrence is idiosyncratic,

⁶Specific compositional semantic procedures are assigned to syntactic structures deemed regular (Steedman, 2000).

then the whole type is irregular. Second, lexical and/or morphosyntactic idiosyncrasy of MWE occurrences and/or types approximates their semantic non-compositionality. Note that the choice of testing defectiveness (of an occurrence) before restrictiveness (of the whole type) is not arbitrary. First, basic observable units in an annotated corpus are occurrences (by contrast, lexicons primarily focus on types). Second, testing irregularity for an occurrence is cognitively easier than regarding the whole type. Third, the definition of a restrictive property is based on the understanding of the literal interpretation of a potential MWE. However, if a token is defective, its literal interpretation is excluded. Finally, the border between defective and restrictive properties is precisely where we would like to ultimately draw the line between UD and PARSEME annotations, i.e. only defective properties would be rendered in the UD layers.

3 Annotation Principles

This section compares the annotation principles of UD and PARSEME, focusing on MWEs.

3.1 Objectives and Principles

The common objective of UD and PARSEME is *universality*, defined as development of cross-linguistically consistent and applicable language descriptions.⁷ Both initiatives aim at representing in a unified way those phenomena which are truly similar, while leaving room for language-specific categories, relations and guidelines. The utility of these descriptions is twofold – meaningful linguistic analysis and useful language processing – in both monolingual and cross-lingual settings.

UD descriptions concern several aspects of language: segmentation, lemmas, morphology and syntax. According to the annotation properties defined by Mathet et al. (2015), these descriptions include *unitizing* (identify sentence and word boundaries) and have a *full covering* (concern all words in a corpus). PARSEME descriptions are mostly semantic (even if largely approximated by morphosyntax, see below). They also require unitizing, but are *sporadic* (only focus on components of MWEs), can be

⁷This is in contrast with the quest for absolute language universals (Greenberg, 1966; Chomsky, 1975; Tallerman, 2009).

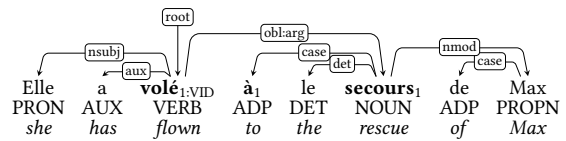


Figure 2: Sentence (8) with main annotations from UD (tree, POS tags) and PARSEME (bolding, subscripts).

nested ([[**let**]₂ **the cat** [**out**]₂ **of the bag**]₁ ‘reveal a secret’) and exhibit *free overlap* (**take**_{1,2} **a walk**₁ and a **shower**₂).

3.2 Notations and Formats

With respect to data formats, UD and PARSEME are largely compatible. Consider the example in sentence (8). Its main UD and PARSEME annotations are visualized in Fig. 2: parts of speech and dependencies are the UD-specific data, while MWEs (highlighted in boldface and subscripts) are tagged by PARSEME. The same example, in more detail, is presented in Fig. 3 in the tabular .cupt format.⁸ Each word is described in a separate line, with 11 tab-separated fields, whose headings are listed in the first line of each file. The first 10 columns are those of the .conllu format used by UD. The 11th column (PARSEME:MWE) is used by PARSEME. Components of MWEs annotated in column 11 are shown in bold.

- (8) Elle a **volé** **au** **secours** de
 She have.3SG fly.PTCP to.the rescue of
 Max (fr)
 Max
 ‘She hurried up to help Max’

3.3 Words and Tokens

Word is a fundamental notion both for UD, since its basic annotation unit is a word, and for PARSEME, since MWEs must contain at least two words. However, defining a word is one of the hardest challenges in UD, due to its fuzzy borders with morphemes on the one hand and with MWEs on the

⁸The .cupt format instantiates the CONLL-U Plus meta-format meant for complementing UD with additional layers: <https://universaldependencies.org/ext-format.html>

#	global.columns	ID	FORM	LEMMA	UPOS	XPOS	FEATS	HEAD	DEPREL	DEPS	MISC	PARSEME:MWE
1	Elle	il	PRON	_	Gender=Fem Number=Sing Person=3			3	nsubj	_	_	*
2	a	avoir	AUX	_	Mood=Ind Number=Sing Person=3 ...			3	aux	_	_	*
3	volé	voler	VERB	_	Gender=Masc Number=Sing Tense=Past ...			0	root	_	_	1:VID
4-5	au	_	_	_	_			_	_	_	_	*
4	à	à	ADP	_	_			6	case	_	_	1
5	le	le	DET	_	Definite=Def Gender=Masc Number=Sing ...			6	det	_	_	*
6	secours	secours	NOUN	_	Gender=Masc Number=Sing			3	obl:arg	_	_	1
7	de	de	ADP	_	_			8	case	_	_	*
8	Max	Max	PROPN	_	_			6	nmod	_	_	*

Figure 3: Annotation of sentence (8) as the first sentence in a corpus, in the .cupt format.

other. In UD, words are defined in morphosyntactic terms as units bearing morphological properties (e.g. a single POS) and entering into syntactic relations. Words do not always coincide with orthographic units called *tokens*.⁹ Therefore, UD defines a 3-fold relationship between words and tokens:

- A token coincides with a word.
- Several tokens build up one *multitoken word* (MTW), as in *20_000*.
- One *multiword token* (MWT) contains several words, as in (fr) *aux* (à+les) ‘in.the’.

The words (not orthographic tokens) form the basic units of analysis and receive integer indices. MWTs are represented as spans over multiple words (e.g. 4–5 in Fig. 3), including cases where words (*à* and *le*) are not retrievable from tokens (*au*). PARSEME conforms to the same definitions of words, MWTs, and MTWs, with implications for MWEs like in Fig. 3. Only the adposition *à* ‘to’ belongs to the MWE;¹⁰ the determiner *le* ‘the’ is excluded. This is possible in PARSEME due to splitting MWTs into words by UD.

Still, PARSEME covers a considerably higher number of MWTs than UD, especially verb-particle constructions written sometimes as 1 and sometimes as 2 tokens as in (9), and orthographically unitary (*closed* or *synthetic*) compounds as in (10).

- (9) **auf-passen, pass auf!** (de)
on-fit.INF, fit.IMP on!
lit. ‘to fit on, fit on!’
‘to be careful, be careful!’

⁹Neither UD nor PARSEME define tokens. We see them as units stemming from segmenting raw text for annotation.

¹⁰As evidenced by variants like (fr) *voler à son secours* (lit. ‘to.fly to his/her rescue’) ‘to hurry up to help him/her’

- (10) **Hauptrolle spielen** (de)
head.role play
‘to play the leading role’
- 2 sollst sollen ... *
3 **aufpassen** aufpassen ... 1:VPC
...
11 **Hauptrolle** Hauptrolle ... 1:LVC.full
12 **spielen** spielen ... 1

Figure 4: PARSEME annotation of unsplit MWTs.

This discrepancy leads to two issues, illustrated in Fig. 4. First, the definition of a word is inconsistent: item 3 is one word for UD but two words for PARSEME. Second, in item 11 only *rolle* ‘role’ belongs to an MWE, since *Haupt* ‘head’ can be freely replaced (*Nebenrolle spielen* ‘play the secondary role’). This cannot be rendered if UD keeps compounds unsplit.

3.4 Morphology and Syntax

In UD, the morphological description of a word employs 17 universal POS tags and over 200 values for morphological features (columns 4 and 6 in Fig. 3), though explicitly admitting that some of them may not be necessary in some languages. Syntactic annotation in UD follows the dependency approach and adopts the *lexicalist* principle. Namely, words are divided into *content words* – typically verbs, nouns, adjectives or adverbs, with referential meaning – and *function words* – determiners, adpositions, auxiliaries, etc. Content words are linked by syntactic relations, while function words attach to the content words they modify. For instance, in Fig. 2, the verb is the head of the auxiliary (items 2–3) and the nouns are the heads of the prepositions (items 4–6 and 7–8) rather than vice versa. A set of 37

syntactic relations considered universal (column 8 in Fig. 3) is defined. More specific relations in a language are accepted as subtypes of the universal ones (e.g. **obl:arg** in line 6 in Fig. 3) and 26 such subtypes are currently found in the UD treebanks. Treebanks are not required to use language-specific extensions, even if they cover phenomena for which such extensions are defined. This leads to significant inconsistencies in the use of subrelations, even among treebanks of the same language.

PARSEME, while modeling idiosyncrasy, tries to remain as independent of a particular linguistic framework as possible. It considers, for instance, that in a prepositional phrase *a preposition directly governs a noun, or the opposite, depending on a particular linguistic theory*. However, PARSEME approximates semantic compositionality by lexical and morphosyntactic flexibility tests that are driven by syntactic structure. Thus, the main PARSEME decision diagram asks questions about the syntactic head of the candidate expression, its dependents, its morphosyntactic category, etc. This implies a strong dependence on the underlying syntactic framework, and UD provides such framework, validated across many languages.

Another advantage for PARSEME is that the lexicalism in UD helps keep the MWE definition relatively simple. Namely, MWE components more easily form a weakly connected dependency graph (Sec. 3.5) if content words head function words than vice versa (Savary and Waszczuk, 2020). One minor disadvantage from lexicalism concerns MWEs with copulas. For instance in (en) *to be somebody* ‘to be important’ the pronoun heads the copula *be*, which prevents PARSEME from saying that a verbal MWE is always headed by a verb.

The universality of UD thus enables universality for PARSEME, which has been increasingly relying on UD. For all 14 languages in version 1.2 of PARSEME, MWE annotations build upon UD-compatible corpora (manually annotated or automatically predicted); and among all 26 PARSEME corpora, 20 are UD-compatible.

3.5 The Notion of MWE

The way UD and PARSEME understand the notion of an MWE is the major source of apparent discrepancies between the two frameworks. UD did not

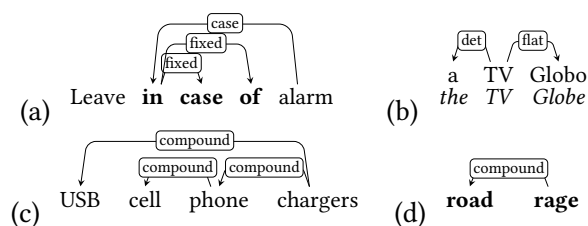


Figure 5: A complex preposition, a proper name (in Portuguese) and a nominal compound.

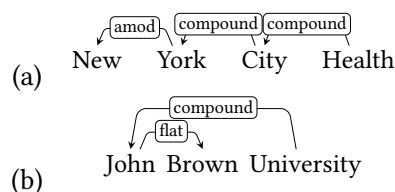


Figure 6: Complex names with mixed dependencies.

attempt to formally define MWEs, using it as an umbrella term for expressions for which other syntactic relations seem useless or inconvenient. UD defines 3 dependency labels in such cases.

[fixed] is used for highly *grammaticalised* expressions, as in Fig. 5a, that typically behave as function words or short adverbials, i.e. belong to *closed* grammatical categories. The name of the label inspired by Sag et al. (2002) signals morphosyntactic fixedness. By convention, all parts of such an expression are attached to the leftmost component, that is, the whole is considered *headless* (even if a head might be identifiable).

[flat] is meant for *headless* semi-fixed expressions, like names or complex numerals, as in Fig. 5b. These belong to *open* categories and are subject to high *productivity*.

[compound] marks any word-level compounding, including nouns, adjectives, and verbs. Compounds are seen as *headed* expressions, i.e. modification relations are rendered, as in Fig. 5c. A compound may or may not be semantically compositional, as in Fig. 5c and 5d, respectively.

This typology concerns dependency relations, not expressions. In particular, various labels can be mixed within one expression, as shown in Fig. 6. Some UD subtypes (e.g. **compound:lvc**, **expl:pv**) are related to MWEs in PARSEME (Sec. 3.6).

For PARSEME, an MWE is a combination of words with at least two *lexicalized components* (always realized by the same lexemes) displaying lex-

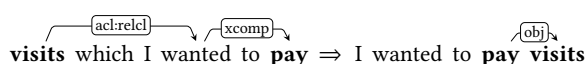


Figure 7: A VMWE candidate and its canonical form.

ical, morphological, syntactic or semantic *idiosyncrasies* (Sec. 2). Even if PARSEME’s ambition is to model MWEs in general, its major efforts were put into *verbal MWEs* (VMWEs). A VMWE is defined as an MWE whose *canonical form* (least syntactically marked form keeping the idiomatic reading) is such that its syntactic head is a verb, its other lexicalized components form phrases directly dependent on this verb (the whole forms a weakly connected graph), and it passes the idiosyncrasy tests defined in the PARSEME guidelines. MWE candidate sequences must be transformed into canonical forms. For instance, the candidate on the left of Fig. 7 does not fulfill the conditions, but transforming it into the canonical form on the right restores graph connectivity and verb-headedness.

3.6 MWE Categories

PARSEME defines 3 quasi-universal categories (the first 3 below, present in many languages but not all), and 2 universal ones (the last 2 below, present in all languages under study).¹¹ Statistics about these annotations in the data are given in Appendix A.

Inherently Reflexive Verbs (IRV) combine a verb *V* and a reflexive clitic *R* such that (i) *V* never occurs without *R*, as in (sv) *gifta sig* (lit. ‘get-married oneself’) ‘get married’, or (ii) *R* distinctly changes the meaning or valency of *V*, as in (es) *recogerse* ‘go home’/ *recoger* ‘gather’. They are contrasted with regular reflexives: true reflexive, reciprocal, middle passive and impersonal, e.g. (ro) *casele se vând bine* (lit. ‘houses sell themselves well’) ‘houses sell well’. In UD, the above uses are divided into two classes, depending on if the reflexive clitic can or cannot be mapped on a semantic argument of the verb. In the former case, the “regular” dependency label corresponding to the role of the clitic is used, e.g. **obj** in (pl) *myć się* ‘wash oneself’. The latter is covered by the **expletive** label. Subrelations can further distinguish these uses, in

¹¹https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.3/?page=030_Categories_of_VMWEs

particular, **expl:pv** covers case (i) above, signaling idiosyncrasy.

Verb-Particle Constructions (VPCs) have two subclasses in PARSEME. In fully non-compositional VPCs (VPC.full), adding the particle considerably changes the meaning of the verb, as in (sv) *Det gick upp för mig* (lit. ‘It went up to me’) ‘It occurred to me’. In semi-compositional VPCs (VPC.semi), the particle adds a partly predictable but non-spatial meaning to the verb, as in (sv) *äta upp* (lit. ‘eat up’) ‘finish eating’. Verb-particle combinations where the particle only adds spatial meaning are not annotated, as in (sv) *gick upp på vinden* ‘went up to the attic’. In UD the subrelation **compound:prt** can be used to connect a particle to its head verb, regardless of idiomaticity, i.e. all 3 examples above fall into this category.

Multi-Verb Constructions (MVCs) are idiomatic combinations of two verbs, e.g. (fr) *laisser tomber* (lit. ‘to let fall’) ‘to abandon’, in particular serial verbs in Asian languages, e.g. (hi) *kar le* (lit. ‘do take’) ‘do for one’s own benefit’. This relates to the UD **compound:svc** subrelation, which however covers serial verbs both in idiomatic and compositional uses, e.g. (ja) *naguri korosi* (lit. ‘punch kill’) ‘kill by punching’.

Light-Verb Constructions (LVCs) are combinations of semantically light verbs and predicative nouns expressing the semantics of the action or state. Two subcategories are defined. In LVC.full the verb’s subject is the noun’s semantic argument as in (sl) *imeti predavanje* ‘give a lecture’. In LVC.cause the verb’s subject is the cause or source of the noun, as in (en) *grant right*. In UD, the same expressions are most often annotated with the “regular” **obj** dependency, even if the scope of the **compound:lvc** subrelation is similar to LVC.full.

Verbal Idioms (VID) is the most diverse category in PARSEME, gathering cases not covered by other categories. The verb’s dependents are unrestricted, including subjects, as in (en) *a little bird told me*, direct objects, as in (6), etc. The verb can have several dependents, as in (en) *cut a long story short*, or combine features from other VMWE categories, as in (sv) *sätta sig upp mot någon* (lit. ‘sit oneself up against someone’) ‘defy someone’. A VID candidate must display lexical or morphosyntactic idiosyncrasy, as in (1)–(6). As VIDs are so diverse, there is no direct correspondence in UD. They are

typically annotated as syntactically regular, possibly with subrelations for particles and reflexives when those are parts of the VID. The UD **fixed** relation cannot be used to signal inflexibility in VIDs since it is limited to functional MWEs.

4 Towards UD-PARSEME Unification

The discrepancies discussed above harm universality, therefore we are taking steps towards unifying UD and PARSEME. The expected advantage lies in a better parallelism in annotating syntactic vs. semantic and regular vs. idiosyncratic properties. Our intuition is that semantic non-compositionality is an intriguing phenomenon and annotators wish to signal it even when annotating morphosyntax. If an MWE has (partly) regular syntax but idiomatic semantics, and if only morphosyntactic labels are available, annotators might prefer to signal idiosyncrasy rather than regularity.¹² Another temptation is to introduce new subtypes such as **obj:lvc**, which could block other useful syntactic distinctions that could be encoded with subtypes (since recursive subtypes are not allowed). Adding the MWE layer to the annotation schema solves these problems.

Another motivation is that both automatic processing of MWEs and parsing benefit from solving the two tasks jointly (Constant et al., 2017; Taslimipoor et al., 2020), therefore aligning morphosyntactic and MWE annotations serves NLP. Here, we lay out a multistage unification roadmap for major issues, summarized in Appendix B.

Note that no re-annotation effort is required on the UD side in the first two stages. This is important for at least three reasons. Firstly, while for PARSEME idiosyncrasy is central, for UD it is only one of the many phenomena to be modeled. It is therefore natural for the PARSEME community to be the main responsible party for changes related to idiosyncrasy. Secondly, the UD community of treebank creators and users is very large. Any change in annotation principles, in order to be widely adopted, should minimize manual re-annotation and should be divided into small, easily achievable steps. Thirdly, as mentioned in Sec-

¹²E.g. in the Romanian Reference Treebank, VIDs with regular syntax like *avea loc* ‘take place’ are marked as **fixed**.

```
# global.columns = ID FORM LEMMA ... PARSEME:MWE
1 die der ... *
2 Hauptrolle Hauptrolle ... 1@6-10:LVC.full
3 spielen spielen ... 1
```

Figure 8: PARSEME tag with a sub-token span

tion 3.1, while PARSEME annotation is sporadic, UD trees fully cover the annotated text, which implies a heavier (re-)annotation workload.

4.1 Words and Tokens

PARSEME’s notion of word is sometimes more granular than UD’s (Sec. 3.3), and the segmentation of tokens into words would need to be reconciled. This is crucial for cases where MWEs cover parts of tokens, as in (10). Another such case occurs in Korean UD treebanks (Chun et al., 2018; Oh et al., 2020), where agglutinative postpositions are considered to form a syntactic word with their stem (segmented only in the lemma), as in (11). The postposition *에* (-*ey*) and following word *대해* (*tayhay*) together mean ‘about’, but because the postposition is not split, we would need to refer to a subword unit.

(11) 언어에 대해 읽다 (ko)
language:POSTP about read
‘read about languages’

Short-term Proposal We propose to supplement existing UD parses with MWE annotations, *without altering tokenization*. MWEs that encompass entire MWTs, as in (9), are already covered by PARSEME. For cases like (10) and (11), the MWE column could specify sub-token spans, as in Fig. 8.

Long-term Proposal Parts of unsplit tokens participating separately in MWEs suggest a deficiency in UD’s implementation of MWTs. We propose that, ultimately, UD syntactically recognize synthetic compounds as productive, regardless of the MWE status. This would require UD treebanks in some languages to systematically split current compound words into MWTs, ensuring each component word has an appropriate lemma and morphological features, and adding a dependency relation (such as **compound** or **compound:prt**) between them. This could also help disambiguate the interpretation of some compounds, as in (sv) *bildrulle: bil+drulle*, *bild+rulle* ‘car maniac (bad driver), picture roll (roll

of film)’.¹³

4.2 Terminology and Guidelines

A common understanding of MWE-related terminology is a basic requirement for UD/PARSEME convergence. This could be achieved progressively.

Short-term proposals Different interpretations of the term “multiword expression” are understandable (Sec. 3.5), since it literally means an expression containing two or more words, with no further restriction. However, the term, as understood by the MWE community (Baldwin and Kim, 2010), has an extra meaning component of idiosyncrasy (it is itself an MWE!), and we propose to adhere to this definition.¹⁴ This would mean, for UD, *not to use the term MWE for phenomena considered regular*, replacing the MWE heading (currently describing **compound**, **fixed** and **flat**) with a more neutral description like “other complex constructions”. This should be easy to achieve, as MWEs do not have a technical definition in UD: the term is used casually in the guidelines, but is not part of the morphological or syntactic labels or their criteria. This proposal is conservative in the sense that it does not, in principle, require modifications of the annotations.¹⁵ On the PARSEME side, the VPC label might be renamed to IVPC (for idiomatic VPC), so as to signal that verb-particle combinations can be both regular and idiomatic, and only the latter are MWEs (Sec. 3.6). Criticism of the current VMWE guidelines (Savary and Waszczuk, 2020; Fotopoulou et al., 2021) should also be addressed.

Mid-term Proposals A major mid-term requirement for PARSEME would be to extend its terminology and guidelines to *all syntactic types* of MWEs, rather than VMWEs only, e.g. based on the foundational work by Schneider et al. (2014) and Candito et al. (2021). Challenges include defining the borders between named entities and MWEs.

Long-term Proposals Most languages contain *productive grammatical subsystems* which yield

¹³Tokenization issues occur not only in compounds. Agglutinative languages may adopt different word segmentation strategies, in spite of similar structure (Han et al., 2020). This must also be addressed (Tyers et al., 2021) but goes beyond this paper’s scope.

¹⁴Even if “idiomatic MWE” would be more precise.

¹⁵One exception, in English, would be to abandon the semantic **compound:prt** vs. **advmod** distinction in VPCs.

expressions with particular syntax and semantics, such as names, numbers, measurements, and dates (Kahane et al., 2017; Schneider and Zeldes, 2021). Their heavy semantic load makes them central units of interest for NLP. They partially overlap with regular syntax and MWEs, e.g. (pl) *Małgorzata Kowalska* is a name with a regular noun-adjective structure, and (en) “*Always Look On The Bright Side Of Life*” is a title containing a VID. However, they also follow specific patterns, such as defective number agreement in (en) *two million*, and nesting (Fig. 6). They call for normalisation standards like TimeML and AMR (Pustejovsky et al., 2003; Banarescu et al., 2013). Annotating subsystems jointly with UD and PARSEME would require new instantiations of CONLL-U Plus, with extra columns, such as ‘NE’ in Fig. 9. Other initiatives are making progress towards adding entity and coreference layers to UD (Nedoluzhko et al., 2022).

4.3 Occurrence vs. Type Encoding

We suggest unification steps towards a better account of the type/occurrence nature of idiosyncrasies.

Mid-term Proposals As soon as PARSEME extends its guidelines to all syntactic MWE types, they should be applied to all PARSEME corpora. The general principle would be:

- UD layers only account for lexical/morphosyntactic idiosyncrasy of MWE *occurrences*, such as irregular syntax in (3). Grammatically regular MWE occurrences would receive “ordinary” annotation, regardless of semantics.
- The PARSEME layer would signal *any* kind of semantic *idiosyncrasy*, i.e. it would flag each expression which is lexically/morphosyntactically irregular, whether at the level occurrences or of types, e.g. for all examples in Sec. 2.

This would require a systematic use of the .cupt format to jointly represent all dimensions of idiosyncrasy. This would also question the utility of UD’s **fixed** label, since fixedness is a property of types rather than occurrences. Maybe this label could be merged with **flat** and both renamed to **headless** to avoid confusion with previous interpretations.

```

# global.columns = ID FORM ... HEAD DEPREL ... MWE NE
1  Leave ... 0  root    ... *          *
2  in   ... 3  case    ... 1:AdvMWE.fixed *
3  case ... 1  obl     ... 1          *
4  of    ... 5  case    ... *          *
...
11 Leave ... 0  root    ... *          *
12 in   ... 15 case    ... 1:AdvMWE.fixed *
13 case ... 12 headless ... 1          *
14 of    ... 12 headless ... *          *

31 a     ... 32 det     ... *          *
32 TV    ... 34 nsubj   ... *          1:ORG
33 Globo ... 32 headless ... *          1

```

Figure 9: Two possible annotations for a multiword preposition; and a headless organization name.

The PARSEME layer might deal with signaling total (rather than partial) fixedness if needed. The example in Fig. 5a would be annotated as in Fig. 9, depending if it is seen as analysable (lines 1–4) or **headless** (lines 11–14). The example in Fig. 5b would also be **headless** (lines 31–33), with a possible named entity type (column 12), if a subsystem layer is added to the schema (Sec. 4.2).

These would be major changes, and authors of some treebanks might not be sufficiently interested in idiomatity to accept the addition of a column. In this case, the previous distinction between **fixed** and **flat** should be kept to distinguish grammaticalized and productive headlessness. Subrelations such as **compound:prt** and **compound:svc** should probably be kept but used more consistently, since they are orthogonal to idiosyncrasy. Subrelations **:lvc** and **:pv** are superfluous: we propose to abandon them and use the 11th column instead.

Long-term Proposals Most optimally, the occurrence-type dichotomy of idiosyncrasy could be modeled in a framework in which corpus and lexicon are interlinked. A corpus would document occurrences, i.e. MWE occurrences would only be annotated for individual properties (including occurrence-wise idiosyncrasy such as irregular syntax). The lexicon would describe types, i.e. all occurrences of the same MWE would be linked to a lexicon entry representing its type and storing its type-wise properties such as categories (LVC, VID, etc.) and a meaning. A similar schema was implemented by Bejcek and Stranák (2010). An MWE lexicon entry could also contain other type-specific

properties such as canonical forms (lemmas), flexibility and agreement constraints, as in (4–6), and links to ontologies (Hajnicz and Bartosiak, 2019).

Finally, an MWE lexicon could be more compliant with a typological perspective. PARSEME’s current MWE typology is driven by annotation needs, i.e. new categories are introduced if specific tests are needed to identify some MWE in texts. An orthogonal, more typologically-driven categorization could use cross-linguistic constructions and language-specific structural types (Koptjevskaja-Tamm, 2002).

4.4 Data Quality

Both UD and PARSEME provide contributors with automatic data quality checkers. These should be unified and extended. PARSEME might enhance its validator to check compliance with guidelines (e.g. a verb in an LVC must have a single lexicalised dependent), and should integrate it with the UD validator, which runs automatically when a new version of a treebank is pushed to the GitHub repository. UD might develop tools inspired by PARSEME’s consistency checks, in which a “vertical” view of the corpus groups annotations of the same MWE. This might help overcome inconsistencies within a treebank or within treebanks for the same language, e.g. due to the optionality of subrelations.

5 Conclusions and Future Work

We have compared how UD and PARSEME capture linguistic idiosyncrasy. Since PARSEME largely agrees with UD’s objectives, it increasingly follows UD on data formats, morphology, (regular) syntax and tokenisation.

We are optimistic about UD and PARSEME joining forces for compatible encoding of regular and idiosyncratic phenomena, as detailed in our roadmap proposal. In the long run, these efforts might benefit from more typological insights. Also, extending the annotation schema to large classes of constructions would enable an even more comprehensive account of idiosyncrasy. The implementation of these suggestions will depend, however, on a delicate balance between existing and upcoming data, automation tools, and—above all—on availability and willingness of contributors.

Acknowledgements

This work started out as a discussion at the Dagstuhl Seminar 21351: Universals of Linguistic Idiosyncrasy in Multilingual Computational Linguistics.¹⁶ We thank all the participants for inspiration. We are grateful to Schloss Dagstuhl, the Leibniz Center for Informatics, as well as the organizers of the event, for bringing us together.

We are also building upon the efforts of UniDive, the CA21167 COST Action: Universality, diversity and idiosyncrasy in language technology.¹⁷ At the time of drafting this paper, UniDive was at the proposal stage and paved the way towards many of the ideas presented here.

Our work has also been partly funded by the French Agence Nationale pour la Recherche, through the SELEXINI project (ANR-21-CE23-0033-01).

We thank Jena Hwang for providing the Korean example (11).

References

- Timothy Baldwin, William Croft, Joakim Nivre, and Agata Savary. 2021. Universals of Linguistic Idiosyncrasy in Multilingual Computational Linguistics (Dagstuhl Seminar 21351). *Dagstuhl Reports*, 11(7):89–138.
- Timothy Baldwin and Su Nam Kim. 2010. Multiword Expressions. In Nitin Indurkha and Fred J. Damerau, editors, *Handbook of Natural Language Processing, Second edition*, pages 267–292. CRC Press, Boca Raton.
- Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for sembanking. In *Proc. of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria.
- Eduard Bejcek and Pavel Stranák. 2010. Annotation of multiword expressions in the Prague dependency treebank. *Language Resources and Evaluation*, 44(1-2):7–21.
- Marie Candito, Mathieu Constant, Carlos Ramisch, Agata Savary, Bruno Guillaume, Yannick Parmentier, and Silvio Cordeiro. 2021. A French corpus annotated for multiword expressions and named entities. *Journal of Language Modelling*, 8(2):415–479.
- Noam Chomsky. 1975. *Reflections on Language*. Pantheon, New York.
- Jayeol Chun, Na-Rae Han, Jena D. Hwang, and Jinho D. Choi. 2018. Building Universal Dependency treebanks in Korean. In *Proc. of LREC*, pages 2194–2202, Miyazaki, Japan.
- Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke van der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Multiword expression processing: A survey. *Computational Linguistics*, 43(4):837–892.
- Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.
- Aggeliki Fotopoulou, Eric Laporte, and Takuya Nakamura. 2021. Where Do Aspectual Variants of Light Verb Constructions Belong? In *Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021)*, pages 2–12, Online. Association for Computational Linguistics.
- Raymond W. Gibbs and Nandini P. Nayak. 1989. Psycholinguistic studies on the syntactic behavior of idioms. *Cognitive Psychology*, 21:100–138.
- Joseph H. Greenberg, editor. 1966. *Universals of language*, 2nd edition. MIT Press, Cambridge, MA.
- Gaston Gross. 1988. Degré de figement des noms composés. *Langages*, 23(90):57–72.
- Elżbieta Hajnicz and Tomasz Bartosiak. 2019. Connections between the semantic layer of *Walenty* valency dictionary and PIWordNet. In *Proceedings of the 10th Global WordNet Conference (GWC 2019)*, pages 99–107, Wrocław. Oficyna Wydawnicza Politechniki Wrocławskiej.

¹⁶<https://drops.dagstuhl.de/opus/volltexte/2021/15591/>

¹⁷<https://www.cost.eu/actions/CA21167/>

- Ji Yoon Han, Tae Hwan Oh, Lee Jin, and Hansaem Kim. 2020. Annotation issues in Universal Dependencies for Korean and Japanese. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pages 99–108, Barcelona, Spain (Online). Association for Computational Linguistics.
- Sylvain Kahane, Marine Courtin, and Kim Gerdes. 2017. Multi-word annotation in syntactic treebanks - Propositions for Universal Dependencies. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories*, pages 181–189, Prague, Czech Republic.
- Maria Koptjevskaja-Tamm. 2002. Adnominal possession in the European languages: form and function. *STUF - Language Typology and Universals*, 55(2):141–172.
- Timm Lichte, Simon Petitjean, Agata Savary, and Jakub Waszczuk. 2019. Lexical encoding formats for multi-word expressions: The challenge of “irregular” regularities. In Yannick Parmentier and Jakub Waszczuk, editors, *Representation and parsing of multiword expressions: Current trends*, pages 1–33. Language Science Press, Berlin.
- Stella Markantonatou, Carlos Ramisch, Victoria Rosén, Mike Rosner, Manfred Sailer, Agata Savary, and Veronika Vincze. 2021. PMWE conventions for examples containing multiword expressions. Technical report, *Phraseology and Multiword Expressions – book series at Language Science Press*.
- Yann Mathet, Antoine Widlöcher, and Jean-Philippe Métévier. 2015. The unified and holistic method gamma (γ) for inter-annotator agreement measure and alignment. *Computational Linguistics*, 41(3):437–479.
- Anna Nedoluzhko, Michal Novák, Martin Popel, Zdeněk Žabokrtský, Amir Zeldes, and Daniel Zeman. 2022. CorefUD 1.0: Coreference meets Universal Dependencies. In *Proc. of LREC*, Marseille, France.
- Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: a multilingual treebank collection. In *Proc. of LREC*, pages 1659–1666, Portorož, Slovenia.
- Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proc. of LREC*, pages 4027–4036, Marseille, France.
- Tae Hwan Oh, Ji Yoon Han, Hyonsu Choe, Seokwon Park, Han He, Jinho D. Choi, Na-Rae Han, Jena D. Hwang, and Hansaem Kim. 2020. Analysis of the Penn Korean Universal Dependency Treebank (PKT-UD): Manual revision to build robust parsing model in Korean. In *Proc. of IWPT*, pages 122–131, Online.
- Adam Przepiórkowski and Agnieszka Patejuk. 2018. Arguments and adjuncts in Universal Dependencies. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3837–3852, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- James Pustejovsky, José M. Castaño, Robert Ingria, Roser Saurí, Robert J. Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R. Radev. 2003. TimeML: Robust specification of event and temporal expressions in text. In *IWCS-5, Fifth International Workshop on Computational Semantics*, Tilburg, Netherlands.
- Carlos Ramisch, Agata Savary, Bruno Guillaume, Jakub Waszczuk, Marie Candito, Ashwini Vaidya, Verginica Barbu Mititelu, Archana Bhatia, Uxoá Iñurrieta, Voula Giouli, Tunga Güngör, Menghan Jiang, Timm Lichte, Chaya Liebeskind, Johanna Monti, Renata Ramisch, Sara Stymne, Abigail Walsh, and Hongzhi Xu. 2020. Edition 1.2 of the PARSEME Shared Task on Semi-supervised Identification of Verbal Multiword Expressions. In *Proc. of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 107–118, Barcelona, Spain (Online).
- Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the 3rd International Conference on*

Intelligent Text Processing and Computational Linguistics (CICLing-2002), pages 1–15, Mexico City, Mexico.

Agata Savary, Marie Candito, Verginica Barbu Mititelu, Eduard Bejček, Fabienne Cap, Slavomír Čéplö, Silvio Ricardo Cordeiro, Gülşen Eryiğit, Voula Giouli, Maarten van Gompel, Yaakov HaCohen-Kerner, Jolanta Kovalevskaite, Simon Krek, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Lonneke van der Plas, Behrang QasemiZadeh, Carlos Ramisch, Federico Sangati, Ivelina Stoyanova, and Veronika Vincze. 2018. PARSEME multilingual corpus of verbal multiword expressions. In Stella Markantonatou, Carlos Ramisch, Agata Savary, and Veronika Vincze, editors, *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, pages 87–147. Language Science Press., Berlin.

Agata Savary, Silvio Cordeiro, and Carlos Ramisch. 2019. Without lexicons, multiword expression identification will never fly: A position statement. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 79–91, Florence, Italy. Association for Computational Linguistics.

Agata Savary and Jakub Waszczuk. 2020. Polish corpus of verbal multiword expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 32–43, online. Association for Computational Linguistics.

Nathan Schneider, Spencer Onuffer, Nora Kazour, Emily Danchik, Michael T. Mordowanec, Henrietta Conrad, and Noah A. Smith. 2014. Comprehensive annotation of multiword expressions in a social web corpus. In *Proc. of LREC*, pages 455–461, Reykjavík, Iceland.

Nathan Schneider and Amir Zeldes. 2021. Mischievous nominal constructions in Universal Dependencies. In *Proc. of the Fifth Workshop on Universal Dependencies (UDW, SyntaxFest 2021)*, pages 160–172, Sofia, Bulgaria.

Mark Steedman. 2000. *The Syntactic Process*. MIT Press.

Maggie Tallerman. 2009. If language is a jungle, why are we all cultivating the same plot? *Behavioral and Brain Sciences*, 32:469–470.

Shiva Taslimipoor, Sara Bahaadini, and Ekaterina Kochmar. 2020. MTLB-STRUCT @PARSEME 2020: Capturing unseen multiword expressions using multi-task learning and pre-trained masked language models. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 142–148, online. Association for Computational Linguistics.

Francis Tyers, Ekaterina Vylomova, Daniel Zeman, and Tim Zingler. 2021. *Working Group 1 (What counts as a word?)*, chapter 4.1. Volume 11 of (Baldwin et al., 2021).

A Statistics of the Use of MWE-related Labels in the UD and PARSEME Corpora

Table 1 shows the statistics and comments about the use of MWE-related labels in the UD treebanks in version 2.9 (with 131 treebanks in total).

Table 2 documents the number of PARSEME languages in which the MWE labels are used.

B Roadmap for UD-PARSEME Unification

Table 3 summarises the proposals from Section 4.

Label	Treebanks	Comments
fixed	109	Limited to functional MWEs
flat	119	Productive headless constructions
compound	99	Productive headed compounds. 10 additional treebanks have the compound relation, but always with a subtype.
expl:pv	20	Inconsistent use in Spanish-AnCora vs. Spanish-GSD, French-GSD vs. other French treebanks
compound:prt	32	In English compound:prt is used when the particle is not spacial, and advmod otherwise. The same distinction is suggested in the universal guidelines. Inconsistently used in Persian-Seraji vs. Persian-PerDT
compound:svc	8	
compound:lvc	11	Most often commuted for obj . Inconsistently used in Turkish-BOUN and Turkish-IMST vs. all other Turkish treebanks

Table 1: Use of MWE-related labels in the UD treebanks in version 2.9 (with 131 treebanks in total)

Label	Corpora	Comments
IRV	8	
VPC.full	6	Greek and Hebrew use only VPC.full
VPC.semi	5	Chinese uses only VPC.semi
MVC	7	
LVC.full	14	Hindi allows adjectives in place of nouns
LVC.cause	13	Not in Turkish
VID	14	

Table 2: Use of MWE-related labels in the PARSEME corpora in version 1.2 (with 14 languages in total)

	Short-term	Mid-term	Long-term
UD	Assume idiosyncrasy of MWEs Don't use <i>MWE</i> as umbrella term for fixed , compound and flat	Use the .cupt format Merge fixed with flat , maybe rename to headless Abandon compound:lvc and expl:pv In new annotations, only flag token idiosyncrasy	Annotate subwords whenever appropriate (e.g. <i>Haupt-rolle</i>) Extend the annotation schema to subsystems
PARSEME	Tag spans for subtokens (<i>Hauptrolle</i>) Rename VPCs to IVPCs	Guidelines for all syntactic types of MWEs, with subtypes for totally fixed MWEs Define the border between named entities and MWEs Annotate MWEs of all syntactic types Flag both token and type idiosyncrasy	Link corpora with MWE lexicons, encode MWE type properties in the lexicons Use orthogonal typology-inspired categories Extend the annotation schema to constructions

Table 3: Roadmap for the UD-PARSEME unification. Actions with white background require no manual (re-)annotation. Actions highlighted in blue will require major annotation effort: those in **dark blue** apply to all languages, whereas those in **light blue** (concerning subword-level annotations) apply to a subset of languages.

Barriers and enabling factors for error analysis in NLG research

Emiel van Miltenburg, Tilburg University, The Netherlands. c.w.j.vanmiltenburg@tilburguniversity.edu*

Miruna Clinciu, Edinburgh Centre for Robotics, Heriot-Watt University, University of Edinburgh, UK

Ondřej Dušek, Charles University, Prague, Czechia

Dimitra Gkatzia, Edinburgh Napier University, UK

Stephanie Inglis, Arria NLG, Aberdeen, UK

Leo Leppänen, University of Helsinki, Finland

Saad Mahamood, trivago N.V., Düsseldorf, Germany

Stephanie Schoch, University of Virginia, USA

Craig Thomson, University of Aberdeen, UK

Luou Wen, Independent Researcher, UK

Abstract Earlier research has shown that few studies in Natural Language Generation (NLG) evaluate their system outputs using an error analysis, despite known limitations of automatic evaluation metrics and human ratings. This position paper takes the stance that error analyses should be encouraged, and discusses several ways to do so. This paper is based on our shared experience as authors as well as a survey we distributed as a means of public consultation. We provide an overview of existing barriers to carrying out error analyses, and propose changes to improve error reporting in the NLG literature.

1 Introduction

Error analysis is a formalised procedure through which researchers identify and categorise errors in system output. In the context of Natural Language Generation (NLG), error identification often means manually annotating the output text, ideally with multiple annotators ([van Miltenburg et al., 2021a](#)). The results of this analysis are often presented in a table, ranking the error categories by their frequency. This goes beyond the more common practice of providing some (strategically) hand-picked examples of ‘cherries’ (showing good model performance) and ‘lemons’ (showing the opposite).¹

While error analysis is relatively labour-intensive, it has some important advantages over commonly used evaluation metrics (see [Celikyilmaz et al. 2020](#) for an

overview) or human ratings ([Howcroft et al., 2020](#); [van der Lee et al., 2021](#)). These metrics only provide overall scores, and they do not explain what aspects of the output show room for improvement. Error analysis *does* provide this information, and as such it is an essential step towards tackling issues with the output. Based on an error analysis, one might for example establish a benchmark that targets common weaknesses of NLG systems. (See [Van Miltenburg et al. 2021a](#) for further discussion.) Moreover, error analyses provide a healthy dose of skepticism with regard to system performance, and as such help avoid the *fallacy of AI functionality* ([Raji et al., 2022](#))². Finally, it is simply not possible to automatically evaluate *all* aspects of NLG output ([Raji et al., 2021](#)). Error analysis is flexible enough to identify

²Briefly, the fallacy of AI functionality is the assumption that AI systems work as advertised, and can readily be deployed to carry out the task they were trained to perform, without any strong evidence to back up this claim. Although neural NLG systems may achieve high scores through automatic metrics on community leaderboards, they may still make surprising mistakes that keep them from being useful. These mistakes can be detected through manual inspection.

*This project was led by the first author. The remaining authors are presented in alphabetical order.

¹For reference on this terminology, see https://en.wikipedia.org/wiki/Cherry_picking and https://en.wikipedia.org/wiki/Lemon_law

the issues that are most salient to the human eye.

Despite the usefulness of error analyses, [Van Miltenburg et al. \(2021a\)](#) have shown in their survey of INLG papers published in 2010, 2015, and 2020 that relatively few NLG papers included them (about 11% of the papers surveyed). [Gehrmann et al. \(2022\)](#) provide a similarly low number (about 23% of papers published at ACL, INLG, or EMNLP 2021). It is unclear why most authors do not report error analyses in their work, or how we might encourage authors to carry out an error analysis. We aim to provide clarity on both counts.

Based on earlier work by [Van Miltenburg et al. \(2021a\)](#) and our own experiences as NLG researchers, we identified nine different factors that might influence authors in their decision (not) to carry out an error analysis. We then carried out a public consultation in the form of a survey among NLG researchers to ask for their opinions on error analysis and to identify additional barriers and enabling factors for carrying out an error analysis. This way, we obtained a validated list to discuss in this position paper, where we take the stance that error analysis should be promoted.

Our findings suggest that NLG researchers generally appreciate error analyses they see in the work of others, but they are held back from carrying out an error analysis themselves for various reasons. We discuss the aspects that could enable the reporting of error analyses and argue for meaningful changes to the publication process, so that future researchers may reap the benefits of a research culture where error analyses are rewarded. The code and data for this research project are freely available online.³

2 Related work

2.1 Evaluation of NLP & NLG systems

Evaluation is a hot topic that is garnering more attention in both NLG and NLP research communities. There is increasing recognition that current automatic and human evaluation practices are insufficient ([Gehrmann et al., 2022](#)). This has resulted, recently, in several evaluation-focused workshops, such as *Eval4NLP*, *EvalNLGEval*, *HumEval*, and *GEM*. This shows a high interest in topics that specifically address the question of evaluation. These workshops are being organised on top of well-established academic conferences and events.

We believe there are several (interconnected) factors that have led to evaluation receiving this increased level of attention:

‘Superhuman’ performance Tasks are becoming saturated more quickly, with systems performing at

³<https://github.com/evanmiltenburg/ErrorAnalysisSurvey>

or above what has been defined as a human level of performance under the given evaluation setup ([Kiela et al., 2021](#)). Current benchmarks have been criticized from two main angles. (1) The decontextualized setup of these tasks tends to make benchmarks less natural, which puts human judges at a disadvantage ([Läubli et al., 2020](#)). (2) More generally, it is questionable whether many of these computational tasks suitably model the broad language tasks that they claim to model ([Raji et al., 2021](#)).

Uninformative metrics There is also an awareness of poor correlation between human judgements and automatic metrics ([Reiter and Belz, 2009](#); [Novikova et al., 2017](#); [Clinciu et al., 2021](#)), as well as the need to move beyond a single number to evaluate the performance of a given system with diverse sets of evaluation suites ([Mille et al., 2021](#)).

Unequal comparisons Recent advances, such as the Transformer architecture ([Vaswani et al., 2017](#)), provided NLP practitioners with new and undoubtedly powerful tools for building NLG systems and metrics. However, these novel advances have not yet led to a flourish of commercial neural NLG systems, which remain largely symbolic ([Dale, 2020](#)).⁴ Neural systems are prone to hallucination; they include extraneous and often factually inaccurate content ([Ji et al., 2022](#)) that metrics either miss or were never designed to detect ([Thomson and Reiter, 2021](#)). [Dušek et al. \(2020\)](#) show that, compared to non-neural data-driven, rule-based, or template-based models, sequence-to-sequence models typically score higher on word-overlap metrics such as BLEU or METEOR, and human ratings for naturalness, but lower in human ratings of overall quality.

Taken together, all of the above factors indicate that our evaluation tasks, metrics, and procedures likely need to be improved so that we can meaningfully compare different systems with each other as well as to humans, simple baselines, or other measures of acceptable performance.

As [Gehrmann et al. \(2022\)](#) note, there are many known issues with evaluation practices in NLG, and many proposals have been made to improve the situation. [Gehrmann et al. \(2022\)](#) looked at the adoption rates of different evaluation techniques, and they show that many current best practices (including error analysis) are not being followed. A recent interview of NLG practitioners ([Zhou et al., 2022](#)) showed that authors tend to prioritise certain types of quality criteria (such as correctness, grammaticality, usefulness, etc.) without a shared full understanding of what these criteria mean, something also observed by [Howcroft et al.](#)

⁴With the exception of machine translation, which may or may not be counted as an NLG task (depending on who you ask).

(2020) and Belz et al. (2020). There are also open questions as to which criteria are sufficient to demonstrate that a system is suitable for purpose.

2.2 Meta-science

This paper is an exercise in *meta-science*. By this term, we mean researchers studying and reflecting on the way scientific research is carried out and subsequently reported. Many people associate meta-science with the open science movement. Following the replication crisis in psychology and other fields, researchers have made different proposals to make our results more open and reproducible (Munafò et al., 2017). In NLP, we have seen initiatives to improve our reporting practices (Dodge et al., 2019) and to pre-register studies before carrying them out (van Miltenburg et al., 2021b).

Next to openness and reproducibility (Belz et al., 2021), one can also look at the incentive structures that exist in the scientific community, and that may boost some kinds of research, while discouraging other kinds of work. ‘The incentives’ constitute a broad header, which includes *citations* (what kinds of papers get cited?), *awards* (what kinds of papers get recognized through best paper awards?), *community standards* (what is seen as a valuable contribution?), and so on. Next to these, there are also restrictions such as *paper length* (how long should papers be?) which disincentivise authors to write lengthy discussions, and thus form barriers to carry out specific kinds of research. This paper looks at the structural properties of the NLP research culture that influence authors’ decisions (not to carry out error analyses).

This is not the first study looking at publication incentives in NLP. Rogers and Augenstein (2020) discuss our reviewing process and publication culture, and Van Miltenburg et al. (2021a) discuss different incentives that may en/discourage the inclusion of error analyses. Of those incentives, Gehrmann et al. (2022) identify accountability to reviewers as the main driver to improve the evaluation quality in published NLG research. This paper aims to find out to what extent these factors influence authors’ decisions.

There is also work outside NLP that studies how to make researchers show desirable behaviour. For example, Ali-Khan et al. (2017) looked into incentives to take part in open science, and Singh et al. (2014) did the same for engagement in public policy. Given the number of variables involved in academic publishing, this is a multifaceted problem with different schools of thought on peer review improvement. Waltman et al. (2022) argue that there are four different perspectives on how to improve peer review (focusing on Quality & Reproducibility, Democracy & Transparency, Equity & Inclusion, Efficiency & Incentives). These categories of schools of thought provide a useful framework for thinking about

the implications of any changes to the review process. For example, the idea to require or reward error analyses as part of the review process aligns with the Quality & Reproducibility school, but may go against the principles of the Efficiency & Transparency school, since it further burdens the reviewers (who already show signs of fatigue).

Regardless of your meta-scientific position, any proposal to improve the field should start by asking the relevant stakeholders about their experiences and ideas. We did this through a questionnaire, which is described in the next section.

3 Method

We asked NLG researchers and practitioners for their opinions about error analysis, as well as factors that affect the likelihood of including one in their work. We purposefully did not posit any hypotheses, since our aim is to describe the current perceptions of error analysis, and to sketch a path towards greater adoption of it in NLG research.

Survey Our survey opens with an information letter describing our study and its main goals, followed by an informed-consent form. Participants were allowed to skip all questions except for the informed consent. Upon their consent, participants were asked some general demographic questions, followed by questions about the following topics (see Appendix C for details):

1. Experience reading error analyses
2. Experience carrying out error analyses
3. Barriers and enabling factors to carry out error analyses
4. Necessity and usefulness of error analyses
5. Reporting practices
6. Other comments

Population of interest Our survey targets researchers and practitioners interested in NLG research. To maximize our reach, we spread our survey through Discord, Slack, Twitter, and the Corpora⁵ and SIGGEN⁶ mailing lists (SIGGEN is the Special Interest Group for ACL researchers working on Natural Language Generation). The SIGGEN community is not very large. For the 2020 SIGGEN board member elections, there were 428 eligible members (i.e., people subscribed to the SIGGEN list, after filtering out any duplicates). Of these, only 92 members cast a ballot.⁷ This puts an upper bound on

⁵<https://mailman.uib.no/listinfo/corpora>

⁶<https://www.jiscmail.ac.uk/cgi-bin/wa-jisc.exe?A0=SIGGEN>

⁷As reported through the SIGGEN mailing list, by Jose M. Alonso (SIGGEN board member at the time of writing): <https://www.jiscmail.ac.uk/cgi-bin/wa-jisc.exe?A2=SIGGEN;5f3966e0.2012>

Experience in NLG		Affiliation	
No response	13	No response	12
Less than 2 years	13	Academia	51
2 - 5 years	23	Industry	8
6 - 10 years	5	Other	1
11 or more years	13		
I don't work in NLG	5		

Table 1: Demographics for our participants.

the number of responses we might reasonably expect to receive (particularly since voting takes less effort than filling in a survey).

Participants We received 72 responses (consenting to the survey and answering at least one question). Of those who indicated their affiliation, 51 were academics, eight were from industry, and one selected “other”. Table 1 provides a general overview of the demographics. Because of the limited number of respondents per category, we did not carry out any subgroup analyses.

Analysis We performed a *quantitative* analysis of the responses to our closed questions. In addition, we performed a *qualitative* analysis of the open question responses, inspired by other qualitative approaches, such as thematic analysis (Braun and Clarke, 2006) and grounded theory (e.g., Strauss and Corbin 1994). We first read the responses for each question, to get a general sense of the answers. Then, we apply *open coding*: we organise the responses using short, descriptive labels (known as *codes*). The coding was done independently by one or two of the authors for each section. We used these codes to develop coherent themes that are reflected in the answers (*axial coding*). In turn, these themes are used to form a narrative about barriers and limitations, and enabling factors and benefits of error analyses.

The goal of obtaining a high inter-annotator agreement (or inter-coder reliability) is often criticized by qualitative researchers because it assumes the positivist idea that an objective interpretation of the data is both possible and desirable (Terry et al., 2017). If the focus on inter-annotator agreement is too strong, we may lose track of insights that cannot be captured by a strictly defined taxonomy. Instead, we can embrace researcher subjectivity in our quest to gain a deeper understanding of the perspectives of our respondents. Through discussions among ourselves, we ensure that the final narrative is both consistent with and supported by the coded responses. For a related discussion in NLP, see Basile et al. (2021) and the *Perspectivist Data Manifesto* (<https://pdai.info>), where the au-

thors argue against aggregated datasets that hide any disagreements between annotators.

Pilot and positionality We acknowledge that our own position on the subject of error analysis is not neutral: all authors are in favor of promoting it. However, since we are all researchers in NLG, we did fill in a preliminary version of the survey, along with some colleagues outside of our project, resulting in 12 complete responses. This enabled us to test the questions, determine the duration of the survey, and substantiate our own stance towards error analysis. In lieu of a pre-registration (since this is not a confirmatory study, see van Miltenburg et al. 2021b), we made sure to analyse our responses before the deployment of our survey, and committed the report to GitHub, so that it would be time-stamped. None of the authors filled in the final survey, so we can compare the final results to our own responses.

IRB approval Before carrying out our study, we obtained ethical approval from the Institutional Review Board (IRB) at the lead author’s university. See §8 for more details on our ethical considerations.

4 Results

Our results are generally organised by the topics identified above in Section 3, but there are several themes (such as the importance of resources such as time and money) that recur throughout this section.

4.1 Experience reading error analyses

Of the 49 participants that answered this part of the survey, the majority (33) recalled having read an error analysis in an NLG paper. Most respondents found reading an error analysis at least moderately useful, and no respondent found it not useful. We also asked these participants what they found useful about the error analyses they have read. Their answers will be discussed in Section 4.4.

Sixteen participants indicated that they have not previously read a published error analysis. We asked these participants whether they found it surprising they had not seen any published error analyses. Seven participants responded to this question. Of these respondents, three participants agreed with this statement. One surprised respondent reasons that NLG errors are evident to daily users of NLG systems, while another observed that without understanding errors properly “it is quite hard to correctly develop a system”, contrasting to a blind hyperparameter optimization effort for neural nets.

Participants who did not find the lack of published error analyses surprising highlighted that error analysis is time-consuming, tedious, and that the lack of standards for error analyses prevents useful comparisons even if the analysis is conducted. We also anticipated that these issues would form barriers to the broader adoption of error analyses, and will return to them in Section 4.3.

4.2 Experience running error analyses

A total of 37 respondents answered a question regarding whether they had ever carried out an error analysis, with 25 indicating they had and 12 indicating they had not. The respondents who had carried out an error analysis indicated in their free-text answers that the primary challenge and difficulty in carrying out an error analysis is resources. By this, they chiefly meant time, but the responses also mention tooling, scale, annotators and other similar factors. Error analyses were also seen as difficult to conduct, both in terms of developing a high-quality categorization scheme and in ensuring high inter-annotator agreement. The latter aspect plays into the resource cost, as iterative development is needed to ensure high inter-annotator agreement. This is further exacerbated by the lack of a standard methodology.

Experienced participants Among the 23 respondents that had carried out an error analysis, 13 participants reported having felt that there had not been enough resources or reference material for them to carry out an error analysis. At the same time, almost all of the participants (22 out of 23) that have conducted an error analysis would consider conducting another error analysis again in the future.

When asked why they were likely to carry out an error analysis in the future, the respondents generally indicate a belief in the analyses being useful. Some explicitly state that analyses allow for improved results in the future and provide insights beyond those provided by standard evaluation metrics. Some of the other respondents viewed error analyses as required, some for intrinsic reasons, with one answer being unclear with regard to whether the requirement is an intrinsic one or an extrinsic one. A few responses highlight that their ability to conduct error analyses is limited by resources or collaborator views on their necessity. One respondent viewed error analyses as unnecessary for academic publishing, but as a standard operating procedure for industry work.

Other participants For the participants that have not carried out an error analysis (12), seven have considered doing so, or plan to do so in the future, with

only four respondents reporting never having even considered conducting one. Asked for the reason why they had not carried out an error analysis, a few respondents had simply not considered conducting an error analysis. Some lacked the resources, most commonly time, to do so. Multiple respondents indicated that they were conducting, or had conducted, research into rule-based NLG, and as such had ensured their systems did not make any errors before evaluating them.

When queried whether they would be willing to carry out an error analysis, seven respondents would consider conducting an error analysis, four respondents were uncertain, and one respondent answered with ‘probably not.’ We conclude that our community could potentially publish more error analyses (after all: most are willing to do so), given the right publishing environment. This brings us to the next section.

4.3 Barriers and enabling factors

Quantitative results. Before carrying out this survey, we identified nine factors that may influence the authors’ decision (not) to carry out an error analysis. These factors were based on work by [Van Miltenburg et al. \(2021a\)](#), and our experiences as NLG researchers:

1. Page limits: if there is not enough space to present an error analysis, authors may be hesitant to include it or prioritise other aspects of their work.
2. Error taxonomy: if there is no established error taxonomy, authors may find it hard to categorize errors in the output of their system.
3. Annotation tools: if there are annotation tools dedicated to error analysis, it would make the process easier.
4. Crowdsourcing template: if there is no template, there is a higher barrier to carry out an error analysis, because the authors need to design a task by themselves.
5. Appreciation from reviewers: if reviewers do not ask for error analyses, or they do not reward them enough, authors are less tempted to carry out an error analysis.
6. Availability of annotators: if there are no annotators (other than the authors themselves), then carrying out an error analysis may be considered too much work to carry out alone.
7. Time: error analysis can be time-consuming. If researchers don’t have enough time to carry out an error analysis, they will not do it.
8. Money: if researchers do not have the money to hire annotators/crowd workers, they need to carry out the full error analysis themselves.
9. Collaborators: error analysis may be considered too much work to be carried out alone.

Figure 1 provides an indication of which factors

I would be more likely to carry out an error analysis in a conference/journal paper if...	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree
... there was a higher page limit.	3	3	9	12	4
... there would be an existing error taxonomy that I could use.	1	2	6	11	12
... there would be dedicated annotation tools for error analysis that I could use.	1	4	7	10	10
... there would be a crowdsourcing template for carrying out error analyses.	1	4	8	11	8
... reviewers paid more attention to error analyses.	0	2	6	9	15
... there were an available pool of annotators or crowd workers.	3	3	6	13	7
... I had more time.	0	4	2	10	16
... I had more money.	1	4	3	9	15
... I had more collaborators.	0	3	7	10	12

Figure 1: Heat map table showing our participants’ (dis)agreement with nine statements about factors that make them more likely to carry out an error analysis. Numbers are absolute, i.e., counts of participants (dis)agreeing. Darker cells contain higher numbers.

make it more likely for our participants to carry out an error analysis. For all nine factors, the results skew positive, with participants recognising all the identified factors act as barriers to completing error analyses. Three of these stand out: time, money, and recognition from reviewers seem to be the most important. These results are also confirmed by the qualitative results.

Qualitative results. We further surveyed participants regarding other barriers that prevent them from carrying out an error analysis and what factors would instead enable them. The participants confirmed that resources are the premier barrier: time (including the time that could be allocated for improving the NLG systems), funds, tools to help with error analyses including a taxonomy of errors, access to experts that could help with error annotation as well as lack of system outputs in literature which could be used for comparison. Similarly, Zhou et al. (2022) also found that time limitations, especially for industry teams, constrained the use of qualitative or participatory evaluation approaches. As expected, access to these resources was identified as an enabler that helps researchers focus their effort on performing error analyses.

A number of participants mentioned that the current research culture does not reward such analyses, which prevents them from performing and reporting them. In fact, most participants identified culture change towards error analysis as an important factor for adopting it. Specifically, the participants proposed making error analysis a requirement for papers and explicitly recognising it in review forms; this should highlight its importance both for research and industrial/commercial applications.

15 participants responded that they are more likely to include an error analysis in a journal article, motivated by the benefits of publishing in a journal article, such as a higher page limit, increased time to publish, and higher demands on details. However, 14 partici-

pants responded that is equally likely to include an error analysis in a journal article, as well as in a conference publication as NLG research is heavily conference-focused.

When asked if there are currently enough resources to support error analysis, the majority of respondents to this question reported that error analysis resources are still missing (20), while a few participants stated that there are some resources available (10). Participants suggested that a well-documented error analysis taxonomy and procedures and standards, as well as annotation tools, are missing. Also, funding plays an important role in performing error analysis.

4.4 Necessity & usefulness

Quantitative results. Figure 2 shows the participants’ attitude towards error analyses. The respondents overwhelmingly agree that error analyses are useful and provide insight into system performance. At the same time, we find that our participants have mixed feelings about carrying out an error analysis themselves. When asked whether they find it enjoyable or boring/tedious, there is a slight majority agreeing with both statements. Although some respondents responded positively to only one of the two statements, nine participants somewhat agreed with error analysis being both “enjoyable” and “tedious.” Based on this observation, we might say that carrying out an error analysis is like eating broccoli or Brussels sprouts; we all know it is good for you (and there certainly are long-term health benefits), but not everyone enjoys the taste, and it may be difficult to finish your plate.⁸

Should both journal and conference papers include error analyses? Developing our questionnaire, we ex-

⁸Continuing the analogy: in our experience, it is generally more enjoyable to eat (annotate) together, than having dinner alone, even if you’re not having the same meal.

	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree
There should be more error analyses in the NLG literature	0	1	1	10	19
Error analyses are a valuable part of a paper.	0	0	2	4	25
Carrying out an error analysis is enjoyable.	0	7	6	14	3
Carrying out an error analysis is boring/tedious.	3	4	6	17	0
Error analyses are necessary to fully evaluate the performance of an NLG system.	1	0	1	5	23
Knowing what errors a system makes is helpful for future research.	0	0	0	9	21
Knowing what errors a system makes is helpful for practitioners/NLG in industry.	0	0	1	5	24
If you publish at a conference, and you present an NLG system as one of your main contributions, you should include an error analysis.	0	0	5	13	12
If you publish in a journal, and you present an NLG system as one of your main contributions, you should include an error analysis.	0	0	2	10	18

Figure 2: Heat map table showing the distribution of responses to a question where participants were asked to indicate their (dis)agreement with nine statements about the desirability/usefulness of error analyses. Numbers are absolute, i.e., counts of participants (dis)agreeing. Darker cells contain higher numbers.

pected that there would be a difference in standards between journals and conferences; journal papers might be seen as definitive products of research, while conference papers are still work-in-progress. The preliminary nature of conference papers might make our participants more lenient. Surprisingly, the majority of our participants agreed for both journal and conference papers that they should include an error analysis (if applicable). Admittedly, the agreement is less strong for conference papers than for journal papers, but these results do show that error analysis is important to readers of NLG papers.

Qualitative results. We asked the participants who have read error analyses in the past about the usefulness of those error analyses. By far the most common answer was that error analysis could help identify remaining challenges and direct future work, both at a high level, and in terms of improving individual systems. Several responses also mentioned researchers' bias and noted that a thorough error analysis is better than cherry-picked examples more commonly seen in a qualitative analysis section. Some respondents indicated that error analysis was a good complement to imperfect metrics, and could detect overlooked errors. The usage of error analysis to gauge whether a system was suitable for its purpose was also mentioned, along with gaining a better understanding of system limitations.

We received 27 responses in total to our question on what kinds of papers error analyses may be useful for. Most replies (16) mentioned experimental papers or papers presenting a new system. Five more respondents even implied that *all* papers should include error analysis; this probably still applies mostly to experimental

papers as they are the most common type. Nine respondents mentioned various specific sub-fields or system types (e.g. end-to-end systems, dialogue systems). Three participants mentioned evaluation-related papers specifically. We also received multiple general remarks arguing in favour of error analysis and/or complaining about the lack thereof in current works.

4.5 Reporting practices

What should be included in reports containing an error analysis? Common themes underlying the responses were reporting practices that could enable replicability, reliability, and usefulness of both methodology and results. Table 2 provides an overview of the responses that were given in our pilot study, the main survey, or both.

Of the 16 respondents who answered this question, seven focused on reporting descriptive details such as the annotator training process, annotation process, and actual annotator details, expressing that this would better enable replicability of results as well as enable comparisons across studies via replicable methodology. This also includes reporting details that ensure the reliability of the methodology and results, such as reporting inter-annotator agreement and evidence of annotator quality or sampling method (specifically arguing for statistically-driven sampling).^{9,10} At the same time, one participant warned against over-formalising error analyses.

Seven respondents explicitly argued for reporting

⁹In a recent publication, Shimorina and Belz (2022) provide a useful template for reporting these details.

¹⁰Also see Popović and Belz 2022 for a discussion of reporting scores and agreement for error annotation tasks.

Source	Recommendation
B	Provide the annotation guidelines, with an explanation of how these were created (e.g. as an appendix).
R	Provide details on how annotators were trained.
B	Provide details about the background of the annotators.
R	Provide inter-annotator agreement scores, to assess the reliability of the annotation process.
R	If using an existing error taxonomy, ensure it is appropriate for your system.
R	If possible, provide a comparison between different systems.
A	If comparing different systems, use appropriate statistics (e.g. Chi-square tests comparing the distribution of particular kinds of errors).
R	Provide a reflection on the potential sources of the errors.
R	Provide correlation scores between different types of errors to see which ones co-occur.
B	Provide details on how the outputs were sampled (e.g. stratified sampling).
A	Provide actual examples of system output.

Table 2: List of reporting practices suggested in the responses to our questionnaire by either the current authors (A), our respondents (R), or both (B).

practices related to error taxonomies and compared systems. The goal here is to increase the usefulness of the analyses for both aiding researchers and understanding systems: reporting of (potentially customized) error categories with definitions, justifications, and limitations to enable use in other works, and explicitly reporting system comparisons and observations (such as identifying commonalities across systems and the system impacts or correlations of errors).

Two participants also left suggestions in the ‘other comments’ field. One noted that “Error analysis should focus on language features, text genre characteristics and adequacy to the task, not a mere statistical analysis.” The other participant highlighted the importance of sentence structure and the manual labour that goes into an error analysis. We may interpret this in light of the fact that humans can pick up nuances that (thus far) NLP systems have not been able to detect.

5 Discussion

5.1 Incentives and social dynamics

As noted in Section 4.3, most participants thought a culture change is necessary to make error analysis a common practice. One promising idea in this direction seems to be to explicitly reward researchers with badges for exemplary behaviour, such as preregistering confirmatory studies and publishing research code and data.¹¹ This idea has been proven to work in psychology (Kidwell et al., 2016), where open science practices increased among published papers after the introduction of badges displayed alongside each paper.¹² Building on the badges from the ACM (2020), NAACL 2022 also offered reproducibility badges.¹³ Over 25% of accepted submissions earned at least one badge. Relat-

edly, as program chairs of COLING 2018, Derczynski and Bender (2021) introduced awards for specific parts of papers (best evaluation, most reproducible, best challenge, best error analysis) instead of having an overall best paper award. On top of that, only papers with published code and data were eligible for any of these awards. Following this initiative, the conference saw about one-third of all papers with full code. One other innovation from Derczynski and Bender (2021) was to introduce *paper types*: categories of papers with associated review forms that are tailored to the kind of contribution that authors want to make.¹⁴ These review forms are public, so authors can prepare their work accordingly. Having specific review forms may nudge authors to include different kinds of information in their submissions, which they perhaps would not have included otherwise.¹⁵

It is still hard to gauge the impact of these initiatives on the NLP community, but at least open science badges help make our community norms and values explicit. However, following Yarkoni (2018), we have to acknowledge that scholarly behavior is also just a matter of personal responsibility. If you believe that it is important to highlight the limitations of your approach, then the time and effort needed to carry out an error analysis should be included in the planning of your project.

The carrot and the stick Incentives generally come in two forms: the carrot and the stick. The initiatives discussed above are an example of the former, rewarding authors for good behavior. What about the latter? Can we require authors to carry out an error analysis, *or else...*? This is not without precedent. NLP conferences have recently started requiring the inclusion of *Limita-*

¹⁴NEJLT also uses the same paper types. See: <https://www.nejlt.org/authorinfo/>

¹⁵We are not aware of any studies that look into the effects of reviewing forms on the form or content of the submitted work. Future research could study e.g. the content of NLP papers before and after introducing (new criteria on) checklists for conference submissions.

¹¹See: <https://www.cos.io/initiatives/badges>

¹²Though see Crüwell et al. 2023 for a critical evaluation of the *open data badge* policy in the *Psychological Science* journal.

¹³See: <https://naacl2022-reproducibility-track.github.io>

tions and *Ethical Considerations* sections for all papers where such sections are appropriate (i.e., most NLP papers). Moreover, one might argue that an evaluation of an NLG system is not complete without an error analysis, especially given the unreliable nature of automatic metrics and the reductive nature of summary scores. It is simply good scholarship to provide an error analysis.

When should error analyses be required? Almost all of our respondents agreed that journal submissions should include an error analysis, and the majority of our respondents also agreed that the same should hold for conference papers. In hindsight, it is probably not the *venue* that counts, but the *state of completion* of the project. If you report on a finished project, then the final publication is the end product, regardless of the venue. At this point, the project should be fully documented, including an overview of all the limitations of the end product. This prevents *technical debt* (Sculley et al., 2015) from building up in the NLG community.¹⁶

Based on our observations, we would like to posit the following rule: *if* a paper presents a final result (as opposed to work-in-progress), *and* the paper presents both an automatic and a human evaluation, *then* the paper should also contain an error analysis.

Getting there A priori, the carrot is preferable to the stick. Without any hard requirements, there is more room for exceptions, i.e. papers that do not fit the traditional mould of NLG publications. Furthermore, encouragement policies are less likely to run into resistance from the community, compared to hard requirements. We do not necessarily need *everyone* to provide an error analysis; if we can encourage a critical mass of researchers to provide error analyses, then this will just grow to become the norm.

5.2 Making space for error analyses

Although page limits do not seem to be the main barrier for carrying out error analyses, it is also clear that additional content takes up space. We have recently seen this with limitations and ethical considerations sections, which for many conferences are now allowed to be put on an additional page following the conclusion (even though ethical considerations are an integral part of research design). EMNLP also features a reproducibility checklist, the authors of which suggest that researchers may want to provide important technical details in the appendix.¹⁷ From these observations, it seems that our community is struggling to put all relevant information in the four-to-eight pages that are

¹⁶Epstein et al. (2018) make a similar point, but using a different framing than Sculley et al. (2015). They talk about the *AI knowledge gap*, where studies on new systems are published faster than studies characterizing the behaviour of those new systems.

¹⁷See: <https://2020.emnlp.org/blog/2020-05-20-reproducibility>

currently allotted to conference papers. *The medium is the message* (McLuhan, 1964); if conference papers remain the main publication venue for NLP research, then it is important that our values are reflected in the submission types. All relevant information should fit in the main body of the paper. We discuss two options to improve the situation.

Option 1: increase paper length The first option is to simply increase paper length (e.g. moving from 4/8 pages for short/long papers to 5/10 pages), or to add another length tier (resulting in papers of either 4, 8, or 12 pages).¹⁸ This creates additional space to include relevant information, without introducing any new requirements. Over time, we should see the community converge on the type and amount of content that is required for papers in each tier to be publishable. The main attraction of this proposal is its simplicity, requiring little to no extra administration. The downside of this proposal is that it is unconstrained, so without any additional requirements it is not clear whether authors would actually carry out more error analyses.

Option 2: reserve space for error analyses Continuing the previous section (§5.1), the *ACL main conferences in NLP have not just required authors to include *limitations* and *ethical considerations* sections; they have also given authors additional space to provide these sections. Typically this space is provided *after* the conclusion, to ensure that authors do not cheat the page limit by using the additional space for other purposes. One way to stimulate error reporting would be to do the same for error analyses as well. On the one hand, this initiative adds more administrative burden, and it prevents authors from integrating the relevant content into the narrative of the paper (at least at submission time), but it does guarantee that authors actually include an error analysis, and it helps to normalise the idea that every paper should have sections detailing limitations, ethical considerations, and error analyses.

5.3 Error taxonomies & standardization

Recent work in the NLG community has aimed to provide an overview of our evaluation practices, and move towards standardising our terminology and assessment materials (Belz et al., 2020; Howcroft et al., 2020). There have been similar efforts in the areas of Explainable AI (Nauta et al., 2022) and Intelligent Virtual Agents (Fitrianie et al., 2019, 2020). The majority of our respondents indicated that they would be more likely to carry out an error analysis if there were an existing taxonomy of

¹⁸Of course there are many other possibilities, including the option to let go of page limits altogether, or to only set an upper bound for conference submissions (based on the reviewing timeline).

errors that they could use. However, is it even possible to establish a standardised error taxonomy for NLG output? As one participant noted: it is “better to use a sensible characterization of errors that actually occur [...] than trying to shoehorn them into an existing taxonomy.”

Several taxonomies have been proposed for different NLG/NLP tasks and some are used for evaluation by annotation, an approach that readily lends itself to error analysis. For machine translation, [Popović \(2020\)](#) asked annotators in separate experiments to mark comprehensibility and adequacy errors, also distinguishing major errors (those which alter the meaning) from minor errors (grammar or style). [Freitag et al. \(2021\)](#) asked annotators to mark up to five of the most severe errors within a segment, these were then assigned both a category and a severity. [Costa et al. \(2015\)](#) proposed a linguistically motivated and hierarchical taxonomy, and [He et al. \(2021\)](#) proposed a taxonomy and then used it to create the TGEA annotated dataset. For factual accuracy in data-to-text generation, [Thomson and Reiter \(2020\)](#) asked annotators to mark non-overlapping spans of text and assign them one of six categories. For prompted generation, [Dou et al. \(2022\)](#) asked annotators to mark all errors from a wide range of categories,¹⁹ allowing multiple overlapping annotations and with some subjectivity between categories (*Encyclopedic* for one person could be *Needs Google* for another). These taxonomies could be used as-is, or they can be developed further to provide a more detailed analysis.²⁰

NLG is difficult to define as a field ([Gatt and Kraemer, 2018](#)) and despite sharing some commonality (the generation of text), the purpose of any generated text is key to how we interact with it ([Evans et al., 2002](#)). This makes it difficult to form a “one size fits all” definition of NLG and, similarly, an error taxonomy. However, there are some high-level considerations when selecting or adapting a taxonomy:

Evaluation criterion: Humans are known to miss some errors when reading ([Huang and Staub, 2021](#)), and whether their annotations for one criterion might affect their subsequent reading and annotation of the remaining text is unknown. Asking annotators to consider multiple criteria simultaneously could compound this problem, increasing both disagreement and the volume of missed errors. In line with more general best practices for NLG evaluation ([van der Lee et al., 2021](#)), annotators should consider one criterion at a time.

¹⁹Grammar and Usage, Off-Prompt, Redundant, Self-Contradiction, Incoherent, Bad math, Encyclopedic, Commonsense, Needs Google, Technical Jargon.

²⁰For more examples, [Huidrom and Belz \(2022\)](#) provide a further survey of existing error taxonomies, which they plan to use to develop a taxonomy of semantic errors in NLG output.

Annotator agreement: Very low inter-annotator agreement might be indicative of an annotation procedure issue, but disagreement between annotators does not necessarily mean that some of the annotations must be flawed ([Popović, 2021](#)). [Thomson and Reiter \(2021\)](#) noted that even within a single criterion, two annotators could provide sets of errors that only partially overlap, yet can both be considered valid representations of the same complex underlying problem. In addition to calculating agreement, annotators could check each other’s annotations and indicate whether they consider them one valid way of describing the underlying problems [Thomson et al. \(2023\)](#).

Distinct categories: Principles from both taxonomy and close-response survey design are also relevant to annotation; categories should be mutually exclusive and as exhaustive as is practical ([Fowler and Cosenza, 2008](#)). If there are too many categories (making it hard for annotators to keep all distinctions in mind), it may be beneficial to use more coarse-grained taxonomy.

Error instance vs cause: Hallucination is commonly considered a core error type in NLG but [Van Miltenburg et al. \(2021a\)](#) argue that errors should not be defined in the first instance by the process that caused them. An error in generated text can be defined in terms of how it fails to meet its purpose, a grammatical error, factual mistake, etc. The reason for this failure can then (optionally) be determined. Process errors should be recorded separately from text errors, i.e., we could mark an error as being an incorrect named entity, then indicated that this was caused by hallucination. Different types of hallucination, such as intrinsic versus extrinsic ([Ji et al., 2022](#)), can be considered at this second stage.

Error severity: Different errors may have a different impact on readers ([van Miltenburg et al., 2020b](#)). Similarly to error causes, severity can be assessed after the error is identified and categorised ([Popović, 2020](#); [Freitag et al., 2021](#)), although this may be done immediately as part of recording the error. In such cases, annotators are following a sequential procedure where they first find the error span and assign a category, then consider how severe the error is.

Although there are still many (context-dependent) decisions for authors to make about the design of a suitable error analysis, these considerations do constrain the space of possible approaches. Moreover, it should be possible for researchers to agree on a standard error analysis taxonomy and format for common NLG tasks. These could be decided upon during the development of new tasks, or with new iterations of existing shared

tasks, e.g. WebNLG (Castro Ferreira et al., 2020) or the surface realization shared task (Mille et al., 2020).

Another useful step may be the development of guidelines for what the output should look like. This is mostly a problem for neural data-driven NLG systems, which are commonly trained and evaluated on crowd-sourced data, where annotators are asked to write an output text for a given input. If the guidelines for writing those texts are underspecified, then there will (1) be a high degree of variation in the human-authored texts (see, e.g., van Miltenburg et al. 2017),²¹ and (2) the decision of what the output should look like is essentially delegated to the crowd, meaning that the standard for comparison is only extensionally defined by the training corpus (van Miltenburg et al., 2020a; Schlangen, 2021). Without any clearly defined standards, it is more difficult to judge the quality of automatically generated output. With standards in place, it is also possible to define deviations from the norm, which we can then more easily flag as errors.

Finally, any taxonomy is better than no taxonomy at all. If there is no existing set of error categories, then we encourage authors to develop a taxonomy of their own. Once established, error taxonomies can have a big impact on future work in two ways:

1. They facilitate future error analyses and make it easier to compare different systems,
2. They may steer future research by highlighting specific issues in system output that should be resolved.

5.4 Resources: time, money, and tools

Time and money were considered by our respondents to be the main barriers to carrying out error analyses. These two factors are also clearly correlated: time-consuming tasks can be outsourced by paying someone else to do them, and vice-versa. You can save money by doing everything yourself. So what if you have neither time nor money to spend on error analysis?

Using student annotators. The go-to option for cheap annotation in academia is to have students carry out the work. We do not think it is ethical to have students annotate large amounts of data for free, but at least small batches of error analysis could be incorporated in education. We suggest the following guidelines for ethical data collection:

²¹This variation is not necessarily bad (users may sometimes appreciate diversity), but it has been shown for use cases such as professional weather forecasting that users appreciate consistency in the output (Sripada et al., 2004). Either way, we do need to ensure that the texts are congruent with the purpose of the task. If the purpose is not made clear to the crowd-workers, the human-authored texts may be sub-optimal with regard to the communicative situation that the NLG system is embedded in.

1. The exercise should support the end-goals of the course.
2. The amount of items to annotate should not be excessive. Once the learning goals have been achieved, it is not necessary to continue to exercise.
3. The data should be anonymised such that it is not possible to identify which student contributed the annotations.
4. Students should have the opportunity to opt-out of their data being used for research purposes (without this having any negative effect on their grades). Or even better: use an opt-in procedure where students may (anonymously) submit their results.
5. As a corollary of the previous points: grades should not be contingent on data quality.
6. Researchers should check with their colleagues or their institutional review board (IRB) whether this form of data collection is appropriate, given the power differential between teachers and students.

In short: ‘free’ annotation should not come at the cost of students’ well-being. It requires dedication, and an up-front investment to responsibly integrate the exercise in an educational context.

(Lack of) time is an illusion. Many researchers have internalized the corporate values of *speed* and *efficiency*, prioritizing them over the slow contemplation that has traditionally been the hallmark of academia (Berg and Seeber, 2018). As a result, it often *feels* like we are just living from deadline to deadline, without any time to sit down and thoroughly analyze our results. But this is a *choice*; there are other options! In his (2018) COLING keynote, Min-Yen Kan promoted the idea of ‘slow research’ in NLP, as a counterpart to the fast-paced style of research that has grown popular in recent years. We would argue that a publication with a slow, deliberate error analysis may over time be more impactful than a paper lacking such in-depth information. (One might respond that slower research risks being scooped, but this overlooks the fact that error analyses and other time-consuming methods are substantial contributions in and of themselves.)

Of course, fast-paced research is there for a reason; many researchers believe they are expected to live up to the aphorism that they should *publish or perish*. Not publishing enough papers may reduce your chances of success in academia.²² But, again following Yarkoni (2018), we shouldn’t sacrifice good scholarship based on these incentives. At this point we should

²²And as Rahal et al. (2023) note: “Quality research needs good working conditions.” With more permanent positions, researchers may find themselves better able to focus on long-term research goals.

ask ourselves: how long does an error analysis *really* take? Granted, an extensive error analysis can be quite labour-intensive, but we should not let perfect be the enemy of good. Including a systematic error analysis of any kind is already much better than randomly picking some cherries and lemons to include in the appendix.

Just do it yourself. As with any annotation task, it is important to at least carry out some portion of the analysis yourself. There is no replacement for getting familiar with the output of your system, or with the process of identifying potential errors. This *dogfooding*²³ ensures that the task is feasible, and decreases the odds of overlooking important properties of the generated data. Although the majority of our participants found error analyses to be boring/tedious, there are clear benefits to this method, and an equal majority found the process to be enjoyable as well. As [Sambasivan et al. \(2021\)](#) note, data work is considered to be much less glamorous than modeling, but it is essential that we do it anyway.

Trade-offs are inevitable. Some NLG tasks are more time-consuming to evaluate than others. For example, manually assessing the quality of longer texts (e.g. summaries, stories, or news articles) takes longer than the assessment of shorter texts (e.g. image captions, product descriptions). In a multilingual setting, evaluation is also going to be more involved: one may want to have a universal set of error categories that work across different languages, or a large enough sample size for outputs in each language under consideration. Given time and money constraints, it may not be feasible to carry out a large-scale error analysis. As noted above: any error analysis is better than none, but the authors also need to be clear about their considerations and the limitations of their analysis. Example trade-offs include:

1. Coverage versus specificity: Carry out an in-depth analysis of a specific subset of the outputs, or a more superficial analysis of all the outputs?
2. Coverage versus reliability: Annotate more outputs with fewer annotators per output, or fewer outputs with more annotators?

There is no one-size-fits-all recommendation with regard to the trade-offs that authors should make. This process is guided by the research question, hypotheses, and the claims that the authors would like to make about their system. The strength of the error analysis influences the extent to which any claims about system performance can be substantiated.

²³For lack of a better term, although *dogfooding* is typically used to refer to developers using their own software rather than just inspecting the results. See: https://en.wikipedia.org/wiki/Eating_your_own_dog_food

Optimisation and tools It may be possible to develop tools to carry out error analyses more efficiently. For example, after developing a dedicated app or mobile website, error analyses could be carried out *on the go* in brief sessions (e.g., waiting for the bus, or on the train). This is an interesting avenue for future research, although following Section 5.3 one might wonder whether it is feasible to develop universal tools for supporting error analysis, given the challenges of standardisation.

5.5 Collaboration

The majority of our participants indicated that they would be more likely to carry out an error analysis if they had more collaborators. How can we address this issue?

Shared tasks One proposal is to copy successful evaluation practices from other subfields of NLP. The Workshop on Machine Translation (WMT) asks all of its participants to rate a collection of translations “proportional to the number of tasks they entered” ([Barrault et al., 2020](#)).²⁴ This approach has been proposed in the NLG community as well, for the GEM shared task ([Gehrmann et al., 2021](#), p. 109). Next to providing ratings, participants of shared tasks could also conduct error analyses. Once the outputs of all systems are submitted, the participants could analyse a subset of the outputs of all systems using an agreed-upon error taxonomy and annotation methodology. This has at least three distinct advantages: (1) Authors would be intimately familiar with the different kinds of mistakes that systems could potentially make, (2) system labels would be hidden so that participants are not biased in their judgments, (3) each shared task would produce richly annotated datasets (potentially further enriched with human and automatic evaluation scores).

Sharing resources Researchers in Psychology have proposed StudySwap ([Chartier et al., 2018](#)): a dedicated platform to share resources, such as equipment, participants, expertise, and so on.²⁵ The NLG/NLP community lacks such a platform. Of course, researchers may informally help each other out, but this is always easier for established researchers with a bigger network. It is tempting to suggest a centralised platform for collaborative NLP/NLG research, but this may not be feasible to sustain.

²⁴These judgments are further complemented by those from crowd-workers, and a dedicated pool of linguists.

²⁵Unfortunately the platform is currently dormant, but it has resulted in fruitful collaborations in the past.

6 Limitations of this study

Because our participants are volunteers, we run the risk of possible self-selection bias: only people that are interested in error analysis may have taken the time to respond to our survey. This means that our survey may overestimate the support for error analysis in our community. This issue is inherent to any voluntary survey. (For example, [Jakobsen and Rogers \(2022\)](#) report this limitation as well). Given this limitation, we are still able to make existential claims about the barriers that exist for researchers wanting to carry out and publish error analyses; at least some researchers are held back by the barriers listed above.

Another limitation is that our sample size is relatively small, with 72 participants. As we discussed above, this is not very surprising, given the limited size of the NLG community. Our participants were also allowed to skip as many questions as they liked in our survey. As a result, several questions were answered by less than half of our 72 participants. This may be seen as a limitation of our study, because a small group of researchers may not be representative of the larger research community. But our study does serve its original purpose: to consult other researchers about potential barriers and enabling factors for the use of error analysis in NLG, and to ensure that our list of barriers and enabling factors does not have any glaring omissions.

Two participants indicated that they were not familiar with the concept of error analysis before this study. One of them also noted that, because of this, they would have liked to see an “I don’t know” option for the Likert scale questions (although it was possible to leave these questions blank).

7 Conclusion & Future Work

We have carried out a survey among NLG researchers and practitioners. Our respondents were generally positive about error analysis, but they did see multiple barriers to the general adoption of this approach. By removing or minimizing these barriers (as discussed in Section 4.3) and motivating authors to include error analyses in their work (section 5.1), we may see greater adoption of error analysis in the future.

In the future, we would like to focus on developing tools and resources, such as error taxonomies, annotation tools, and clear guidelines that would help to encourage more routine and robust error analyses. In addition to development of resources, there also needs to be a structural change in the incentives around research publication that encourages prospective authors to conduct such analyses. More work is still needed to help enable error analyses by researchers and practitioners, but we are optimistic about the future of eval-

uation within NLG.

8 Ethical considerations

8.1 Positionality and transparency

We are aware that our position as authors is not neutral: we are all proponents of error analysis, and many of us have enough job stability to not have to worry about publishing as much. This gives us the time and space needed to publish longer studies, potentially with detailed error analyses. We have attempted to explicitly capture our opinions about error analysis before distributing our survey. This information is also available through our GitHub repository, both in raw form as well as in a short report.

8.2 IRB approval

Before carrying out our study, we obtained IRB approval from the lead author’s university. This process separately considers the treatment of our participants, and the treatment of our research data. Our considerations for the IRB are detailed below.

8.2.1 Participants

Invitations: We sent out the invitation to take part in our study through social media and two mailing lists (SIGGEN and Corpora). These mailing lists are explicitly set up for the purpose of sending each other news (e.g. about upcoming conferences) and questions. People voluntarily subscribe to these mailing lists, and the invitation for our study falls within the expected use of those lists.

Information letter and informed consent: Our study starts with an information letter, describing the goal of the study, the expected duration, and potential risks/benefits of the study. The letter provides the names of the researchers involved, as well as an email address to contact for more information. The information letter is followed by a separate informed consent form, which specifies explicitly what participants agree to, when they take part in our study. They are also reminded of their rights: participation is fully anonymous, and participants are always free to quit the survey or withdraw their consent at any time, without any negative consequences.

Demographics and survey length: We aimed to minimize the amount of data collected about each participant. We only collected their general affiliation (Academia, Industry, Other) and their amount of experience (expressed in broad ranges, so as not to make people identifiable by the exact number of years). The rest of the survey has been streamlined to reduce the

burden as much as possible, and should be doable in about 10-15 minutes.

8.2.2 Data

IP-addresses: By default, our survey platform (Qualtrics) is set to store the IP addresses of all participants. Because this may be identifying information, we turned this setting off.

Data management: Because the data is fully anonymous, and participants have consented to the publication of the data, we are free to publish the responses to our survey. Before doing so, we checked the responses to the open questions for any identifying information that may need to be removed to protect the identity of our participants. All code and data have been shared through GitHub, and submitted along with this paper, thus providing maximal transparency.

8.3 Intended use of our results

Our proposals should be seen as part of the broader and ongoing discussions on publication and peer review in NLP (Rogers and Augenstein, 2020), and the state and quality of evaluations in NLG (Howcroft et al., 2020). As such, our proposals are not final, but are meant to be discussed further.

Although our policy proposals are grounded in the responses from the general NLG community, we do not know whether they are broadly supported by the community. Workshop and conference chairs may experiment with minor changes, but bigger changes may need to be put to a vote.

Acknowledgements

We would like to thank Emma Manning, who helped prepare this study and provided feedback on the survey questions. We would also like to thank members of SIGGEN, Corpora List, and everyone else who responded to our survey. We also appreciate the suggestions made by the anonymous reviewers of this paper (both for NEJLT, and for an earlier version we submitted to EMNLP). This project was supported by multiple different grants. Ondřej Dušek's work was funded by the European Union (ERC, Grant agreement No. 101039303 NG-NLG). Leo Leppänen's work was supported by the European Union's Horizon 2020 research and innovation program under grant agreement No. 825153, project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media). Dimitra Gkatzia's work is supported under the EPSRC projects NLG for low-resource domains (EP/T024917/1) and CiViL (EP/T014598/1). Craig Thomson's work was supported under an EPSRC NPIF stu-

dentship grant (EP/R512412/1) and the ReproHum EP-SRC grant (EP/V05645X/1).

References

- ACM, Association for Computing Machinery. 2020. Artifact Review and Badging, Version 1.1. Online policy document, retrieved June 2022.
- Ali-Khan, Sarah E, Liam W Harris, and E Richard Gold. 2017. Point of View: Motivating Participation in Open Science by Examining Researcher Incentives. *Elife*, 6:e29319.
- Barraut, Loïc, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.
- Basile, Valerio, Federico Cabitza, Andrea Campagner, and Michael Fell. 2021. Toward a Perspectivist Turn in Ground Truthing for Predictive Computing. In *Conference of the Italian Chapter of the Association for Intelligent Systems*.
- Belz, Anya, Shubham Agarwal, Anastasia Shimorina, and Ehud Reiter. 2021. A systematic review of reproducibility research in natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 381–393, Online. Association for Computational Linguistics.
- Belz, Anya, Simon Mille, and David M. Howcroft. 2020. Disentangling the properties of human evaluation methods: A classification system to support comparability, meta-evaluation and reproducibility testing. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 183–194, Dublin, Ireland. Association for Computational Linguistics.
- Berg, Maggie and Barbara K. Seeber. 2018. *Challenging the Culture of Speed in the Academy*. University of Toronto Press, Toronto.
- Braun, Virginia and Victoria Clarke. 2006. Using Thematic Analysis in Psychology. *Qualitative Research in Psychology*, 3(2):77–101.
- Castro Ferreira, Thiago, Claire Gardent, Nikolai Ilinykh, Chris van der Lee, Simon Mille, Diego Moussallem,

- and Anastasia Shimorina. 2020. The 2020 bilingual, bi-directional WebNLG+ shared task: Overview and evaluation results (WebNLG+ 2020). In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 55–76, Dublin, Ireland (Virtual). Association for Computational Linguistics.
- Celikyilmaz, Asli, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of Text Generation: A Survey. *CoRR*, abs/2006.14799.
- Chartier, Christopher R., Amy Riegelman, and Randy J. McCarthy. 2018. Studyswap: A platform for inter-lab replication, collaboration, and resource exchange. *Advances in Methods and Practices in Psychological Science*, 1(4):574–579.
- Clinciu, Miruna-Adriana, Arash Eshghi, and Helen Hastie. 2021. A study of automatic metrics for the evaluation of natural language explanations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2376–2387, Online. Association for Computational Linguistics.
- Costa, Ângela, Wang Ling, Tiago Luís, Rui Correia, and Luísa Coheur. 2015. A linguistically motivated taxonomy for Machine Translation error analysis. *Mach. Transl.*, 29(2):127–161.
- Crüwell, Sophia, Deborah Apthorp, Bradley J. Baker, Lincoln Colling, Malte Elson, Sandra J. Geiger, Sebastian Lobentanzer, Jean Monéger, Alex Patterson, D. Samuel Schwarzkopf, Mirela Zaneva, and Nicholas J. L. Brown. 2023. What’s in a badge? a computational reproducibility investigation of the open data badge policy in one issue of psychological science. *Psychological Science*. PMID: 36730433 (First published online February 2, 2023; issue information not known yet).
- Dale, Robert. 2020. Natural language generation: The commercial state of the art in 2020. *Natural Language Engineering*, 26(4):481–487.
- Derczynski, Leon and Emily M. Bender. 2021. Towards Better Interdisciplinary Science: Learnings From COLING 2018. Technical report, IT University of Copenhagen. TR-2021-208. Available through: <https://en.itu.dk/Research/Technical-Reports/Technical-Reports-Archive/2021/TR-2021-208>.
- Dodge, Jesse, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. 2019. Show your work: Improved reporting of experimental results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2185–2194, Hong Kong, China. Association for Computational Linguistics.
- Dou, Yao, Maxwell Forbes, Rik Koncel-Kedziorski, Noah A. Smith, and Yejin Choi. 2022. Is GPT-3 Text Indistinguishable from Human Text? Scarecrow: A Framework for Scrutinizing Machine Text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, *ACL 2022*, pages 7250–7274, Dublin, Ireland.
- Dušek, Ondřej, Jekaterina Novikova, and Verena Rieser. 2020. Evaluating the state-of-the-art of End-to-End Natural Language Generation: The E2E NLG challenge. *Comput. Speech Lang.*, 59:123–156.
- Epstein, Ziv, Blakeley H. Payne, Judy Hanwen Shen, Abhimanyu Dubey, Bjarke Felbo, Matthew Groh, Nick Obradovich, Manuel Cebrián, and Iyad Rahwan. 2018. Closing the AI knowledge gap. *CoRR*, abs/1803.07233.
- Evans, Roger, Paul Piwek, and Lynne Cahill. 2002. What is NLG? In *Proceedings of the International Natural Language Generation Conference*, pages 144–151, Harriman, New York, USA. Association for Computational Linguistics.
- Fitrianie, Siska, Merijn Bruijnes, Deborah Richards, Amal Abdulrahman, and Willem-Paul Brinkman. 2019. What are We Measuring Anyway?: - A Literature Survey of Questionnaires Used in Studies Reported in the Intelligent Virtual Agent Conferences. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents, IVA 2019*, pages 159–161, Paris, France.
- Fitrianie, Siska, Merijn Bruijnes, Deborah Richards, Andrea Bönsch, and Willem-Paul Brinkman. 2020. The 19 Unifying Questionnaire Constructs of Artificial Social Agents: An IVA Community Analysis. In *IVA ’20: ACM International Conference on Intelligent Virtual Agents*, pages 21:1–21:8, Virtual Event, Scotland, UK.
- Fowler, Floyd J. and Carol Cosenza. 2008. Writing effective questions. In *International Handbook of Survey Methodology*, pages 136–160. Lawrence Erlbaum Associates, New York, NY, US.
- Freitag, Markus, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

- Gatt, Albert and Emiel Kraemer. 2018. Survey of the State of the Art in Natural Language Generation: Core tasks, applications and evaluation. *J. Artif. Intell. Res.*, 61:65–170.
- Gehrmann, Sebastian, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Chinenye Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Mihir Kale, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi Narayan, Vitaly Nikolaev, Andre Niyongabo Rubungo, Salomey Osei, Ankur Parikh, Laura Perez-Beltrachini, Niranjan Ramesh Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, João Sedoc, Thibault Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo, Hendrik Strobelt, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, and Jiawei Zhou. 2021. The GEM benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 96–120, Online. Association for Computational Linguistics.
- Gehrmann, Sebastian, Elizabeth Clark, and Thibault Sellam. 2022. Repairing the Cracked Foundation: A Survey of Obstacles in Evaluation Practices for Generated Text. *CoRR*, abs/2202.06935.
- He, Jie, Bo Peng, Yi Liao, Qun Liu, and Deyi Xiong. 2021. TGEA: An error-annotated dataset and benchmark tasks for TextGeneration from pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6012–6025, Online. Association for Computational Linguistics.
- Howcroft, David M., Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182, Dublin, Ireland. Association for Computational Linguistics.
- Huang, Kuan-Jung and Adrian Staub. 2021. Why do readers fail to notice word transpositions, omissions, and repetitions? A review of recent evidence and theory. *Lang. Linguistics Compass*, 15(7).
- Huidrom, Rudali and Anya Belz. 2022. A survey of recent error annotation schemes for automatically generated text. In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 383–398, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Jakobsen, Terne Sasha Thorn and Anna Rogers. 2022. What Factors Should Paper-Reviewer Assignments Rely On? Community Perspectives on Issues and Ideals in Conference Peer-Review. *CoRR*, abs/2205.01005.
- Ji, Ziwei, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2022. Survey of Hallucination in Natural Language Generation. *CoRR*, abs/2202.03629.
- Kan, Min-Yen. 2018. "research fast and slow". Keynote presented at COLING 2018, Santa Fe, NM, USA. Slides available through <http://bit.ly/kan-coling18>.
- Kidwell, Mallory C., Ljiljana B. Lazarević, Erica Baranski, Tom E. Hardwicke, Sarah Piechowski, Lina-Sophia Falkenberg, Curtis Kennett, Agnieszka Slowik, Carina Sonnleitner, Chelsey Hess-Holden, Timothy M. Errington, Susann Fiedler, and Brian A. Nosek. 2016. Badges to Acknowledge Open Practices: A Simple, Low-Cost, Effective Method for Increasing Transparency. *PLOS Biology*, 14(5):1–15.
- Kiela, Douwe, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online. Association for Computational Linguistics.
- Läubli, Samuel, Sheila Castilho, Graham Neubig, Rico Sennrich, Qinlan Shen, and Antonio Toral. 2020. A Set of Recommendations for Assessing Human-Machine Parity in Language Translation. *J. Artif. Intell. Res.*, 67:653–672.

- van der Lee, Chris, Albert Gatt, Emiel van Miltenburg, and Emiel Kraemer. 2021. Human evaluation of automatically generated text: Current trends and best practice guidelines. *Comput. Speech Lang.*, 67:101151.
- McLuhan, Marshall. 1964. *Understanding Media: The Extensions of Man*. McGraw-Hill. ISBN 81-14-67535-7.
- Mille, Simon, Anya Belz, Bernd Bohnet, Thiago Castro Ferreira, Yvette Graham, and Leo Wanner. 2020. The third multilingual surface realisation shared task (SR'20): Overview and evaluation results. In *Proceedings of the Third Workshop on Multilingual Surface Realisation*, pages 1–20, Barcelona, Spain (Online). Association for Computational Linguistics.
- Mille, Simon, Kaustubh D. Dhole, Saad Mahamood, Laura Perez-Beltrachini, Varun Gangal, Mihir Kale, Emiel van Miltenburg, and Sebastian Gehrmann. 2021. Automatic Construction of Evaluation Suites for Natural Language Generation Datasets. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021*.
- van Miltenburg, Emiel, Miruna Clinciu, Ondřej Dušek, Dimitra Gkatzia, Stephanie Inglis, Leo Leppänen, Saad Mahamood, Emma Manning, Stephanie Schoch, Craig Thomson, and Luou Wen. 2021a. Underreporting of errors in NLG output, and what to do about it. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 140–153, Aberdeen, Scotland, UK. Association for Computational Linguistics.
- van Miltenburg, Emiel, Desmond Elliott, and Piek Vossen. 2017. Cross-linguistic differences and similarities in image descriptions. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 21–30, Santiago de Compostela, Spain. Association for Computational Linguistics.
- van Miltenburg, Emiel, Chris van der Lee, Thiago Castro-Ferreira, and Emiel Kraemer. 2020a. Evaluation rules! on the use of grammars and rule-based systems for NLG evaluation. In *Proceedings of the 1st Workshop on Evaluating NLG Evaluation*, pages 17–27, Online (Dublin, Ireland). Association for Computational Linguistics.
- van Miltenburg, Emiel, Chris van der Lee, and Emiel Kraemer. 2021b. Preregistering NLP research. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 613–623, Online. Association for Computational Linguistics.
- van Miltenburg, Emiel, Wei-Ting Lu, Emiel Kraemer, Albert Gatt, Guanyi Chen, Lin Li, and Kees van Deemter. 2020b. Gradations of error severity in automatic image descriptions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 398–411, Dublin, Ireland. Association for Computational Linguistics.
- Munafò, Marcus R, Brian A Nosek, Dorothy VM Bishop, Katherine S Button, Christopher D Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J Ware, and John Ioannidis. 2017. A Manifesto for Reproducible Science. *Nature human behaviour*, 1(1):1–9.
- Nauta, Meike, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. 2022. From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. *CoRR*, abs/2201.08164.
- Novikova, Jekaterina, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.
- Popović, Maja. 2020. Informative manual evaluation of machine translation output. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5059–5069, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Popović, Maja. 2021. Agree to disagree: Analysis of inter-annotator disagreements in human evaluation of machine translation output. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 234–243, Online. Association for Computational Linguistics.
- Popović, Maja and Anya Belz. 2022. On reporting scores and agreement for error annotation tasks. In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 306–315, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Rahal, Rima-Maria, Susann Fiedler, Adeyemi Adetula, Ronnie P. A. Berntsson, Ulrich Dirnagl, Gordon B. Feld, Christian J. Fiebach, Samsad Afrin Himi, Aidan J. Horner, Tina B. Lonsdorf, Felix Schönbrodt, Miguel Alejandro A. Silan, Michael Wenzler, and Flávio Azevedo. 2023. Quality research needs good working conditions. *Nature Human Behaviour*.

- Raji, Inioluwa Deborah, Emily Denton, Emily M. Bender, Alex Hanna, and Amandalynne Paullada. 2021. AI and the Everything in the Whole Wide World Benchmark. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021*.
- Raji, Inioluwa Deborah, I. Elizabeth Kumar, Aaron Horowitz, and Andrew Selbst. 2022. The fallacy of ai functionality. In *2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22*, page 959–972, New York, NY, USA. Association for Computing Machinery.
- Reiter, Ehud and Anja Belz. 2009. An Investigation into the Validity of Some Metrics for Automatically Evaluating Natural Language Generation Systems. *Comput. Linguistics*, 35(4):529–558.
- Rogers, Anna and Isabelle Augenstein. 2020. What can we do to improve peer review in NLP? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1256–1262, Online. Association for Computational Linguistics.
- Sambasivan, Nithya, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. “everyone wants to do the model work, not the data work”: Data cascades in high-stakes ai. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, CHI '21*, New York, NY, USA. Association for Computing Machinery.
- Schlangen, David. 2021. Targeting the benchmark: On methodology in current natural language processing research. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 670–674, Online. Association for Computational Linguistics.
- Sculley, D., Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-François Crespo, and Dan Denison. 2015. Hidden technical debt in machine learning systems. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.
- Shimorina, Anastasia and Anya Belz. 2022. The human evaluation datasheet: A template for recording details of human evaluation experiments in NLP. In *Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 54–75, Dublin, Ireland. Association for Computational Linguistics.
- Singh, Gerald G, Jordan Tam, Thomas D Sisk, Sarah C Klain, Megan E Mach, Rebecca G Martone, and Kai MA Chan. 2014. A More Social Science: Barriers and Incentives for Scientists Engaging in Policy. *Frontiers in Ecology and the Environment*, 12(3):161–166.
- Sripada, Somayajulu G., Ehud Reiter, Ian Davy, and Kristian Nilssen. 2004. Lessons from deploying nlg technology for marine weather forecast text generation. In *Proceedings of the 16th European Conference on Artificial Intelligence, ECAI'04*, page 760–764, NLD. IOS Press.
- Strauss, Anselm and Juliet Corbin. 1994. Grounded Theory Methodology: An Overview. In N. K. Denzin and Y. S. Lincoln, editors, *Handbook of qualitative research*, pages 273–285. Sage Publications, Inc.
- Terry, Gareth, Nikki Hayfield, Victoria Clarke, and Virginia Braun. 2017. Chapter 2: Thematic Analysis. In C. Willig and W. Rogers, editors, *The SAGE Handbook of Qualitative Research in Psychology*, pages 17–36. SAGE Publications Ltd.
- Thomson, Craig and Ehud Reiter. 2020. A gold standard methodology for evaluating accuracy in data-to-text systems. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 158–168, Dublin, Ireland. Association for Computational Linguistics.
- Thomson, Craig and Ehud Reiter. 2021. Generation challenges: Results of the accuracy evaluation shared task. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 240–248, Aberdeen, Scotland, UK. Association for Computational Linguistics.
- Thomson, Craig, Ehud Reiter, and Barkavi Sundararajan. 2023. Evaluating factual accuracy in complex data-to-text. *Computer Speech & Language*, 80:101482.
- Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.
- Waltman, Ludo, Wolfgang Kaltenbrunner, Stephen Pinfield, and Helen B Woods. 2022. How to Improve Scientific Peer Review: Four Schools of Thought.
- Yarkoni, Tal. 2018. No, It's Not the Incentives, It's You. Published on [citation needed], personal blog of Tal Yarkoni.
- Zhou, Kaitlyn, Su Lin Blodgett, Adam Trischler, Hal Daumé III, Kaheer Suleman, and Alexandra Olteanu. 2022. Deconstructing NLG Evaluation: Evaluation Practices, Assumptions, and Their Implications. *CoRR*, abs/2205.06828.

A Information letter

What is this study about?

This research project aims to understand the status of error analysis in NLG. We aim to answer three questions:

- What do researchers think about error analysis?
- In what circumstances are researchers willing and able to carry out an error analysis?
- What are the barriers to carrying out an error analysis?

This study builds on an earlier position paper about error analysis, which shows that relatively few NLG papers provide an error analysis, and which provide a how-to guide for carrying out error analyses. You can read the paper [here](#).

What does participating in the study entail?

For this study, we ask you to answer a short series of questions. We expect this to take about 10 minutes. Most of these questions are multiple-choice, but there are also some open questions. Your answers will be completely anonymous, and it is impossible for us to trace back the answers to you.

Disadvantages, consequences & risks

- You will be asked to answer a series of questions, which takes time. We tried to make the questionnaire as short as possible, so as to minimise any possible inconvenience.
- Although we tried to prevent any question from offending any participants, it may still be the case that you take offense to some of the questions. In this case, feel free to leave a comment at the end of the survey, or to contact either us or the ethics committee directly. Contact details are at the bottom of this page.
- Some questions might be controversial. We record minimal personal information, so that you are free to speak your mind, without any consequences. The only personal information we collect is whether you work in industry or in academia, and how experienced you are.
- We do not foresee any other risks connected to your taking part in this study.

Advantages

There are no direct advantages to taking part in this study. The indirect advantage is that your contribution will help us understand how NLG researchers feel

about error analysis, and we aim to publish a full report through one of the many open-access venues in our field (e.g. INLG).

Rights

Under the main applicant's University's code of ethics, you are entitled to a number of rights:

- Your participation is completely voluntary, and you have the right to decline to participate and withdraw from the research once participation has begun, without any negative consequences, and without providing any explanation.
- You have the right, in principle, to request access to and rectification, erasure, restriction of or object to the processing of personal data. For more information, please see: [URL](#). Do note that, because all data is fully anonymised, it may be impossible for us to delete or alter your responses.
- Your participation is fully confidential, meaning that your answers will be fully anonymised. We have configured Qualtrics such that it will also not collect your IP address.
- Your consent to participate only lasts for the duration of the study, and may be withdrawn at any time.

What does consent mean?

By consenting, you indicate that you are voluntarily taking part in this study, and that you allow for your data to be processed. This means that:

- You agree that your answers may be used to publish a research article on this topic.
- The data will be stored on the computers of the research team, with both local (hard drive) and online (protected cloud drive) backups.
- The data will be made public upon completion of this study.
- You acknowledge that there is no financial compensation for taking part in this study.

The actual consent form is on the next page.

Contact details

This study has been approved by the Research Ethics and Data Management Committee (REDC) of the DEPARTMENT. If you have any questions about this study, you may contact the principal investigator via email: [EMAIL](#). If you have any remarks or complaints regarding this research, you may also contact the REDC via: [EMAIL](#).

Full list of the researchers involved: [NAMES](#)

B Informed consent form

This is the consent form for our study about the status of error analysis in NLG. Full details about this study were provided on the previous page. If you want to read this information again, you can go back to the previous page. If anything is still unclear about this study, please contact: EMAIL.

Consent

By consenting, you indicate that you have read the description on the previous page, that you are voluntarily taking part in this study, and that you allow for your data to be processed. This means that:

- You agree to your responses being anonymously recorded.
- Your answers will be used to study the status of error analysis in NLG, and may be used in future publications pertaining to this topic.
- The data will be shared with our research team, with both local (hard drive) and online (protected cloud drive) backups. This data will be stored indefinitely, and made public upon completion of our research. Note again that none of your answers can be traced back to you.
- You acknowledge that there is no financial compensation for taking part in this study.

Note that you may still withdraw your consent after completing this form, without any negative consequences. We will delete all incomplete forms from our study.

Do you consent?

Do you agree to take part in this study? If you consent, please indicate this below by clicking “Yes”. If you click “No”, you will be directed to the end of this questionnaire. You may also close this page to stop participating in this study.

- Yes, I consent.
- No, I do not consent.

C Survey questions

These are all the questions we have asked our participants to answer. Due to the display logic, participants always see a subset of the questions, based on their earlier answers. We have reproduced this display logic below with conditional statements (*if * was selected for question **). If the statement is true, then the question immediately following the statement is displayed. Otherwise, questions with false conditionals

are hidden.

Start of survey

1. Are you in academia or in industry? (If you have a dual affiliation, please respond with your dominant affiliation in mind.)

- Academia
- Industry
- Other

2. How many years have you been working in NLG?

- Less than 2 years
- 2-5 years
- 6-10 years
- 11 or more years
- I don't work in NLG

Definition of “error analysis”

Before continuing, we need to agree on the definition of error analysis. For the purposes of this questionnaire:

- We define “error analysis” as a formalised procedure (similar to annotation) in which errors in the output of an NLG system are identified and categorised, after which the frequencies for the different kinds of errors are reported.
- Error analyses are different from “error mentions”, which give an impression of the kinds of errors that are made by an NLG system, but are less formal and don't quantify the distribution of errors.

Example

Below is an excerpt from Table 3 of Barros & Lloret (2015, ENLG). The authors “manually analysed all the generated sentences and classified these errors attending to frequent grammatical errors and frequent drafting errors.” The table shows how often each type of error occurs in their data.

Error types	Number of sentences
Grammatical concordance: Nominal	2
Verbal	7
Non words semantic relations	36
Missing main verb	7
Incorrect syntactic order	38

3. Do you remember reading any NLG papers that include an error analysis?

- Yes
- No

If positive answer to question 3:

4. Did you find the error analyses to be useful?

- Not at all useful
- Slightly useful

- Moderately useful
- Very useful
- Extremely useful

If *not at all useful* was *not* selected for question 4:

5. What did you find useful about the error analyses you've seen?

(Open question)

If *not at all useful* was selected for question 4:

6. Why didn't you find the error analyses to be useful?

(Open question)

If negative answer to question 3:

7. Is it surprising to you that you haven't seen any published error analyses?

- Yes, because ...
- No, because ...

8. Have you ever carried out an error analysis?

- Yes
- No

If positive answer to question 8:

9. What did you find challenging or difficult about carrying out an error analysis?

(Open question)

If positive answer to question 8:

10. Did you feel like there were enough resources/reference material for you to carry out an error analysis?

- Yes
- No

If positive answer to question 8:

11. Do you think you'll carry out an error analysis again in the future?

- Definitely not
- Probably not
- Might or might not
- Probably yes
- Definitely yes

If positive answer to question 8:

12. Could you explain your answer to the previous question?

(Open question)

If negative answer to question 8:

13. Have you ever considered carrying out an error analysis?

- Never
- Once or twice

Regularly

I'm planning to carry out an error analysis in the future

If negative answer to question 8:

14. What is the reason you haven't carried out an error analysis?

(Open question)

If negative answer to question 8:

15. Are you willing to carry out an error analysis?

- Definitely not
- Probably not
- Might or might not
- Probably yes
- Definitely yes

16. For what kinds of papers do you think error analyses may be useful?

(Open question)

17. I would be more likely to carry out an analysis in a conference/journal paper if...

(Closed question with multiple statements. Answer options: Strongly disagree, Somewhat disagree, Neither agree nor disagree, Somewhat agree, Strongly agree)

- There was a higher page limit.
- There would be an existing error taxonomy that I could use.
- There would be dedicated annotation tools for error analysis that I could use.
- There would be a crowdsourcing template for carrying out error analyses.
- Reviewers paid more attention to error analyses.
- There were an available pool of annotators or crowd workers
- I had more time.
- I had more money.
- I had more collaborators.

18. Are there any other barriers that prevent you from carrying out an error analysis?

(Open question)

19. Please indicate whether you agree or disagree with the following statements

(Closed question with multiple statements. Answer options: Strongly disagree, Somewhat disagree, Neither agree nor disagree, Somewhat agree, Strongly agree)

- There should be more error analyses in the NLG literature
- Error analyses are a valuable part of a paper.
- Carrying out an error analysis is enjoyable.
- Carrying out an error analysis is boring/tedious.

- Error analyses are necessary to fully evaluate the performance of an NLG system.
- Knowing what errors a system makes is helpful for future research.
- Knowing what errors a system makes is helpful for practitioners/NLG in industry.
- If you publish at a **conference**, and you present an NLG system as one of your main contributions, you should include an error analysis.
- If you publish in a **journal**, and you present an NLG system as one of your main contributions, you should include an error analysis.

20. I am ... likely to include an error analysis in a journal article than/as I would be for a conference publication.

- More
- Less
- Equally

21. Please explain your answer to the previous question (Open question)

22. Are there currently enough resources to support error analysis?

- Yes
- No, I am still missing: ...

23. Besides resources, are there any other factors that would make it more likely for you to carry out an error analysis?

(Open question)

We believe that it is essential for authors of error analyses to include a table with the distribution of errors in the output of their system. This data should be based on a formalised annotation procedure, with at least two annotators, so that the paper can also report inter-annotator agreement to gauge the reliability of the analysis.

24. What else would you recommend that authors should include in an error analysis?

(Open question)

25. This is the final question. Is there anything you would like to add or comment on?

Benchmark for Evaluation of Danish Clinical Word Embeddings

Martin S. Laursen^{*}, University of Southern Denmark, Odense, Denmark msla@mmmi.sdu.dk

Jannik S. Pedersen^{*}, University of Southern Denmark, Odense, Denmark jasp@mmmi.sdu.dk

Pernille Just Vinholt, Odense University Hospital, Denmark pernille.vinholt@rsyd.dk

Rasmus Søgaard Hansen, Odense University Hospital, Denmark rasmus.sogaard.hansen@rsyd.dk

Thiusius Rajeeth Savarimuthu, University of Southern Denmark, Odense, Denmark trs@mmmi.sdu.dk

Abstract In natural language processing, benchmarks are used to track progress and identify useful models. Currently, no benchmark for Danish clinical word embeddings exists. This paper describes the development of a Danish benchmark for clinical word embeddings. The clinical benchmark consists of ten datasets: eight intrinsic and two extrinsic. Moreover, we evaluate word embeddings trained on text from the clinical domain, general practitioner domain and general domain on the established benchmark. All the intrinsic tasks of the benchmark are publicly available¹.

1 Introduction

Word embeddings are real-valued vectors that are trained to represent words based on the context in which they appear. Based on the distributional hypothesis (Harris, 1954), which suggests that words with similar contexts have similar meaning, embeddings of semantically similar words are expected to appear close to each other in vector space.

Since their introduction, word embeddings have been ubiquitous in natural language processing (NLP) due to their ability to represent word meaning. Typically, word embeddings are trained on a general text corpus such as Wikipedia. Afterwards, word embeddings are used as stand-alone features or as input to neural networks to perform a wide variety of NLP tasks such as text classification, named entity recognition (NER) and machine translation.

In specialized domains, such as the clinical, word embeddings are also widely used to e.g. extract information from electronic health records (EHRs). However, the text in clinical EHRs differs significantly from the general domain. Clinical EHRs include rare words, domain specific abbreviations and a mix of languages (for example Latin, English and Danish). The text is often non-narrative and very concise, free of syntactic rules, sometimes consisting of a sequence of keywords. Moreover, it contains many spelling errors, and the se-

mantic meaning of words can differ from that of the general domain (Leaman et al., 2015). In the clinical domain, word embeddings are, therefore, often trained on an in-domain corpus to better capture the vocabulary and the semantic meaning of words. After being trained on an in-domain corpus, they are used for e.g. clinical NER, International Classification of Diseases coding, clinical event detection, de-identification and patient similarity estimation with improved performance over general word embeddings (Zhao et al., 2018; Wang et al., 2018; Chen et al., 2019).

For evaluating word embeddings, two different methods are typically used: intrinsic and extrinsic evaluation (Wang et al., 2019c). In intrinsic evaluation, word embeddings are evaluated based on their inherent information, e.g. by exploring the syntactic or semantic relationship between words. In extrinsic evaluation, word embeddings are evaluated based on their ability to solve a downstream task, e.g. by using them as input to a neural network. While word embeddings can be evaluated using extrinsic benchmarks by holding the network architecture fixed while varying the set of word embeddings, intrinsic benchmarks provide an intermediate evaluation of the embeddings' properties before being used as input to a larger system. This supports the need for intrinsic evaluation.

Word embeddings for the general domain are publicly available in many languages (Grave et al., 2018). However, publicly available embeddings for the clinical domain are scarce (Khattak et al., 2019). This is

^{*}Both authors contributed equally to this paper.

¹www.github.com/jannikskyt/DaClinWordEmbeddings

most likely due to strict regulations around clinical data which contain sensitive information making them unsuitable for sharing. Therefore, researchers in clinical NLP are often forced to create their own word embeddings in order not to expose sensitive information (Abdalla et al., 2020).

Clinical intrinsic benchmark datasets do not necessarily contain sensitive information and can, in that case, be shared openly, benefitting researchers producing clinical word embeddings. For the English language, both intrinsic and extrinsic benchmarks exist, e.g. University of Minnesota Medical Residents Similarity / Relatedness Set (UMNSRS) (Pakhomov et al., 2010) for word similarity and relatedness, and BLUE (Peng et al., 2019), which includes both clinical and biomedical datasets, for extrinsic evaluation. For Danish, though, no clinical benchmark exists.

In this paper, we introduce a clinical word embedding benchmark for the Danish language. Moreover, we produce clinical word embeddings and use the benchmark to compare them to embeddings trained on the general domain and embeddings trained on the general practitioner (GP) domain.

The benchmark is specifically constructed to evaluate static word embeddings such as GloVe (Pennington et al., 2014), Continuous Bag-of-Words (Mikolov et al., 2013a), Skip-gram (Mikolov et al., 2013a) and FastText (Bojanowski et al., 2017). It is therefore not suitable for evaluation of contextual word embeddings produced by transformer models like BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019) and ELECTRA (Clark et al., 2019).

Although transformer models achieve state-of-the-art results (Wang et al., 2019b; Wang et al., 2019a), static word embeddings are still useful as input to NLP pipelines. Some advantages are that they require less compute to train and at inference time, and they work better on limited data (Peng et al., 2021). This is relevant for research within specialized domains, such as clinical NLP, where researchers must often train their own word embeddings on limited data and the hardware to train and run a transformer model is not necessarily available. Static word embeddings are also relevant in time-critical tasks in clinical practice such as expanding single-word searches in the EHR using the nearest neighbors of the search term. Expanding single-word searches is especially relevant in the clinical domain where many different terms can be used about the same basic symptom or disease. Another advantage is their ease of use for medical doctors (MDs) and clinical researchers who are not machine learning scientists compared to contextual word embeddings.

The remainder of this paper first introduces the benchmark including the methods for creating each intrinsic and extrinsic dataset. It then describes the train-

ing methods of the produced Danish clinical word embeddings and those from the general and GP domains which they will be benchmarked against. Finally, the benchmark results are presented and discussed.

2 Establishing Benchmark

The benchmark consists of an intrinsic and extrinsic part. In this paper, intrinsic performance is evaluated based on the quality of the semantic and syntactic inherent information using analogy and similarity tasks. We produce datasets for three different intrinsic evaluation methods: analogy tasks, similarity and relatedness tasks, and an equality task.

Analogy tasks, introduced by Mikolov et al. (2013b), take tuples of four words (A, B, C and D) and evaluate ‘what is to C as B is to A’ by selecting the nearest neighbour to the calculated vector in the embedding space, excluding the words forming the analogy:

$$\vec{C} + \vec{B} - \vec{A} = \vec{D}$$

If the nearest neighbor to the calculated vector is D, the analogy task is correct. The task is evaluated on the percentage of correct predictions in the dataset.

Similar to Pennington et al. (2014), the similarity tasks take a tuple of two words and their similarity score, in our case, produced by one or more MDs. The similarity score of a word pair is compared to the cosine similarity of the pair’s word embeddings. The cosine similarity is calculated as:

$$similarity(\vec{v}, \vec{u}) = \frac{\vec{v} \cdot \vec{u}}{\|\vec{v}\| \cdot \|\vec{u}\|}$$

The correlation between MD scores and cosine similarities for the dataset of word pairs is evaluated using the Spearman’s rank correlation coefficient. Relatedness tasks are identical to similarity tasks except the MDs produce a relatedness score instead of a similarity score. Relatedness refers to one word calling to mind another word (e.g., needle–thread), while similarity reflects the degree of semantic feature overlap between words (e.g., whale–dolphin) (Pakhomov et al., 2010).

Equality tasks take a tuple of two terms with the exact same meaning. As the similarity score of a pair is 1 for a perfect match, the objective is maximization of the cosine similarity between terms. The task is evaluated as the mean of the cosine similarities for all pairs in the dataset.

The extrinsic part consists of two different text classification tasks in the clinical domain with the word embeddings as input. The quality of the word embeddings is evaluated based on the evaluation metric of the classification task.

An overview of all datasets can be seen in Table 1.

Task	Description	Example
Intrinsic tasks		
Clinical analogy	Evaluate "what is to C as B is to A"	$\vec{joint} + \vec{colonoscopy} - \vec{colon} = \vec{arthroscopy}$
Clinical similarity UMNSRS similarity UMNSRS relatedness	Compare the human similarity/relatedness score of a word pair to the cosine similarity of the pair's word embeddings	uterus, cervix
Clinical abbreviation equality	Compare the similarity of a word and its abbreviation	cm, centimeter
Verb, adjective, and noun inflection analogy datasets	Evaluate "what is to C as B is to A"	$\vec{adjusted} + \vec{remove} - \vec{removed} = \vec{adjust}$
Extrinsic tasks		
Bleeding classification	Classify a paragraph as either positive or negative for bleeding	15-year-old girl hospitalized with bleeding tendency and anemia symptoms
Hospital department classification	Classify a paragraph into one of six hospital departments	Clinical contact. Prepared by clinic. Conclusion and plan: As agreed and as a follow-up to the note on 10.2.99, I have contacted pt. However, pt. is hospitalized due to ...

Table 1: Overview of the datasets used in the benchmark. Examples are translated to English.

2.1 Intrinsic Datasets

The intrinsic part consists of the following semantic tasks: clinical analogy, clinical similarity, clinical abbreviation equality, and UMNSRS similarity and relatedness; and the following syntactic tasks: verb inflection analogy, adjective inflection analogy, and noun inflection analogy. The intrinsic syntactic tasks are evaluating the syntactic properties of word embeddings in general rather than specifically for clinical use cases. As good clinical word embeddings must also contain syntactic information, the syntactic tasks are constructed to specifically evaluate the inherent syntactic information on words from the clinical domain.

The development of each intrinsic task consisted of 1) selecting the terms to use for the task and 2) creating the evaluation dataset. This is described for each task below. All intrinsic datasets are supplied in the supplementary material.

2.1.1 Clinical Analogy Dataset

Two MDs, in agreement, created 41 distinct clinical analogies such as (translated from Danish)

$$\vec{colonoscopy} - \vec{colon} = \vec{arthroscopy} - \vec{joint}$$

where the word pairs on each side of the equation have the same one-to-one relationship. For the example above with the one-to-one relationship 'is telescopic examination of', it means that colonoscopy is a telescopic examination of only the colon, and that the colon has only one telescopic examination: a colonoscopy. Some other common relationships were 'treats', 'is indicator for', 'is disease in anatomy', 'is test for', 'is examination

of', 'leads to' and 'is symptom of'. We relaxed the one-to-one relationship condition in a few cases: if for example a symptom is predominant for one disease but also minorly associated with another, we accepted the word pair. We augmented each distinct analogy to form four analogies by changing the order of the words inside the word pairs and by changing the order of the word pairs. This means that, for the analogy example above, we predicted each of 'colonoscopy', 'colon', 'arthroscopy', and 'joint' from the remaining three words. We performed this augmentation because the analogy tasks are based on evaluating the nearest neighbour to the calculated vector. Since the surrounding embedding space for each of the four calculated vectors may vary in distance to neighbours, the result may vary depending on which of the four words is predicted.

The clinical analogy dataset consists of 164 analogies.

2.1.2 Clinical Similarity Dataset

For the clinical similarity dataset, we predefined the following goals for achieving a diverse set of word pairs:

1. The selected words should be of different categories, e.g. they should not all be diseases.
2. The selected words should appear with varying frequency in clinical EHRs.
3. Word pairs should be matched within and across the categories and frequencies.
4. Words should not be selected based on an existing clinical EHR database because it could introduce bias to the dataset, e.g. the frequency of

words in our clinical EHR database might differ from other databases.

To achieve this, we predefined five clinical categories: anatomy, symptom/finding, disease, treatment, and diagnostic; and three frequency categories indicating how frequently a word appears in clinical EHRs: infrequent, occasional, and frequent. Then, two MDs selected words from a reference work on internal medicine (Schaffalitzky de Muckadell et al., 2009) by turning to approximately every fifth page, randomly selecting words, and subjectively assigning them categories until all three frequency categories per five clinical categories had 36 words each. This generated a total of 108 words per clinical category and 540 words overall.

We defined 270 word pairs by pairing 36 words from each clinical category with 36 words from the same category and 36 words evenly distributed on the four other clinical categories. We opted to use more words per group for intra-category-pairs than inter-category-pairs because we expected it would decrease the overrepresentation of pairs with low similarity. The pairings were distributed evenly across frequency categories. Finally, to further decrease overrepresentation of pairs with low similarity, the MDs subjectively defined 19 extra pairs with high similarity by pairing any two words from the word pool, resulting in a total of 289 word pairs.

Ten MDs with 2 to 17 years (mean: 7.5 years) of clinical experience used between 17 and 45 minutes (mean: 30.5 minutes) to rate the 289 pairs. Nine MDs had clinical biochemistry as speciality and one had pathology. The pairs were rated for similarity on a scale from 0 to 6 with 0 being lowest similarity and 6 being highest similarity. It was emphasized that the MDs should rate for similarity and not relatedness. If a word pair was unknown to the MDs, they did not rate it. One pair was rated by eight MDs and the rest were rated by at least nine. The similarity score for each pair is the mean rating. The mean ratings span from 0 to 6 with a minimum similarity score of 0.3, a mean of 1.1, and a maximum of 5.4. The standard deviations range from 0.3 to 1.6 with a mean of 0.7.

2.1.3 Clinical Abbreviation Equality Dataset

A list of 319 clinical abbreviations and their corresponding words was collected from online sources (supplementary material). Only abbreviations of single words were collected to simplify the evaluation of word embeddings, which usually represent single words. Ambiguous abbreviations and the abbreviations deemed unlikely to appear in clinical EHRs by an MD were removed. For example, the abbreviation ‘all’ is ambiguous because it could both mean ‘allergy’ or ‘acute lymphocytic leukemia’. The final dataset comprises 195

abbreviation–word pairs with the same meaning.

2.1.4 UMNSRS Similarity and Relatedness Datasets

The UMNSRS consists of 566 English term pairs rated for semantic similarity and 587 for semantic relatedness on a continuous scale from 0 to 1600. One MD translated the datasets into Danish. Pairs consisting of a term that translates into a multi-word expression were removed. As were terms that do not exist in Danish, for example a non-traded drug. In cases where a Danish counterpart drug exists, for example ‘betalaktam’ for ‘cefexitin’, this term was used as a translation. The Danish translation of the UMNSRS consists of 528 similarity pairs and 557 relatedness pairs.

2.1.5 Verb Inflection Analogy Dataset

A list of all verbs was extracted from the Danish orthographic dictionary (Danish Language Council, 2012). One MD selected verbs from the list that were deemed would occasionally or frequently occur in a clinical EHR. Next, verbs were conjugated in the following inflections: infinitive, present/future (same form in Danish), past tense, and present/past perfect. If a verb did not exist in all four inflections or had the same form in multiple inflections, it was removed from the list as it would cause analogy tasks involving the zero-vector. The final list contained 92 words, each in four inflections.

For each verb, six types of inflection pairs were made, for example infinitive–past, by pairing each inflection with the three other inflections. Next, we randomly combined each verb with 20 other verbs, evenly distributed on types of inflection pairs except for the remainder after equal division. This produced 1,840 analogies like the following of type infinitive–past (translated from Danish):

$$\overrightarrow{\text{remove}} - \overrightarrow{\text{removed}} = \overrightarrow{\text{adjust}} - \overrightarrow{\text{adjusted}}$$

2.1.6 Adjective Inflection Analogy Dataset

The same method as described for the verb inflection analogy dataset was used to develop the adjective inflection analogy dataset. Adjectives were declined in the following inflections: common positive, neuter positive, plural positive, comparative and superlative. The final list contained 43 words, each in five inflections.

For each adjective, we made seven types of inflection pairs by pairing each of the three positive inflections with comparative and superlative and finally, the comparative with the superlative.

We combined each adjective with all other adjectives to produce 1,806 analogies.

2.1.7 Noun Inflection Analogy Dataset

We created a list from the 180 frequent words from the combined five clinical categories of the clinical similarity dataset. We removed words which were not nouns and declined the remaining in the following inflections: indefinite singular, definite singular, indefinite plural and definite plural. If a noun did not exist in all four inflections or had the same form in multiple inflections, it was removed from the list. The final list contained 138 words, each in four inflections. For each noun, we made six types of inflection pairs by pairing each inflection with the three other inflections. Next, we randomly combined each noun with 13 other nouns, evenly distributed on types of inflection pairs except for the remainder after equal division, to produce 1,794 analogies.

2.2 Extrinsic Datasets

The extrinsic part consists of a hospital department classification task and a bleeding classification task. All datasets were obtained according to each dataset's respective data usage policy. The datasets are described below.

2.2.1 Bleeding Classification

For the bleeding classification dataset, we used that of Pedersen et al. (2021). It consists of 9,430 training sentences, 1,178 validation sentences, and 1,178 test sentences which are evenly distributed on the two classes: 'indicates bleeding' and 'does not indicate bleeding'. The latter class consists of 50% sentences that were deemed by the MDs to be at high risk of being misinterpreted by the deep learning model. The other 50% were random negative sentences. The classification objective is to predict if a sentence indicates bleeding.

The data came from 300 EHRs corresponding to 88,477 notes from the EHR system of the Region of Southern Denmark between 2015 and 2020. The sentences were annotated by splitting the annotation of EHRs between twelve MDs.

2.2.2 Hospital Department Classification

The hospital department classification dataset was constructed without the need of human annotators by using the department associated with each note as a label. This approach is an advantage since the task of annotating clinical records is time consuming and expensive.

The hospital department classification dataset consists of 42,000 clinical EHR notes evenly distributed on the following six Odense University Hospital departments: Cardiology; Cardiac, Thoracic and Vascular Surgery; Orthopaedic Surgery; Rheumatology;

Surgery; and Medical Gastrointestinal Diseases. Danish clinical EHR notes have a tree structure consisting of many generic node headlines. MDs only fill out the end-nodes manually. To avoid node headlines or text passages specific to one department making the classification a simple task, each note was preprocessed by only keeping the lowercased end-node texts. Furthermore, end-nodes which were duplicates based only on their words, disregarding all but letters, were removed across the whole dataset. The notes are between 51 and 220 tokens. The dataset contains 7,000 notes from each department in a class-balanced train:validation:test ratio of 5:1:1. The classification objective is to predict the hospital department.

3 Word Embedding Evaluation

This section describes an evaluation of word embedding models, trained on data from different domains, using the established benchmark. We make a clinical-general domain comparison using a FastText (Bojanowski et al., 2017) model as it has the best performance on Danish text according to benchmark results (Brogaard Pauli et al., 2021). We make a clinical-GP domain comparison using a GloVe (Pennington et al., 2014) model as it is the only available type of embeddings trained on Danish GP data. We describe how the benchmark can be used to show strengths and weaknesses of different word embeddings.

We trained two sets of clinical word embeddings using the FastText and GloVe methods. The embeddings were trained on 299,718 Danish EHRs from Odense University Hospital. The text was preprocessed by lowercasing and removing headlines, subheadings, phone numbers, social security numbers, emails, URLs, dates and time stamps. Samples were defined as text from the same subheading. After removal of duplicates and samples with less than 3 words, the corpus consisted of 1.4 billion tokens.

For the clinical-general domain comparison, the clinical FastText embeddings were trained with the default settings from the FastText API (www.fasttext.cc) except from a vector size of 300, 10 negative samples and 10 epochs. The hyperparameters were chosen to be able to compare the produced embeddings with the FastText word embeddings from Grave et al. (2018) pre-trained on a general domain, specifically Wikipedia and Common Crawl. The FastText models can generate out-of-vocabulary (OOV) words from subwords which e.g. makes it capable of representing unknown spelling errors. For clarity, only the results without OOV generation are reported here while the results with OOV generation are found in Appendix A.

For the clinical-GP domain comparison, the clinical GloVe embeddings are 100-dimensional embeddings

trained with the default settings from the code and paper by Pennington et al. (2014) except for a min-count of 3. The hyperparameters were chosen to be able to compare with the GloVe word embeddings from Rasmussen et al. (2019) trained on 323,122 GP EHRs.

The word embedding models are benchmarked on the established intrinsic and extrinsic datasets. For each intrinsic task, we show the performance of the embeddings on the part of the evaluation dataset which is in-vocabulary (IV), ignoring the word pairs or analogies containing OOV words. We also produce the IV rate as the proportion of word pairs or analogies which are in the vocabulary of the embeddings. Additionally, Appendix B contains the IV intersection results which show the performance of the embeddings on the intersection of all embeddings' IV dataset for that task.

For the extrinsic tasks, the word embeddings are used as input to a recurrent neural network which is initialized and trained three times with the same set of standard hyperparameters. No hyperparameter tuning is performed. A bidirectional gated recurrent unit (Cho et al., 2014) with 128 units followed by a dropout layer with probability 0.3 is trained with the Adam optimizer with a learning rate of $5e-4$ for a maximum of 100 epochs using early stopping. The best model, based on the validation loss, is evaluated on the test set. The test set accuracy is reported as the evaluation result.

3.1 Intrinsic Results

We present the intrinsic semantic and syntactic benchmark results.

3.1.1 Semantic Results

Table 2 shows the intrinsic semantic results. The clinical FastText embeddings achieve better performance than the general FastText embeddings on the abbreviation equality task, clinical similarity task, UMNSRS similarity task and UMNSRS relatedness task. The clinical analogy task shows different results with the general FastText embeddings performing better with an IV accuracy of 0.14 while the clinical FastText embeddings have an IV accuracy 0.05. The clinical GloVe embeddings perform better than the GP GloVe embeddings on all intrinsic semantic tasks.

The word embeddings trained on the clinical domain show the highest IV rates, followed by the GP domain and then the general domain. The two clinical models have an IV rate equal to or higher than 0.83 for all semantic tasks. The GP GloVe embeddings have IV rates between 0.57 and 0.75 while the general FastText embeddings have IV rates between 0.54 and 0.61.

Appendix C presents the correct clinical analogy predictions for all word embedding models. Moreover, Appendix D shows the results on the clinical analogy

task where a prediction is considered correct if the correct term is in the top 1, 5 and 10 nearest neighbours to the calculated vector.

3.1.2 Syntactic Results

Table 3 shows the intrinsic syntactic results. The results show that the general FastText embeddings achieve better performance than the clinical FastText embeddings on all syntactic tasks with an IV accuracy of 0.69 on verbs, 0.60 on nouns and 0.41 on adjectives. The clinical FastText embeddings perform at IV accuracies of 0.28, 0.19 and 0.16, respectively. The clinical GloVe embeddings perform better than the GP GloVe embeddings on the verb and noun inflection tasks with IV accuracies of 0.21 and 0.04, and 0.09 and 0.01, respectively. The GP GloVe embeddings perform best on the adjective inflection task with an IV accuracy of 0.04 contra 0.03 for the clinical GloVe embeddings.

The clinical domain embeddings have the highest IV rates for the verb and noun inflection tasks at 0.99 and 0.39, respectively. The general FastText embeddings have the highest IV rate for the adjective inflection task at 0.65, followed by the clinical GloVe embeddings at 0.47.

3.2 Extrinsic Results

Table 4 shows the extrinsic results. For both the FastText and GloVe models, the clinical domain embeddings achieve higher performances than their respective general domain and GP domain counterparts.

4 Discussion

In this paper, we have presented the first benchmark for evaluating Danish clinical word embeddings. Although the clinical word embeddings cannot be shared due to privacy concerns, having a publicly available benchmark will allow researchers to compare and evaluate locally available clinical word embeddings. Below, we discuss the capability of the benchmark to compare word embedding performance in the clinical domain.

As the intrinsic benchmark tasks consist of words which are typically, and in some cases, exclusively, used in the clinical domain, we expected higher IV rates from clinical domain embeddings. In concurrence, the results show that the clinical word embeddings, in general, have higher IV rates than those trained on the GP and general domain. An exception is that the general FastText embeddings have the highest IV rate for the adjective inflection analogy task. One explanation could be that clinical written language does not use as many inflections of adjectives as the general. Interestingly, when comparing the GP and general word em-

	FastText (300d)	
	Clinical	General
Clinical analogy, accuracy (IV)	0.05 (0.88)	<u>0.14</u> (0.54)
Abbreviation equality, similarity (IV)	<u>0.53</u> (0.84)	0.27 (0.58)
Clinical similarity, ρ (IV)	<u>0.64</u> (0.93)	0.43 (0.61)
UMNSRS similarity, ρ (IV)	<u>0.60</u> (0.88)	0.30 (0.59)
UMNSRS relatedness, ρ (IV)	<u>0.54</u> (0.83)	0.32 (0.56)
	GloVe (100d)	
	Clinical	GP
Clinical analogy, accuracy (IV)	<u>0.08</u> (0.88)	0.06 (0.61)
Abbreviation equality, similarity (IV)	<u>0.49</u> (0.85)	0.24 (0.57)
Clinical similarity, ρ (IV)	<u>0.56</u> (0.96)	0.34 (0.75)
UMNSRS similarity, ρ (IV)	<u>0.41</u> (0.89)	0.18 (0.74)
UMNSRS relatedness, ρ (IV)	<u>0.41</u> (0.84)	0.21 (0.70)

Table 2: Semantic benchmark results on the in-vocabulary (IV) dataset for each task by model type (FastText, GloVe) and domain (clinical, general, general practitioner (GP)). The accuracy metric is the accuracy on the dataset. The similarity metric is the average cosine similarity on the dataset. The ρ metric is the Spearman’s rank correlation coefficient on the dataset. IV rates are reported in parenthesis. We underline the best results per task by model type.

	FastText (300d)	
	Clinical	General
Verb inflection analogy, accuracy (IV)	0.28 (0.99)	<u>0.69</u> (0.92)
Noun inflection analogy, accuracy (IV)	0.19 (0.36)	<u>0.60</u> (0.13)
Adjective inflection analogy, accuracy (IV)	0.16 (0.36)	<u>0.41</u> (0.65)
	GloVe (100d)	
	Clinical	GP
Verb inflection analogy, accuracy (IV)	<u>0.21</u> (0.99)	0.09 (0.83)
Noun inflection analogy, accuracy (IV)	<u>0.04</u> (0.39)	0.01 (0.18)
Adjective inflection analogy, accuracy (IV)	0.03 (0.47)	<u>0.04</u> (0.25)

Table 3: Syntactic benchmark results on the in-vocabulary (IV) dataset for each task by model type (FastText, GloVe) and domain (clinical, general, general practitioner (GP)). The accuracy metric is the accuracy on the dataset. IV rates are reported in parenthesis. We underline the best results per task by model type.

beddings on the semantic tasks, the GP embeddings, in four out of five tasks, have higher IV rates but lower accuracy. This result shows that the GP embeddings have seen more clinical domain words than the general embeddings during training, but the general embeddings capture higher quality information for the words that it has seen. This could be due to the size and quality of the dataset, differences between model types or the dimensionality of the embeddings. Future work should investigate these claims further.

The benchmark shows that the clinical embeddings surpass the general and GP embeddings in all semantic tasks except for the clinical analogy task where the general FastText embeddings performed better than the clinical FastText embeddings. This discrepancy may be caused by the clinical analogy dataset only containing 164 analogies of which only 54% are IV for the general FastText model.

The general embeddings surpass the clinical embeddings on the syntactic tasks which shows that it

has captured higher quality syntactic information for the words that it has seen during training. This is most likely due to Wikipedia and Common Crawl, which it was trained on, containing a higher quality of syntactic information than clinical EHRs.

The fact that the general embeddings achieve the highest IV rate on the adjective inflection task suggests that the task consists of more inflections specific to the general domain than our clinical dataset. On the contrary, clinical domain embeddings achieve the highest IV rates on the verb and noun inflection tasks which suggests that these syntactic tasks do contain inflections specific to the clinical domain.

Similar to earlier work (Zhao et al., 2018; Wang et al., 2018; Chen et al., 2019), we found that the clinical word embeddings perform better than the GP and general domain embeddings on extrinsic tasks. It is notable that for the extrinsic tasks, the GP GloVe embeddings are closer to the performance of the clinical GloVe embeddings than the general FastText embeddings are to

	FastText (300d)	
	Clinical	General
Bleeding classification, accuracy	<u>0.93</u>	0.84
Department classification, accuracy	<u>0.83</u>	0.65
	GloVe (100d)	
	Clinical	GP
Bleeding classification, accuracy	<u>0.90</u>	0.87
Department classification, accuracy	<u>0.76</u>	0.66

Table 4: Extrinsic benchmark results by model type (FastText, GloVe) and domain (clinical, general, general practitioner (GP)). We report accuracies on the class-balanced bleeding and department classification tasks using the word embeddings as input. We underline the best results per task by model type.

that of the clinical FastText embeddings. This could be explained by the fact that there is some similarity between the GP and clinical domains, both being subdomains of the healthcare domain.

Considering that the general embeddings perform well on syntactic tasks and clinical embeddings perform well on semantic and extrinsic tasks, future work should explore training word embeddings from the general FastText checkpoint on clinical data. This might provide word embeddings that better capture both clinical and general syntactic and semantic properties.

4.1 Limitations

This study compared GloVe and FastText word embeddings. While FastText performed best on some benchmarks, other word embedding methods might perform better. We leave these investigations to future work.

Future use of the presented resources relies on the assumption that the words in the intrinsic datasets also appear in the user’s vocabulary. In section 2.1.2 we described how we tried to mitigate this shortcoming.

The clinical similarity dataset would benefit from including more pairs with high similarity and decreasing the mean standard deviation, e.g. by including more raters from different specialities. To alleviate MD rating disagreement, we have included in the supplementary material the clinical similarity ratings for each MD with information about the standard deviation of each word pair, which can be used to set a threshold of maximum allowed disagreement. Appendix E shows the results on the clinical similarity dataset consisting of pairs with standard deviations at or below 1.

The extrinsic department classification task might as well classify the writing styles of specific MDs in a department, thus not necessarily generalizing to other MDs. This can be remedied by having unique authors in the test split.

It is a limitation to the extrinsic results that no hyperparameter tuning was performed. Results from a model trained with a standard set of hyperparameters can rank the word embeddings but the results are not

indicative of the best performance of each embedding.

We have shown a discrepancy between the clinical analogy task and all other semantic tasks. We believe it would be beneficial to include more analogies in the clinical analogy dataset as the result is based on few IV analogies.

The syntactic results suggest that the adjective inflection task consists of more inflections specific to the general domain than the clinical domain. Many of the inflections do exist in the clinical domain but it is a limitation for the evaluation of clinical word embeddings that not enough inflections are specific for the clinical domain.

The tasks were designed to evaluate static word embeddings using only single-word expressions which limits the use of the benchmark for contextual word embeddings such as transformer models and word embeddings trained on n-grams.

It is a limitation to our benchmark that it only provides two extrinsic tasks, and in general, that there are no Danish clinical extrinsic datasets publicly available. Due to privacy concerns, we cannot publish the extrinsic datasets, but we provide a method for creating an extrinsic test that leverages already existing labels in the form of the department of the clinical note. This method does not need any labeling but still requires access to EHRs. We encourage interested researchers to contact us for the possibility of sharing the extrinsic datasets.

Future work should focus on developing more diverse extrinsic tasks such as named entity recognition, relation extraction and question answering.

5 Conclusion

In this paper, we presented a benchmark for Danish clinical word embeddings. The benchmark consists of two extrinsic tasks, five intrinsic semantic tasks and three intrinsic syntactic tasks. We developed clinical word embeddings and compared them with word embeddings trained on a general and general practitioner

domain. The benchmark showed that the word embeddings trained on clinical data performed better on the extrinsic and semantic tasks, except for the clinical analogy task. On the syntactic tasks, the FastText word embeddings trained on a general domain performed better than those trained on a clinical domain.

Acknowledgements

The authors thank Anne Bryde Alnor, Charlotte Gils, Eline Sandvig Andersen, Ida Stangerup, Jesper Dupont Ewald, Jesper Farup Revsholm, Katrine Sølling Borlund Madsen and Kristina Bjerg Appel for the rating work for the clinical similarity task.

References

- Abdalla, Mohamed, Moustafa Abdalla, Graeme Hirst, and Frank Rudzicz. 2020. Exploring the privacy-preserving properties of word embeddings: Algorithmic validation study. *J Med Internet Res*, 22(7):e18055.
- Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.
- Brogaard Pauli, Amalie, Maria Barrett, Ophélie Lacroix, and Rasmus Hvingelby. 2021. DaNLP: An open-source toolkit for danish natural language processing. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa 2021)*.
- Chen, Qingyu, Yifan Peng, and Zhiyong Lu. 2019. Biosentvec: creating sentence embeddings for biomedical texts. In *2019 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 1–5.
- Cho, Kyunghyun, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.
- Clark, Kevin, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2019. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.
- Danish Language Council. 2012. *Retskrivningsordbogen*, 4 edition. Danish Language Council. Including 8 digital issues (2017).
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Grave, Edouard, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Harris, Zellig S. 1954. Distributional structure. *Word*, 10(2-3):146–162.
- Khattak, Faiza Khan, Serena Jebblee, Chloé Pou-Prom, Mohamed Abdalla, Christopher Meaney, and Frank Rudzicz. 2019. A survey of word embeddings for clinical text. *Journal of Biomedical Informatics*, 100:100057. Articles initially published in *Journal of Biomedical Informatics*: X 1-4, 2019.
- Leaman, Robert, Ritu Khare, and Zhiyong Lu. 2015. Challenges in clinical natural language processing for automated disorder normalization. *Journal of Biomedical Informatics*, 57:28–37.
- Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- Mikolov, Tomas, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.
- Schaffalitzky de Muckadell, Ove B., Stig Haunsø, and Hendrik Vilstrup. 2009. *Medicinsk kompendium*, 17 edition. Nyt Nordisk Forlag.
- Pakhomov, Serguei, Bridget McInnes, Terrence Adam, Ying Liu, Ted Pedersen, and Melton Genevieve B. 2010. Semantic similarity and relatedness between clinical terms: An experimental study. *AMIA annual symposium proceedings*, 2010:572–576.
- Pedersen, Jannik S., Martin S. Laursen, Thiusius Rajeeth Savarimuthu, Rasmus Søgaard Hansen, Anne Bryde Alnor, Kristian Voss Bjerre, Ina Mathilde Kjær, Charlotte Gils, Anne-Sofie Faarvang Thorsen,

Eline Sandvig Andersen, Cathrine Brødsgaard Nielsen, Lou-Ann Christensen Andersen, Søren Andreas Just, and Pernille Just Vinholt. 2021. Deep learning detects and visualizes bleeding events in electronic health records. *Research and Practice in Thrombosis and Haemostasis*, 5(4):e12505.

Peng, X, Y Zheng, C Lin, and A Siddharthan. 2021. Summarising historical text in modern languages. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3123–3142. Association for Computational Linguistics (ACL).

Peng, Yifan, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65, Florence, Italy. Association for Computational Linguistics.

Pennington, Jeffrey, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Rasmussen, Mathias, Nichlas Berggrein, and Leon Derzycynski. 2019. Named entity recognition and disambiguation in danish electronic health records. Master’s thesis, IT University of Copenhagen.

Wang, Alex, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019a. Superglue: a stickier benchmark for general-purpose language understanding systems. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 3266–3280.

Wang, Alex, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019b. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019*.

Wang, Bin, Angela Wang, Fenxiao Chen, Yuncheng Wang, and C.-C. Jay Kuo. 2019c. Evaluating word embedding models: methods and experimental results. *APSIPA Transactions on Signal and Information Processing*, 8:e19.

Wang, Yanshan, Sijia Liu, Naveed Afzal, Majid Rastegar-Mojarad, Liwei Wang, Feichen Shen, Paul Kingsbury, and Hongfang Liu. 2018. A comparison of word embeddings for the biomedical natural language processing. *Journal of Biomedical Informatics*, 87:12–20.

Zhao, Mengnan, Aaron J. Masino, and Christopher C. Yang. 2018. A framework for developing and evaluating word embeddings of drug-named entity. In *Proceedings of the BioNLP 2018 workshop*, pages 156–160, Melbourne, Australia. Association for Computational Linguistics.

A Benchmark Results Including Models With OOV Generation

Table 5 shows the semantic benchmark results for all models, including the FastText models with OOV generation.

Table 6 shows the syntactic benchmark results for all models, including the FastText models with OOV generation.

Table 7 shows the extrinsic benchmark results for all models, including the FastText models with OOV generation.

B In-Vocabulary Intersection Results

We report the IV intersection results which show the performance of the embeddings on the intersection of all embeddings’ IV dataset for that task. We also report relative coverage (RC) for each model as the proportion that the IV words of a model constitute out of the union of all models’ IV words.

Table 8 shows the semantic benchmark results on the intersection of IV datasets of all embeddings.

Table 9 shows the syntactic benchmark results on the intersection of IV datasets of all embeddings.

C Correctly Predicted Semantic Analogies

Table 10 shows the correctly predicted semantic analogies for all models, including the FastText models with OOV generation.

D Clinical Analogy Task Top N Accuracies

Table 11 shows the results on the clinical analogy task where a prediction is considered correct if the correct term is in the top 1, 5 and 10 nearest neighbours to the calculated vector.

E Results on Clinical Similarity Dataset With Standard Deviation at or Below 1

Table 12 shows the results on the clinical similarity dataset with standard deviation at or below 1.

F Supplementary Material

All datasets are txt files with tab-separated values. Each row has one word pair or analogy. Some datasets are divided into parts. A headline of a part is in all caps and introduced with ‘ ’.

The following datasets are attached:

- Clinical analogy dataset (txt)
- Abbreviation equality dataset (txt)
- Clinical similarity dataset (txt)
- Clinical similarity SD1 dataset (txt)
- UMNSRS similarity dataset (txt)
- UMNSRS relatedness dataset (txt)
- Verb inflection analogy dataset (txt)
- Noun inflection analogy dataset (txt)
- Adjective inflection analogy dataset (txt)

The clinical similarity ratings and their standard deviations are found in file:

- Clinical similarity ratings (xlsx)

The online sources of clinical abbreviations are found in file:

- Abbreviation sources (xlsx)

	FastText (300d)			
	Clinical		General	
	No gen.	OOV gen.	No gen.	OOV gen.
Clinical analogy, acc (IV)	0.05 (0.88)	0.04 (1.0)	<u>0.14</u> (0.54)	0.07 (1.0)
Abbreviation equality, sim (IV)	<u>0.53</u> (0.84)	0.52 (1.0)	0.27 (0.58)	0.30 (1.0)
Clinical similarity, ρ (IV)	<u>0.64</u> (0.93)	0.62 (1.0)	0.43 (0.61)	0.32 (1.0)
UMNSRS similarity, ρ (IV)	<u>0.60</u> (0.88)	0.58 (1.0)	0.30 (0.59)	0.25 (1.0)
UMNSRS relatedness, ρ (IV)	<u>0.54</u> (0.83)	<u>0.54</u> (1.0)	0.32 (0.56)	0.27 (1.0)
	GloVe (100d)			
	Clinical (No gen.)		GP (No gen.)	
Clinical analogy, acc (IV)	<u>0.08</u> (0.88)		0.06 (0.61)	
Abbreviation equality, sim (IV)	<u>0.49</u> (0.85)		0.24 (0.57)	
Clinical similarity, ρ (IV)	<u>0.56</u> (0.96)		0.34 (0.75)	
UMNSRS similarity, ρ (IV)	<u>0.41</u> (0.89)		0.18 (0.74)	
UMNSRS relatedness, ρ (IV)	<u>0.41</u> (0.84)		0.21 (0.70)	

Table 5: Semantic benchmark results by model type (FastText, GloVe), domain (clinical, general, general practitioner (GP)), and out-of-vocabulary (OOV) generation (no gen., OOV gen.). The acc metric is the accuracy on the in-vocabulary (IV) dataset. The sim metric is the average cosine similarity on the IV dataset. The ρ metric is the Spearman’s rank correlation coefficient on the IV dataset. IV rates are reported in parenthesis. We underline the best results per task by model type.

	FastText (300d)			
	Clinical		General	
	No gen.	OOV gen.	No gen.	OOV gen.
Verb inflection, acc (IV)	0.28 (0.99)	0.28 (1.0)	<u>0.69</u> (0.92)	0.66 (1.0)
Noun inflection, acc (IV)	0.19 (0.36)	0.11 (1.0)	<u>0.60</u> (0.13)	0.20 (1.0)
Adjective inflection, acc (IV)	0.16 (0.36)	0.07 (1.0)	<u>0.41</u> (0.65)	0.29 (1.0)
	GloVe (100d)			
	Clinical (No gen.)		GP (No gen.)	
Verb inflection, acc (IV)	0.21 (0.99)		0.09 (0.83)	
Noun inflection, acc (IV)	0.04 (0.39)		0.01 (0.18)	
Adjective inflection, acc (IV)	0.03 (0.47)		<u>0.04</u> (0.25)	

Table 6: Syntactic benchmark results by model type (FastText, GloVe), domain (clinical, general, general practitioner (GP)), and out-of-vocabulary (OOV) generation (no gen., OOV gen.). The acc metric is the accuracy on the in-vocabulary (IV) dataset. IV rates are reported in parenthesis. We underline the best results per task by model type.

	FastText (300d)			
	Clinical		General	
	No gen.	OOV gen.	No gen.	OOV gen.
Bleeding classification, acc	<u>0.93</u>	0.92	0.84	0.84
Department classification, acc	<u>0.83</u>	<u>0.83</u>	0.65	0.64
	GloVe (100d)			
	Clinical (No gen.)		GP (No gen.)	
Bleeding classification, acc	<u>0.90</u>		0.87	
Department classification, acc	<u>0.76</u>		0.66	

Table 7: Extrinsic benchmark results by model type (FastText, GloVe), domain (clinical, general, general practitioner (GP)), and out-of-vocabulary (OOV) generation (no gen., OOV gen.). We report accuracies on the bleeding and department classification task using the word embeddings as input. We underline the best results per task by model type.

	FastText (300d)	
	Clinical	General
Clinical analogy, accuracy (RC)	0.06 (1.0)	<u>0.14</u> (0.61)
Abbreviation equality, similarity (RC)	<u>0.55</u> (0.97)	0.27 (0.67)
Clinical similarity, ρ (RC)	<u>0.67</u> (0.98)	0.44 (0.63)
UMNSRS similarity, ρ (RC)	<u>0.57</u> (0.98)	0.28 (0.66)
UMNSRS relatedness, ρ (RC)	<u>0.52</u> (0.97)	0.29 (0.66)
	GloVe (100d)	
	Clinical	GP
Clinical analogy, accuracy (RC)	<u>0.13</u> (1.0)	0.08 (0.69)
Abbreviation equality, similarity (RC)	<u>0.60</u> (0.98)	0.25 (0.66)
Clinical similarity, ρ (RC)	<u>0.60</u> (1.0)	0.35 (0.79)
UMNSRS similarity, ρ (RC)	<u>0.40</u> (0.99)	0.15 (0.82)
UMNSRS relatedness, ρ (RC)	<u>0.40</u> (0.98)	0.23 (0.82)

Table 8: Semantic benchmark results on the intersection of IV datasets for each task by model type (FastText, GloVe) and domain (clinical, general, general practitioner (GP)). The accuracy metric is the accuracy on the dataset. The similarity metric is the average cosine similarity on the dataset. The ρ metric is the Spearman’s rank correlation coefficient on the dataset. Relative coverage (RC) is reported in parenthesis. We underline the best results per task by model type.

	FastText (300d)	
	Clinical	General
Verb inflection analogy, accuracy (RC)	0.29 (0.99)	<u>0.71</u> (0.92)
Noun inflection analogy, accuracy (RC)	0.17 (0.92)	<u>0.63</u> (0.54)
Adjective inflection analogy, accuracy (RC)	0.18 (0.55)	<u>0.54</u> (0.98)
	GloVe (100d)	
	Clinical	GP
Verb inflection analogy, accuracy (RC)	<u>0.23</u> (0.99)	0.09 (0.83)
Noun inflection analogy, accuracy (RC)	<u>0.08</u> (0.98)	0.03 (0.64)
Adjective inflection analogy, accuracy (RC)	<u>0.05</u> (0.71)	0.04 (0.38)

Table 9: Syntactic benchmark results on the intersection of IV datasets for each task by model type (FastText, GloVe) and domain (clinical, general, general practitioner (GP)). The accuracy metric is the accuracy on the dataset. Relative coverage (RC) is reported in parenthesis. We underline the best results per task by model type.

	FastText (300d)				GloVe (100d)	
	Clinical		General		Clinical	GP
	No gen.	OOV gen.	No gen.	OOV gen.	No gen.	No gen.
hoftealloplastik + knæ - knæalloplastik = hofte (hip replacement + knee - knee replacement = hip)	✓	✓	✓	✓	✓	✓
hofte + knæalloplastik - knæ = hoftealloplastik (hip + knee replacement - knee = hip replacement)	✓	✓	✓	✓	✓	✓
knæalloplastik + hofte - hoftealloplastik = knæ (knee replacement + hip - hip replacement = knee)	✓	✓	✓	✓	✓	✓
knæ + hoftealloplastik - hofte = knæalloplastik (knee + hip replacement - hip = knee replacement)	✓	✓			✓	✓
ovarier + mand - testikler = kvinde (ovaries + man - testicles = woman)	✓	✓	✓	✓	✓	✓
testikler + kvinde - ovarier = mand (testicles + woman - ovaries = man)			✓	✓		
trombocyt pool + anæmi - sag-m = trombocytopeni (thrombocyte pool + anemia - sag-m = thrombocytopenia)	✓	✓				
sag-m + trombocytopeni - trombocyt pool = anæmi (sag-m + thrombocytopenia - thrombocyte pool = anemia)	✓	✓			✓	
høretab + øjne - synstab = ører (hearing loss + eyes - visual obscuration = ears)			✓	✓	✓	
synstab + ører - høretab = øjne (visual obscuration + ears - hearing loss = eyes)			✓	✓		
mad + tørst - væske = sult (food + thirst - liquid = hunger)			✓	✓	✓	
milt + gastrektomi - mavesæk = splenektomi (spleen + Gastrectomy - stomach = splenectomy)			✓	✓	✓	
aids + borrelia - neuroborreliose = hiv (aids + borreliosis - neuroborreliosis = hiv)			✓	✓		
levothyroxin + hyperthyroidisme - thiamazol = hypothyroidisme (levothyroxine + lperthyroidism - thiamazole = hypothyroidism)			✓	✓		
respirator + nyresvigt - dialyse = respirationssvigt (respirator + renal failure - dialysis = respiratory failure)			✓	✓		
virus + dyrkning - bakterie = pcr (virus + cultivation - bacteria = pcr)					✓	
tarm + hæmoptyse - lunger = melæna (intestine + hemoptysis - lung = melena)					✓	
Kreatinin + knoglemarvsfunktion - differentialtælling = nyrefunktion (creatinine + bone marrow function - differential count = renal function)					✓	
nyrefunktion + differentialtælling - knoglemarvsfunktion = kreatinin (renal function + differential count - bone marrow function = creatinine)						✓

Table 10: Overview of the correctly predicted semantic analogies by model type (FastText, GloVe), domain (clinical, general, general practitioner (GP)), and out-of-vocabulary (OOV) generation (no gen., OOV gen.). Each analogy is presented in Danish, and the English translation is parenthesis. A mark signifies a correctly predicted analogy.

	FastText (300d)			
	Clinical		General	
	No gen.	OOV gen.	No gen.	OOV gen.
Top 1, acc	0.05	0.04	<u>0.14</u>	0.07
Top 5, acc	0.10	0.09	<u>0.27</u>	0.16
Top 10, acc	0.13	0.11	<u>0.41</u>	0.28
IV rate	0.88	1.0	0.54	1.0
	GloVe (100d)			
	Clinical (No gen.)		GP (No gen.)	
	Top 1, acc	<u>0.08</u>		
Top 5, acc	<u>0.15</u>			0.08
Top 10, acc	<u>0.21</u>			0.13
IV rate	0.88			0.61

Table 11: Top n accuracies and IV rate on the semantic clinical analogy dataset by model type (FastText, GloVe), domain (clinical, general, general practitioner (GP)), and out-of-vocabulary (OOV) generation (no gen., OOV gen.). A prediction is considered correct if the correct term is in the top n nearest neighbours to the calculated vector. The IV rate is the proportion of word pairs or analogies which are in-vocabulary. We underline the best results by model type.

	FastText (300d)			
	Clinical		General	
	No gen.	OOV gen.	No gen.	OOV gen.
Clinical similarity SD1, ρ	<u>0.60</u>	0.57	0.44	0.35
IV rate	0.94	1.0	0.61	1.0
	GloVe (100d)			
	Clinical (No gen.)		GP (No gen.)	
	Clinical similarity SD1, ρ	<u>0.54</u>		
IV rate	0.96			0.75

Table 12: Results on the clinical similarity dataset with standard deviation at or below 1 by model type (FastText, GloVe), domain (clinical, general, general practitioner (GP)), and out-of-vocabulary (OOV) generation (no gen., OOV gen.). The ρ metric is the Spearman's rank correlation coefficient on the IV dataset. The IV rate is the proportion of word pairs or analogies which are in-vocabulary. The dataset contains 255 word pairs. We underline the best result by model type.

NL-Augmenter 🐼 → 🐼

A Framework for Task-Sensitive Natural Language Augmentation

Kaustubh D. Dhole^{*†}, Varun Gangal[†], Sebastian Gehrmann[†], Aadesh Gupta[†], Zhenhao Li[†], Saad Mahamood[†], Abinaya Mahendiran[†], Simon Mille[†], Ashish Shrivastava[†], Samson Tan[†], Tongshuang Wu[†], Jascha Sohl-Dickstein[†], Jinho D. Choi[†], Eduard Hovy[†], Ondrej Dusek[†], Sebastian Ruder[†], Sajant Anand, Nagender Aneja, Rabin Banjade, Lisa Barthe, Hanna Behnke, Ian Berlot-Attwell, Connor Boyle, Caroline Brun, Marco Antonio Sobrevilla Cabezudo, Samuel Cahyawijaya, Emile Chapuis, Wanxiang Che, Mukund Choudhary, Christian Clauss, Pierre Colombo, Filip Cornell, Gautier Dagan, Mayukh Das, Tanay Dixit, Thomas Dopierre, Paul-Alexis Dray, Suchitra Dubey, Tatiana Ekeinhor, Marco Di Giovanni, Tanya Goyal, Rishabh Gupta, Louanes Hamla, Sang Han, Fabrice Harel-Canada, Antoine Honoré, Ishan Jindal, Przemyslaw K. Joniak, Denis Kleyko, Venelin Kovatchev, Kalpesh Krishna, Ashutosh Kumar, Stefan Langer, Seungjae Ryan Lee, Corey James Levinson, Hualou Liang, Kaizhao Liang, Zhexiong Liu, Andrey Lukyanenko, Vukosi Marivate, Gerard de Melo, Simon Meoni, Maxime Meyer, Afnan Mir, Nafise Sadat Moosavi, Niklas Muennighoff, Timothy Sum Hon Mun, Kenton Murray, Marcin Namysl, Maria Obedkova, Priti Oli, Nivranshu Pasricha, Jan Pfister, Richard Plant, Vinay Prabhu, Vasile Pais, Libo Qin, Shahab Raji, Pawan Kumar Rajpoot, Vikas Raunak, Roy Rinberg, Nicholas Roberts, Juan Diego Rodriguez, Claude Roux, Vasconcellos P. H. S., Ananya B. Sai, Robin M. Schmidt, Thomas Scialom, Tshephisho Sefara, Saqib N. Shamsi, Xudong Shen, Yiwen Shi, Haoyue Shi, Anna Shvets, Nick Siegel, Damien Sileo, Jamie Simon, Chandan Singh, Roman Sitelew, Priyank Soni, Taylor Sorensen, William Soto, Aman Srivastava, KV Aditya Srivatsa, Tony Sun, Mukund Varma T, A Tabassum, Fiona Anting Tan, Ryan Teehan, Mo Tiwari, Marie Tolkiehn, Athena Wang, Zijian Wang, Zijie J. Wang, Gloria Wang, Fuxuan Wei, Bryan Wilie, Genta Indra Winata, Xinyi Wu, Witold Wydmanski, Tianbao Xie, Usama Yaseen, Michael A. Yee, Jing Zhang, Yue Zhang

Abstract Data augmentation is an important method for evaluating the robustness of and enhancing the diversity of training data for natural language processing (NLP) models. In this paper, we present NL-Augmenter, a new participatory Python-based natural language (NL) augmentation framework which supports the creation of transformations (modifications to the data) and filters (data splits according to specific features). We describe the framework and an initial set of 117 transformations and 23 filters for a variety of NL tasks annotated with noisy descriptive tags. The transformations incorporate noise, intentional and accidental human mistakes, socio-linguistic variation, semantically-valid style, syntax changes, as well as artificial constructs that are unambiguous to humans. We demonstrate the efficacy of NL-Augmenter by using its transformations to analyze the robustness of popular language models. We find different models to be differently challenged on different tasks, with quasi-systematic score decreases. The infrastructure, datacards, and robustness evaluation results are publicly available on [GitHub](#) for the benefit of researchers working on paraphrase generation, robustness analysis, and low-resource NLP.

📄 El aumento de datos es un método importante para evaluar la solidez y mejorar la diversidad del entrenamiento de datos para modelos de procesamiento de lenguaje natural (NLP). 🇮🇳 इस लेख में, हम एनएल-ऑगमेंटर का प्रस्ताव करते हैं - एक नया भागीदारी पूर्वक, पायथन में बनाया गया, लैंग्वेज (एनएल) ऑगमेंटेशन फ्रेमवर्क जो ट्रांसफॉर्मेशन (डेटा में बदलाव करना) और फिल्टर (फीचर्स के अनुसार डेटा का भाग करना) के निरमान का समर्थन करता है। 🇨🇳 我们描述了NL-Augmenter框架及其初步包含的117种转换和23个过滤器,并大致标注分类了一系列可适配的自然语言任务 🇮🇳 این دگرگونی ها شامل نویز, اشتباهات عمدی و تصادفی انسانی, تنوع اجتماعی-زبانی, سبک معنایی معتبر, تغییرات نحوی و همچنین ساختارهای مصنوعی است که برای انسان ها مبهم است. 🇮🇳 NL-Augmenterpa allin kaynintam qawachiyku, tikrakuyininkunata servichikuspayku, chaywanmi qawariyku modelos de lenguaje popular nisqapa allin takyasqa kayninta. 🇮🇳 Kami menemukan model yang berbeda ditantang secara berbeda pada tugas yang berbeda, dengan penurunan skor kuasi-sistematis. Infrastruktur, kartu data, dan hasil evaluasi ketahanan dipublikasikan tersedia secara gratis di [GitHub](#) untuk kepentingan para peneliti yang mengerjakan pembuatan parafrase, analisis ketahanan, dan NLP sumber daya rendah.

*Corresponding author: kdhole@emory.edu

1 Introduction

Data augmentation, the act of creating new datapoints by slightly modifying copies or creating synthetic data based on existing data, is an important component in the robustness evaluation of models in natural language processing (NLP) and in enhancing the diversity of their training data. Most data augmentation techniques create examples through transformations of existing examples which are based on prior task-specific knowledge (Feng et al., 2021; Chen et al., 2021). Such transformations seek to disrupt model predictions or can be used as training candidates for improving regularization and denoising models, for example through consistency training (Xie et al., 2020). Figure 1 illustrates a number of possible transformations for a sample sentence.

However, the vast majority of transformations do not alter the structure of examples in drastic and meaningful ways, rendering them qualitatively less effective as potential training or test examples. Moreover, different NLP tasks may benefit from transforming different linguistic properties. Changing the word “happy” to “very happy” in an input is more relevant for sentiment analysis than for summarization. Despite this, many transformations are universally useful, for example changing places to ones from different geographic regions, or changing names to those from different cultures. Hence, a single repository that aggregates both task-specific and task-independent transformations will lower the barrier to entry for creating appropriate augmentation suites for any task.

Another advantage of supporting a broad range of transformations is the ability to capture the long-tailed nature and high diversity of surface forms of natural language (Bamman, 2017). The current paradigm of testing models on data drawn i.i.d. from long-tailed distribution results in the head of the distribution being emphasized even in the test dataset and rare phenomena implicitly ignored by aggregate performance numbers. Researchers have thus argued for more fine-grained breakdowns of results in ways that capture these under-represented groups (Mitchell et al., 2019). However, the identification of these groups depends on and benefits from different cultural backgrounds and expertise. To capture a wide range of backgrounds, we thus capitalize on the “wisdom-of-researchers” and develop NL-Augmenter in a participatory framework.

NL-Augmenter is a Python-based natural language (NL) augmentation framework that aims to enable more diverse and better characterized data during testing and training.¹ Drawing upon researchers from computational linguistics, NLP, and other related fields, we collect 117 different ways to augment data for NL tasks.

¹<https://github.com/GEM-benchmark/NL-Augmenter>

To encourage task-specific implementations, we link each transformation to a widely-used data format (e.g. text pair, a question-answer pair, etc.) along with the task types (e.g. entailment, tagging, etc.) that they support. NL-Augmenter also provides more than 23 different filters, which can be used to create input subpopulations, according to features such as input complexity, input size, etc. Unlike a transformation, the output of a filter is a boolean value, indicating whether the input meets the filter criterion, e.g., whether the input text is classified as toxic. We evaluate the robustness of four common pre-trained language models on four different tasks by testing their performance on perturbed test sets. The results demonstrate how NL-Augmenter can easily corroborate prior findings that current pre-trained models are strongly affected by small perturbations in texts. Additionally, we expect NL-Augmenter to be an effective tool for training data augmentation to develop models that are robust to diverse language characteristics.

2 Related Work

Participatory Benchmarks & Wisdom-of-Researchers Addressing the problem of under-resourced African languages in machine translation, Masakhane adopted a participatory approach to construct benchmarks for over thirty languages (Nekoto et al., 2020). Such collaborative approaches are becoming increasingly common (Cahyawijaya et al., 2022) in NLP to keep up with the rapid pace of NLP progress via benefitting from collaboration. The Generation Evaluation and Metrics benchmark (Gehrmann et al., 2021, 2022), which started the development of NL-Augmenter, is a participatory project to document and improve evaluation processes in natural language generation. BIG-bench² is a collaborative framework to collect few-shot tasks that gauge the abilities of large, pretrained language models. DynaBench (Kiela et al., 2021) iteratively evaluates models in a human-in-the-loop fashion by enabling humans to construct challenging examples. SyntaxGym (Gauthier et al., 2020) provides a platform for researchers to contribute and use evaluation sets with a focus on targeted syntactic evaluation of Language Models (LMs), particularly psycho-linguistically motivated ones. The collaboration process for NL-Augmenter is inspired by these projects allowing us to reach for a much broader scope and to collect transformations that operate on a larger variety of tasks and model types. Through our participatory approach, the lived experiences of a diverse group of individuals enable identifying and codifying an extensive list dimensions of variation

²<https://github.com/google/BIG-bench>

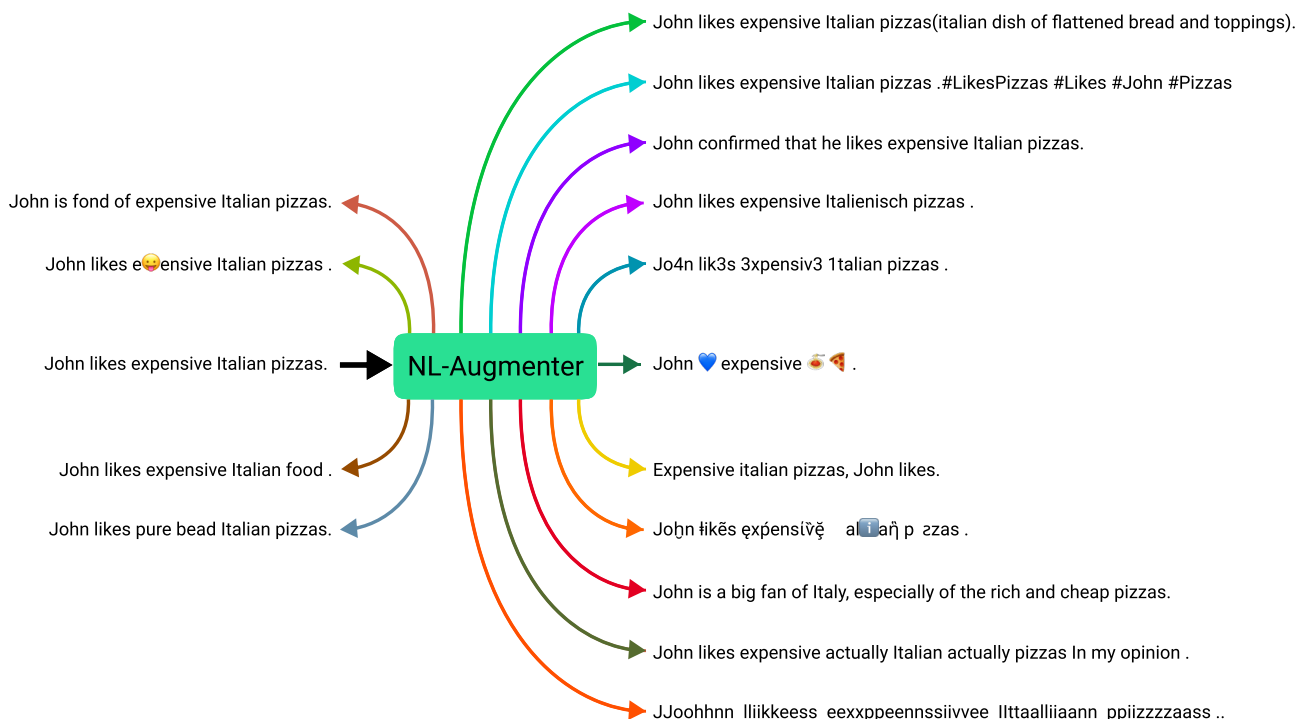


Figure 1: A few randomly chosen transformations of NL-Augmenter for the original sentence *John likes expensive pizzas*. While the meaning (almost) always remains the same and identifiable by humans, models can have a much harder time representing the transformed sentences.

which are encoded as executable transformations (Tan et al., 2021b). Leveraging the wisdom-of-the-crowd (Galton, 1907; Yi et al., 2010) is common in our field of NLP, often through the use of crowdsourcing platforms like Amazon Mechanical Turk that provide access to many raters, although not representative of the broader population (Fort et al., 2011). To harness the wisdom-of-researchers instead, we follow the example by BIG-bench which is hosted on GitHub and offers co-authorship in exchange for task contribution.

Robustness Evaluation Tools There are many projects with similar goals that inspired NL-Augmenter. For example Gardner et al. (2020) create “contrast” sets of perturbed test examples. In their approach, each example is manually perturbed, which may lead to higher-quality results but is costly to replicate for each new task due to scale and annotator cost. TextAttack (Morris et al., 2020) and TextFlint (Wang et al., 2021a) are libraries to conduct adversarial evaluations of English and Chinese models. They cover linguistic and task-specific transformations, adversarial attacks, and subpopulation analyses. In contrast, while the majority of transformations are focused on English, NL-Augmenter supports many more languages and each contribution can specify a set of supported languages.

Robustness Gym (Goel et al., 2021) unifies four different types of robustness tests — subpopulations, transformations, adversarial attacks, and evaluation sets — in a single interface in their released library. While conceptually similar, the design of NL-Augmenter puts an emphasis on modularity to enable a low barrier of entry for contributors, which is reflected in its size and diversity. Checklist (Ribeiro et al., 2020) argues for the need to go beyond simple accuracy and evaluate the model on basic linguistic capabilities, for example their response to negations. Polyjuice (Wu et al., 2021) perturbs examples using GPT-2 — though this is automatic and scalable, it offers limited control over type of challenging examples generated, making fine-grained analysis beyond global challenge-set level difficult. In contrast, our method offers a richer taxonomy with 117 (and growing) transformations for extensive analysis and comparison. Tan et al. (2021b) propose decomposing each real world environment into a set of dimensions before using randomly sampled and adversarially optimized transformations to measure the model’s average- and worst-case performance along each dimension. NL-Augmenter can be used, out-of-the-box, to measure average-case performance and we plan to extend it to support worst-case evaluation. See Table 1 for a comparison of the different libraries.

Library	#Transform.	Task-specific?	Filters?	Diversity of Resources
TextAttack	*19	✗	✗	WordNet (WD), Language Models (LM)
OpenAttack	15	✗	✗	WN, LM
NLPAug	16	✗	✗	WN, LM, PPDB
Checklist	12	✗	✗	WN, LM, Wikidata
Robustness Gym	< 20	✗	✓	WN
TextFlint	80	✓	✓	LM
NL-Augmenter	*117	✓	✓	WN, LM, Wiki, Geographies, Abbreviations, NeoPro-nouns, PropBank, Implicatives, Emojis, etc.

Table 1: Comparison of NL-Augmenter with other data augmentation and robustness evaluation libraries. *These are configurable transformations with multiple child transformations.

3 NL-Augmenter 🐛 → 🐞

NL-Augmenter is a crowd-sourced suite to facilitate rapid augmentation of data for NLP tasks to assist in training and evaluating models. NL-augmenter was introduced in (Mille et al., 2021) in the context of the creation of evaluation suites for the GEM benchmark (Gehrmann et al., 2021, 2022); three types of evaluation sets were proposed: (i) transformations, i.e. original test sets are perturbed in different ways (e.g. back-translation, introduction of typographical errors, etc.), (ii) subpopulations, i.e. test subsets filtered according to features such as input complexity, input size, etc.; and (iii) data shifts, i.e. new test sets that do not contain any of the original test set material.

In this paper, we present a participant-driven repository for creating and testing **transformations** and **filters**, and for applying them to all dataset splits (training, development, evaluation) and to all NLP tasks (NLG, labeling, question answering, etc.). As shown by Mille et al. (2021), applying filters and transformations to development/evaluation data splits allows for testing the robustness of models and for identifying possible biases; on the other hand, applying transformations and filters to training data (data augmentation) allows for possibly mitigating the detected robustness and bias issues (Wang et al., 2021b; Pruksachatkun et al., 2021; Si et al., 2021).

A majority of the augmentations that the framework supports are transformations of single sentences that aim to paraphrase these sentences in various ways. NL-Augmenter loosens the definition of “transformations” from the logic-centric view of strict equivalence to the more descriptive view of linguistics, closely resembling Bhagat and Hovy (2013)’s “quasi-paraphrases”. We extend this to accommodate noise, intentional and accidental human mistakes, socio-linguistic variation, semantically-valid style, syntax changes, as well as artificial constructs that are unambiguous to humans (Tan et al., 2021b). Some transformations vary the socio-linguistic perspective permitting a crucial source of variation wherein language

goals span beyond conveying ideas and content.

In this section, we provide organizational details, list the transformations and filters that the repository currently contains, and we present the list of tags we associated to transformations and filters and how we introduced them.

3.1 Participatory Workshop on GitHub

A workshop was organized towards constructing this full-fledged participant-driven repository. Unlike a traditional workshop wherein people submit papers, participants were asked to submit python implementations of transformations to the GitHub repository. Organizers of this workshop created a base repository extending Mille et al. (2021)’s NLG evaluation suite and incorporated a set of *interfaces*, each of which catered to popular NL example formats. This formed the backbone of the repository. A sample set of transformations and filters along with evaluation scripts were provided as starter code. Figure 2 shows an annotated code snippet of a submission. Following the format of BIG-bench’s review process, multiple review criteria were designed for accepting contributions. The review criteria (see Appendix C) guided participants to follow a style guide, incorporate test cases in JSON format, and encouraged novelty and specificity. Apart from the general software development advantages of test cases, they made reviewing simpler by providing an overview of the transformation’s capability and scope of generations.

3.2 Review Process

Each participant was expected to follow the review criteria mentioned in Figure 3 (see Appendix C). Rule-based transformations depending on well-studied lexical resources like WordNet, Wikipedia, PropBank, Implications were almost always selected due to their high precision as well as their ability to offer diverse synonymy. Machine Learning-based transformations (e.g. Transformers fine-tuned on paraphrase datasets) were

encouraged if they included either previously reported or newly measured metrics. ML-based transformations based on previously published work were thus also accepted. Duplicate submissions were rejected.

Format of a Transformation

The name of the transformation, `ReplaceFinancialAmount` followed by the interface `SentenceOperation`.

The tasks that the transformation is applicable to. The languages for which transformations are generated. And the relevant keywords which categorise the transformation.

```
class ReplaceFinancialAmount(SentenceOperation):
    tasks = [
        TaskType.TEXT_CLASSIFICATION,
        TaskType.TEXT_TO_TEXT_GENERATION,]
    languages = ["en"]
    keywords = [
        "lexical",
        "rule-based",
        "external-knowledge-based",
        "possible-meaning-alteration",
        "high-precision"]

    def __init__(self, seed: int = 0, max_outputs: int = 1):
        super().__init__(seed=seed, max_outputs=max_outputs)

    def generate(self, sentence: str) -> List[str]:

        """
        The actual logic of the transformation. The
        'generate' method takes in a sentence and returns
        multiple transformed sentences.
        """

        return transformed_sentences
```

Figure 2: Participants were expected to write their python class adhering to the above format.

Those transformations which resulted in immeasurable meaning change or untracked label changes were rejected. During the peer review, reviewers examined example outputs to decide whether a transformation had immeasurable meaning change. Reviewers were asked to instigate constructive discussions and suggest improvements to the code and the transformations. As each transformation was paired with at least 2 reviewers³ and the submissions were discussed publicly, most of these transformations had to improve & resubmit modified versions. The discussions between reviewers and participants leading up to acceptances or rejections are available publicly to encourage transparency and reproducibility as well as foster ancillary projects.

Since reviewers were the main guarantors of quality, it was imperative to provide a fair and qualitative review to participants and hence submissions were scrutinised by both participants as well as the organizers. From our initial advertising on relevant mailing lists and personally emailing authors of the relevant papers (i.e. papers focused on paraphrasing, augmentation, adversarial learning and robustness analysis) helped us in obtaining a diverse pool of volunteers. The reviewers were affiliated to about 90 organisations during the

³Some submissions also received up to 5-6 reviews.

course of review out of which approximately two-thirds were academic and the rest were industrial in nature. To ensure that the submissions adhere to the larger goals of the project we let organizers have the final say of acceptance, much like meta-reviewers in conferences.

3.3 Transformations and filters

We received a total of 170 submissions out of which 117 transformations and 23 filters were accepted and merged. They have been listed in Tables 2 and 3 respectively (and alphabetically ordered according to the submission name in the repository). For each transformation/filter, a link to the corresponding Appendix subsection is provided, where a detailed description, illustrations and an external link to the implementation in the NL-Augmenter repository can be found.

3.4 Tags for the classification of perturbations

We defined a list of tags which are useful for an efficient navigation in the pool of existing perturbations and for understanding the performance characteristics of the contributed transformations and filters (see e.g. the robustness analysis presented in Section 4). There are three main categories of tags: (i) General properties tags, (ii) Output properties tags, and (iii) Processing properties tags.

General properties tags are shown in Table 4, and cover the type of the augmentation, i.e. whether it is a transformation or a filter (*Augmented set type*), its general purpose, i.e. whether it is intended for augmentation, robustness, etc. (*General purpose*), for which NLP tasks the created data will be useful (*Task type*), to which languages it has been applied (*Language(s)*), and on which linguistic level of representation it operates, i.e. semantic, syntactic, lexical, etc. (*Linguistic level*).

Output properties tags, shown in Table 5, apply to transformations only; they provide indications about how the data was affected during the respective transformations. There are currently six properties in this category: one to capture the number of different outputs that a transformation can produce (*Output/Input ratio*), one to capture in which aspect the input and the output are alike (*Input/Output similarity*), and four to capture intrinsic qualities of the produced text or structured data, namely how were the meaning, the grammaticality, the readability and the naturalness affected by the transformation (respectively *Meaning preservation*, *Grammaticality preservation*, *Readability preservation* and *Naturalness preservation*). Note that apart from Output/Input ratio, the output properties tags need to be specified manually for each transformation/filter (see Section 3.5), and are thus subject to the interpretation of the annotator.

Transformation	App.	Transformation	App.
Abbreviation Transformation	A.1	Mix transliteration	A.60
Add Hash-Tags	A.2	MR Value Replacement	A.61
Adjectives Antonyms Switch	A.3	Multilingual Back Translation	A.62
Americanize/Britishize English	A.4	Multilingual Dictionary Based Code Switch	A.63
Antonyms Substitute	A.5	Multilingual Lexicon Perturbation	A.64
Auxiliary Negation Removal	A.6	Causal Negation and Strengthening	A.65
AzertyQwertyCharsSwap	A.7	Question Rephrasing transformation	A.66
BackTranslation	A.8	English Noun Compound Paraphraser [N+N]	A.67
BackTranslation for Named Entity Recognition	A.9	Number to Word	A.68
Butter Fingers Perturbation	A.10	Numeric to Word	A.69
Butter Fingers Perturbation For Indian Languages	A.11	OCR Perturbation	A.70
Change Character Case	A.12	Add Noun Definition	A.71
Change Date Format	A.13	Pig Latin Cipher	A.72
Change Person Named Entities	A.14	Pinyin Chinese Character Transcription	A.73
Change Two Way Named Entities	A.15	SRL Argument Exchange	A.74
Chinese Antonym and Synonym Substitution	A.16	ProtAugment Diverse Paraphrasing	A.75
Chinese Pinyin Butter Fingers Perturbation	A.17	Punctuation	A.76
Chinese Person NE and Gender Perturbation	A.18	Question-Question Paraphraser for QA	A.77
Chinese (Simplified and Traditional) Perturbation	A.19	Question in CAPS	A.78
City Names Transformation	A.20	Random Word Deletion	A.79
Close Homophones Swap	A.21	Random Upper-Case Transformation	A.80
Color Transformation	A.22	Double Context QA	A.81
Concatenate Two Random Sentences (Bilingual)	A.23	Replace Abbreviations and Acronyms	A.82
Concatenate Two Random Sentences (Monolingual)	A.24	Replace Financial Amounts	A.83
Concept2Sentence	A.25	Replace Numerical Values	A.84
Contextual Meaning Perturbation	A.26	Replace Spelling	A.85
Contractions and Expansions Perturbation	A.27	Replace nouns with hyponyms or hypernyms	A.86
Correct Common Misspellings	A.28	Sampled Sentence Additions	A.87
Country/State Abbreviation	A.29	Sentence Reordering	A.88
Decontextualisation of the main Event	A.30	Emoji Addition for Sentiment Data	A.89
Diacritic Removal	A.31	Shuffle Within Segments	A.90
Disability/Differently Abled Transformation	A.32	Simple Ciphers	A.91
Discourse Marker Substitution	A.33	Slangificator	A.92
Diverse Paraphrase Generation	A.34	Spanish Gender Swap	A.93
Dislexia Words Swap	A.35	Speech Disfluency Perturbation	A.94
Emoji Icon Transformation	A.36	Paraphrasing through Style Transfer	A.95
Emojify	A.37	Subject Object Switch	A.96
English Inflectional Variation	A.38	Sentence Summarization	A.97
English Mention Replacement for NER	A.39	Suspecting Paraphraser for QA	A.98
Filler Word Augmentation	A.40	Swap Characters Perturbation	A.99
Style Transfer from Informal to Formal	A.41	Synonym Insertion	A.100
French Conjugation Substitution	A.42	Synonym Substitution	A.101
Gender And Culture Diversity Name Changer	A.43	Syntactically Diverse Paraphrasing	A.102
Neopronoun Substitution	A.44	Subsequence Substitution for Seq. Tagging	A.103
Gender Neutral Rewrite	A.45	Tense	A.104
Gender Swapper	A.46	Token Replacement Based on Lookup Tables	A.105
GeoNames Transformation	A.47	Transformer Fill	A.106
German Gender Swap	A.48	Added Underscore Trick	A.107
Grapheme to Phoneme Substitution	A.49	Unit converter	A.108
Greetings and Farewells	A.50	Urban Thesaurus Swap	A.109
Hashtagify	A.51	Use Acronyms	A.110
Insert English and French Abbreviations	A.52	Visual Attack Letter	A.111
Leet Transformation	A.53	Weekday Month Abbreviation	A.112
Lexical Counterfactual Generator	A.54	Whitespace Perturbation	A.113
Longer Location for NER	A.55	Context Noise for QA	A.114
Longer Location Names for testing NER	A.56	Writing System Replacement	A.115
Longer Names for NER	A.57	Yes-No Question Perturbation	A.116
Lost in Translation	A.58	Yoda Transformation	A.117
Mixed Language Perturbation	A.59		

Table 2: List of transformations and link to their detailed descriptions in Appendix A

Filter	App.	Filter	App.
Code-Mixing Filter	B.1	Polarity Filter	B.13
Diacritics Filter	B.2	Quantitative Question Filter	B.14
Encoding Filter	B.3	Question type filter	B.15
Englishness Filter	B.4	Repetitions Filter	B.16
Gender Bias Filter	B.5	Phonetic Match Filter	B.17
Group Inequity Filter	B.6	Special Casing Filter	B.18
Keyword Filter	B.7	Speech-Tag Filter	B.19
Language Filter	B.8	Token-Amount filter	B.20
Length Filter	B.9	Toxicity Filter	B.21
Named-entity-count Filter	B.10	Universal Bias Filter	B.22
Numeric Filter	B.11	Yes/no question filter	B.23
Oscillatory Hallucinations Filter	B.12		

Table 3: List of filters and link to their detailed descriptions in Appendix B

Property	Definition	Tags
Augmented set type	Transformation or Filter (Subpopulation)?	Filter, Transformation, Multiple (specify), Unclear, N/A
General purpose	What will the data be used for? Augmenting training data? Testing robustness? Finding and fixing biases? Etc.	Augmentation, Bias, Robustness, Other (specify), Multiple (specify), Unclear, N/A
Task type	For which NLP task(s) will the perturbation be beneficial?	Quality estimation, Question answering, Question generation, RDF-to-text, Table-to-text generation, Sentiment analysis, Text classification, Text tagging, Text-to-text generation
Language(s)	To which language(s) is the perturbation applied?	*
Linguistic level	On which linguistic level does the perturbation operate?	Discourse, Semantic, Style, Lexical, Syntactic, Word-order, Morphological, Character, Other (specify), Multiple (specify), Unclear, N/A

Table 4: Criteria and possible tags for **General Properties** of perturbations

Processing properties tags, shown in Table 6, capture information related to the type of processing applied on the input (*Input data processing*), the type of algorithm used (*Algorithm type*), how it is implemented (*Implementation*), its estimated precision and recall (*Precision/recall*) and computational complexity (*Computational complexity / Time*), and whether an accelerator is required to apply the transformation/filter (*GPU required?*).

3.5 Tag retrieval and assignment

Transformation and filters are assigned tags for each of the properties listed in Tables 4-6. There are two sources for the tags: (i) assigning them manually, and (ii) using existing metadata embedded in the respective source code implementations of each given transformation and filter. The in-code metadata provides descriptions for each one identifiable aspects such as the language(s) supported, the type of task that the transformation or filter is applicable for, and other charac-

teristical keywords. The specification and type of this metadata was pre-defined as a requirement for all contributors to the NL-Augmenter project to enable identification of the type of transformation or filter being written by their respective author(s). Having a language tag separately was crucial to emphasize and encourage multi-lingual transformations and filters.

This metadata was initially collected through the creation of an automated script which programmatically iterated through each transformation and filter and gathered all stated metadata. The metadata was then mapped by the script into discrete property groups as defined in Tables 4-6. All contributing authors were invited to review the initially collected metadata and, where possible, add additional data.

Property	Definition	Tags
Output/input ratio	Does the transformation generate one single output for each input, or a few, or many?	=1, >1 (Low), >1 (High), Multiple (specify), Unclear, N/A
Input/output similarity	On which level are the input and output similar (if applicable)?	Aural, Meaning, Visual, Other (specify), Multiple (specify), Unclear, N/A
Meaning preservation	If you compare the output with the input, how is the meaning affected by the transformation?	Always-preserved, Possibly-changed, Always-changed, Possibly-added, Always-added, Possibly-removed, Always-removed, Multiple (specify), Unclear, N/A
Grammaticality preservation	If you compare the output with the input, how is the grammatical correctness affected by the transformation?	Always-preserved, Possibly-impaired, Always-impaired, Possibly-improved, Always-improved, Multiple (specify), Unclear, N/A
Readability preservation	If you compare the output with the input, how is the easiness of read affected by the transformation?	Always-preserved, Possibly-impaired, Always-impaired, Possibly-improved, Always-improved, Multiple (specify), Unclear, N/A
Naturalness preservation	If you compare the output with the input, how is the naturalness of the text affected by the transformation?	Always-preserved, Possibly-impaired, Always-impaired, Possibly-improved, Always-improved, Multiple (specify), Unclear, N/A

Table 5: Criteria and possible tags for **Output Properties** of perturbations (applicable to transformations only)

Property	Definition	Tags
Input data processing	What kind of NL processing is applied to the input?	Addition, Chunking, Paraphrasing, Parsing, PoS-Tagging, Removal, Segmentation, Simplification, Stemming, Substitution, Tokenisation, Translation, Other (specify), Multiple (specify), Unclear, N/A
Implementation	Is the perturbation implemented as rule-based or model-based?	Model-based, Rule-based, Both, Unclear, N/A
Algorithm type	What type of algorithm is used to implement the perturbation?	API-based, External-knowledge-based, LSTM-based, Transformer-Based, Other (specify), Multiple (specify), Unclear, N/A
Precision/recall	To what extent does the perturbation generate what it intends to generate (precision)? To what extent does the perturbation return an output for any input (recall)?	High-precision-High-recall, High-precision-Low-recall, Low-precision-High-recall, Low-precision-Low-recall, Unclear, N/A
GPU Required?	Is GPU needed to run the perturbation?	No, Yes, Unclear, N/A
Computational complexity / Time	How would you assess the computational complexity of running the perturbation? Does it need a lot of time to run?	High, Medium, Low

Table 6: Criteria and possible tags for **Processing Properties** of perturbations

4 Robustness Analysis

All authors of the accepted perturbations were asked to provide the task performance scores for each of their respective transformations or filters. In Section 4.1 we provide details on how the scores were obtained, and in Section 4.2 we provide a first analysis of these scores.

4.1 Experiment

The perturbations are currently split into three groups, according to the task(s) they will be evaluated on: text classification tasks, tagging tasks, and question-answering tasks. For experiments we focus on text classification and its relevant perturbations. We compare the models' performance on the original test data and on the perturbed data. The percentage of sentences be-

Tag	# _{All}	SST-2 Roberta-base			QQP BERT-base-unc.			MNLI Roberta-large			IMDB Roberta-base		
		# _{Evl}	R _T	Var _S	# _{Evl}	R _T	Var _S	# _{Evl}	R _T	Var _S	# _{Evl}	R _T	Var _S
Augmentation	34	20	0.63	-13.25	20	0.75	-6	18	0.74	-8.89	17	0.73	-4.41
Bias	3	1	0.5	-5	2	0.52	-11.5	2	0.53	-16	1	0.71	0
Robustness	15	8	0.82	-9.38	7	0.59	-8.14	7	0.65	-12.14	7	0.88	-13.71
Other*	1	1	0.5	-38	1	0.5	-23	1	0.5	-44	1	0.6	1
Multiple*	21	13	0.72	-4.15	13	0.64	-5.08	12	0.68	-4.08	11	0.92	-5.64
Total	74	43			43			40			37		

Table 7: Results of the robustness evaluation from the perspective of the **General purpose** criterion (#_{All} = Total number of tags, #_{Evl} Total number of evaluations collected, R_T = Transformation rate, Var_S = Score variation)

Tag	# _{All}	SST-2 Roberta-base			QQP BERT-base-unc.			MNLI Roberta-large			IMDB Roberta-base		
		# _{Evl}	R _T	Var _S	# _{Evl}	R _T	Var _S	# _{Evl}	R _T	Var _S	# _{Evl}	R _T	Var _S
Qual. estim.	2	2	0.52	-2.5	2	0.51	-6	2	0.53	-6.5	1	0.56	0
Question ans.	3	2	0.7	-0.5	2	0.89	-1.5	2	0.77	-1	2	0.98	-4
Question gen.	2	1	0.41	0	1	0.77	-1	1	0.54	-2	1	0.97	-5
RDF to text	1	1	0.01	0	1	0.02	0	1	0.04	0	1	0.21	0
Sentiment ana.	4	1	0.99	-12	1	0.99	-14	1	0.93	-18	1	1	-15
Table to text	1	1	0.01	0	1	0.02	0	1	0.04	0	1	0.21	0
Text class.	95	52	0.71	-9.27	52	0.68	-6.21	49	0.69	-8.33	43	0.83	-5.74
Text tagging	25	17	0.79	-10.94	17	0.64	-6.82	16	0.66	-9.75	13	0.84	-9.23
Text to text gen.	92	49	0.69	-8.86	49	0.66	-5.86	46	0.68	-7.57	40	0.79	-5.62
Total	231	126			126			119			103		

Table 8: Results of the robustness evaluation from the perspective of the **Task type** criterion (#_{All} = Total number of tags, #_{Evl} Total number of evaluations collected, R_T = Transformation rate, Var_S = Score variation)

Tag	# _{All}	SST-2 Roberta-base			QQP BERT-base-unc.			MNLI Roberta-large			IMDB Roberta-base		
		# _{Evl}	R _T	Var _S	# _{Evl}	R _T	Var _S	# _{Evl}	R _T	Var _S	# _{Evl}	R _T	Var _S
Semantic	3	1	1	-35	1	1	-20	1	1.0	-42	1	1	-3
Lexical	44	30	0.67	-5.83	30	0.61	-5	30	0.64	-4.4	25	0.73	-2.44
Syntactic	3	1	1	-8	1	0.74	-7	1	0.85	-15	1	1	0
Word-order	2	2	0.6	-1.5	2	0.61	-1	2	0.63	-2	1	1	0
Morphological	3	2	0.75	-25.5	2	0.75	-21.5	2	0.75	-28.5	2	0.8	-4.5
Character	6	2	1	-16.5	2	1.0	-12.5	1	0.95	-31	2	1	-26
Other*	1	1	0	0	1	0.7	-4	0			1	1	-1
Multiple*	25	9	0.74	-11.22	9	0.71	-7	9	0.74	-12.56	8	0.8	-14.5
Unclear	1	1	1	-46	1	0.79	-2	0			0		
Total	92	49			49			46			41		

Table 9: Results of the robustness evaluation from the perspective of the **Linguistic level** criterion (#_{All} = Total number of tags, #_{Evl} Total number of evaluations collected, R_T = Transformation rate, Var_S = Score variation)

ing changed by a transformation (*transformation rate*) and the percentage of performance drop on the perturbed data compared to the performance on the original data (*score variation*) are reported.

Tasks. We choose four evaluation datasets among three English NLP tasks: (1) sentiment analysis on both short sentences (SST-2 (Socher et al., 2013)) and full paragraphs (IMDB Movie Review (Maas et al.,

2011)), (2) Duplicate question detection (QQP) (Wang et al., 2019a), and (3) Natural Language Inference (MNLI) (Williams et al., 2017). These tasks cover both classifications on single sentences, as well as pairwise comparisons, and have been widely used in various counterfactual analysis and augmentation experiments (Wu et al., 2021; Kaushik et al., 2019; Gardner et al., 2020; Ribeiro et al., 2020).

Tag	# <i>All</i>	SST-2 Roberta-base			QQP BERT-base-unc.			MNLI Roberta-large			IMDB Roberta-base		
		# <i>Evl</i>	<i>R_T</i>	<i>Var_S</i>	# <i>Evl</i>	<i>R_T</i>	<i>Var_S</i>	# <i>Evl</i>	<i>R_T</i>	<i>Var_S</i>	# <i>Evl</i>	<i>R_T</i>	<i>Var_S</i>
Aural	5	3	1	-4.33	3	0.7	-6.67	2	0.7	-6.5	3	0.85	-3.67
Meaning	51	31	0.6	-8.58	32	0.64	-5.72	31	0.64	-7.52	28	0.74	-5.75
Visual	12	7	0.86	-15.29	6	0.8	-10.17	5	0.8	-12.8	5	0.92	-1
Other*	5	1	0.83	0	1	0.55	-4	1	0.69	-2	0		
Multiple*	2	1	1	-34	1	1	-20	1	1.0	-38	2	1	-23
N/A	2	2	0.92	-1	2	0.67	-6	2	0.77	-5	0		
Total	77	45			45			42			38		

Table 10: Results of the robustness evaluation from the perspective of the **Input/output similarity** criterion (#*All* = Total number of tags, #*Evl* Total number of evaluations collected, *R_T* = Transformation rate, *Var_S* = Score variation)

Evaluation models. We represent each dataset/task with its corresponding most downloaded large model hosted on Huggingface (Wolf et al., 2020), resulting in four models for evaluation: [roberta-base-sst-2](#), [roberta-base-imdb](#), [roberta-large-mnli](#), and [bert-base-uncased-qqp](#).

Perturbation strategy. For each task, we perturb a random sample of 20% of the validation set. Since all the transformations are on single text snippets, for datasets with sentence pairs, i.e., QQP and MNLI, we perturb the first question and the premise sentence, respectively.

4.2 Results and Analysis

Tables 7 to 17 show the results of the robustness analysis performed on the four datasets described in Section 4.1 and presented according to the tags introduced in Section 3.4. As we will see further, many of the tags relay interesting qualitative assessments while in some cases there is no direct correlation.

General purpose (Table 7): Transformations designed with a “robustness testing” objective displayed mean performance drops between 9% and 13.7% across models. Interestingly, 34 sentence transformations designed for “augmentation” tasks showed similar mean robustness drops ranging between 4% and 13%, emphasizing the need to draw on the paraphrasing literature to improve robustness testing.

Task type (Table 8): The results table shows that there is not necessarily a correlation between which task a transformation is marked to be relevant for and which task it actually challenges the robustness of the models on.

Linguistic level (Table 9): Transformations making character level and morphological changes were able to show drastic decreases in the level of performance compared to those making lexical or syntactic changes. These drops in performance were consistent across all four models. [roberta-large](#) finetuned on

the MNLI dataset was the most brittle - character-level transformations on an average dropped performance by over 31% and morphological changes dropped it by 28% while those which made lexical changes displayed a mean drop of 4.4%. The [visual_attack_letters \(A.111\)](#) transformation, which replaces characters with similarly looking ones (like *y* and *v*), shows a large accuracy drop from 94% to 56% on the ‘roberta-base’ model fine tuned on SST. ‘bert-base-uncased’ fine-tuned on the QQP dataset drops from 92 to 69. [roberta-large-mnli](#) drops from 91 to 47. In the case of [visual_attack_letters](#), one can easily conceive a scenario in which a model is applied to OCR text which likely exhibit similar properties. In this case, one may expect similarly poor performance, arguably attributed to a narrow set of characters that the models have been exposed to. This drop could potentially be alleviated by adversarial training. As is shown in previous work (Si et al., 2021), training on augmented data improves the performance on the test set with same perturbations.

Meaning preservation (Table 11): 22 transformations which were marked as highly meaning preserving surprisingly showed a larger average performance drop as compared to 20 of those which were marked as possibly meaning changing. Not discounting the possibility of the noisiness of the transformation’s logic, we believe further investigation could help understand whether models focus on the meaning of words or sentences or take shortcuts by focusing on commonly occurring surface forms associated with a particular prediction, as was already shown for some phenomena by [McCoy et al. \(2019\)](#), among others.

Grammaticality preservation (Table 12): Preserving grammaticality did not correlate with high robustness. Transformations marked as grammaticality always-preserved showed significant average drops of 10.6%, 8.1% and 4.6% across [roberta-base-sst-2](#), [roberta-large-mnli](#) and [bert-base-uncased-qqp](#) respectively. For example, the [grapheme_to_phoneme](#) transformation showed drastic drops in performance: 13%, 20% and 13% respectively.

Tag	# _{All}	SST-2 Roberta-base			QQP BERT-base-unc.			MNLI Roberta-large			IMDB Roberta-base		
		# _{Evl}	R _T	Var _S	# _{Evl}	R _T	Var _S	# _{Evl}	R _T	Var _S	# _{Evl}	R _T	Var _S
Alw. preserved	40	22	0.65	-9.77	22	0.63	-7.36	22	0.61	-11.23	19	0.72	-9.89
Poss. changed	33	20	0.78	-5.45	20	0.73	-5.15	17	0.75	-4.76	18	0.87	-1.5
Alw. changed	12	5	0.7	-4	5	0.54	-5.4	5	0.61	-6.8	3	0.78	-7.33
Alw. added	2	1	0	-94	1	0.7	-4	1	0.78	0	1	0.99	-1
Poss. removed	2	2	1	-18	2	1	-13	2	0.88	-23.5	1	1	-3
Total	89	50			50			47			42		

Table 11: Results of the robustness evaluation from the perspective of the **Meaning preservation** criterion (#_{All} = Total number of tags, #_{Evl} Total number of evaluations collected, R_T = Transformation rate, Var_S = Score variation)

Tag	# _{All}	SST-2 Roberta-base			QQP BERT-base-unc.			MNLI Roberta-large			IMDB Roberta-base		
		# _{Evl}	R _T	Var _S	# _{Evl}	R _T	Var _S	# _{Evl}	R _T	Var _S	# _{Evl}	R _T	Var _S
Alw. preserved	31	19	0.59	-10.58	19	0.52	-4.63	18	0.53	-8.11	17	0.76	-4.94
Poss. impaired	36	20	0.69	-3.15	20	0.69	-4.55	19	0.72	-4.21	18	0.81	-2.11
Alw. impaired	2	1	0.93	-7	1	0.94	-20	1	0.92	-16	1	1	-1
Poss. improved	6	6	0.83	-16.33	6	0.8	-8.17	5	0.79	-14.8	2	0.52	-1.5
Unclear	1	1	1	-34	1	1	-20	1	1.0	-38	1	1	-45
N/A	2	2	1	-23.5	2	1	-22	2	1	-27	2	1	-36.5
Total	79	49			49			46			41		

Table 12: Results of the robustness evaluation from the perspective of the **Grammaticality preservation** criterion (#_{All} = Total number of tags, #_{Evl} Total number of evaluations collected, R_T = Transformation rate, Var_S = Score variation)

Tag	# _{All}	SST-2 Roberta-base			QQP BERT-base-unc.			MNLI Roberta-large			IMDB Roberta-base		
		# _{Evl}	R _T	Var _S	# _{Evl}	R _T	Var _S	# _{Evl}	R _T	Var _S	# _{Evl}	R _T	Var _S
Alw. preserved	25	15	0.66	-3	15	0.54	-3.47	15	0.56	-5.53	12	0.83	-2.33
Poss. impaired	38	24	0.64	-10.67	24	0.69	-6.25	22	0.69	-6.59	22	0.79	-2.41
Alw. impaired	9	4	1	-25.25	4	1.0	-17.25	3	0.98	-36.67	4	1	-40
Poss. improved	4	4	0.75	-11.75	4	0.75	-8.75	4	0.75	-16.25	2	0.52	-1.5
Alw. improved	2	1		-1	1		-6	1	0.77	-5	0		
Unclear	1	1	1	0	1	0.06	0	1	0.15	0	1	0.32	0
Total	79	49			49			46			41		

Table 13: Results of the robustness evaluation from the perspective of the **Readability preservation** criterion (#_{All} = Total number of tags, #_{Evl} Total number of evaluations collected, R_T = Transformation rate, Var_S = Score variation)

Readability and Naturalness (Tables 13-14): In general, as expected, the transformations tagged as modifying the readability or naturalness show large drops across all tasks and models, in particular the ones tagged as “always impairing” the input.

Unsurprisingly, many of the injected perturbations, despite being artificial would not distract human readers from the actual meaning and intent of the text (e.g. `simple_ciphers` transformation (A.91)). Character-level perturbations might not distract human readers as much as compared to word-level perturbations but the above language models on the other hand behaved con-

trarily. Such departure from learning meaningful abstractions is further validated with the low correlation of grammaticality preservation and robustness. These results further re-question how we can expand these models from being just pure statistical learners to those which can incorporate meaning and surface-level abstraction, both across natural as well as artificial constructs. The large drops in performance of such perturbations necessitate looking at expanding training sets with even artificial data sources as well expand our definitions of text similarity from pure linguistic ones to those which abstract morphological, visual and other

Tag	# _{All}	SST-2 Roberta-base			QQP BERT-base-unc.			MNLI Roberta-large			IMDB Roberta-base		
		# _{Evl}	R _T	Var _S	# _{Evl}	R _T	Var _S	# _{Evl}	R _T	Var _S	# _{Evl}	R _T	Var _S
Alw. preserved	18	9	0.59	-3.33	10	0.52	-3.5	9	0.51	-7.44	9	0.75	-2.56
Poss. impaired	45	29	0.66	-8.48	29	0.64	-5.38	27	0.67	-5.15	24	0.79	-1.75
Alw. impaired	8	4	1.0	-20.5	4	1.0	-16.25	4	0.97	-23.25	4	1	-32.25
Poss. improved	4	4	0.75	-11.75	4	0.75	-8.75	4	0.75	-16.25	2	0.52	-1.5
Unclear	1	1	1	-34	1	1	-20	1	1.0	-38	1	1	-45
Total	77	47			48			45			40		

Table 14: Results of the robustness evaluation from the perspective of the **Naturalness preservation** criterion (#_{All} = Total number of tags, #_{Evl} Total number of evaluations collected, R_T = Transformation rate, Var_S = Score variation)

Tag	# _{All}	SST-2 Roberta-base			QQP BERT-base-unc.			MNLI Roberta-large			IMDB Roberta-base		
		# _{Evl}	R _T	Var _S	# _{Evl}	R _T	Var _S	# _{Evl}	R _T	Var _S	# _{Evl}	R _T	Var _S
Addition	1	1	0	-94	1	0.7	-4	1	0.78	0	1	0.99	-1
Paraphrasing	5	5	0.79	-1.8	5	0.74	-5.6	4	0.77	-6.25	3	0.77	-0.67
Parsing	1	1	0.02	0	1	0.16	-1	1	0.15	0	1	0.59	0
PoS-Tagging	5	3	0.44	-11.67	3	0.54	-6.67	3	0.54	-14.33	2	0.98	-1.5
Removal	2	2	1	-4.5	2	0.74	-6.5	2	0.81	-10	1	1	0
Segmentation	3	1	1	-4	1	0.93	-6	1	0.94	-5	1	1	-4
Substitution	17	13	0.63	-8.08	14	0.61	-8	14	0.64	-9.36	13	0.67	-5
Tokenisation	23	9	0.67	-4.89	9	0.5	-4.22	9	0.54	-4.56	10	0.76	-3.8
Translation	3	2	0.99	-11	2	0.99	-13.5	2	0.97	-18.5	1	1	-15
Other*	3	2	1	-17	2	1.0	-10	1	0.95	-38	2	1	-23
Multiple*	13	6	0.69	-1.33	5	0.6	-2.2	5	0.58	-4.8	3	0.72	-2
Unclear	1	1	1	-46	1	0.79	-2	0			0		
N/A	3	2	0.85	-18.5	2	0.9	-14	2	0.89	-20.5	2	1	-32
Total	81	48			48			45			40		

Table 15: Results of the robustness evaluation from the perspective of the **Input data processing** criterion (#_{All} = Total number of tags, #_{Evl} Total number of evaluations collected, R_T = Transformation rate, Var_S = Score variation)

Tag	# _{All}	SST-2 Roberta-base			QQP BERT-base-unc.			MNLI Roberta-large			IMDB Roberta-base		
		# _{Evl}	R _T	Var _S	# _{Evl}	R _T	Var _S	# _{Evl}	R _T	Var _S	# _{Evl}	R _T	Var _S
Model-based	19	11	0.95	-11.27	11	0.93	-7.64	9	0.93	-11.78	7	0.81	-2.43
Rule-based	66	38	0.65	-9.24	38	0.61	-6.26	37	0.64	-8.14	34	0.79	-6.5
Both	6	2	0.31	0	2	0.5	-0.5	2	0.42	-1.5	1	0.97	-5
Unclear	1	1	1	-7	1	0.84	-4	1	0.9	-2	1	1	-1
Total	103	52			52			49			43		

Table 16: Results of the robustness evaluation from the perspective of the **Implementation** criterion (#_{All} = Total number of tags, #_{Evl} Total number of evaluations collected, R_T = Transformation rate, Var_S = Score variation)

errors which can be unambiguous to humans.

Tables 10, 15, 16 and 17 show the robustness scores for **Input/Output similarity**, **Input processing**, **Implementation** and **Algorithm type** respectively. The score drops for these criteria may not be easily interpretable; e.g. that model-based implementations showed comparatively larger average drops as compared to rule-based implementations may not be due to the difference in implementation, but rather to which transformations were implemented that way.

5 Discussion and Broader Impact

Limitations In Section 4.2, we analyze the results of applying some of the transformations on existing datasets and running models on the perturbed data. Even though it was not possible to test all of the currently existing perturbations due to time constraints, the overall results show that the tested perturbations do pose a challenge to different models on different tasks, with quasi-systematic score drops. However, with so many transformations applied to four different

Tag	# _{All}	SST-2 Roberta-base			QQP BERT-base-unc.			MNLI Roberta-large			IMDB Roberta-base		
		# _{Evl}	R _T	Var _S	# _{Evl}	R _T	Var _S	# _{Evl}	R _T	Var _S	# _{Evl}	R _T	Var _S
API-based	22	14	0.78	-7.86	14	0.67	-7	13	0.73	-9.23	11	0.88	-11.45
Ext. K.-based	33	19	0.47	-11	19	0.55	-6.95	19	0.55	-7.89	20	0.68	-4.45
LSTM-based	1	1	1	0	1	1.0	0	0	0.9		1	1	-1
Transf.-based	15	7	0.89	-9.57	7	0.85	-5.29	6	0.87	-7.17	1	1	-4
Multiple*	3	1	0.41	0	1	0.77	-1	1	0.54	-2	1	0.97	-5
Unclear	1	0			0			0			1		-1
N/A	24	4	1.0	-13.25	4	0.77	-8.5	4	0.75	-18.75	3	0.89	-6
Total	103	46			46			43			38		

Table 17: Results of the robustness evaluation from the perspective of the **Algorithm type** criterion (#_{All} = Total number of tags, #_{Evl} Total number of evaluations collected, R_T = Transformation rate, Var_S = Score variation)

datasets, the presented robustness analysis can only be shallow, and a separate analysis of each transformation would be needed in order to get more informative insights. Second, our superficial analysis above relies on tags which were in many cases annotated by hand, and some of the surprising results (e.g. meaning-preserving are more challenging than non-meaning-preserving transformations) may reflect a lack of consistency in the annotations. We believe that assessing the quality of the tag assignment so as to ensure a high inter-annotator agreement will be needed for reliable analyses in the future. Finally, the current robustness analysis only shows that the perturbations are effective for detecting a possible weakness in a model; further experiments are needed to demonstrate that the perturbations can also help mitigating the weaknesses they bring to light.

Dilution of Contributions While this is not our intent, there is a risk in large scale collections of work like this that individual contributions are being less appreciated than releasing them as a standalone project. This risk is a trade-off with the advantage that it becomes much easier to switch between different transformations, which can lead to a better adoption of introduced methods. To proactively give appropriate credit, each transformation has a data card in the form of a standard README file mentioning the contributors and all participants are listed as co-authors of this paper. We further encourage all users of our repository to cite the work that a specific implementation builds on, if appropriate. The relevant citations are listed on the respective data cards and in the description in the appendix. In the same vein, there is a risk of NL-Augmenter as a whole to monopolize the augmentation space due to its large scope, leading to less usage of related work which may cover additional transformations or filters. While this is not our intention and we actively worked with contributors to related repositories to integrate their work, we encourage researchers to try other solutions as well.

Participatory Setup Conducting research in environments with a shared mission, a low barrier of entry, and directly involving affected communities was popularized by [Nekoto et al. \(2020\)](#). This kind of participatory work has many advantages, most notably that it changes the typically prescriptive research workflow toward a more inclusive one. Another advantage is that through open science, anyone can help shape the overall mission and improve the end result. Following the related BIG-bench ([Srivastava et al., 2022](#)) project, we aimed to design NL-Augmenter in a similar spirit – by providing the infrastructure, the participation barrier is reduced to filling a templated interface and providing test example. By making the interface as flexible as possible, the contributions range from filters for subpopulations with specific protected attributes to transformations via neural style transfer. Through this wide range, we hope that researchers can apply a wider range of augmentation and evaluations strategies to their data and models.

6 Conclusion

In this paper, we introduced NL-Augmenter, a framework for text transformations and filters with the goal of assisting in robustness testing and data augmentation tasks. We demonstrated that through an open participation strategy, NL-Augmenter can cover a substantially wider set of languages, tasks, transformations, and filters than existing work, without a loss of focus. Our repository provides >117 transformations and >23 filters that have been documented and tested. We used these transformations to conduct robustness evaluations of popular transformer-based models and found that they are not robust, even to randomly (i.e., non-adversarially) sampled perturbations. Although our analyses have revealed some aspects in which NL-Augmenter can be improved, we showed how it can be beneficial to efforts in evaluating the robustness of NLP models. NL-Augmenter can serve as a crucial resource for data augmentation especially for low-

resource domains and task-specific language processing. We welcome future contributions to improve its coverage of the augmentation space and to address its current shortcomings. Investigating the effect on model robustness with larger-scale experiments is a potential direction for future work.

7 Organization

NL-Augmenter is an effort organized by researchers and developers ranging across different niches of NLP. To acknowledge everyone's contributions, we list the contribution statements below for all.

Steering Committee: Kaustubh Dhole, Varun Gangal, Sebastian Gehrmann, Aadesh Gupta, Zhenhao Li, Saad Mahmood, Simon Mille, Jascha Sohl-Dickstein, Ashish Shrivastava, Samson Tan, Tongshuang Wu and Abinaya Mahendiran make up the steering committee. Jinho Choi, Eduard Hovy & Sebastian Ruder provided guidance and feedback. Kaustubh Dhole coordinates and leads the NL-Augmenter effort. All others provide feedback and discuss larger decisions regarding the direction of NL-Augmenter and act as organizers and reviewers.

Repository: Kaustubh, Aadesh, Zhenhao, Tongshuang, Ashish, Saad, Varun & Abinaya created the interfaces and the base repository NL-Augmenter for participants to contribute. This was also a continuation of the repository developed for creating challenge sets (Mille et al., 2021) for GEM (Gehrmann et al., 2021). All the other authors expanded this repository with their implementations.

Reviewers: Kaustubh, Simon, Zhenhao, Sebastian, Varun, Samson, Abinaya, Saad, Tongshuang, Aadesh, Ondrej were involved in reviewing the submissions of participants of the first phase. In the 2nd phase, all other authors performed a cross-review, in which participants were paired with 3 other participants. This was followed by a meta review by the organizers.

Robustness Evaluation: Ashish, Tongshuang, Kaustubh & Zhenhao created the evaluation engine. Simon, Kaustubh, Saad, Abinaya & Tongshuang performed the robustness analysis.

Website: Aadesh and Sebastian created the web-pages for the project.

The abstract has been written in English, Spanish, Hindi, Chinese, Persian, Quechua, and Indonesian.

References

2006. Respectful Disability Language: Heres Whats Up! https://www.aucd.org/docs/add/sa_summits/Language%20Doc.pdf.
- Bamman, David. 2017. Natural language processing for the long tail. In *DH*.
- Berard, Alexandre, Ioan Calapodescu, and Claude Roux. 2019. Naver labs europe's systems for the wmt19 machine translation robustness task. *arXiv preprint arXiv:1907.06488*.
- Bhagat, Rahul and Eduard Hovy. 2013. What is a paraphrase? *Computational Linguistics*, 39(3):463–472.
- Bhatt, Abhinav and Kaustubh D. Dhole. 2020. Benchmarking biorelex for entity tagging and relation extraction.
- Bird, Steven. 2006. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 69–72.
- Board, Smorga's. 2021. Frequently misspelled word list for dyslexia.
- Bonial, Claire, Jena Hwang, Julia Bonn, Kathryn Conger, Olga Babko-Malaya, and Martha Palmer. 2012. English propbank annotation guidelines. *Center for Computational Language and Education Research Institute of Cognitive Science University of Colorado at Boulder*, 48.
- Cahyawijaya, Samuel, Holy Lovenia, Alham Fikri Aji, Genta Indra Winata, Bryan Wilie, Rahmad Mahendra, Christian Wibisono, Ade Romadhony, Karissa Vincentio, Fajri Koto, Jennifer Santoso, David Moeljadi, Cahya Wirawan, Frederikus Hudi, Ivan Halim Parmonangan, Ika Alfina, Muhammad Satrio Wicaksono, Ilham Firdausi Putra, Samsul Rahmadani, Yulianti Oenang, Ali Akbar Septiandri, James Jaya, Kaustubh D. Dhole, Arie Ardiyanti Suryani, Rifki Afina Putri, Dan Su, Keith Stevens, Made Nindyatama Nityasya, Muhammad Farid Adilazuarda, Ryan Ignatius, Ryandito Diandaru, Tiezheng Yu, Vito Ghifari, Wenliang Dai, Yan Xu, Dyah Damapusita, Cuk Tho, Ichwanul Muslim Karo Karo, Tirana Noor Fatyanosa, Ziwei Ji, Pascale Fung, Graham Neubig, Timothy Baldwin, Sebastian Ruder, Herry Sujaini, Sakriani Sakti, and Ayu Purwarianti. 2022. Nusacrowd: Open source initiative for indonesian nlp resources.
- Chen, Jiaao, Derek Tam, Colin Raffel, Mohit Bansal, and Diyi Yang. 2021. An empirical survey of data augmentation for limited data learning in nlp.

- Dai, Xiang and Heike Adel. 2020. An analysis of simple data augmentation for named entity recognition. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3861–3867, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Damodaran, Prithiviraj. Styleformer.
- Deorowicz, Sebastian and Marcin G Ciura. 2005. Correcting spelling errors by modelling their causes. *International journal of applied mathematics and computer science*, 15:275–285.
- Dhole, Kaustubh D. 2020. Resolving intent ambiguities by retrieving discriminative clarifying questions.
- Dinan, Emily, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2020. Queens are powerful too: Mitigating gender bias in dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8173–8188, Online. Association for Computational Linguistics.
- Dolan, William B and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.
- Dopierre, Thomas, Christophe Gravier, and Wilfried Logerais. 2021. Protaugment: Unsupervised diverse short-texts paraphrasing for intent detection meta-learning. *CoRR*, abs/2105.12995.
- Eger, Steffen and Yannik Benz. 2020. From hero to zéro: A benchmark of low-level adversarial attacks.
- Eger, Steffen, Gözde Gül Şahin, Andreas Rücklé, Ji-Ung Lee, Claudia Schulz, Mohsen Mesgar, Krishnkant Swarnkar, Edwin Simpson, and Iryna Gurevych. 2019a. Text processing like humans do: Visually attacking and shielding NLP systems. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1634–1647, Minneapolis, Minnesota. Association for Computational Linguistics.
- Eger, Steffen, Gözde Gül Sahin, Andreas Rücklé, Ji-Ung Lee, Claudia Schulz, Mohsen Mesgar, Krishnkant Swarnkar, Edwin Simpson, and Iryna Gurevych. 2019b. Text processing like humans do: Visually attacking and shielding NLP systems. *CoRR*, abs/1903.11508.
- Eisner, Ben, Tim Rocktäschel, Isabelle Augenstein, Matko Bošnjak, and Sebastian Riedel. 2016. emoji2vec: Learning emoji representations from their description. In *Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media*, pages 48–54, Austin, TX, USA. Association for Computational Linguistics.
- Fadaee, Marzieh, Arianna Bisazza, and Christof Monz. 2017. Data augmentation for low-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 567–573, Vancouver, Canada. Association for Computational Linguistics.
- Fan, Angela, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.
- Fan, Angela, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Çelebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation. *ArXiv*, abs/2010.11125.
- Feng, Steven Y, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for nlp. *arXiv preprint arXiv:2105.03075*.
- Fort, Karën, Gilles Adda, and K. Bretonnel Cohen. 2011. Amazon Mechanical Turk: Gold Mine or Coal Mine? *Computational Linguistics*, 37(2):413–420.
- Galton, Francis. 1907. Vox populi (the wisdom of crowds). *Nature*, 75(7):450–451.
- Gangal, Varun, Steven Y Feng, Eduard Hovy, and Teruko Mitamura. 2021. Nareor: The narrative re-ordering problem. *arXiv preprint arXiv:2104.06669*.
- Gardner, Matt, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, et al. 2020. Evaluating models local decision boundaries via contrast sets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 1307–1323.
- Gauthier, Jon, Jennifer Hu, Ethan Wilcox, Peng Qian, and Roger Levy. 2020. Syntaxgym: An online platform for targeted evaluation of language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 70–76.

- Gehrmann, Sebastian, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Chinenye Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Mihir Kale, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi Narayan, Vitaly Nikolaev, Andre Niyongabo Rubungo, Salomey Osei, Ankur Parikh, Laura Perez-Beltrachini, Niranjan Ramesh Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, João Sedoc, Thibault Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo, Hendrik Strobelt, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, and Jiawei Zhou. 2021. The GEM benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 96–120, Online. Association for Computational Linguistics.
- Gehrmann, Sebastian, Abhik Bhattacharjee, Abinaya Mahendiran, Alex Wang, Alexandros Papanagelis, Aman Madaan, Angelina McMillan-Major, Anna Shvets, Ashish Upadhyay, Bingsheng Yao, Bryan Wilie, Chandra Bhagavatula, Chaobin You, Craig Thomson, Cristina Garbacea, Dakuo Wang, Daniel Deutsch, Deyi Xiong, Di Jin, Dimitra Gkatzia, Dragomir Radev, Elizabeth Clark, Esin Durmus, Faisal Ladhak, Filip Ginter, Genta Indra Winata, Hendrik Strobelt, Hiroaki Hayashi, Jekaterina Novikova, Jenna Kanerva, Jenny Chim, Jiawei Zhou, Jordan Clive, Joshua Maynez, João Sedoc, Juraj Juraska, Kaustubh Dhole, Khyathi Raghavi Chandu, Laura Perez-Beltrachini, Leonardo F. R. Ribeiro, Lewis Tunstall, Li Zhang, Mahima Pushkarna, Mathias Creutz, Michael White, Mihir Sanjay Kale, Moussa Kamal Eddine, Nico Daheim, Nishant Subramani, Ondrej Dusek, Paul Pu Liang, Pawan Sasanka Ammanamanchi, Qi Zhu, Ratish Puduppully, Reno Kriz, Rifat Shahriyar, Ronald Cardenas, Saad Mahamood, Salomey Osei, Samuel Cahyawijaya, Sanja tajner, Sebastien Montella, Shailza, Shailza Jolly, Simon Mille, Tahmid Hasan, Tianhao Shen, Tosin Adewumi, Vikas Raunak, Vipul Raheja, Vitaly Nikolaev, Vivian Tsai, Yacine Jernite, Ying Xu, Yisi Sang, Yixin Liu, and Yufang Hou. 2022. Gemv2: Multilingual nlg benchmarking in a single line of code.
- Gildea, Daniel and Martha Stone Palmer. 2002. The necessity of parsing for predicate argument recognition. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 239–246. ACL.
- Goel, Karan, Nazneen Rajani, Jesse Vig, Samson Tan, Jason Wu, Stephan Zheng, Caiming Xiong and Mohit Bansal, and Christopher Ré. 2021. Robustness Gym: Unifying the NLP evaluation landscape. *arXiv preprint arXiv:2101.04840*.
- Goldberg, Yoav. 2017. Neural network methods for natural language processing. *Synthesis lectures on human language technologies*, 10(1):1–309.
- Goyal, Tanya and Greg Durrett. 2020. Neural syntactic preordering for controlled paraphrase generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 238–252, Online. Association for Computational Linguistics.
- Guntuku, Sharath Chandra, Mingyang Li, Louis Tay, and Lyle H Ungar. 2019. Studying cultural differences in emoji usage across the east and the west. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 226–235.
- Gupta, Aadesh, Kaustubh D. Dhole, Rahul Tarway, Swetha Prabhakar, and Ashish Shrivastava. 2021. Candle: Decomposing conditional and conjunctive queries for task-oriented dialogue systems.
- Harel-Canada, Fabrice. 2021. Sibyl.
- Hendrickx, Iris, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Stan Szpakowicz, and Tony Veale. 2013. Semeval-2013 task 4: Free paraphrases of noun compounds. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 138–143.
- hyperreality@GitHub. American british english translator. <https://github.com/hyperreality/American-British-English-Translator>.
- Jalalzai, Hamid, Pierre Colombo, Chloé Clavel, Eric Gaussier, Giovanna Varni, Emmanuel Vignon, and Anne Sabourin. 2020. Heavy-tailed representations, text polarity classification & data augmentation. In *Advances in Neural Information Processing Systems*, volume 33, pages 4295–4307. Curran Associates, Inc.
- Jia, Robin and Percy Liang. 2017a. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017*,

- Copenhagen, Denmark, September 9-11, 2017, pages 2021–2031. Association for Computational Linguistics.
- Jia, Robin and Percy Liang. 2017b. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.
- Jindal, Ishan, Ranit Aharonov, Siddhartha Brahma, Huaiyu Zhu, and Yunyao Li. 2020. Improved semantic role labeling using parameterized neighborhood memory adaptation. *arXiv preprint arXiv:2011.14459*.
- Kaushik, Divyansh, Eduard Hovy, and Zachary C Lipton. 2019. Learning the difference that makes a difference with counterfactually-augmented data. *arXiv preprint arXiv:1909.12434*.
- Khachatrian, Hrant, Lilit Nersisyan, Karen Hambarzumyan, Tigran Galstyan, Anna Hakobyan, Arsen Arakelyan, A. Rzhetsky, and A. G. Galstyan. 2019. Biorelex 1.0: Biological relation extraction benchmark. In *BioNLP@ACL*.
- Kiela, Douwe, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online. Association for Computational Linguistics.
- Kingsbury, Paul R. and Martha Palmer. 2002. From treebank to propbank. In *Proceedings of the Third International Conference on Language Resources and Evaluation, LREC 2002, May 29-31, 2002, Las Palmas, Canary Islands, Spain*. European Language Resources Association.
- Kočišký, Tomáš, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.
- Kovatchev, Venelin, Phillip Smith, Mark Lee, and Rory Devine. 2021. Can vectors read minds better than experts? comparing data augmentation strategies for the automated scoring of children’s mindreading ability. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1196–1206, Online. Association for Computational Linguistics.
- Krishna, Kalpesh, John Wieting, and Mohit Iyyer. 2020. Reformulating unsupervised style transfer as paraphrase generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 737–762, Online. Association for Computational Linguistics.
- Kumar, Ashutosh, Satwik Bhattamishra, Manik Bhandari, and Partha Talukdar. 2019. Submodular optimization-based diverse paraphrasing and its effectiveness in data augmentation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3609–3619, Minneapolis, Minnesota. Association for Computational Linguistics.
- Lample, Guillaume, Alexis Conneau, Marc’Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *International Conference on Learning Representations*.
- Laserna, Charlyn M, Yi-Tai Seih, and James W Pennebaker. 2014. Um... who like says you know: Filler word use as a function of age, gender, and personality. *Journal of Language and Social Psychology*, 33(3):328–338.
- Lauer, Mark. 1995. *Designing Statistical Language Learners: Experiments on Noun Compounds*. Ph.D. thesis.
- Lee, Kenton, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 687–692.
- Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.
- Lhoest, Quentin, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario ako, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas

- Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander M. Rush, and Thomas Wolf. 2021a. Datasets: A community library for natural language processing.
- Lhoest, Quentin, Albert Villanova del Moral, Patrick von Platen, Thomas Wolf, Mario ako, Yacine Jernite, Abhishek Thakur, Lewis Tunstall, Suraj Patil, Mariama Drame, Julien Chaumond, Julien Plu, Joe Davison, Simon Brandeis, Victor Sanh, Teven Le Scao, Kevin Canwen Xu, Nicolas Patry, Steven Liu, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Nathan Raw, Sylvain Lesage, Anton Lozhkov, Matthew Carrigan, Théo Matussière, Leandro von Werra, Lysandre Debut, Stas Bekman, and Clément Delangue. 2021b. `huggingface/datasets`: 1.14.0.
- Li, Dianqi, Yizhe Zhang, Hao Peng, Liqun Chen, Chris Brockett, Ming-Ting Sun, and Bill Dolan. 2020a. Contextualized perturbation for textual adversarial attack. *arXiv preprint arXiv:2009.07502*.
- Li, Dianqi, Yizhe Zhang, Hao Peng, Liqun Chen, Chris Brockett, Ming-Ting Sun, and Bill Dolan. 2020b. Contextualized perturbation for textual adversarial attack. *CoRR*, abs/2009.07502.
- Li, Zhenhao and Lucia Specia. 2019. Improving neural machine translation robustness via data augmentation: Beyond back-translation. *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*.
- Lin, Bill Yuchen, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. 2019. Commongen: A constrained text generation challenge for generative commonsense reasoning. *arXiv preprint arXiv:1911.03705*.
- Liu, Zihan, Genta Indra Winata, and Pascale Fung. 2021. Continual mixed-language pre-training for extremely low-resource neural machine translation. *arXiv preprint arXiv:2105.03953*.
- Liu, Zihan, Genta Indra Winata, Zhaojiang Lin, Peng Xu, and Pascale Fung. 2020. Attention-informed mixed-language training for zero-shot cross-lingual task-oriented dialogue systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8433–8440.
- Logeswaran, Lajanugen, Honglak Lee, and Samy Bengio. 2018. Content preserving text generation with attribute controls. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 5108–5118.
- Lu, Kaiji, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2019. Gender bias in neural natural language processing.
- Ma, Edward. 2019. Nlp augmentation. <https://github.com/makcedward/nlpaug>.
- Maas, Andrew, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150.
- Marcus, Mitchell P., Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.
- Marivate, Vukosi and Tshephisho Sefara. 2020. Improving short text classification through global augmentation methods. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, pages 385–399. Springer.
- McCoy, Tom, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.
- Merriam-Webster. What is a diacritic, anyway?
- Mille, Simon, Kaustubh Dhole, Saad Mahamood, Laura Perez-Beltrachini, Varun Prashant Gangal, Mihir Kale, Emiel van Miltenburg, and Sebastian Gehrmann. 2021. Automatic construction of evaluation suites for natural language generation datasets. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. Curran.
- Miller, George A. 1998. *WordNet: An electronic lexical database*. MIT press.
- Mishra, Shubhanshu, Sijun He, and Luca Belli. 2020. Assessing demographic bias in named entity recognition. *CoRR*, abs/2008.03415.
- Mitchell, Margaret, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*, pages 220–229. ACM.

- Morris, John, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 119–126.
- Namysl, Marcin, Sven Behnke, and Joachim Köhler. 2020. NAT: Noise-aware training for robust neural sequence labeling. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1501–1517, Online. Association for Computational Linguistics.
- Namysl, Marcin, Sven Behnke, and Joachim Köhler. 2021. Empirical error modeling improves robustness of noisy neural sequence labeling. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 314–329, Online. Association for Computational Linguistics.
- Nekoto, Wilhelmina, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohunge, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Seling, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elshar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Basse, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. 2020. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online. Association for Computational Linguistics.
- Nguyen, Toan Q, Kenton Murray, and David Chiang. 2021. Data augmentation by concatenation for low-resource translation: A mystery and a solution. In *Proceedings of the International Workshop on Spoken Language Translation*, Online. Association for Computational Linguistics.
- Pais, Vasile Florian. 2019. *Contributions to semantic processing of texts; Identification of entities and relations between textual units; Case study on Romanian language*. Ph.D. thesis.
- Palmer, Martha, Paul R. Kingsbury, and Daniel Gildea. 2005. The proposition bank: An annotated corpus of semantic roles. *Comput. Linguistics*, 31(1):71–106.
- Parikh, Soham, Ananya B. Sai, Preksha Nema, and Mitesh M. Khapra. 2019. Eliminet: A model for eliminating options for reading comprehension with multiple choice questions. *CoRR*, abs/1904.02651.
- Park, Kyubyong and Seanie Lee. 2020. g2pm: A neural grapheme-to-phoneme conversion package for mandarin chinese based on a new open benchmark dataset. *CoRR*, abs/2004.03136.
- Pierse, Charles. 2021. Transformers Interpret.
- Piktus, Aleksandra, Necati Bora Edizel, Piotr Bojanowski, Edouard Grave, Rui Ferreira, and Fabrizio Silvestri. 2019. Misspelling oblivious word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3226–3234, Minneapolis, Minnesota. Association for Computational Linguistics.
- Pitler, Emily, Mridhula Raghupathy, Hena Mehta, Ani Nenkova, Alan Lee, and Aravind Joshi. 2008. Easily identifiable discourse relations. In *Coling 2008: Companion volume: Posters*, pages 87–90, Manchester, UK. Coling 2008 Organizing Committee.
- Ponkiya, Girishkumar, Rudra Murthy, Pushpak Bhattacharyya, and Girish Palshikar. 2020. Looking inside noun compounds: Unsupervised prepositional and free paraphrasing using language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 4313–4323.
- Ponkiya, Girishkumar, Kevin Patel, Pushpak Bhattacharyya, and Girish Palshikar. 2018. Treat us like the sequences we are: Prepositional paraphrasing of noun compounds using lstm. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1827–1836.
- Prasad, Rashmi, Nikhil Dinesh, Alan Lee, Eleni Milt-sakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).
- Pruksachatkun, Yada, Satyapriya Krishna, Jwala Dhamala, Rahul Gupta, and Kai-Wei Chang. 2021. Does robustness improve fairness? approaching fairness with word substitution robustness methods for

- text classification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3320–3331, Online. Association for Computational Linguistics.
- Qin, Libo, Minheng Ni, Yue Zhang, and Wanxiang Che. 2020. Cosda-ml: Multi-lingual code-switching data augmentation for zero-shot cross-lingual nlp. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3853–3860. International Joint Conferences on Artificial Intelligence Organization. Main track.
- Radford, Alec, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.
- Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer.
- Raffo, Julio. 2021. WGND 2.0.
- Raunak, Vikas, Arul Menezes, and Marcin Junczys-Dowmunt. 2021. The curious case of hallucinations in neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1172–1183, Online. Association for Computational Linguistics.
- Ravichander, Abhilasha, Siddharth Dalmia, Maria Ryskina, Florian Metze, Eduard Hovy, and Alan W Black. 2021. NoiseQA: Challenge Set Evaluation for User-Centric Question Answering. In *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Online.
- Regina, Mehdi, Maxime Meyer, and Sébastien Goutal. 2020. Text data augmentation: Towards better detection of spear-phishing emails. *CoRR*, abs/2007.02033.
- Ribeiro, Marco Tulio, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.
- Shi, Haoyue, Karen Livescu, and Kevin Gimpel. 2021. Substructure substitution: Structured data augmentation for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3494–3508, Online. Association for Computational Linguistics.
- Shi, Peng and Jimmy Lin. 2019a. Simple bert models for relation extraction and semantic role labeling. *arXiv preprint arXiv:1904.05255*.
- Shi, Peng and Jimmy Lin. 2019b. Simple BERT models for relation extraction and semantic role labeling. *CoRR*, abs/1904.05255.
- Shrivastava, Ashish, Kaustubh Dhole, Abhinav Bhatt, and Sharvani Raghunath. 2021. Saying No is An Art: Contextualized Fallback Responses for Unanswerable Dialogue Queries. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 87–92, Online. Association for Computational Linguistics.
- Shwartz, Vered and Ido Dagan. 2018. Paraphrase to explicit: Revealing implicit noun-compound relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1200–1211.
- Si, Chenglei, Zhengyan Zhang, Fanchao Qi, Zhiyuan Liu, Yasheng Wang, Qun Liu, and Maosong Sun. 2021. Better robustness by more coverage: Adversarial and mixup data augmentation for robust finetuning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1569–1576, Online. Association for Computational Linguistics.
- Smith, R. 2007. An overview of the tesseract OCR engine. In *9th International Conference on Document Analysis and Recognition (ICDAR 2007), 23-26 September, Curitiba, Paraná, Brazil*, pages 629–633. IEEE Computer Society.
- Socher, Richard, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.
- Srivastava, Aarohi, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.
- Sugiyama, Amane and Naoki Yoshinaga. 2019. Data augmentation using back-translation for context-aware neural machine translation. In *Proceedings of*

- the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 35–44, Hong Kong, China. Association for Computational Linguistics.
- Sun, Tony, Kellie Webster, Apurva Shah, William Yang Wang, and Melvin Johnson. 2021. They, them, theirs: Rewriting with gender-neutral english. *CoRR*, abs/2102.06788.
- Tan, Fiona Anting, Devamanyu Hazarika, See-Kiong Ng, Soujanya Poria, and Roger Zimmermann. 2021a. Causal augmentation for causal sentence classification. In *Proceedings of the First Workshop on Causal Inference and NLP*, pages 1–20, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Tan, Samson and Shafiq Joty. 2021. Code-mixing on sesame street: Dawn of the adversarial polyglots. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3596–3616, Online. Association for Computational Linguistics.
- Tan, Samson, Shafiq Joty, Kathy Baxter, Araz Taeihagh, Gregory A. Bennett, and Min-Yen Kan. 2021b. Reliability testing for natural language processing systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4153–4169, Online. Association for Computational Linguistics.
- Tan, Samson, Shafiq Joty, Min-Yen Kan, and Richard Socher. 2020. It’s morphin’ time! Combating linguistic discrimination with inflectional perturbations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2920–2935, Online. Association for Computational Linguistics.
- Tiedemann, Jörg and Santhosh Thottingal. 2020. OPUS-MT Building open translation services for the World. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.
- Vijayakumar, Ashwin, Michael Cogswell, Ramprasaath Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2018. Diverse beam search for improved description of complex scenes.
- Vijayakumar, Ashwin K., Michael Cogswell, Ramprasaath R. Selvaraju, Qing Sun, Stefan Lee, David J. Crandall, and Dhruv Batra. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models. *CoRR*, abs/1610.02424.
- Wang, Alex, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019a. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.
- Wang, Xiao, Qin Liu, Tao Gui, Qi Zhang, Yicheng Zou, Xin Zhou, Jiacheng Ye, Yongxin Zhang, Rui Zheng, Zexiong Pang, Qinzhuo Wu, Zhengyan Li, Chong Zhang, Ruotian Ma, Zichu Fei, Ruijian Cai, Jun Zhao, Xingwu Hu, Zhiheng Yan, Yiding Tan, Yuan Hu, Qiyuan Bian, Zhihua Liu, Shan Qin, Bolin Zhu, Xiaoyu Xing, Jinlan Fu, Yue Zhang, Minlong Peng, Xiaoping Zheng, Yaqian Zhou, Zhongyu Wei, Xipeng Qiu, and Xuanjing Huang. 2021a. TextFlint: Unified multilingual robustness evaluation toolkit for natural language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 347–355, Online. Association for Computational Linguistics.
- Wang, Yuxuan, Wanxiang Che, Jiang Guo, Yijia Liu, and Ting Liu. 2019b. Cross-lingual bert transformation for zero-shot dependency parsing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5721–5727.
- Wang, Yuxuan, Wanxiang Che, Ivan Titov, Shay B. Cohen, Zhilin Lei, and Ting Liu. 2021b. A closer look into the robustness of neural dependency parsers using better adversarial examples. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2344–2354, Online. Association for Computational Linguistics.
- Wei, Jason W. and Kai Zou. 2019. EDA: easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 6381–6387. Association for Computational Linguistics.
- Wieting, John and Kevin Gimpel. 2017. Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In *arXiv preprint arXiv:1711.05732*.
- Wieting, John, Jonathan Mallinson, and Kevin Gimpel. 2017. Learning paraphrastic sentence embeddings


- from back-translated bitext. In *Proceedings of Empirical Methods in Natural Language Processing*.
- Williams, Adina, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.
- Wilson, Steven, Walid Magdy, Barbara McGillivray, Kiran Garimella, and Gareth Tyson. 2020. Urban dictionary embeddings for slang NLP applications. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4764–4773, Marseille, France. European Language Resources Association.
- Wiseman, Sam and Alexander M. Rush. 2016. Sequence-to-sequence learning as beam-search optimization. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1296–1306, Austin, Texas. Association for Computational Linguistics.
- Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Wu, Tongshuang, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel S Weld. 2021. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*.
- Xie, Qizhe, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*, 33.
- Xu, Liang, Qianqian Dong, Cong Yu, Yin Tian, Weitang Liu, Lu Li, and Xuanwei Zhang. 2020. Cluener2020: Fine-grained name entity recognition for chinese. *arXiv preprint arXiv:2001.04351*.
- Yaseen, Usama and Stefan Langer. 2021. Data augmentation for low-resource named entity recognition using backtranslation. *CoRR*, abs/2108.11703.
- Yi, Sheng Kung, Mark Steyvers, Michael Lee, and Matthew Dry. 2010. Wisdom of the crowds in minimum spanning tree problems. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 32.
- Yorke, Alex. butter-fingers. <https://github.com/alexeyorke/butter-fingers>.
- Yunfei. Chinese-Names-Corpus. <https://github.com/wainshine/Chinese-Names-Corpus>.
- Zhang, Jing, Bonggun Shin, Jinho D Choi, and Joyce C Ho. 2021. Smat: An attention-based deep learning solution to the automation of schema matching. In *European Conference on Advances in Databases and Information Systems*, pages 260–274. Springer.
- Zhang, Wei Emma, Quan Z. Sheng, and Ahoud Abdulrahmn F. Alhazmi. 2019a. Generating textual adversarial examples for deep learning models: A survey. *CoRR*, abs/1901.06796.
- Zhang, Yuan, Jason Baldridge, and Luheng He. 2019b. PAWS: paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 1298–1308. Association for Computational Linguistics.
- Zhao, Zhe, Hui Chen, Jinbin Zhang, Xin Zhao, Tao Liu, Wei Lu, Xi Chen, Haotang Deng, Qi Ju, and Xiaoyong Du. 2019. Uer: An open-source toolkit for pre-training models. *EMNLP-IJCNLP 2019*, page 241.

A Transformations

The following is the list of all accepted transformations to NL-Augmenter. Many of the transformations tokenize sentences using SpaCy⁴ or NLTK (Bird, 2006) tokenizers. We discuss each implementation along with their limitations. The title of each transformation subsection is clickable and redirects to the actual python implementation. Many of the transformations use external libraries and we urge readers to look at each implementation and its corresponding ‘requirements.txt’ files.

A.1 Abbreviation Transformation

This transformation replaces a word or phrase with its abbreviated counterpart “homework” -> “hwk” using a web-scraped slang dictionary.⁵

 You → **yu** driving at **80 miles per hour** → **mph** is why insurance **is** → **tis** so **freaking** → **friggin** expensive.

⁴<https://spacy.io/>

⁵Scraped from <https://www.noslang.com/dictionary>

A.2 Add Hash-Tags

This transformation uses words in the text to generate hashtags. These hashtags are then appended to the original text. Using the same words appearing in the sentence to generate the hashtags acts as redundant noise that models should learn to ignore. Hashtags are widespread in social media channels and are used to draw attention to the source text and also as a quick stylistic device.

☞ I love domino's pizza. →
#LovePizza #Love #I #Pizza

A.3 Adjectives Antonyms Switch

This transformation switches English adjectives in a sentence with their WordNet (Miller, 1998) antonyms to generate new sentences with possibly different meanings and can be useful for tasks like Paraphrase Detection, Paraphrase Generation, Semantic Similarity, and Recognizing Textual Entailment.

☞ Amanda's mother was very beautiful → ugly .

A.4 AmericanizeBritishizeEnglish

This transformation takes a sentence and tries to convert it from British English to American English and vice-versa. A select set of words have been taken from [hyperreality@GitHub](#).

☞ I love the pastel colours → colors

A.5 AntonymsSubstitute

This transformation introduces semantic diversity by replacing an even number of adjective/adverb antonyms in a given text. We assume that an even number of antonyms transforms will revert back sentence semantics; however, an odd number of transforms will revert the semantics. Thus, our transform only applies to the sentence that has an even number of revertible adjectives or adverbs. We called this mechanism double negation.

☞ Steve is able → unable to recommend movies that depicts the lives of beautiful → ugly minds.

A.6 Auxiliary Negation Removal

This is a low-coverage transformation which targets sentences that contain negations. It removes negations in English auxiliaries and attempts to generate new sentences with the opposite meaning.

☞ Ujjal Dev Dosanjh was not → Ujjal Dev Dosanjh was the 1st Premier of British Columbia from 1871 to 1872.

A.7 AzertyQwertyCharsSwap

☞ Preferably use the above download link, as the release tarballs are generated deterministically → qre generqted deterministicqllly whereas GitHub's are not.

A.8 BackTranslation

This transformation translates a given English sentence into German and back to English. This transformation acts like a light paraphraser. Multiple variations can be easily created via changing parameters like the language as well as the translation models which are available in plenty. Backtranslation has been quite popular now and has been a quick way to augment examples (Li and Specia, 2019; Sugiyama and Yoshinaga, 2019).

☞ Andrew finally returned → eventually gave Chris the French book the French book I bought last week.

A.9 BackTranslation for Named Entity Recognition

This transformation splits the token sequences into segments of entity mention(s) and "contexts" around the entity mention(s). Backtranslation is used to paraphrase the contexts around the entity mention(s), thus resulting in a different surface form from the original token sequence. The resultant tokens are also assigned new tags. Exploiting this transformation has shown to empirically benefit named entity tagging (Yaseen and Langer, 2021) and hence could arguably benefit other low-resource tagging tasks (Bhatt and Dhole, 2020; Khachatryan et al., 2019; Gupta et al., 2021).

A.10 Butter Fingers Perturbation

This perturbation adds noise to all types of text sources (sentence, paragraph, etc.) proportional to noise erupting from keyboard typos making common spelling errors. Few letters picked at random are replaced with letters which are at keyboard positions near the source letter. The implementation has been borrowed from here (Yorke) as used in (Mille et al., 2021). There has also been some recent work in NoiseQA (Ravichander et al., 2021) to mimick keyboard typos.

☞ Sentences → Senhences with gapping, such as Paul likes coffee → coffwe and Mary tea, lack an overt predicate to indicate → indicatx the relation → relauion between two or more arguments → argumentd .

A.11 Butter Fingers Perturbation For Indian Languages

This implements the butter fingers perturbation as used above for 7 Indian languages: Bangla, Gujarati, Hindi, Kannada, Malayalam, Oriya, Punjabi, Tamil, and Telugu. The implementation considers the InScript keyboard ⁶ which is decreed as a standard for Indian scripts.

A.12 Change Character Case

This transformation acts like a perturbation and randomly swaps the casing of some of the letters. The transformation's outputs will not work with uncased models or languages without casing.

☞ Alice in Wonderland is a 2010 American live- action → actIon / animated → anImated dark fantasy → faNtasy adventure film.

A.13 Change Date Format

This transformation changes the format of dates.

☞ The first known case of COVID-19 was identified in Wuhan, China in December → Dec 2019.

A.14 Change Person Named Entities

This perturbation changes the name of the person from one name to another by making use of the lexicon of person names in Ribeiro et al. (2020).

☞ Andrew → Nathaniel finally returned the French book to Chris that I bought last week

A.15 Change Two Way Named Entities

This perturbation also changes the name of the person but also makes a parallel change in the label or reference text with the same name making it useful for text-to-text generation tasks.

☞ He finally returned the French book to Chris → Austin that I bought last week

A.16 Chinese Antonym and Synonym Substitution

This transformation substitutes Chinese words with their synonyms or antonyms by using the Chinese dic-

⁶https://en.wikipedia.org/wiki/InScript_keyboard

tionary⁷ and NLP Chinese Data Augmentation dictionary⁸.

A.17 Chinese Pinyin Butter Fingers Perturbation

This transformation implements the Butter Fingers Perturbation for Chinese characters. Few Chinese words and characters that are picked at random will be substituted with others that have similar pinyin (based on the default Pinyin keyboards in Windows and Mac OS). It uses a database of 16142 Chinese characters⁹ and its associated pinyins to generate the perturbations for Chinese characters. A smaller database of 3500¹⁰ more frequently seen Chinese characters are also used in the perturbations with a higher probability of being used compared to less frequently seen Chinese characters. It also uses a database of 575173 words¹¹ that are combined from several sources¹² in order to generate perturbations for Chinese words.

A.18 Chinese Person Named Entities and Gender Perturbation

This perturbation adds noise to all types of text sources containing Chinese names (sentence, paragraph, etc.) by swapping a Chinese name with another Chinese name whilst also allowing the possibility of gender swap. CLUENER (Xu et al., 2020; Zhao et al., 2019) is used for tagging named entities in Chinese. The list of names is taken from the Chinese Names Corpus! (Yunfei). It can provide assistance in detecting biases present in language models and the ability to infer implicit gender information when presented with gender-specific names. This can also be useful in mitigating representation biases in the input text.

A.19 Chinese (Simplified & Traditional) Perturbation

This perturbation adds noise to all types of text sources containing Chinese words and characters (sentence, paragraph, etc.) by changing the words and characters between Simplified and Traditional Chinese as well as other variants of Chinese Characters such as Japanese Kanji, character-level and phrase-level conversion, character variant conversion and regional idioms among Mainland China, Taiwan and Hong

⁷Chinese Dictionary: https://github.com/guotong1988/chinese_dictionary

⁸NLP Chinese Data Augmentation: <https://github.com/425776024/nlpda>

⁹<https://github.com/pwxcoo/chinese-xinhua>

¹⁰<https://github.com/elephantnose/characters>

¹¹<http://thuoc1.thunlp.org/>

¹²https://github.com/fighting411love/Chinese_from_dongxiexidian

Kong, all available as configurations originally in the OpenChineseConvert project ¹³.

A.20 City Names Transformation

This transformation replaces instances of populous and well-known cities in Spanish and English sentences with instances of less populous and less well-known cities to help reveal demographic biases (Mishra et al., 2020) prevalent in named entity recognition models. The choice of cities have been taken from the World Cities Dataset ¹⁴.

☞ The team was established in **Dallas** → **Viera West** in 1898 and was a charter member of the NFL in 1920.

A.21 Close Homophones Swap

Humans are generally guided by their senses and are unconsciously robust against phonetic attacks. Such types of attacks are highly popular in languages like English which has an irregular mapping between pronunciation and spelling (Eger and Benz, 2020). This transformation mimics writing behaviors where users swap words with similar homophones either intentionally or by accident. This transformation acts like a perturbation to test robustness. Few words picked at random are replaced with words with similar homophones which sound similar or look similar. Some of the word choices might not be completely natural to normal human behavior, since humans "prefer" some words over others even they sound exactly the same. So it might not be fully reflecting the natural distribution of intentional or unintentional swapping of words.

☞ Sentences with gapping, such as Paul likes coffee and Mary **tea** → **Tee**, lack an overt predicate to indicate **the** → **Thee** relation between two or **more** → **Morr** arguments.

A.22 Color Transformation

This transformation augments the input sentence by randomly replacing mentioned colors with different ones from the 147 extended color keywords specified by the World Wide Web Consortium (W3C) ¹⁵. Some of the colors include “dark sea green”, “misty rose”, “burly wood”.

☞ Tom bought 3 apples, 1 **orange** → **misty rose**, and 4 bananas and paid \$10.

¹³<https://github.com/BYVoid/OpenCC>

¹⁴<https://www.kaggle.com/juanmah/world-cities>

¹⁵<https://www.w3.org/TR/2021/REC-css-color-3-20210805/>

A.23 Concatenate Two Random Sentences (Bilingual)

Given a dataset, this transformation concatenates a sentence with a previously occurring sentence as explained in (Nguyen et al., 2021). A monolingual version is mentioned in the subsequent subsection below. This concatenation would benefit all text tasks that use a transformer (and likely other sequence-to-sequence architectures). Previously published work (Nguyen et al., 2021) has shown a large gain in performance of low-resource machine translation using this method. In particular, the learned model is stronger due to being able to see training data that has context diversity, length diversity, and (to a lesser extent) position shifting.

A.24 Concatenate Two Random Sentences (Monolingual)

This is the monolingual counterpart of the above.

☞ I am just generating a very very very long sentence to make sure that the method is able to handle it. It does not even need to be a sentence. Right? This is not splitting on punctuation... I am just generating a very very very long sentence to make sure that the method is able to handle it. It does not even need to be a sentence. Right? This is not splitting on punctuation...

A.25 Concept2Sentence

This transformation intakes a sentence, its associated integer label, and (optionally) a dataset name that is supported by huggingface/datasets (Lhoest et al., 2021a,b). It works by extracting keyword concepts from the original sentence, passing them into a BART (Lewis et al., 2020) transformer trained on CommonGen (Lin et al., 2019) to generate a new, related sentence which reflects the extracted concepts. Providing a dataset allows the function to use transformers-interpret (Pierse, 2021) to identify the most critical concepts for use in the generative step. Underneath the hood, this transform makes use of the Sibyl tool (Harel-Canada, 2021), which is capable of also transforming the label as well. However, this particular implementation of C2S generates new text that is invariant (INV) with respect to the label. Since the model is trained on CommonGen, which is focussed on image captioning, the style of the output sentence would be geared towards scenic descriptions and might not necessarily adhere to the syntax of the original sentence. Besides, it can be hard to argue that a handful subset of keywords could provide a complete description of the original sentence.

A.26 Contextual Meaning Perturbation

This transformation was designed to model the "Chinese Whispers" or "Telephone" children's game: The transformed sentence appears fluent and somewhat logical, but the meaning of the original sentence might not be preserved. To achieve logical coherence, a pre-trained language model is used to replace words with alternatives that match the context of the sentence. Grammar mistakes are reduced by limiting the type of words considered for changes (based on POS tagging) and replacing adjectives with adjectives, nouns with nouns, etc. where possible.

This transformation benefits users who seek perturbations that preserve fluency but not the meaning of the sentence. For instance, it can be used in scenarios where the meaning is relevant to the task, but the model shows a tendency to over-rely on simpler features such as the grammatical correctness and general coherence of the sentence. A real-world example would be the training of quality estimation models for machine translation (does the translation maintain the meaning of the source?) or for text summarisation (does the summary capture the content of the source?).

Word substitution with pre-trained language models has been explored in different settings. For example, the augmentation library `nlpaug` (Ma, 2019) and the adversarial attack library `TextAttack` (Morris et al., 2020) include contextual perturbation methods. However, their implementations do not offer control over the type of words that should be perturbed and introduce a large number of grammar mistakes. If the aim is to change the sentence's meaning while preserving its fluency, this transformation can help to get the same effect with significantly fewer grammatical errors. Li et al. (2020a) propose an alternative approach to achieve a similar objective.

A.27 Contractions and Expansions Perturbation

This perturbation substitutes the text with popular expansions and contractions, e.g., "I'm" is changed to "I am" and vice versa. The list of commonly used contractions & expansions and the implementation of perturbation has been taken from Checklist (Ribeiro et al., 2020).

☞ He often does **n't** → **not** come to school.

A.28 Correct Common Misspellings

This transformation acts like a lightweight spell-checker and corrects common misspellings appearing in text by looking for words in Wikipedia's Lists of Common Misspellings.

☞ Andrew **andd** → **and** Alice finally **returnd** → **returned** the French book that I bought **lastr** → **last** week

A.29 Country/State Abbreviation

This transformation replaces country and state names with their common abbreviations¹⁶. Abbreviations can be common across different locations: ☞ "MH" can refer to Country Meath in Ireland as well as the state of Maharashtra in India and hence this transformation might result in a slight loss of information, especially if the surrounding context doesn't have enough signals.

☞ One health officer and one epidemiologist have boarded the ship in San Diego, **CA** → **California** on April 13, 2015 to conduct an environmental health assessment.

A.30 Decontextualisation of the main Event

Semantic Role Labelling (SRL) is a powerful shallow semantic representation to determine who did what to whom, when, and where (and why and how etc). The core arguments generally talk about the participants involved in the event. Additionally, contextual arguments on the other hand provide more specific information about the event. After tagging a sentence with an appropriate semantic role labels using an SRL labeller (Jindal et al., 2020; Shi and Lin, 2019a). This transformation crops out contextual arguments to create a new sentence with a minimal description of the event. Helping to generate textual pairs for entailment.

A.31 Diacritic Removal

"Diacritics are marks placed above or below (or sometimes next to) a letter in a word to indicate a particular pronunciation in regard to accent, tone, or stress as well as meaning, especially when a homograph exists without the marked letter or letters." Merriam-Webster. This transformation removes these diacritics or accented characters, and replaces them with their non-accented versions. It can be common for non-native or inexperienced speakers to miss out on any accents and specify non-accented versions.

☞ She **lookèd** → **looked** east an she **lookèd** → **looked** west.

¹⁶Countries States Cities Database: <https://github.com/dr5hn/countries-states-cities-database>

A.32 Disability/Differently Abled Transformation

Disrespectful language can make people feel excluded and represent an obstacle towards their full participation in the society (Res, 2006). This low-coverage transformation substitutes outdated references to references of disabilities with more appropriate and respectful ones which avoid negative connotations. A small list of inclusive words and phrases have been taken from a public article on [inclusive communication](#), Wikipedia's list of [disability-related terms](#) with negative connotations, [terms to avoid while writing about disability](#).

☞ They are **deaf** → **person or people with a hearing disability**.

A.33 Discourse Marker Substitution

This perturbation replaces a discourse marker in a sentence by a semantically equivalent marker. Previous work has identified discourse markers that have low ambiguity (Pitler et al., 2008). This transformation uses the corpus analysis on PDTB 2.0 (Prasad et al., 2008) to identify discourse markers that are associated with a discourse relation with a chance of at least 0.5. Then, a marker is replaced with a different marker that is associated to the same semantic class.

☞ It has plunged 13% **since** → **inasmuch as** July to around 26 cents a pound. A year ago ethylene sold for 33 cents

A.34 Diverse Paraphrase Generation Using SubModular Optimization and Diverse Beam Search

This transformation generates multiple paraphrases of a sentence by employing 4 candidate selection methods on top of a base set of backtranslation models. 1) DiPS (Kumar et al., 2019) 2) Diverse Beam Search (Vijayakumar et al., 2018) 3) Beam Search (Wiseman and Rush, 2016) 4) Random. Unlike beam search which generally focusses on the top-k candidates, DiPS introduces a novel formulation of using submodular optimisation to focus on generating more diverse paraphrases and has been proven to be an effective data augementer for tasks like intent recognition and paraphrase detection (Kumar et al., 2019). Diverse Beam Search attempts to generate diverse sequences by employing a diversity promoting alternative to the classical beam search (Wiseman and Rush, 2016).

A.35 Dislexia Words Swap

This transformation acts like a perturbation by altering some words of the sentences with abberations (Board, 2021) that are likely to happen in the context of dyslexia.

☞ Biden hails **your** → **you're** relationship with Australia just days after new partnership drew ire from France.

A.36 Emoji Icon Transformation

This transformation converts emojis into their equivalent keyboard format (e.g., 😊 -> ":)") and vice versa (e.g., ":)" -> 😊).

A.37 Emojify

This transformation augments the input sentence by swapping words with emojis of similar meanings. Emojis, introduced in 1997 as a set of pictograms used in digital messaging, have become deeply integrated into our daily communication. More than 10% of tweets¹⁷ and more than 35% of Instagram posts¹⁸ include one or more emojis in 2015. Given the ubiquitousness of emojis, there is a growing body of work researching the linguistic and cultural aspects of emojis (Guntuku et al., 2019) and how we can leverage the use of emojis to help solve NLP tasks (Eisner et al., 2016).

☞ Apple is looking at buying U.K. startup for \$132 billion. → 🍏 is 🙄 at 🛍️ 🇬🇧 startup for \$ **1** **3** **2**.

A.38 English Inflectional Variation

This transformation adds inflectional variation to English words and can be used to test the robustness of models against inflectional variations. In English, each inflection generally maps to a Part-Of-Speech tag¹⁹ in the Penn Treebank (Marcus et al., 1993). For each content word in the sentence, it is first lemmatised before randomly sampling a valid POS category and reflecting the word according to the new category. The sampling process for each word is constrained using its POS tag to maintain the original sense for polysemous words. This has been adapted from the Morpheus (Tan et al., 2020) adversarial attack.

☞ Ujjal Dev Dosanjh **served** → **serve** as 33rd **Premier** → **Premiers** of British Columbia from **2000** to **2001**

¹⁷https://blog.twitter.com/en_us/a/2015/emoji-usage-in-tv-conversation

¹⁸<https://instagram-engineering.com/>

¹⁹Penn TreeBank POS

A.39 English Mention Replacement for NER

This transformation randomly swaps an entity mention with another entity mention of the same entity type. Exploiting this transformation as a data augmentation strategy has been empirically shown to improve the performance of underlying (NER) models (Dai and Adel, 2020).

A.40 Filler Word Augmentation

This augmentation adds noise in the form of colloquial filler phrases. 23 different phrases are chosen across 3 different categories: general filler words and phrases ("uhm", "err", "actually", "like", "you know"...), phrases emphasizing speaker opinion/mental state ("I think/believe/mean", "I would say"...) & phrases indicating uncertainty ("maybe", "perhaps", "probably", "possibly", "most likely"). The latter two categories had shown promising results Kovatchev et al. (2021) when they were concatenated at the beginning of the sentence unlike this implementation which perform insertions at any random positions. Filler words are based on the work of Laserna et al. (2014) but have not been explored in the context of data augmentation.

A.41 Style Transfer from Informal to Formal

This transformation transfers the style of text from formal to informal and vice versa. It uses the implementation of Styleformer (Damodaran).

☞ What you upto → currently doing ?

A.42 French Conjugation Substitution

This transformation change the conjugation of verbs for simple french sentences with a specified tense. It detects the pronouns used in the sentence in order to conjugate accordingly whenever a sentence contains differents verbs. This version only works for indicative tenses. It also only works for simple direct sentences (subject, verb, COD/COI), which contains a pronoun as subject (il, elle, je etc.). It does not detect when the subject is a couple of nouns ("les enfants" or "la jeune femme").

A.43 Gender And Culture Diversity Name Changer (1-way and 2-way)

Corpora exhibits many representational biases and this transformation focuses on one particular mediator, the personal names. It diversifies names in the corpora along two critical dimensions, gender and cultural background. Technically, the transformation samples

a (country, gender) pair and then randomly draws a name from that (country, gender) pair to replace the original name. We collected 42812 distinct names from 141 countries. They are primarily from the World Gender Name Dictionary (Raffo, 2021).

Common name augmentations do not consider their gender and cultural implication. Thus, they do not necessarily mitigate biases or promote the minority's representation because the augmented name may be from the same gender and cultural background. This is the case, for example in the CheckList's (Ribeiro et al., 2020) implemented name augmentation. Taking the interaction of the names therein with ours, 34.0%, 33.5%, 31.9%, 30.8% of them are popular names in US, Canada, Australia, and UK, respectively. Only 0.4%, 0.4%, 0.5%, 2.1% of them are from India, Korea, China, and Kazakhstan.

☞ Rachel → Charity Green, a sheltered but friendly woman, flees her wedding day and wealthy yet unfulfilling life.

A.44 Neopronoun Substitution

This transformation performs grammatically correct substitution from English to English of the gendered pronouns, he/she, in a given sentence with their neopronoun counterparts, based on a list compiled by UNC Greensboro and LGBTA WIKI²⁰. NLP models, such as those for neural machine translation, often fail to recognize the neopronouns and treat them as proper nouns. This transformation seeks to render the training data used in NLP pipelines more neopronoun aware to reduce the risk of trans-erasure. The reason why a simple look-up-table approach might not work is due to the fact that the case may differ depending on the context.

☞ She → They had her → their friends tell her → them about the event.

A.45 Gender Neutral Rewrite

This transformation involves rewriting an English sentence containing a single gendered entity with its gender-neutral variant. One application is machine translation, when translating from a language with gender-neutral pronouns (e.g. Turkish) to a language with gendered pronouns (e.g. English). This transformation is based on the algorithm proposed by Sun et al. (2021).

☞ His → Their dream is to be a fireman → firefighter when he → they grows → grow up.

²⁰<https://intercultural.uncg.edu/wp-content/uploads/Neopronouns-Explained-UNCG-Intercultural-Engagement.pdf>

A.46 GenderSwapper

This transformation introduces gender diversity to the given data. If used as data augmentation for training, the transformation might mitigate gender bias, as shown in [Dinan et al. \(2020\)](#). It also might be used to create a gender-balanced evaluation dataset to expose the gender bias of pre-trained models. This transformation performs lexical substitution of the opposite gender. The list of gender pairs (shepherd \leftrightarrow shepherdess) is taken from [Lu et al. \(2019\)](#). Genderwise names used from [Ribeiro et al. \(2020\)](#) are also randomly swapped.

A.47 GeoNames Transformation

This transformation augments the input sentence with information based on location entities (specifically cities and countries) available in the GeoNames database²¹. E.g., if a country name is found, the name of the country is appended with information about the country like its capital city, its neighbouring countries, its continent, etc. Some initial ideas of this nature were explored in [Pais \(2019\)](#).

A.48 German Gender Swap

This transformation replaces the masculine nouns and pronouns with their female counterparts for German sentences from a total of 2226 common German names.²²

☞ Er → Sie ist ein Arzt → eine Ärztin
und mein Vater → meine Mutter .

A.49 Grapheme to Phoneme Substitution

This transformation adds noise to a sentence by randomly converting words to their phonemes. Grapheme-to-phoneme substitution is useful in NLP systems operating on speech. An example of grapheme to phoneme substitution is “permit” → P ERØ M IH1 T’.

A.50 Greetings and Farewells

This transformation replaces greetings (e.g. “Hi”, “Howdy”) and farewells (e.g. “See you”, “Good night”) with their synonymous equivalents.

☞ Hey → Hi everyone. It’s nice → Pleased to meet you. How have → are you been ?

²¹<http://download.geonames.org/export/dump/>

²²<https://de.wiktionary.org/wiki/Verzeichnis:Deutsch/Namen>

A.51 Hashtagify

This transformation modifies an input sentence by identifying named entities and other common words and turning them into hashtags, as often used in social media.

A.52 Insert English and French Abbreviations

This perturbation replaces in texts some well known English and French words or expressions with (one of) their abbreviations. Many of the abbreviations covered here are quite common on social medias platforms, even though some of them are quite generic. This implementation is partly inspired by recent work in Machine Translation ([Berard et al., 2019](#)).

A.53 Leet Transformation

Visual perturbations are often used to disguise offensive comments on social media (e.g., !d10t) or as a distinct writing style (1337 in leet speak) ([Eger et al., 2019a](#)), especially common in scenarios like video gaming. Humans are unconsciously robust to such visually similar texts. This perturbation replaces letters with their visually similar “leet” counterparts.²³

☞ Ujjal Dev Dosanjh served →
U7jal Øev D0san74 serv3d as 33rd
Premier of British Columbia from →
Pr33i3r Of 8ritis4 00lu36ia fr0m 2000 to →
t0 2001

A.54 Lexical Counterfactual Generator

This transformation generates counterfactuals by simply substituting negative words like “not”, “neither” in one sentence of a semantically similar sentence pair. The substituted sentence is then backtranslated in an attempt to correct for grammaticality. This transformation would be useful for tasks like entailment and paraphrase detection.

A.55 Longer Location for NER

This transformation augments data for Named Entity Recognition (NER) tasks by augmenting examples which have a Location Tag. Names of locations are expanded by appending them with cardinal directions like “south”, “N”, “northwest”, etc. The transformation ensures that the tags of the new sentence are accordingly modified.

²³<https://simple.wikipedia.org/wiki/Leet>

A.56 Longer Location Names for testing NER

This transformation augments data for Named Entity Recognition (NER) tasks by augmenting examples that have a Location (LOC) Tag. Names of location are expanded by inserting random prefix or postfix word(s). The transformation also ensures that the labels of the new tags are accordingly modified.

A.57 Longer Names for NER

This transformation augments data for Named Entity Recognition (NER) tasks by augmenting examples which have a Person Tag. Names of people are expanded by inserting random characters as initials. The transformation also ensures that the labels of the new tags are accordingly modified.

A.58 Lost in Translation

This transformation is a generalization of the Back-Translation transformation to any sequence of languages supported by the Helsinki-NLP OpusMT models (Tiedemann and Thottingal, 2020).

☞ Andrew finally returned → brought Chris back the French book the French book I bought last week I bought last week

A.59 Mixed Language Perturbation

Mixed language training has been effective for cross-lingual tasks (Liu et al., 2020), to help generate data for low-resource scenarios (Liu et al., 2021) and for multilingual translation (Fan et al., 2021). Two transformations translate randomly picked words in the text from English to other languages (e.g., German). It can be used to test the robustness of a model in a multilingual setting.

☞ Andrew finally returned the → die Comic book to Chris that I bought last week → woche

A.60 Mix transliteration

This transformation transliterates randomly picked words from the input sentence (of given source language script) to a target language script. It can be used to train/test multilingual models to improve/evaluate their ability to understand complete or partially transliterated text.

A.61 MR Value Replacement

This perturbation adds noise to a key-value meaning representation (MR) (and its corresponding sentence) by randomly substituting values/words with their synonyms (or related words). This transformation uses a simple strategy to align values of a MR and tokens in the corresponding sentence inspired by how synonyms are substituted for tasks like machine translation (Fadaee et al., 2017). This way, there could be some problems in complex sentences. Besides, the transformation might introduce non-grammatical segments.

A.62 Multilingual Back Translation

This transformation translates a given sentence from a given language into a pivot language and then back to the original language. This transformation is a simple paraphraser that works on 100 different languages. Back Translation has been quite popular now and has been a quick way to augment (Li and Specia, 2019; Sugiyama and Yoshinaga, 2019; Fan et al., 2020).

☞ Being honest → Honesty should be one of our most important character traits → characteristics

A.63 Multilingual Dictionary Based Code Switch

This transformation generates multi-lingual code-switching data to fine-tune encoders of large language models (Qin et al., 2020; Tan and Joty, 2021; Wang et al., 2019b) by making use of bilingual dictionaries of MUSE (Lample et al., 2018).

A.64 Multilingual Lexicon Perturbation

This perturbation helps to create code-mixed sentences for both high-resource and low-resource languages by randomly translating words with a specified probability from any supported languages (e.g., English) to other supported languages (e.g., Chinese) by using a multilingual lexicon. Thus, it can be used to generate code-mixed training data to improve models for multilingual and cross-lingual settings. As of now 100 languages are supported and 3000 common English words listed on ef.com²⁴ are supported. The lexicon implementation is also 160x faster than its model based counterpart.

A.65 Causal Negation & Strengthening

This transformation is targeted at augmenting Causal Relations in text and adapts the code from the pa-

²⁴<https://www.ef.com/wwen/english-resources/english-vocabulary/top-3000-words/>

per Causal Augmentation for Causal Sentence Classification (Tan et al., 2021a). There are two operations: 1. Causal Negation: Negative words like "not, no, did not" are introduced into sentences to unlink the causal relation. 2. Causal Strengthening: Causal meaning is strengthened by converting weaker modal words into stronger ones like "may" to "will" to assert causal strength.

The implementation provides users with the option to amend causal meaning automatically from the root word of the sentence, or by explicitly highlighting the index of the word they wish to amend. Additionally, we include WordNet (Miller, 1998) synonyms and tense matching to allow for more natural augmentations.

🔊 The rs7044343 polymorphism **could be** → **was** involved in regulating the production of IL-33.

A.66 Question Rephrasing transformation

This implementation rephrases questions for sentence tasks by using the T5 model used in A.75 for Question Answering tasks.

A.67 English Noun Compound Paraphraser [N+N]

This transformation replaces two-word noun compounds with a paraphrase, based on the compound paraphrase dataset from SemEval 2013 Task 4 (Hendrickx et al., 2013). It currently only works for English. Any two-word compound that appears in a dataset of noun compound paraphrases will be replaced by a paraphrase. If more than one two-word compound appears, then all combinations of compound paraphrases (including no paraphrase at all) will be returned. For example, the paraphrases of "club house" include "house for club activities", "house for club members", "house in which a club meets", etc. We start with replacing paraphrases with the highest score (the specified frequency in the annotated dataset), and paraphrases with the same score (ties) are sorted randomly. This transformation currently only checks for noun compounds from Hendrickx et al. (2013) and therefore has low coverage. To improve it, other datasets could be added, e.g., from Ponkiya et al. (2018) or Lauer (1995). To attain even wider-coverage (at the expense of lower precision), machine learning approaches such as Schwartz and Dagan (2018) or Ponkiya et al. (2020) could be considered. In addition, some of the the paraphrases in Hendrickx et al. (2013) sound a little odd (e.g., "blood cell" -> "cell of blood") and may not fit well in context.

A.68 Number to Word

This transformation acts like a perturbation to improve robustness on processing numerical values. The perturbed sentence contains the same information as the initial sentence but with a different representation of numbers.

A.69 Numeric to Word

This transformation translates numbers in numeric form to their textual representations. This includes general numbers, long numbers, basic math characters, currency, date, time, phone numbers, etc.

A.70 OCR Perturbation

This transformation directly induces Optical Character Recognition (OCR) errors into the input text. It renders the input sentence as an image and recognizes the rendered text using the OCR engine Tesseract 4 (Smith, 2007). It works with text in English, French, Spanish, and German. The implementation follows previous work by Namysl et al. (2021).

A.71 Add Noun Definition

This transformation appends noun definitions onto the original nouns in a sentence. Definitions of nouns are collected from Wikidata ²⁵.

A.72 Pig Latin Cipher

This transformation translates the original text into pig latin. Pig Latin is a well-known deterministic transformation of English words, and can be viewed as a cipher which can be deciphered by a human with relative ease. The resulting sentences are completely unlike examples typically used in language model training. As such, this augmentation change the input into inputs which are difficult for a language model to interpret, while being relatively easy for a human to interpret.

A.73 Pinyin Chinese Character Transcription

This transformation transcribes Chinese characters into their Mandarin pronunciation using the Pinyin romanization scheme. The Character-to-Pinyin converter at the core of this transformation is a neural model by Park and Lee (2020).

²⁵https://www.wikidata.org/wiki/Wikidata:Main_Page

A.74 SRL Argument Exchange

This perturbation adds noise to all types of English text sources (sentence, paragraph, etc.) proportional to the number of arguments identified by SRL BERT (Shi and Lin, 2019b). Different rules are applied to deterministically modify the sentence in a meaning-preserving manner. Rules look as follows: *if ARGM-LOC and ARGM-TMP both present, exchange them.*

Example: [ARG0: Alex] [V: left] [ARG2: for Delhi] [ARGM-COM: with his wife] [ARGM-TMP: at 5 pm] . → Alex left for Delhi at 5 pm with his wife.

The transformation relies on propbank annotations (Bonial et al., 2012; Kingsbury and Palmer, 2002; Palmer et al., 2005; Gildea and Palmer, 2002).

A.75 ProtAugment Diverse Paraphrasing

This transformation utilizes the PROTAUGMENT method by Dopierre et al. (2021). The paraphrase generation model is a BART model (Lewis et al., 2020), finetuned on the paraphrase generation task using 3 datasets: Google-PAWS (Zhang et al., 2019b), MSR (Dolan and Brockett, 2005), Quora²⁶.

When paraphrasing a sentence, the transformation uses Diverse Beam Search (Vijayakumar et al., 2016) to generate diverse outputs. The diversity penalty term is by default set to 0.5 but can be set to custom values. Additionally, the transformation can use the following generation constraints: (1) A fraction of the words in the input sentence are forbidden in the paraphrase (default 0.7). (2) All bi-grams in the input sentence are forbidden in the paraphrase. This means the paraphrase cannot contain any bi-gram that are in the input sentence. This constraint enforces the paraphrase generation model to change the sentence structure.

A.76 Punctuation

This transformation removes/adds punctuation from an English sentence. This transformation was first introduced by Mille et al. (2021) and used as an example implementation for NL-Augmenter.

A.77 Question-Question Paraphraser for QA

This transformation creates new QA pairs by generating question paraphrases from a T5 model fine-tuned on Quora Question pairs²⁷. Generated questions can have a very different surface form from the original

²⁶<https://quoradata.quora.com/First-Quora-Dataset-Release-Question-Pairs>

²⁷Quora Question Pairs

question making it a strong paraphrase generator. A T5 model (Raffel et al., 2019; Wolf et al., 2020) fine tuned²⁸ on the Quora Question Pairs dataset was being used to generate question paraphrases. This transformation would benefit Question Answering, Question Generation as well as other tasks which could indirectly benefit eg. for dialog tasks (Shrivastava et al., 2021; Dhole, 2020).

A.78 Question in CAPS

This transformation upper-cases the context of a question answering example. It also adds upper-cased versions of the original answers to the set of acceptable model responses.

A.79 Random Word Deletion

This transformation randomly removes a word with a given probability p (by default 0.25). The transformation relies on whitespace tokenization and thus only works for English and other languages that are segmented via whitespace. Due to the destructive nature of the transformation, it is likely that the meaning of a sequence may be changed as a result of the change. A similar transformation was suggested by Wei and Zou (2019). Word dropout (Goldberg, 2017) has been common to help models understand unknown words encountered during evaluation by exposing them to this unknown-word condition during training itself.

A.80 Random Upper-Case Transformation

This perturbation adds noise to all types of text sources (sentence, paragraph, etc.) by randomly adding upper cased letters. With a default probability of 0.1, each character in a sequence is upper-cased. This transformation does not rely on a tokenizer and thus works with all languages that have upper and lower-case letters. One limitation of this transformation is that it will not affect a tokenizer that does lower case for all input. A similar transformation was suggested by Wei and Zou (2019). Further improvement of this transformation exists by potentially relying on extreme value theory (Jalalzai et al., 2020).

A.81 Double Context QA

This transformation repeats the context of a question answering example. This should not change the result in any way.

²⁸https://huggingface.co/ramsriouthamg/t5_paraphraser

A.82 Replace Abbreviations and Acronyms

This transformation changes abbreviations and acronyms appearing in an English text to their expanded form and respectively, changes expanded abbreviations and acronyms appearing in a text to their shorter form. For example, “send this file asap to human resources” might be changed to “send this file as soon as possible to HR”. The list of abbreviations and acronyms used in this transformation were manually gathered focusing on common abbreviations present in business communications. When abbreviations are context-dependent or highly specific, the induced change may change the meaning of a text, or an abbreviation may not be available in the lookup. The transformation was first introduced by Regina et al. (2020).

A.83 Replace Financial Amounts

This transformation replaces financial amounts throughout a text with the same value in a different currency. The replacement changes the amount, the writing format as well as the currency of the financial amount. For example, the sentence “I owe Fred 20 and I need 10 for the bus.” might be changed to “I owe Fred 2 906.37 Yen and I need 1 453.19 Yen for the bus.” The transformation was first introduced by Regina et al. (2020).

A.84 Replace Numerical Values

This transformation looks for numerical values in an English text and replaces it with another random value of the same cardinality. For example, “6.9” may be replaced by “4.2”, or “333” by “789”. The transformation was first introduced by Mille et al. (2021).

A.85 Replace Spelling

This transformation adds noise to all types of English text sources (sentence, paragraph, etc.) using corpora of common spelling errors introduced by Deorowicz and Ciura (2005). Each word with a common misspelling is replaced by the version with mistake with a probability p which by default is set to 0.2.

A.86 Replace nouns with hyponyms or hypernyms

This transformation replaces common nouns with other related words that are either hyponyms or hypernyms. Hyponyms of a word are more specific in meaning (such as a sub-class of the word), eg: ‘spoon’ is a hyponym of ‘cutlery’. Hypernyms are related words with a broader

meaning (such as a generic category /super-class of the word), eg: ‘colour’ is a hypernym of ‘red’. Not every word will have a hypernym or hyponym.

A.87 Sampled Sentence Additions

This transformation adds generated sentence to all types of English text sources (sentence, paragraph, etc.) by passing the input text to a GPT-2 model (Radford et al., 2019). By default, GPT-XL is used, together with the prompt “*paraphrase:*” appended to the original text, after which up to 75 tokens are sampled. Since the additional text is sampled from a model, the model may introduce harmful language or generate text that contradicts the earlier text or changes its meaning. The idea to sample one or more additional sentences was first introduced by Jia and Liang (2017a).

A.88 Sentence Reordering

This perturbation adds noise to all types of text sources (paragraph, document, etc.) by randomly shuffling the order of sentences in the input text (Lewis et al., 2020). Sentences are first partially decontextualized by resolving coreference (Lee et al., 2018).

This transformation is limited to input text that has more than one sentence. There are still cases where coreference can not be enough for decontextualization. For example, there could be occurrences of ellipsis as demonstrated by Gangal et al. (2021) or events could be mentioned in a narrative style which makes it difficult to perform re-ordering or shuffling (Kočíský et al., 2018) while keeping the context of the discourse intact.

A.89 Emoji Addition for Sentiment Data

This transformation adds positive emojis and smileys to positive sentiment data and negative emojis to negative sentiment data. For non-labelled data, it adds neutral smileys.

A.90 Shuffle Within Segments

In this transformation, a token sequence, for example BIO-tagged, is split into coherent segments. Thus, each segment corresponds to either a mention or a sequence of out-of-mention tokens. For example, a sentence “*She did not complain of headache or any other neurological symptoms.*” with tags O O O O O B-problem O B-problem I-problem I-problem I-problem O is split into five segments: [*She did not complain of*], [*headache*], [*or*], [*any other neurological symptoms*], [*.*]. Then for each segment, a binomial distribution ($p=0.5$) is used to decide whether it should be shuffled. If yes, the order of the tokens within the segment is shuffled while the

label order is kept unchanged. This transformation is inspired by [Dai and Adel \(2020\)](#).

A.91 Simple Ciphers

This transformation modifies the text in ways that a human could rapidly decipher, but which make the input sequences almost completely unlike typical input sequences which are used during language model training. This transformation includes the following text modifications: double the characters, double the words, add spaces between the characters, reverse all characters in the text, reverse the characters within each word, reverse the order of the words in the text, substitute homographs, rot13 cipher.

A.92 Slangificator

This transformation replaces some of the words (in particular, nouns, adjectives, and adverbs) of an English text with their corresponding slang. The replacement is done with the subset of the "Dictionary of English Slang & Colloquialisms".²⁹ The amount of replacement is proportional to the corresponding probabilities of replacement (by default, 0.5 for nouns, adjectives, and adverbs each).

A.93 Spanish Gender Swap

This transformation changes the gender of all animate entities (mostly referring to people, and some animals) in a given Spanish sentence from masculine to feminine. This includes masculine nouns with feminine equivalents (e.g., *doctor doctora*), nouns with a common gender ("sustantivos comunes en cuanto al género", e.g., *el violinista la violinista*), personal pronouns, and (optionally) given names often used with a given gender (e.g., *Pedro Alicia*). Epicene nouns are excluded. In addition, the gender of adjectives, determiners, pronouns and participles are modified in order to maintain the grammatical agreement.

A.94 Speech Disfluency Perturbation

This perturbation randomly inserts speech disfluencies in the form of filler words into English texts. With a given probability (0.2 by default), a speech disfluency is inserted between words. The default disfluencies are "um", "uh", "erm", "ah", and "er". At least one filler word is always inserted by this transformation.

²⁹<http://www.peevish.co.uk/slang/index.htm>

A.95 Paraphrasing through Style Transfer

This transformation provides a range of possible styles of writing English language. The following styles can be chosen:

- Shakespeare - Trained on written works by Shakespeare.
- Switchboard - Trained on a collection of conversational speech transcripts.
- Tweets - Trained on 5.2M English tweets.
- Bible - Trained on texts from the Bible.
- Romantic poetry - Trained on romantic poetry.
- Basic - A light, basic paraphraser with no specific style.

The transformation follows the models and formulations by [Krishna et al. \(2020\)](#).

A.96 Subject Object Switch

This transformation switches the subject and object of English sentences to generate new sentences with a very high surface similarity but very different meaning. This can be used, for example, for augmenting data for models that assess Semantic Similarity.

A.97 Sentence Summarization

This transformation compresses English sentences by extracting subjects, verbs, and objects of the sentence. It also retains any negations. For example, "*Stillwater is not a 2010 American live-action/animated dark fantasy adventure film*" turns into "*Stillwater is film*". [Zhang et al. \(2021\)](#) used a similar idea to this transformation.

A.98 Suspecting Paraphraser for QA

This paraphraser transforms a yes/no question into a declarative sentence with a question tag³⁰, which helps to add more question specific informality to the dataset. Example: "Did the American National Shipment company really break its own fleet?" -> "The American National Shipment company really broke its own fleet, didn't it".

A.99 Swap Characters Perturbation

This perturbation randomly swaps two adjacent characters in a sentence or a paragraph with a default probability ([Zhang et al., 2019a](#)).

³⁰<https://www.englishclub.com/grammar/tag-questions.htm>

A.100 Synonym Insertion

This perturbation adds noise to all types of text sources (sentence, paragraph, etc.) by randomly inserting synonyms of randomly selected words excluding punctuations and stopwords (Marivate and Sefara, 2020).

A.101 Synonym Substitution

This perturbation randomly substitutes some words in an English text with their WordNet (Miller, 1998) synonyms.

A.102 Syntactically Diverse Paraphrasing using Sow Reap models

This transformation is capable of generating multiple syntactically diverse paraphrases for a given sentence based on the work of Goyal and Durrett (2020). The model paraphrases inputs using a two step framework: 1) SOW (Source Order reWriting): This step enumerates multiple feasible syntactic transformations of the input sentence. 2) REAP (REarrangement Aware Paraphrasing): This step conditions on the multiple reorderings/ rearrangements produced by SOW and outputs diverse paraphrases corresponding to these reorderings. The transformation is designed to work only on single-sentence inputs. Multi-sentence inputs results in an empty string/no transformation. The model are trained on the ParaNMT-50M dataset (Wieting and Gimpel, 2017; Wieting et al., 2017), which can be argued to be a bit noisy.

A.103 Subsequence Substitution for Sequence Tagging

This transformation performs same-label subsequence substitution for the task of sequence tagging, which replaces a subsequence of the input tokens with another one that has the same sequence of tags (Shi et al., 2021). This is done as follows: (1) Draw a subsequence A from the input (tokens, tags) tuple. (2) Draw a subsequence B within the whole dataset, with the same tag subsequence. (3) Substitute A with B in the input example.

A.104 Change English Tense

This transformation converts English sentences from one tense to the other, for example simple present to simple past. This transformation was introduced by Logeswaran et al. (2018).

A.105 Token Replacement Based on Lookup Tables

This transformation replaces input tokens with their perturbed versions sampled from a given lookup table of replacement candidates. Lookup tables containing OCR errors and misspellings from prior work are given as examples. Thus, by default, the transformation induces plausible OCR errors and human typos to the input sentence.

The transformation is an adapted and improved version of the lookup table-based noise induction method from Namysl et al. (2020). The OCR lookup table is from Namysl et al. (2021) and the misspellings from Piktus et al. (2019).

A.106 Transformer Fill

This perturbation replaces words based on recommendations from a masked language model. The transformation can limit replacements to certain POS tags (all enabled by default). Many previous papers have used this technique for data augmentation (Ribeiro et al., 2020; Li et al., 2020b, inter alia).

A.107 Underscore Trick

This perturbation adds noise to the text sources like sentence, paragraph, etc. This transformation acts like a perturbation to test robustness. It replaces some random spaces with underscores (or even other selected symbols). This perturbation would benefit all tasks which have a sentence/paragraph/document as input like text classification and text generation, especially on tasks related to understanding/generating scripts.

A.108 Unit converter

This transformation converts length and weight measures to different units (e.g., kilometers to miles) picking at random the new unit but converting accurately the quantity. The transformation conserves the format of the original quantity: "100 pounds" is converted to "1600 ounces" but "one-hundred pounds" is converted to "one thousand, six hundred ounces". Generated transformations display high similarity to the source sentences.

A.109 Urban Thesaurus Swap

This perturbation randomly picks nouns from the input source to convert to related terms drawn from the Urban Dictionary ³¹ resource. It can be applied to an input text to produce semantically-similar output texts

³¹<https://www.urbandictionary.com/>

in order to generate more robust test sets. We first select nouns at random, then query the Urban Thesaurus website ³² to obtain a list of related terms to swap in (Wilson et al., 2020).

A.110 Use Acronyms

This transformation changes groups of words for their equivalent acronyms. It's a simple substitution of groups of words for their acronyms. It helps to increase the size of the dataset as well as improving the understanding of acronyms of models trained on data augmented with this transformation. This transformation works to increase the data for any task that has input texts. It is specially interesting for tasks on semantic similarity, where models should be aware of the equivalence between a set of words and their acronym. The quality of the transformation depends on the list of acronyms. As of now, this list was scraped from wikipedia's List of Acronyms ³³ and naively filtered, which leaves space for improvement .

A.111 Visual Attack Letter

This perturbation replaces letters with visually similar, but different, letters. Every letter was embedded into 576-dimensions. The nearest neighbors are obtained through cosine distance. To obtain the embeddings the letter was resized into a 24x24 image, then flattened and scaled. This follows the Image Based Character Embedding (ICES) (Eger et al., 2019a).

The top neighbors from each letter are chosen. Some were removed by judgment (e.g. the nearest neighbors for 'v' are many variations of the letter 'y') which did not qualify from the image embedding (Eger et al., 2019b).

A.112 Weekday Month Abbreviation

This transformation abbreviates or expands the names of months and weekdays, e.g. Mon. -> Monday. Generated transformations display high similarity to the source sentences and does not alter the meaning and the semantic of the original texts. It does not abbreviate plural names, e.g. Sundays, and does not influence texts without names of weekdays or months.

A.113 Whitespace Perturbation

This perturbation adds noise to text by randomly removing or adding whitespaces.

³²<https://urbanthesaurus.org/>

³³https://en.wikipedia.org/wiki/Lists_of_acronyms

A.114 Context Noise for QA

This transformation chooses a set of words at random from the context and the question and forms a sentence out of them. The sentence is then prepended or appended to the context to create a new QA pair. The transformation is inspired by the the **AddAny** method described in Adversarial SQUAD (Jia and Liang, 2017b). However, instead of probing the model to generate adversaries, random words from the context and the question are simply selected and joined together into a sentence, ignoring grammaticality. The transformation attempts to create novel QA pairs assuming that the introduction of random words to the context is less likely to change the answer choice to an asked question.

A.115 Writing System Replacement

This transformation replaces the writing system of the input with another writing system. We use CJK Unified Ideographs³⁴ as the source of characters for the generated writing systems. The transformation would benefit text classification tasks, especially in the cases where the input writing system is undeciphered.

A.116 Yes-No Question Perturbation

This transformation turns English non-compound statements into yes-no questions. The generated questions can be answered by the statements that were used to generate them. The text is left largely unchanged other than the fronted/modified/added auxiliaries and be-verbs.

The transformation works by getting dependency parse and POS tags from a machine learning model and applying human-engineered, rule-based transformations to those parses/tags. This transformation would particularly benefit question-answering and question-generation tasks, as well as providing surplus legal text for language modeling and masked language modeling.

A.117 Yoda Transformation

This perturbation modifies sentences to flip the clauses such that it reads like "Yoda Speak". For example, "Much to learn, you still have". This form of construction is sometimes called "XSV", where "the X being a stand-in for whatever chunk of the sentence goes with the verb", and appears very rarely in English normally. The rarity of this construction in ordinary language makes it particularly well suited for NL augmentation and serves as a relatively easy but potentially powerful test of robustness.


³⁴https://en.wikipedia.org/wiki/CJK_Unified_Ideographs

B Filters

The following is the list of all submitted filters to NL-Augmenter. Filters are used to filter data and create subpopulations of given inputs, according to features such as input complexity, input size, etc. Therefore, the output of a filter is a boolean value, indicating that whether the input meet the filter criterion. We discuss the implementations of each filter alongwith their limitations. The title of each filter subsection is clickable and redirects to the actual python implementation.


B.1 Code-Mixing Filter

This filter identifies whether the input text is code-mixed. It checks that there is at least one sentence in the text where there are tokens representing at least 'k' unique languages (with at least a 'threshold' level of confidence that the token is of that language). It is useful for collecting code-mixed data to test the model's performance on multilingual tasks. The filter relies on `ftlid`³⁵ for language detection, therefore, this filter might be limited by the performance of the language detection tool.

 (containing code-mixing) Yo estaba con Esteban yesterday, he was telling me about lo que su esposa vio en los Estados Unidos. →✓True


B.2 Diacritics Filter

This filter checks whether any character in the sentence has a diacritic. It can be used to create splits of the dataset where the sentences have diacritics. Accented characters are typically among the rarer characters and checking the model performance on such a split might help investigate model robustness.

 (containing diacritics) She lookèd east an she lookèd west. →✓True

B.3 Encoding Filter


This filter filters examples which contain characters outside a given encoding. It can be used to find examples containing e.g. non-ASCII Unicode characters. Filtering out and testing examples that contain these characters can provide feedback on how to improve the models accordingly, since most models are trained with plain English text, which contains mostly ASCII characters. Sometimes non-ASCII character are even explicitly stripped away.

 (containing non-ASCII characters) That souvenir sure was expensive at 60č.. or was it 60? →✓True

³⁵<https://pypi.org/project/ftlid/>


B.4 Englishness Filter

This filter identifies texts that contain uniquely British spellings, vocabulary, or slang. The filter uses a vocabulary of common British words/phrases and checks the number of occurrence of British words in the given texts. The text is selected if the number exceeds a pre-defined threshold.

 (containing British spellings) Colour is an attribute of light that is perceived by the human eye. →✓True


B.5 Gender Bias Filter

This filter filters a text corpus to measure gender fairness with respect to a female gender representation. It supports four languages (i.e. English, French, Polish and Russian) and can be used to define whether the female gender is sufficiently represented in a tested subset of sentences. The filter uses a list of lexicals, which includes filter categories such as personal pronouns, words defining the relation, titles and names, corresponding to the female and male genders accordingly.

 (texts with unbalanced representation) "He went home", "He drives a car", "She has returned" →✓True


B.6 Group Inequity Filter

This is a bilingual filter (for English and French languages), which helps to discover potential group inequity issues in the text corpus. It is a topic agnostic filter which accepts user-defined parameters, consisting of keywords inherent to minor group (which potentially might suffer from the discrimination), major group, minor factor and major factor. The filter first flags the sentences as belonging to the minor, and the major groups, and then, the sentences from each of the groups are used to define the intersection with both factors. The filter then compares whether the percentage of major factors exceeds that of the minor factors to determine if the sentences have group inequity issues.

 (containing group inequity issues) "He is a doctor", "She is a nurse", "She works at the hospital" →✓True

B.7 Keyword Filter

This is a simple filter, which filters examples based on a pre-defined set of keywords. It can be useful in creating splits for a specific domain.

 (containing keyword "at") Andrew played cricket in India →✓True

B.8 Language Filter

This filter selects texts that match any of a given set of ISO 639-1 language codes (the default language being English). Language matching is performed using a pre-trained `langid.py` model instance. The model provides normalized confidence scores. A minimum threshold score needs to be set, and all sentences with confidence scores above this threshold are accepted by the filter.

☞ (is English texts) Mein Luftkissenfahrzeug ist voller Aale →**XFalse**

B.9 Length Filter

This filter filters data with the input text length matching a specified threshold. It can be useful in creating data with different length distributions.

☞ (containing more than 3 words) Andrew played cricket in India →**✓True**

B.10 Named-entity-count Filter

This filter filters data where the number of Named Entities in the input match a specified threshold (based on the supported conditions).

☞ (containing more than 1 named entity) Novak Djokovic is the greatest tennis player of all time. →**✓True**

B.11 Numeric Filter

This filter filters example which contain a numeric value. In the tasks like textual entailment, question answering etc., a quantity (number) could directly affect the final label/response. This filter can be used to create splits to measure the performance separately on texts containing numeric values.

☞ (containing numbers in texts) John bought a car worth dollar twenty five thousand . →**✓True**

B.12 Oscillatory Hallucinations Filter

This filter is designed to operate in text generation systems' outputs, with the purpose of extracting oscillatory hallucinations. Oscillatory hallucinations are one class of hallucinations characterized by repeating bigram structure in the output (Raunak et al., 2021). Typically, these behaviors are observed in models trained on noisy corpora. The filter counts the frequency of bigrams in both source and target texts, and compare the frequency difference with a pre-set threshold to determine whether the texts includes oscillatory hallucinations.

☞ (containing hallucinations in target texts) Source: "Community, European Parliament common

regional policy, Mediterranean region", Target: "Arbeitsbedingungen, berufliche Bildung, berufliche Bildung, berufliche Bildung" →**✓True**

B.13 Polarity Filter

This filter filters a transformed text if it does not retain the same polarity as an original text. This filter helps not to distort training data during augmentation for sentiment analysis-related tasks. While generating new data for a sentiment analysis task, it is important to make sure that generated data is labelled correctly.

☞ (texts retaining polarity) "Hotel is terrible", "Hotel is great" →**XFalse**

B.14 Quantitative Question Filter

This is a simple rule-based filter that can be used to identify quantitative questions. It can help to analyse models' performance on questions which require numerical understanding. It is also useful to study possible biases in question generation.

☞ (being quantitative question) How long does the journey take? →**✓True**

B.15 Question type filter

This filter helps identify the question category of a question answering example based on the question word or the named entity type of the answer. Knowledge of the question type can help in the development of question answering systems (Parikh et al., 2019) as well as for assessing performance on individual splits.

☞ (being where question) Where is Delhi located ? →**✓True**

B.16 Repetitions Filter

This filter finds texts with repetitions with simple heuristic rules. It might be helpful in finding repetitions that frequently occur in the spoken language data.

☞ (containing repetitions in texts) I I want to sleep →**✓True**

B.17 Phonetic Match Filter

This filter selects texts that contain matching entries to a list of supplied keywords. It first transform the input sentence and the keywords into phonetic units and then compare whether the two phonetic unit sets have overlap.

☞ (containing homophones of keyword "trombone") I left my trombno on the train →**✓True**

B.18 Special Casing Filter

This filter checks if the input sentence has a special casing, i.e. the string is either all lowercased, all uppercased or has title casing. It might be useful for creating splits that contain texts with unusual casing, e.g. misspellings.

☞ (text being uppercased/lowercased/titlecased) let's go to chipotle → ✓True

B.19 Speech-Tag Filter

This filter filters an example text based on a set of speech tags and identifies whether the count of selected POS tags meet the pre-defined conditions (e.g. above the threshold).

☞ (containing 1 verb and 2 numbers in texts) It all happened between November 2007 and November 2008. → ✓True

B.20 Token-Amount filter

This filter filters an example text based on whether certain keywords are present in a specified amount.

☞ (containing 2 occurrences of "in") Andrew played cricket in a soccer stadium in India at 9pm → ✓True

B.21 Toxicity Filter

This filter filters an example text which has a toxicity value matching a particular threshold. It uses a pre-trained toxicity detector, which can provide 7 toxicity scores. All the 7 types of toxicity scores can be used as criteria for the filtering.

☞ (text being toxic) I disagree. It is not supposed to work that way. → ✗False

B.22 Universal Bias Filter

This filter works the same way as the Gender Bias Filter, but measures balance of representation for more categories (religion, race, ethnicity, gender, sexual orientation, age, appearance, disability, experience, education, economic status). The lexical seeds representing these categories are currently available in English only, however the pool of languages can be extended by a simple addition of the lexical seeds in a desired language to the lexicals.json file.

☞ (texts being biased) "He is going to make a cake.", "She is going to program", "Nobody likes washing dishes", "She agreed to help him" → ✗False

B.23 Yes/no question filter

This filter allows to select questions that can be correctly answered with either 'yes' or 'no'. Since it is rule-

based, the limitation of this filter is that questions that are ambiguous might not be recognized.

☞ (text being yes/no question) Wasn't she angry when you told her about the accident? → ✓True

C Review criteria for submission evaluation

Figure 3 shows the review criteria used for evaluating the transformation and filters submissions.

Correctness: Transformations must be valid Python code and must pass tests.

Interface: Participants should ensure that they use the correct interface. The complete list is mentioned [here](#). E.g., for tasks like machine translation, a transformation which changes the value of a named entity (Andrew->Jason) might need parallel changes in the output too. And hence, it might be more appropriate to use `SentenceAndTargetOperation` or `SentenceAndTargetsOperation` rather than `SentenceOperation`. Similarly, if a transformation changes the label of a sentence, the interface's generate method should take as input the label too - eg. if your transformation reverses the sentiment, `SentenceAndTargetOperation` would be more appropriate than `SentenceOperation`. If you wish to add transformations for input formats other than those specified, you should add an interface [here](#).

Applicable Tasks & Keywords: We understand that transformations can vary across tasks as well as a single transformation can work for multiple tasks. Hence all the tasks where the transformation is applicable should be specified in the list "tasks". The list of tasks has been specified [here](#). The relevant keywords for the [transformation](#) should also be specified.

```
class ButterFingersPerturbation(SentenceOperation):
    tasks = [TaskType.TEXT_CLASSIFICATION, TaskType.TEXT_TO_TEXT_GENERATION, TaskType.TEXT_TAGGING]
    languages = ["en"]
    keywords = ["morphological", "noise", "rule-based", "high-coverage", "high-precision"]
```

Specificity: While this is not a necessary criterion, it is highly encouraged to have a specific transformation. E.g., a perturbation which changes gendered pronouns could give insights about gender bias in models.

Novelty: Your transformation must improve the coverage of NL-Augmenter in a meaningful way. The idea behind your transformation need not be novel, but its contribution to the library **must be different from the contributions of earlier submissions**. If you are unsure if your idea would constitute a new contribution, please email the organizers at nl-augmenter@googlegroups.com and we are happy to help.

Adding New Libraries: We welcome addition of libraries which are light and can be installed via `pip`. Every library should specify the version number associated and be added in a new `requirements.txt` in the transformation's own folder. However, we discourage the use of heavy libraries for a few lines of code which could be manually written instead. Please ensure that all libraries have MIT, Apache 2, BSD, or other permissive license. GPL-licensed libraries are not approved for NL-Augmenter. If you are unsure, please email the organizers at nl-augmenter@googlegroups.com.

Description: The `README.md` file should clearly explain what the transformation is attempting to generate as well as the importance of that transformation for the specified tasks. Here is a [sample README](#).

Data and code source: The `README.md` file should have a subsection titled "Data and code provenance", which should describe where data or code came from, or that it was fully created by the author. This section should also disclose the license that any external data or code is released under.

Paraphrasers and Data Augmenters: Besides perturbations, we welcome transformation methods that act like paraphrasers and data augmenters. For non-deterministic approaches, we encourage you to specify metrics which can provide an estimate of the generation quality. We prefer high precision transformation generators over low accuracy ones. And hence it's okay if your transformation selectively generates.

Test Cases: We recommend you to add at least 5 examples in the file `test.json` as test cases for every added transformation. These examples serve as test cases and provide reviewers a sample of your transformation's output. The format of `test.json` can be borrowed from the sample transformations [here](#). A good set of test cases would include good as well as bad generation. Addition of the the test cases is **not mandatory** but is encouraged.

Evaluating Robustness: To make a stronger PR, a transformation's potential to act as a robustness tool should be tested via executing `evaluate.py` and the corresponding performance should be mentioned in the README. Evaluation should only be skipped in case there is no support in the `evaluation_engine`.

Languages other than English: We strongly encourage multilingual perturbations. All applicable languages should be specified in the list of "languages".

Decent Programming Practise: We recommend adding docstrings to help others follow your code with ease. Check the [PEP 257 Docstring Conventions](#) to get an overview. If you are using spacy, we suggest you use the common global version like [this](#).

All of the above criteria extend to [filters](#) too.

Figure 3: Participants and reviewers were provided with a set of review criteria.

D Contributor Affiliations

Kaustubh D. Dhole^{3,18†}, Varun Gangal^{7†}, Sebastian Gehrmann^{23†}, Aadesh Gupta^{3†}, Zhenhao Li^{32†}, Saad Mahmood^{90†}, Abinaya Mahendiran^{45†}, Simon Mille^{53†}, Ashish Shrivastava^{2†}, Samson Tan^{91†}, Tongshuang Wu^{7†}, Jascha Sohl-Dickstein^{22†}, Jinho D. Choi^{18†}, Eduard Hovy^{7†}, Ondrej Dusek^{10†}, Sebastian Ruder^{13†}, Sajant Anand⁶⁸, Nagender Aneja⁷⁴, Rabin Banjade⁷⁷, Lisa Barthe¹⁹, Hanna Behnke³², Ian Berlot-Attwell⁸⁰, Connor Boyle⁸¹, Caroline Brun⁴⁹, Marco Antonio Sobrevilla Cabezudo⁷⁹, Samuel Cahyawijaya²⁶, Emile Chapuis⁵², Wanxiang Che²⁴, Mukund Choudhary³⁷, Christian Clauss³³, Pierre Colombo⁵², Filip Cornell⁴¹, Gautier Dagan⁸⁴, Mayukh Das⁶³, Tanay Dixit³⁰, Thomas Dopierre³⁹, Paul-Alexis Dray⁸⁹, Suchitra Dubey¹, Tatiana Ekeinhorn⁸⁶, Marco Di Giovanni⁵¹, Tanya Goyal⁴, Rishabh Gupta²⁹, Louanes Hamla¹⁹, Sang Han⁷³, Fabrice Harel-Canada⁷⁰, Antoine Honoré⁸⁶, Ishan Jindal²⁷, Przemyslaw K. Joniak⁶⁶, Denis Kleyko⁷⁵, Venelin Kovatchev⁶⁵, Kalpesh Krishna⁷¹, Ashutosh Kumar³⁴, Stefan Langer⁵⁹, Seungjae Ryan Lee⁵⁵, Corey James Levinson³³, Hualou Liang¹⁵, Kaizhao Liang⁷⁶, Zhexiong Liu⁷⁸, Andrey Lukyanenko⁴³, Vukosi Marivate¹⁴, Gerard de Melo²⁵, Simon Meoni³³, Maxime Meyer⁸⁶, Afnan Mir⁴, Nafise Sadat Moosavi⁶², Niklas Muennighoff⁵⁰, Timothy Sum Hon Mun⁶⁴, Kenton Murray⁴⁰, Marcin Namysl²⁰, Maria Obedkova³³, Priti Oli⁷⁷, Nivranshu Pasricha⁴⁶, Jan Pfister⁸³, Richard Plant¹⁷, Vinay Prabhu⁷³, Vasile Pais⁵⁷, Libo Qin²⁴, Shahab Raji⁵⁸, Pawan Kumar Rajpoot⁵⁶, Vikas Raunak⁴⁴, Roy Rinberg¹¹, Nicholas Roberts⁸², Juan Diego Rodriguez⁷², Claude Roux⁴⁹, Vasconcellos P. H. S.⁵⁴, Ananya B. Sai³⁰, Robin M. Schmidt¹⁶, Thomas Scialom⁸⁹, Tshephisho Sefara¹², Saqib N. Shamsi⁸⁸, Xudong Shen⁴⁸, Yiwen Shi¹⁵, Haoyue Shi⁶⁷, Anna Shvets¹⁹, Nick Siegel⁴, Damien Sileo⁴², Jamie Simon⁶⁸, Chandan Singh⁶⁸, Roman Sitelew³³, Priyank Soni³, Taylor Sorensen⁶, William Soto⁶¹, Aman Srivastava⁸⁵, KV Aditya Srivatsa³⁷, Tony Sun⁶⁹, Mukund Varma T³⁰, A Tabassum⁴⁷, Fiona Anting Tan³⁶, Ryan Teehan⁹, Mo Tiwari⁶⁰, Marie Tolkiehn⁸, Athena Wang⁴, Zijian Wang³³, Zijie J. Wang²¹, Gloria Wang³¹, Fuxuan Wei²⁴, Bryan Wilie³⁵, Genta Indra Winata⁵, Xinyi Wu⁸¹, Witold Wydmanski³⁸, Tianbao Xie²⁴, Usama Yaseen⁵⁹, Michael A. Yee⁹², Jing Zhang¹⁸, Yue Zhang⁸⁷

¹ACKO, ²Agara, ³Amelia R&D, New York, ⁴Applied Research Laboratories, The University of Texas at Austin, ⁵Bloomberg, ⁶Brigham Young University, ⁷Carnegie Mellon University, ⁸Center for Data and Computing in Natural Sciences, Universität Hamburg, ⁹Charles River Analytics, ¹⁰Charles University, Prague, ¹¹Columbia University, ¹²Council for Scientific and Industrial Research, ¹³DeepMind, ¹⁴Department of Computer Science, University of Pretoria, ¹⁵Drexel University, ¹⁶Eberhard Karls University of Tübingen, ¹⁷Edinburgh Napier University, ¹⁸Emory University, ¹⁹Fablab by Inetum in Paris, ²⁰Fraunhofer IAIS, ²¹Georgia Tech, ²²Google Brain, ²³Google Research, ²⁴Harbin Institute of Technology, ²⁵Hasso Plattner Institute / University of Potsdam, ²⁶Hong Kong University of Science and Technology, ²⁷IBM Research, ²⁸IIIT Delhi, ²⁹IIT Delhi, ³⁰IIT Madras, ³¹Illinois Mathematics and Science Academy, ³²Imperial College, London, ³³Independent, ³⁴Indian Institute of Science, Bangalore, ³⁵Institut Teknologi Bandung, ³⁶Institute of Data Science, National University of Singapore, ³⁷International Institute of Information Technology, Hyderabad, ³⁸Jagiellonian University, Poland, ³⁹Jean Monnet University, ⁴⁰Johns Hopkins, ⁴¹KTH Royal Institute of Technology, ⁴²KU Leuven, ⁴³MTS AI, France, ⁴⁴Microsoft, Redmond, WA, ⁴⁵Mphasis NEXT Labs, ⁴⁶National University of Ireland Galway, ⁴⁷National University of Science and Technology, Pakistan, ⁴⁸National University of Singapore, ⁴⁹Naver Labs Europe, ⁵⁰Peking University, ⁵¹Politecnico di Milano and University of Bologna, ⁵²Polytechnic Institute of Paris, ⁵³ADAPT/Dublin City University, ⁵⁴Pontifical Catholic University of Minas Gerais, Brazil, ⁵⁵Princeton University, ⁵⁶Rakuten India, ⁵⁷Research Institute for Artificial Intelligence Mihai Drgnescu, Romanian Academy, ⁵⁸Rutgers University, ⁵⁹Siemens AG, ⁶⁰Stanford University, ⁶¹SyNaLP, LORIA, ⁶²TU Darmstadt, ⁶³Technical University of Braunschweig, ⁶⁴The Alan Turing Institute, ⁶⁵The University of Texas at Austin; (University of Barcelona, University of Birmingham), ⁶⁶The University of Tokyo, ⁶⁷Toyota Technological Institute at Chicago, ⁶⁸UC Berkeley, ⁶⁹UC Santa Barbara / Google, ⁷⁰UCLA, ⁷¹UMass Amherst, ⁷²UT Austin, ⁷³UnifyID, ⁷⁴Universiti Brunei Darussalam, ⁷⁵University of California, Berkeley and Research Institutes of Sweden, ⁷⁶University of Illinois, Urbana Champaign, ⁷⁷University of Memphis, ⁷⁸University of Pittsburgh, ⁷⁹University of São Paulo, ⁸⁰University of Toronto, ⁸¹University of Washington, ⁸²University of Wisconsin/Madison, ⁸³University of Würzburg, ⁸⁴University of Edinburgh, ⁸⁵VMware, ⁸⁶Vade, ⁸⁷Westlake Institute for Advanced Study, ⁸⁸Whirlpool Corporation, ⁸⁹reciTAL, ⁹⁰trivago N.V., ⁹¹AWS AI Research & Education, ⁹²University of Michigan, ⁹¹ Work done independent of AWS tenure.

On the Relationship between Frames and Emotionality in Text

Enrica Troiano,* Vrije Universiteit Amsterdam, The Netherlands e.troiano@vu.nl

Roman Klinger, IMS, University of Stuttgart, Germany klinger@ims.uni-stuttgart.de

Sebastian Padó, IMS, University of Stuttgart, Germany pado@ims.uni-stuttgart.de

Abstract Emotions, which are responses to salient events, can be realised in text implicitly, for instance with mere references to facts (e.g., “*That was the beginning of a long war*”). Interpreting emotions thus relies on the readers’ background knowledge, but that is hardly modeled in computational emotion analysis. Much work in the field is focused on the word level and treats individual lexical units as the fundamental emotion cues in written communication. We shift our attention to the event knowledge they evoke. We leverage frame semantics, a prominent theory for the description of event meanings, and show it is well-suited for the study of emotions: frames build on a “semantics of understanding” whose assumptions rely precisely on people’s world knowledge. Our overarching question is if the events that are represented by frames possess an emotion dimension. We hypothesise that they do, and that such a dimension can be distinguished qualitatively for different groups of frames.

To carry out a large corpus-based correspondence analysis, we automatically annotate texts with emotions as well as with FrameNet frames and roles, and we analyse the correlations between them. Our main finding is that substantial groups of frames have an emotional import. With an extensive qualitative analysis, we show that they capture several properties of emotions that are purported by theories from psychology. These observations contribute to advancing the two strands of research that we combine: emotion analysis can profit from the event-based perspective of frame semantics; in return, frame semantics gains a better grip of its position vis-à-vis emotions, an integral part of word meanings.

1 Introduction

Human life is interwoven with emotions. They echo in our brain, body, behaviors, and attract for this reason a diverse range of disciplines (Barrett et al., 2016, Part I). Psychology, among others, has entered a century-long endeavor to explain how emotions arise, with appraisal theories (Smith and Ellsworth, 1985, i.a.) providing a viewpoint that is widely accepted today: emotions are *responses to (internal or external) events*, specifically to circumstances *evaluated* as salient by their experiencers (Scarantino, 2016). Understanding how humans evaluate events is thus fundamental to discuss this affective phenomenon, and appraisal theories offer many fertile insights on the matter. They spell out, for instance, some human reactions to events, like neurophysiological changes, motor expressions and motivational tendencies (Scherer, 1989). From the perspective of an observer, these hint at what other people feel: the blushing on one’s cheeks might reveal an episode of shame, the raising of a brow could indicate disappointment.

* The work was carried out while the first author was affiliated with IMS, University of Stuttgart, Germany.

Emotions also pervade the sphere of verbal communication, where an observer infers the mental state of others by interpreting their utterances. Decoding emotions from words is key to successful communication, since emotions represent an important aspect of the meaning that speakers and writers intend to convey (Scheff, 1973). This is the idea that fuels (computational) emotion analysis in natural language processing (Canales and Martínez-Barco, 2014), a research field geared towards the creation of systems that sense emotions like humans do. Emotion analysis mainly approaches its task as text classification. It models the import of verbal expressions either as discrete categories, like anger and joy, or through scalar features such as valence and arousal (Nandwani and Verma, 2021). A central challenge in this regard is that emotions are expressed in language in a myriad of ways. At times they emerge explicitly, with words that point to an emotion state by definition (e.g., “*I’m happy*”). Other times, however, emotions can be expressed without unequivocal cues, mental states or evaluative attitudes: writers can describe a stimulus event (e.g., “*my granddad died*”, “*my team won the match*”, which likely spark sad-

ness and joy), or their reaction to it (e.g., “*I cried*”, “*I smiled*”), trusting that the correct emotional interpretation of their production will be drawn by the readers via pragmatic inference (Grice, 1975).

How can emotions be associated with such factual statements? Psychology explains the link via empathy and affective role taking (Mehrabian and Epstein, 1972; Eisenberg and Miller, 1987; Omdahl, 1995), and natural language processing connects emotion decoding more directly to world knowledge. Its starting point is that words possess specific connotations in the collective imagination (Clore et al., 1987) – e.g., *die*: sadness, *win*: joy, *ghost*: fear. Accordingly, it stores such connotations as dictionaries of word-to-emotion associations (Strapparava and Valitutti, 2004; Mohammad and Turney, 2013).

Word-level dictionaries leverage the assumption that individual words are the crucial, emotion-revealing linguistic units. This view is practically useful, but it neglects an important point, namely the impact of the context in which words occur, and thus the paradigmatic and syntagmatic information that allow people to infer emotion meanings. For instance, the surrounding verbal context of “*boiling*” helps disambiguate if this predicate refers to a heat reaction with a nonemotional tone (“*the water is boiling*”) or to an emotional turmoil (“*she is boiling with anger*”). Much work in emotion analysis disregards this type of background knowledge. Approaches that embed emotion meanings into latent vector spaces (Felbo et al., 2017; Li et al., 2017, i.a.) capture contextual information, but they are less transparent to investigation than lexical methods.

In this article, we consider *frame semantics* (Fillmore, 1982) as a source of lexical abstractions that is appropriate for specifying emotions in a dictionary. Frame semantics proposes a formalism (viz., frames) and a practical resource (Berkeley FrameNet, (Baker et al., 1998)) to describe linguistic meanings with a combination of predicates (i.e., frames) and arguments. This “semantics of understanding” or U-semantics (Fillmore, 1985) explains the difference in meaning between “*the water is boiling*” and “*she was boiling with anger*” in terms of reference to two different frames that are evoked by the sentences, respectively. This frame-level disambiguation arguably makes use of knowledge about how the world is organised that is necessary to recognise which of the two sentences is emotional. It also suggests that frames bear a potential value for studies in emotion analysis, even though they are usually dismissed in the computational study of emotions.

We believe that there are many affinities between emotions and frames. Not only does FrameNet dedicate multiple frames to emotions (e.g., `EMOTION_DIRECTED` and `EMOTION_OF_MENTAL_ACTIVITY`), but it pays attention to events, similar to appraisal theories. Figure 1 il-

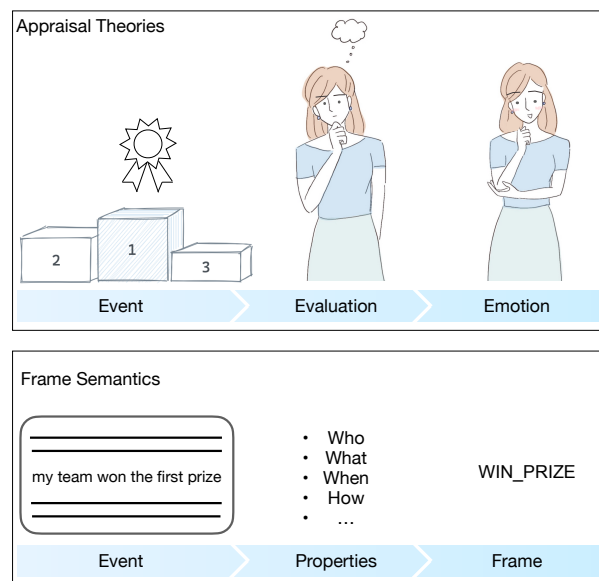


Figure 1: A comparison between the two fields we tap on. Frame semantics studies texts by focusing on events and their characteristics. Appraisal theories, interested in how emotions emerge in humans, also start from the consideration of events; they pay further attention to how event characteristics, as evaluated by individuals, lead to specific emotion reactions.

lustrates this point: frame semantics focuses on abstractions of real-life situations (frames) determined by the structural properties of an event portrayed in text; appraisal theories study emotions as responses to events, whose properties are evaluated by the event participants. The primacy of events in both domains implies that verbal descriptions of emotion-triggering events (e.g., “*my team won*”) can be represented by frames. Other emotion expressions can report (frame-evoking) events as well, from the assessment of the stimuli (“*that’s great*”), to the occurrence of affective experiences and related reactions (“*I’m happy*”, “*I’m all steamed up!*”).

Based on this parallel between the two blocks in Figure 1, this article investigates the relationship between frames and emotions. As a first step, we refrain from analysing different emotions, and concentrate our attention on *emotionality*¹, i.e., whether a text has an emotion content, irrespective of what it is. We ask: are FrameNet frames associated with emotionality? Our expectation is that emotionality represents an integral part of more frames than those indicated in FrameNet as emotion-related ones. By borrowing the definition of emotional experiences from appraisal theories (i.e., emotions as processes engaging events, event evaluations, personal reactions), and assuming that all such

¹We will use “emotion” and “emotionality” interchangeably.

diagnostic features can be communicated via language (e.g., events: “*my team won the match*”, evaluations: “*victory was well deserved*”, reactions: “*I’m happy*”, “*I’m all steamed up!*”), one can conjecture that many frames that are apparently affect-less correspond in fact to the conceptualization of some emotion components. Verifying our conjecture is relevant from two complementary perspectives. For researchers in emotion analysis, we put FrameNet up to scrutiny as a suitable tool to tackle the emotional import of sentences. This could provide insights into the linguistic level at which an affective meaning comes to actualise (e.g., in the relation between words rather than words in isolation), and guide the field towards better automatic text interpretations. For frame semanticists, on the other hand, we inspect whether emotions are an underlying component of the meaning of frames.

At the methodological level, we avoid making assumptions as to which frames are emotional, but exploit an automatic procedure to identify them at scale. We start from a large unlabelled corpus of contemporary American English, on which we add two independent layers of automatic annotation, to label sentences both with binary emotion categories and the frames that they evoke. Then, by investigating the mutual information between the two, we provide evidence that emotionality is an important aspect of frames (by association, not by definition). Besides frames with no emotional import (e.g., `STORING`) and frames that are associated with some degree of emotion (e.g., `CURE`), FrameNet includes a substantial group of strongly emotional frames. Among these are instances evoked by unambiguously emotional predicates (e.g., `EMOTION_DIRECTED`, `FEAR`), and others expressing strongly emotionally loaded events (e.g., `DYING`), bolstering our perspective on the affective dimension of language as described by appraisal theories. As a concrete result of our analysis, we release a resource² with frames-to-emotion associations that can be employed in alternative to typical word-to-emotion lexicons.

The paper starts with an overview of relevant fields. Section 2 introduces emotions, with a focus on how they are studied in text, and Section 3 describes FrameNet, in relation to emotions and the task of semantic role labelling. Section 4 presents the experimental setting used to address our research question. Our main contribution is presented in Section 5, which also elaborates on a possible grouping of the FrameNet emotion vocabulary with a qualitative analysis grounded in appraisal theories, followed by an extensive discussion of its implications in Section 6. We conclude with a summary of the present work and indicate viable ventures for future research.

²Available at <http://www.ims.uni-stuttgart.de/data/FrameEmotionalityMapping>.

2 Emotion in Language

Emotions in Psychology. The body of psychological literature on emotions is extensive and controversial. The field has long established that these states can be investigated systematically (cf. Dixon, 2012, p. 338), but it has reached little consensus on the details, specifically concerning what emotions are, and whether (and which) can be considered cross-cultural universals. Several theories focus indeed on diverse sets of emotions, motivated by specific views on their evolutionary relevance (Ekman, 1992; Plutchik, 2001), or on their underlying dimensions (Russell and Mehrabian, 1977). Ultimately, however, different research lines agree on one point. There exists a handful of “diagnostic features” which indicate that an emotion is taking place (Scarantino, 2016): typically, a starting cause is there (e.g., an event happens); it is evaluated by its experiencers; and it sparks in them some concrete effects, like changes in their voice and posture.

To organise these observations, appraisal theories study emotions in terms of sets of evaluations (Moors et al., 2013; Scherer, 1984). When a stimulus presents itself to an individual, it is evaluated (i.e., appraised) in relation to the individual’s goals, beliefs and desires. For this reason, an appraisal corresponds to specific effects – if I win the competition, I might smile and feel a pleasant sensation because winning supports my well-being; my opponent likely does not have the same reaction. Such effects involve various subsystems, all of which are engaged in an emotion process together with the cognitive appraisal. They consist of a neurophysiological component (i.e., bodily symptoms, like heart beating faster), a motor component (i.e., facial and vocal expressions), a motivational component (i.e., action dispositions), and a subjective feeling component (e.g., winning the competition feels good) (Scherer, 2005).

From psychological research, we retain the idea that an emotion episode involves at least three aspects that can mirror in language: emotion stimuli (i.e., what happens), evaluations (how that is assessed in the light, e.g., of who initiates or is affected by the stimulus), and reactions (e.g., bodily manifestations of emotions).

Emotions in Linguistics. Since emotions are not a primarily linguistic phenomenon, they have remained outside the scope of much work in theoretical linguistics (Kiefer, 1988). Searle’s pragmatic framework (1976), for instance, touches upon expressive acts that convey feelings and attitudes, but it lumps emotions together with multiple other aspects of social interaction.

A more direct account of this phenomenon is given by Martin and White (2003). Tapping into the framework of Systemic Functional Linguistics, they analyse emotion expressions in language, and conclude

that evaluations play a central role. Such evaluations emerge from descriptions of qualities of entities, through modal adjuncts that reflect the position of writers towards an event (e.g., “*sadly, ...*”), through communication of behavioural processes (e.g., “*he smiled at him*”), as well as mental (e.g., “*he liked him*”) and relational ones (e.g., “*he felt angry at him*”). Hence, theories of appraisal, both in psychology and in language, converge on the consideration of embodied manifestations of emotions – either in real life or through language.

Emotions in NLP. The examples above illustrate the data of interest for computational emotion analysis, whose chief task is to classify emotions from text. Works in the field face the choice of following one psychological theory. The selection is usually based on both the textual domain under consideration, as well as its match to the emotions documented by the considered theory. Some opt for dimensional models. Accordingly, they map linguistic data into a continuous space (Preoțiuc-Pietro et al., 2016; Yu et al., 2016; Buechel and Hahn, 2017), like the space comprising the dimensions of valence, arousal and dominance (Russell and Mehrabian, 1977). Others rely on discrete emotion models (e.g., Ekman, 1992; Plutchik, 2001). They associate text to categories like *anger*, *disgust*, *sadness*, either at the sentence-level (Felbo et al., 2017; Li et al., 2017; Schuff et al., 2017) or at the word-level (Mohammad and Turney, 2013; Strapparava and Valitutti, 2004). The latter strand of research leverages the idea that part of a language vocabulary can be described in terms of its emotional meaning (Clore et al., 1987; Hobbs and Gordon, 2011) in order to create affect-oriented lexicons, i.e., resources that formalise the link between emotions and a specific language (Buechel et al., 2020; Chen and Skiena, 2014), encompassing words with an emotion denotation (e.g., the noun *joy*) as well as words with an emotion connotation (e.g., *party*→*joy*).

Only a few works have brought psychological concepts to bear on NLP on a more fundamental level than the acquisition of sets of labels that should be looked for in text (Balahur and Tanev, 2016; Shaikh et al., 2009; Udochukwu and He, 2015, i.a.), and they have rarely relied on a concept of emotions as processes involving complex evaluations (exceptions are Hofmann et al., 2020; Stranisci et al., 2022; Troiano et al., 2023).

Our work differs from previous studies in emotion analysis (Abdul-Mageed and Ungar, 2017; Felbo et al., 2017; Demszky et al., 2020, e.g.) in various respects. We study emotionality instead of a fine-grained set of emotions; we analyse if the emotion information is contained in a well-established resource for semantic role labelling; and we bring together for the first time in the field a theory of emotions (appraisals) with a theory of semantics (frames).

3 Frame Semantics

FrameNet. The theory of frame semantics fundamentally assumes that utterances are understood via frames (Fillmore, 1982). A frame represents a situation fragment that serves to match a word (or a group thereof) to the bundle of knowledge it presupposes (Ruppenhofer et al., 2016). For instance, the term “*abandon*” evokes a conceptual category instantiated by different events (e.g., leaving a membership group, or metaphorically, quitting a bad habit) which comprise a series of participants (e.g., the group being left, the person dropping out of it). The corresponding frame, ABANDONMENT, binds together these bits of knowledge.

For English, the Berkeley FrameNet project (Baker et al., 1998) has been curating the lexical resource FrameNet. It provides an inventory of predicates (lexical units), roles (arguments), and frames. Its latest release (FrameNet 1.7) counts over 13k lexical units and 1.2k frames, which connect to one another via specific frame-to-frame (f2f) relations such as INHERITANCE, SUBFRAME, or USING (Fillmore et al., 2004).

An example for the frame ABANDONMENT from the database³ is in Table 1. ABANDONMENT can be evoked by verbs (boldfaced in the example sentences (1), (2) and (3)) but also by other lexical units such as adjectives and nouns. It has the roles of AGENT and THEME representing the “frame elements” that participate in the situation, where the former expresses the entity leaving the latter. Moreover, this frame links to INTENTIONALLY_AFFECT via an INHERITANCE relation. That is, it inherits properties from this broader conceptual class, and can thus be considered a specific kind of INTENTIONALLY_AFFECT situations.

Frame Identification and Emotion Analysis. In addition to the frame database, FrameNet comprises sentence annotations, like the examples (1), (2), and (3) in Table 1. Such annotations have been used for semantic role labelling (SRL), a task aimed at identifying and labelling the semantic roles that the arguments of a predicate (operationalised as word spans) fill with respect to the event expressed by the predicate (Gildea and Jurafsky, 2002; Màrquez et al., 2008). The specific set of roles depends on the adopted model. Other than FrameNet, PropBank (Palmer et al., 2005) and Abstract Meaning Representation (Banarescu et al., 2013) are commonly used options.

A number of such systems for FrameNet-based SRL have been made available as off-the-shelf tools. Among them are the role labeller that leverages sentence and discourse context by Roth and Lapata (2015), the probabilistic models of Das et al. (2010) which use latent

³Frame definitions can be found at:
<https://framenet.icsi.berkeley.edu/fndrupal/frameIndex>.

Frame: ABANDONMENT	
Definition	An Agent leaves behind a Theme effectively rendering it no longer within their control or of the normal security as one’s property.
Lexical Units	abandon.v, abandoned.a, abandonment.n, forget.v, leave.v
Elements	Agent, Theme
F2F relations	Inherits from: INTENTIONALLY_AFFECT
Example Sentences	<p>(1) Perhaps [he _{Agent}] left [the key _{Theme}] in the ignition.</p> <p>(2) [She _{Agent}] left [her old ways _{Theme}] behind.</p> <p>(3) Abandonment [of a child _{Theme}] is considered to be a serious crime in many jurisdictions.</p>

Table 1: Example of a FrameNet frame. In the three example sentences, boldfaced words are frame-evoking predicates, bracketed words are arguments.

variables of lexical-semantic features to facilitate frame predictions for unknown predicates, and the labeller of Swayamdipta et al. (2017) that detects FrameNet frames and frame-elements.

Frame-based semantic parsers have proven useful in applications like text-to-scene generation (Coyne et al., 2012) and question answering (Shen and Lapata, 2007). Yet, they have never been fully leveraged to address emotions. For example, Ghazi et al. (2015) annotated 820 FrameNet sentences with emotions, but these were sampled based on their link to only one emotional frame (i.e., EMOTION_DIRECTED). On the other hand, the research line in emotion analysis centered on semantic roles (Mohammad et al., 2014; Oberländer and Klinger, 2020; Oberländer et al., 2020) identifies the portions of texts corresponding to emotion causes, emotion holders, and eventually, the targets towards which an emotion is directed, but it disregards frames.

Being the first study that links frame semantics and emotion analysis, we concentrate on frames and leave roles aside. These have an important function which we use implicitly as means that help identify frames in context. For example, they provide a cue that the conceptual situation evoked by, e.g., the predicate “*treats*” in “*the doctor treats the patient with aspirin*” can be distinguished from that in “*the bully treats the student with disdain*”, but we leave the specific analysis of the relationship between roles and emotions for future work.

Emotions in FrameNet. Frames appear to be a valuable formalism to study emotions because FrameNet has an affective core: a small part of the database is ostentatiously concerned with emotions (e.g., FEAR), and some of the others can be traced back to a relevant emo-

tion frame through the relations present in the database – for instance, FLEEING can be related to the FEAR frame via the USE relation (Ruppenhofer, 2018). Past research has indeed provided qualitative evidence of the emotional quality of various frames (Ruppenhofer et al., 2016), but it has done so by focusing on a limited and pre-defined vocabulary of items. In fact, the exact set belonging to the emotion domain is not spelled out, partly because FrameNet is a database under constant development, and partly because emotional meanings are only one type of the world knowledge inferences that can be made from frames – representing all of them would be unfeasible for the FrameNet curators. Our approach can identify them automatically and at large.

In his manual analysis of the emotion domain in FrameNet, Ruppenhofer (2018) discusses the criteria that guided the allocation of lexical units under specific frames. Some of them are the constraint that the lexical units in a frame should accept the same types and number of syntactic dependents, and the idea that specific frames are differentiated by the role of subject/object that is filled in by an emotion participant (EXPERIENCER.SUBJ/EXPERIENCER.OBJ). According to such criteria, words that indicate different emotions can fall within the same group of predicates. Conversely, words with the same lexical root are allowed to be part of different frames (e.g., the verb “*anger*” belongs to EXPERIENCE.OBJ together with the verb “*please*”, but the adjective *angry* does not). There is in this sense a crucial difference between a dictionary-based approach to emotion analysis and a frame semantics one: The latter organises emotion words by reflecting similarities between their linguistic realisations, more than to account for their glossary characterisation.

While we employ frames as a way of grouping words, one could opt for other semantic word organisations to study the affective dimension of meaning. For example, WordNet (Fellbaum, 1998) arranges words into a large network of relations potentially useful for our goal. However, FrameNet has an important advantage over other lexical databases. Its construction principle is not focused on words per se but on the frames that these evoke, as (interrelated) classes of events (Baker and Fellbaum, 2009). This allows to capture the emotional closeness between words that might be far apart in regards to their grammatical classes and meaning (e.g., the noun “*pleasure*” and the verb “*abhor*”), but which belong to the same event class in FrameNet (e.g., both “*pleasure*” and “*abhor*” are lexical units of EXPERIENCER_FOCUSED_EMOTION).

4 Methods

Our goal is to study (a) to what extent (i.e., quantitatively) the emotionality of texts is mirrored in the frames that the texts evoke; (b) if there is a qualitative difference between the emotionality that frames carry; and whether (c) these aspect can help in starting a discussion of emotions in FrameNet. FrameNet contains a narrow emotion nucleus, but for most of the frames their ‘emotionality status’ (whether or not the situation is emotional) is not specified. This constitutes the core of our investigation.

Accessing data with the two types of information that we need is not straightforward. No resource for emotion analysis is labelled with frame semantics information, except for the dataset by Ghazi et al. (2015), which is limited in size and only includes emotion-bearing texts. Likewise, corpora for frame-semantic parsing do not contain emotion annotations – at least, not for the vast majority of frames. As a solution, we devise a method that combines the use of neural technologies and prior knowledge about language as contained in FrameNet: we correlate the categorical variable of emotionality (obtained through an emotion classifier) with that of frame membership (grasped by a frame identification tool).

We use this correlation to find categories of frames (inherently emotional, inherently nonemotional, and others) and to explain their belonging to one category or another in quantitative terms. Focussing on the emotional frames, we conduct a qualitative discussion based on Scherer’s theory (1984), which explains emotions as processes involving the subsystems of an organism (cognitive, motivational, motor, etc.), and has a theoretical counterpart in linguistics.⁴

⁴There exist also other appraisal-based theories, like the OCC model (Ortony et al., 1988) which describes the eliciting conditions of emotions (i.e., consequences of events, agents’ actions and as-

Data. We base our study on an unlabelled corpus, the 2020 version of COCA⁵ (Davies, 2015), which is much larger than any existing resource for emotion analysis.⁶ Its texts were collected from 1990 to 2020 in different domains, namely blogs, magazines, newspapers, academic texts, spoken interactions, fiction, TV and movie subtitles, and webpages. Except for academic texts, which have an arguably impartial language, we consider all other domains, split their paragraphs into sentences, exclude sentences containing words that are masked for copyright reasons and those with less than 3 tokens (tokenization performed with the python library *nltk*⁷). The preprocessed data that we use comprises ~44M sentences and ~536M tokens.

Bridging Data-driven Learning and Semantic Resources. To obtain frames and emotion information, we bypass the use of human annotation which would be prohibitively expensive. We resort instead to an automatic procedure, adopting a two-step methodology illustrated in Figure 2. First, texts are associated with emotion labels (through an emotion classifier) and frames (via a tool for frame identification); second, we carry out a corpus-based correlation analysis where the association between the two annotation sides is quantified and interpreted.

Because this approach exposes us to the risk of mistakes made by the emotion classifier and the frame identifier, we adopt experimental design strategies that boost the robustness of our empirical observations.⁸ One is to employ a corpus with a considerable number of datapoints, which showcase a variety of linguistic realizations of emotions, and evoke frames across both emotion-bearing and nonemotional expressions. Second, we carry out the emotion annotation with classifiers learnt on multigenre data, a strategy that promotes the generalization ability of emotion detection models (Tafreshi and Diab, 2018); for frame labelling, we use an artificial neural network-based technology that has shown to generalise well over unseen sentences and predicates (Swayamdipta et al., 2017). Third, we evaluate the emotion classifier against a manually-annotated sample of our texts as an additional check of

pects of objects), how these are appraised along binary criteria (e.g., desirability–undesirability), and how specific evaluations cause emotions deterministically (e.g., if a condition holds, a certain reaction follows). Yet, the OCC model does not fit our goal. It sees emotions as descriptive structures of prototypical situations, and its binary evaluations, which are purely conceptual constructs, have little to do with the linguistic expression of events. By contrast, the tool that we use for event representation, frame semantics, is primarily linguistic and might not match the conceptual considerations of the OCC.

⁵<https://www.english-corpora.org/coca/>

⁶An overview of existing resources in computational emotion analysis can be found in Bostan and Klinger (2018).

⁷<https://www.nltk.org>

⁸We discuss the limitations of our approach in Appendix A.

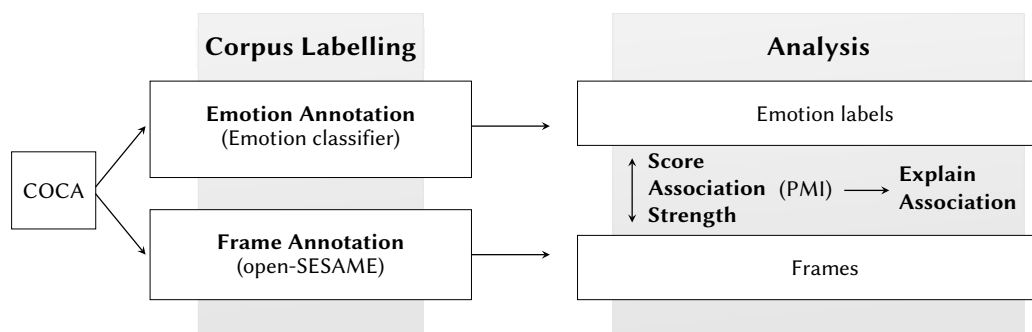


Figure 2: Our two-step experimental setting. **Corpus Labelling**: automatic annotation of sentences extracted from the corpus of contemporary American English with emotions and frames, separately, with the emotion classifier being evaluated on a subset of the corpus previously annotated by human judges, and the tool for frame identification evaluated on a subset of MASC as out-of-domain data. **Analysis**: the two strands of annotations are brought together via PMI, to first score and then explain the association between frames and emotionality.

its reliability, and we do the same for the frame identifier using out-of-domain data. Lastly, we conduct statistical analyses to limit the role of chance in positing frame-emotion associations, and we explain them qualitatively as a safeguard of the quality of our findings.

We now proceed to describe the individual components shown in Figure 2.

4.1 Corpus Labelling

As a first step, we label texts with emotion- and frame-related information. The systems used here are trained separately on different corpora. It is thus necessary to assess their domain independence and get insight into how well they apply to COCA.

Emotion Classification. We start by gathering various resources for emotion analysis that span textual domains similar to those in COCA, from webpages to literary texts: GoEmotion (Demszky et al., 2020), Grounded-Emotions (Liu et al., 2007), Emolnt (Mohammad and Bravo-Marquez, 2017), TEC (Mohammad, 2012), SSEC (Schuff et al., 2017), enISEAR (Troiano et al., 2019), ISEAR (Scherer and Wallbott, 1997), Tales (Alm et al., 2005), DailyDialogs (Li et al., 2017), and Emotion-Stimulus (Ghazi et al., 2015). These datasets feature diverse emotion schemata; we make them consistent to our binary setup by mapping their original labels into the *nonemotional* and *emotional* classes, depending on whether a text was marked as having no emotion, or as having one out of a rich set of alternatives (e.g., *joy*, *fear*, *disgust*, *hope*, *surprise*, *guilt*).

Instead of extracting our test set from this data, we use a portion of COCA. Made available by Troiano et al. (2021), the sample contains 700 texts labelled at the sentence level by three in-lab raters.⁹ They are balanced

⁹<https://www.ims.uni-stuttgart.de/data/emotion-confidence>

across the domains that we consider, and their annotation encompasses the same binary categories of our concern. The nonemotional label corresponds to the absence of any emotion content, the emotional class represents sentences that display either of two qualities: (1) having an emotion as a central component of their meaning, thanks to the presence of an emotion word (as in “*I am so happy to see you*”) or the description of an internal state of an entity (“*And there she was, desperate for her family*”); (2) describing an event, a concept or a state of affairs to which the annotators would personally associate an emotion (“*She was being pretty arrogant to me*”, “*I saw my best friend*”). The annotators were tasked to judge the texts by giving their own emotion reaction, and not to try and reconstruct that of the text authors. Thus, they were allowed to associate similar events to different labels. For instance, the passing away of an unknown entity could be linked to a nonemotional judgment, while that of a person resonating with their own experience (e.g., the mention of a pet) could receive the opposite label.

Next, we train multiple models on the concatenation of the selected (training) resources: we fine-tune BERT (Devlin et al., 2019) models¹⁰, adding a classification layer that outputs the labels *emotion* or *nonemotional*. Different models are obtained by varying the data on which they learn the classification task: the rationale is to identify a subset of training resources that yields a classifier capable of reliably judging out-of-domain data (i.e., COCA). Hence, we evaluate each model on the manually annotated COCA sample, with the majority vote determining the ground truth.¹¹ We

¹⁰<https://huggingface.co/docs/transformers/>

¹¹Associating the 700 sentences to the majority vote resulted in 474 emotional and 226 nonemotional data points. Cohen’s κ (1960) agreement between this ground truth and the three annotators was .6, .8, .6, respectively. The annotators’ decisions were unanimous for 304 emotional and 88 nonemotional instances.

	Frame Id		
	P	R	F1
FrameNet 1.7	.85	.85	.85
MASC	.78	.78	.78

Table 2: Evaluation of the frame identifier provided by Swayamdipta et al. (2017) against FrameNet data and MASC frame-annotated data.

pick the model that performs best on this test set to annotate the rest of the corpus. It reaches a performance of .67 F1 score¹²) Details on model selection are in Appendix B.

Frame Identifier. Models and corpora for semantic role labelling are scarcer than emotion-centered ones. Here, we require a system which, given a sentence, identifies the set of FrameNet frames that are evoked by each of the predicates, as well as the corresponding predicate arguments. To this end, we use open-SESAME¹³. Developed by Swayamdipta et al. (2017), it is a freely available interpreter for SRL with state-of-the-art performance, based on segmental recurrent neural networks (Kong et al., 2016). We re-train the provided implementation¹⁴ using the sentences from the FrameNet release 1.7 (7340 for training, 387 for dev, and 2420 for testing).

We evaluate it on the FrameNet test set as in-domain data, as well as on external data. For that, we use 695 sentences (516 of which are frame-evoking) coming from MASC¹⁵ (Ide et al., 2010), a subset of the Open American National Corpus that provides useful annotations for frame identification. MASC’s texts include emails, essays, fiction, spoken transcripts, and hence, using it as a benchmark illustrates how the frame identifier performs on linguistic expressions similar to those found in COCA.

Precision, recall, and micro-averaged F1 for this frame identification task (Frame Id) are reported across both test sets in Table 2. We obtain these results using the script by Swayamdipta et al. (2017) on the full-text FrameNet annotation. When moving to out-of-domain data, we see a drop in performance (from F1=.85 to F1=.78), which might be partially due to an increase in

¹²This performance is not state of the art in emotion classification. However, systems for emotion detection that work well on existing labelled resources might not perform equally well on COCA. We varied the model architecture and noticed that a model that achieved better results on in-domain data suffered from major performance loss when evaluated on the manually-annotated subsample of COCA. See Appendix B for a discussion of these classification results.

¹³<https://github.com/Noahs-ARK/open-sesame>

¹⁴Training hyperparameters as in Swayamdipta et al. (2017).

¹⁵Downloadable at: <https://www.anc.org/MASC/download/MASC-1.0.3.tgz>

the sentence length (avg. for the FrameNet test = 16.5 tokens, for the MASC test = 23.4 tokens) and in the average number of frames per sentence (2.8 for FrameNet, 6.5 for MASC). Still, we take these numbers to be sufficiently high that the frame identification system can be used to proceed with the annotation.

4.2 Analysis: Investigating Emotionality in Frames

Once COCA is labelled with emotionality and frames, we can finally proceed to our research question: are FrameNet frames associated with emotionality? Estimating the degree of this association requires an appropriate alignment strategy, as the labels we obtained differ in granularity: emotions refer to entire sentences, while the output of the frame parser relates to tokens. We choose the most straightforward alignment strategy: considering each frame in a sentence as having a separate and full-fledged alignment with the sentence-level emotionality label. This choice is a simplification, because the frame parser could identify multiple frames for an input sentence, and emotionality might be attributed to their inter-relation rather than their individual contribution. However, this is a transparent approach, comparable to related work such as aspect-based summarization in sentiment analysis (Hu and Liu, 2004), where multiple aspects identified at the sub-sentence level are grouped under the same sentiment label. The literature offers various weighting schemes to refine such alignments, but not all weighting schemes work equally well for all tasks (Buckley, 1993; Pekar et al., 2004; Ushio et al., 2021).

To identify patterns of frames occurring with emotionality status (emotional/nonemotional), we compute pointwise mutual information (PMI) (Church and Hanks, 1990). This information-theoretic measure quantifies the dependence between the values that two discrete random variables can take, and accounts for their chance co-occurrence. More specifically, PMI compares the probability of observing two variables together, against that of observing them independently, or by chance. In our case, the variables are the output labels of the automatic annotation procedure from the corpus labelling step. For each pair (f, e) consisting of a frame and an emotionality label, we estimate PMI as the number of times that such frame and emotionality label co-occur in the entire corpus, divided by the product of their individual frequencies. Formally, for each f and e, we compute

$$\text{PMI}(f;e) = \log_2 \frac{p(e, f)}{p(e)p(f)} = \log_2 \frac{p(e | f)}{p(e)}.$$

As already mentioned, the number of extracted pairs (frame f, emotionality e) varies from sentence to sen-

	Emotional	Nonemotional
Sent. with frames	19.717.813	16.092.214
Sent. w/o frames	4.194.783	2.141.299
Number of frames	75.889.290	57.517.465

Table 3: Outcome of Corpus Labelling: number of sentences associated with the emotion and nonemotional labels, both with frames and evoking no frames, and number of frames.

tence, depending on whether one or many frames are evoked.

PMI does not have predefined bounds. Positive values indicate that a frame and an emotion connotation are semantically associated: they appear together more than one could expect by considering the two events independently. A PMI=0 indicates that there is no dependency between the two variables (i.e., emotionality and frames). Lastly, negative values indicate that f co-occurs with the considered e with less than chance expectancy and therefore is associated more with the opposite emotion label.

5 Emotionality-Frame Associations

The processing steps described in Section 4.1 result in two independent layers of annotation for the same texts, for which Table 3 shows statistics: the emotion classification module results in ≈ 23 M sentences labelled as emotional and ≈ 18 M as nonemotional. From this total, ≈ 6 M sentences (i.e., ≈ 4 M emotional and ≈ 2 M nonemotional, row “Sents. w/o Frames”) are not associated with any frame by the frame identifier. In our analysis, we do not consider these frameless sentences, which typically consist of short texts like “*That’s what it was*” and “*-No, it’s not a guy*”. For all others (row “Sents. with Frames”), the role labeller identified 133M frames, specifically in the 76M emotional sentences and 57M in the nonemotional counterparts, with an average of 3.7 frames per sentence.

Given these numbers, we focus on the 758 unique frames which appear at least 50 times in either textual domain of COCA and analyse the PMI between those and emotionality¹⁶, as reported in Figure 3. One might argue that emotionality, when expressed in language,

¹⁶In this binary classification setup, the distributions given by $\text{PMI}(f; \text{emotional})$ and $\text{PMI}(f; \text{nonemotional})$ are essentially symmetric: Frames which are positively correlated to one label are negatively correlated to the other. E.g., the frame `MORALITY_EVALUATION` illustrated in Figure 3: $\text{PMI}(f; \text{emotional})=.44$, while $\text{PMI}(f; \text{nonemotional})=-.8$. For this reason, we only report the emotional distribution.

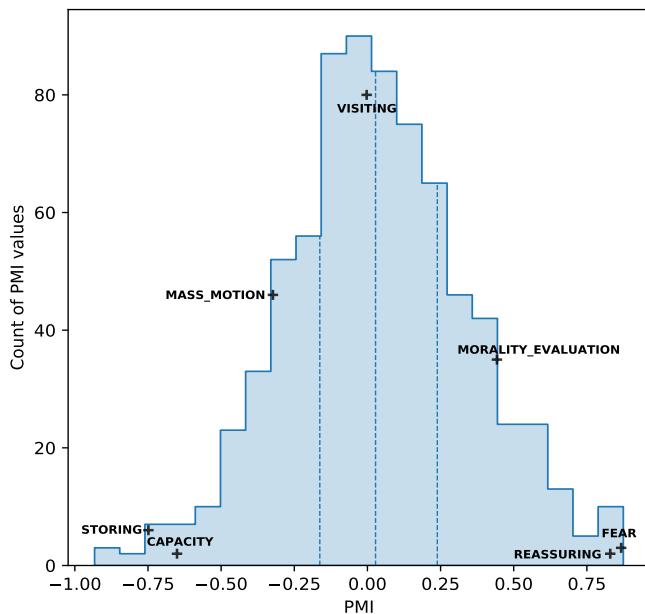


Figure 3: Histogram of $\text{PMI}(f; \text{emotional})$. Dashed lines: beginning of the second, third and fourth quartiles.

falls on a spectrum. Different frames can convey varying degrees of emotion, depending on factors such as context, cultural nuances, and more. But to navigate and discern patterns within this continuum, we leverage the simplifying assumption that frames comprise: a predominantly emotional vocabulary (larger than the one openly designated as emotional in FrameNet); vice versa, a set of frames that count as nonemotional; frames that can be either emotional or not, whose status is determined by the context in which they appear – they basically mirror words that dictionary-based emotion models in computational emotion analysis associate with different emotions (e.g., “*abundance*” in the lexicon of Mohammad and Turney (2013) is mapped to anticipation, disgust, joy, and trust).

We need to find lists of frames belonging to these three groups in order to evaluate our assumption. The distribution of PMI values in Figure 3 does not naturally provide such a tripartition. We could define it in many ways, for instance using $\text{PMI}=0$ to decide on what counts as emotional and what not. However, we adopt the quartiles of the PMI distribution because they represent a good balance between the precision and recall of our findings: as opposed to a binary separation, they shield us from considering as emotional some frames with a minimally positive PMI (due, e.g., to bias in the data or mistakes of either automatic labeller); compared to more restrictive cuts (e.g., taking the top 10% of frames as emotional), they facilitate our analysis of what frames other than those already known to be emotional are so.¹⁷

¹⁷We rely on the thresholds for a structured and clear analysis. This

Hence, we consider the top quartile of the distribution ($PMI \geq .24$) to correspond to frames that are consistently emotional across various contexts, as it identifies the highest 25% of PMI values positively associated with the label *emotional*. Frames in the bottom quartile ($PMI \leq -.16$) will henceforth be treated as *nonemotional*. Both the emotional and nonemotional quartiles encompass 190 frames.

All other frames fall within the second and third quartiles of the distribution: the fact that the PMI values in Figure 3 are approximately normally distributed around 0 does not indicate the absence of a correlation between emotions and frames; it rather tells us that a large group of items are neither strongly associated with nonemotionality nor with emotionality. These 378 frames will be referred to as *contextually determined* for reasons discussed in Section 5.3.

In the following subsections, we characterise the emotional, nonemotional, and contextually-determined frames, validating the findings of the PMI procedure with a detailed qualitative analysis.

5.1 Emotional Frames

The procedure from the previous subsection has provided us with a set of frames which are purportedly emotional. In order to better understand *how* PMI values relate to the emotional aspects of frames, we ask two questions: (a), how do PMI values vary within frames? (b), how can we characterise emotional frames – can we find a clustering that is coherent according to both qualitative and quantitative criteria?

5.1.1 PMI Values across Lexical Units

In order for the notion of an “emotional frame” to have substance, we need to show that emotionality is not just the result of a small number of frequent, highly emotional lexical units in the frame, but that rather (almost) all of the frame’s lexical units are emotional.

To assess whether this is the case, we compute the PMI between the label *emotional* and the lexical units of the 35 most emotional frames. We observe indeed that the frames’ PMI values remain consistent across units, with minor variations. Examples are “*frightened*”, “*afraid*” and “*terror*”, having a PMI score of .86, .85 and .81, all close to the .86 of the corresponding frame FEAR.

heuristic may appear as predefining the three groups ad-hoc. But our goal is not to propose a conclusive categorization of frames in three classes that we assumed to find. Instead, we aim at understanding what brings together frames fallen under one category (see following sections). The sizeable presence of contextually-determined frames, for example, could be dismissed as an influence of the textual genre in which they appear. Our inquiry asks: Is there anything else that makes them more emotionally variable than the others? Future work could explore, e.g., clustering methods that provide a categorization of frames without the quartile-based division.

This tendency holds mainly for the units of frames overtly defined in terms of emotions (like FEAR with a standard deviation across lexical units of .03), but also for others, like the adjectives “*sickening*” and “*troubling*” which have a statistical association to emotionality comparable to that of STIMULUS_FOCUS, that they evoke (.70, .68 and .68, respectively; standard deviation across all units of the frame: .28), as well as “*fiasco*” (PMI= .65) and “*ruin*” (.69), headed by BUNGLING (PMI=.66, standard deviation: .31). Exceptions are lexical units that appear to have little subjective connotation. For example, for STIMULUS_FOCUS, the noun “*relaxation*”, which has a less prominent evaluative undertone than the above-mentioned adjectives, deviates noticeably from the frame’s PMI (.31). These numbers show that emotional frames display a fair degree of internal consistency concerning their emotionality.

5.1.2 Characterising Emotional Frames

Figure 4 (a) illustrates the 35 highest PMI-valued frames – some COCA sentences in which they appear are shown in Table 4. This small subset hints already at the diversity of the 190 emotional frames, which capture situations ranging from circumstances of interpersonal communication (e.g., OPINION, REVEAL_SECRET, WARNING) to actions (e.g., RUN_RISK), from internal motives (e.g., WILLINGNESS, RENUNCIATION) to social circumstances (e.g., HOSTILE_ENCOUNTER, PREVARICATION). A handful of these frames, like FEAR or EMOTION_ACTIVE, has a clear emotional quality. They are treated in FrameNet itself as such. However, for almost all of them (e.g., FAIRNESS_EVALUATION), an emotion content is more opaque and warrants investigation.

This diversity suggests that we need to corroborate the emotionality of the instances in the top quartile of the PMI distribution. We do that by conducting a qualitative analysis to define a few frame clusters that share emotion-related characteristics, followed by a quantitative discussion to validate our findings.

Qualitative Evidence. We build upon a discussion of the emotion vocabulary initiated by Ruppenhofer (2018). Its core idea is that it is instructive to “examine to what extent the notions FrameNet uses for its analysis do match ones found in psychological theories” (p. 96), as a way of relating the experts’ understanding of emotions (formalised in theories and definitions) to the folk’s understanding of their experiences (captured, e.g., by FrameNet), and linking psychological views on emotions to linguistic analyses.

We put this view into practice by manually clustering the 190 frames into different groups that map either to the emotion vocabulary in FrameNet, or to theoretically-motivated emotion properties. The clusters are:

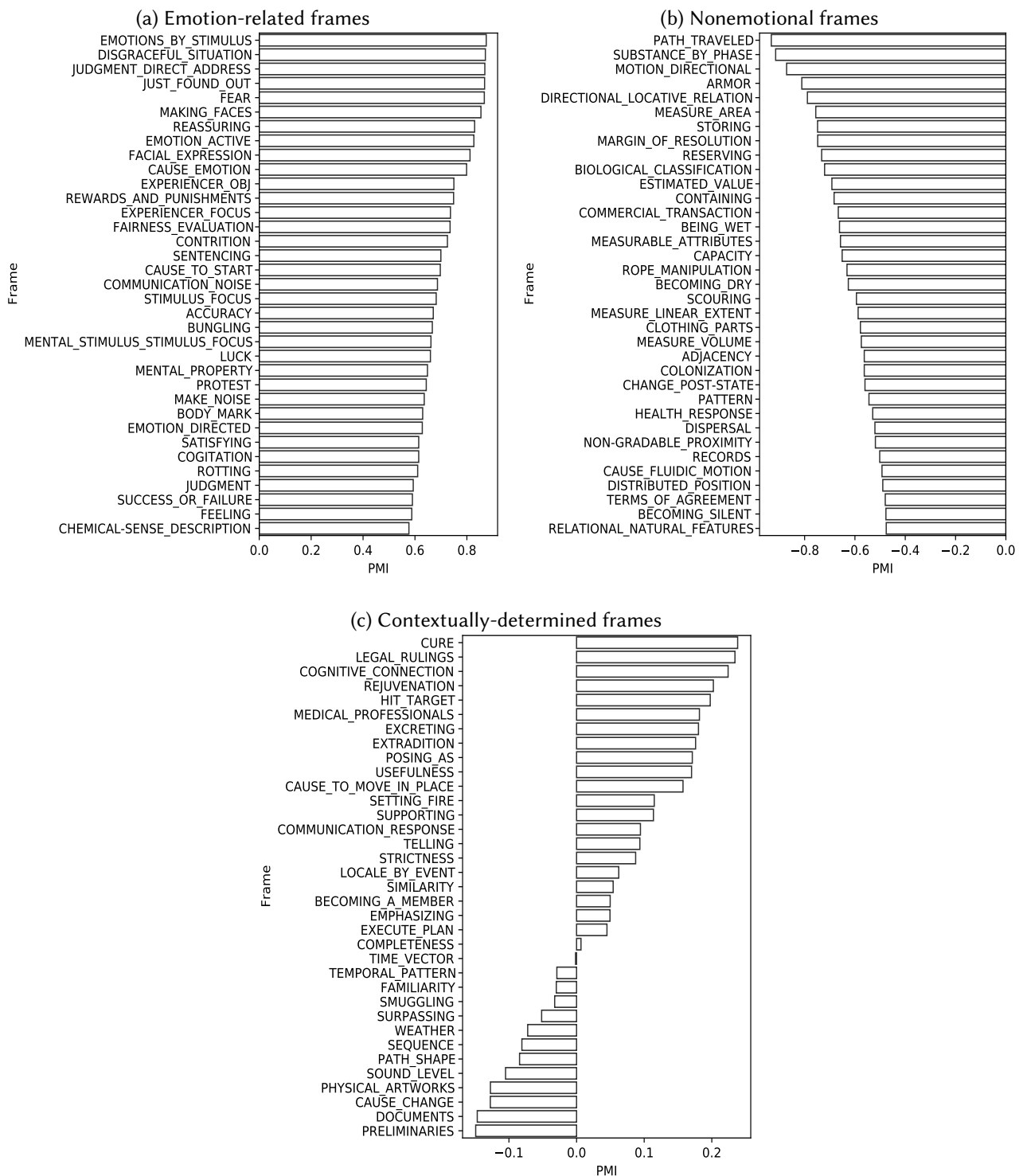


Figure 4: (a) The 35 frames with the highest PMI values in the emotional distribution, in comparison to the frames with the lowest values (b) and in-between these two extremes (c). See Table 4, 10 and 11 for example sentences in which these frames are evoked.

Frame	Text
JUDGMENT_DIRECT_ADDRESS	Oh, thank God, thank God you're not mad at me for pushing you that day.
EMOTIONS_BY_STIMULUS	So glad we're friends .
DISGRACEFUL_SITUATION	This is outright, outrageous, disgraceful, disgusting.
REASSURING	He spoke with a dentist's tone of calm reassurance.
CAUSE EMOTION	The whole thing was quite pathetic, really, and insulting to boot.
EXPERIENCER OBJ	I am surprised the judges bought it.
COMMUNICATION_NOISE	For the first week I cried.
STIMULUS_FOCUS	The silence of the candidates is amazing.
LUCK	Fortunately, adventure found him in college.
PROTEST	He marched, he organized, he protested, he was gassed, he was beaten, he was jailed.
CONTRITION	Blinking furiously, looking furiously guilty, Jimmy Lowe says, "All's I did – Ziggefoos cuts him off."
FAIRNESS_EVALUATION	To the guy who is whining about how this would be so unfair if it were applied to any other social or racial group God, get over yourself.
EMOTION_DIRECTED	And – and she just made you happy.

Table 4: Examples of emotional frames with sentences in which they appear.

	Mean	St.dev
Overtly Emotional	.68	.15
Emotion Stimuli	.42	.14
Appraisal-based	.43	.15
Incidentally Emotional	.32	.07
All Emotional Frames	.43	.16

Table 5: PMI mean and standard deviations over lexical units within each cluster and across all 190 emotional frames.

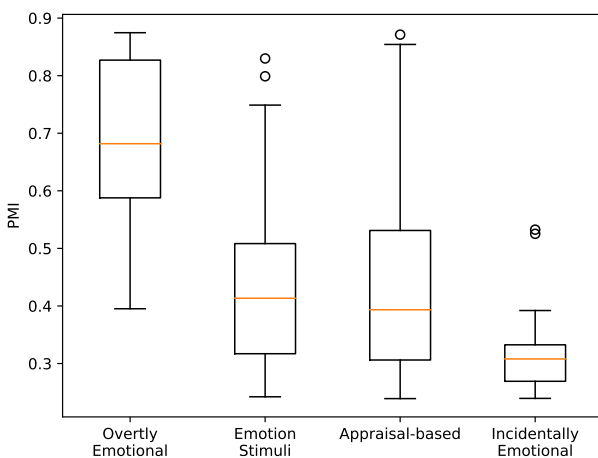


Figure 5: PMI values for each cluster.

(1) **Overtly Emotional** frames;

(2) frames that express events or concepts that might cause an emotion (i.e., **Emotion Stimuli**);

(3) **Appraisal-based** frames, which capture diagnostic features of emotions (e.g., emotion manifestations) or cognitive evaluations of situations (i.e., the factor that appraisal theories see as fundamental for an emotion to occur).

(4) **Incidentally Emotional** – the remaining frames in the top quartile which cannot be given a straightforward interpretation in terms of the first three clusters.

The list of frames classified as either cluster is given below in Tables 6 through 9. We will show that frames in each group share a common affect-laden ground, despite their variety. Before we dive into a qualitative analysis, however, we inspect some quantitative evidence.

Quantitative Evidence. While the guiding principle of our annotation is theoretically driven, the frames' membership in either cluster is our empirical decision. Actually, some items could fit into multiple clusters: HIT_OR_MISS and ATTEMPT, which have to do with the goals and concerns of an experiencer (much in an appraisal-oriented fashion), could also be arranged among the Emotion Stimuli; DESIRABILITY, that we annotated as (1), expresses a positive stance towards a circumstance and could belong to (3). Indeed, there is a large number of frames from separate clusters that are directly related to one another (e.g., a USING relation

holds between MISDEED, which we placed in the Emotion Stimuli, and MORALITY_EVALUATION).

Therefore, we look for quantitative validation of our annotation: Table 5 contains mean and standard deviation for each cluster across lexical units (cf. Section 5.1.1), and Figure 5 reports the per-cluster distribution of PMI values. Both corroborate the observation that items from cluster (1), Overtly Emotional, are clearly separate from the others, and cover the highest PMI values overall. This is likely due to directly emotional frames being less prone to be contextualized in text in a nonemotional manner, because they inherently signify emotion concepts. By contrast, frames in cluster (2), Emotion Stimuli, have the potential to elicit an emotional response but can be more easily contextualised without an emotional tone. For example, FEAR, in (1), *denotes* an emotion concept, while DEATH, in (2), arguably has an emotional *connotation* that could or could not be manifest in text. We perform a Mann Whitney U test between pairs of groups, as a way of controlling if the difference between the PMI values of the corresponding frames is statistically significant. This is (partially) the case: $p\text{-value} < .05$ for each pairwise comparison, except for the difference between clusters (2) and (3). Considering the conceptual overlap of these two categories¹⁸, as well as the fact that their distinction does not reflect linguistic or semantic properties but constructs from psychology, we take this outcome as a confirmation of our initial assumption: some frames are straightforwardly emotional, while the emotionality of many others can be made sense of thanks to appraisal-grounded concepts.

5.1.3 Four Clusters of Emotional Frames

We now discuss the outcome of our manual classification in more detail, and visualise how it identifies coherent clusters in FrameNet.¹⁹

Overtly Emotional. This cluster encompasses 17 frames that are direct children of the node EMOTIONS (or children of its children), and can thus be considered to have an emotional status in FrameNet. Examples are JUDGMENT, EMOTION_DIRECTED and STIMULUS_FOCUS, FEELING and CONTRITION which express the internal state caused by an emotion episode. The whole list of members is in Table 6, together with the definition of this

¹⁸Cluster (3) includes qualities of stimulus events (e.g., OPPORTUNITY) and following reactions (AGREE_OR_REFUSE_TO_ACT), which can also be considered as events themselves.

¹⁹For simplicity, Figure 6, 7 and 7 include only frames among the 100 with the highest positive emotional associations and do not show relations between all frames. Note that the grey nodes are not among the top 100 frames. They are illustrated to reproduce the FrameNet structure and account for how the frames under consideration (text in black) relate to one another through relations (represented by the coloured arrows, each corresponding to a specific type of relation).

group that we used as a guideline for the task. Figure 6 illustrates them. Circled grey frames are frames, such as EMOTIONS, that are not part of the cluster but are necessary to connect the individual frames.²⁰ The figure demonstrates how our PMI-based analysis aligns with the FrameNet database and in particular the frame-to-frame relations. The fact that these frames form an almost connected component in FrameNet corroborates the intrinsic emotionality of its affective vocabulary.

Emotion Stimuli. 72 frames express emotion-inducing circumstances. They are shown in Table 7 and visualised in Figure 7. The frame EVENT is included in the visualisation despite belonging to the Incidentally Emotional cluster, because it delineates a generic super-category from which all other specific events branch out.

Recall that in the view of appraisal theories, events are causes of emotions: they make emotions different from other affective states, such as mood, which are more independent from the environment. Our second group of frames captures precisely this notion. It comprises items that revolve around emotion-stimulating circumstances, like ROTTING and DESTROYING, and therefore, can account for the emotionality assigned to texts that convey an affective content via purely factual descriptions. In this light, this cluster is also close to the idea underlying emotion lexicons, namely, that some words evoke mental representations that have a prototypical affective substrate, somewhat established in the collective knowledge.

For some of them, an emotional attachment might result weak at first glance, but it is clarified by looking at the texts in which they appear. MAKE_COMPROMISE, for instance, is typically evoked by sentences that bring up people sacrificing self principles; CAUSE_TO_FRAGMENT is evoked by texts depicting an entity being “broken” (e.g., being hurt by a breakup). There are also instances that do not indicate events strictly speaking, but kin concepts. Two examples are VIOLENCE and HOSPITALITY, recognised by the frame identifier in sentences that manifest appreciation for conviviality.

Appraisal-based Frames. The third cluster of 76 frames is reported in Table 8, some of which are displayed in Figure 8. This cluster formalises implicitly emotional cases like Emotion Stimuli, but it captures either properties of events, as evaluated emotion experiencers, or other emotion components that manifest in the experiencers’ reactions. Similar to events, these are given a prominent role by appraisal theories: the emotion mechanism involves an experiencer who assesses

²⁰EMOTIONS has only a single lexical unit, the noun *emotion*, which is generally used to refer to, rather than express, emotions.

Definition These frames are direct children of the node EMOTIONS. They must be its immediate derivation, or a derivation of one of its children nodes.

Frames 1. EMOTIONS_BY_STIMULUS, 3. JUDGMENT_DIRECT_ADDRESS, 4. JUST_FOUND_OUT, 5. FEAR, 8. EMOTION_ACTIVE, 11. EXPERIENCER_OBJ, 13. EXPERIENCER_FOCUS, 15. CONTRITION, 19. STIMULUS_FOCUS, 22. MENTAL_STIMULUS_STIMULUS_FOCUS, 28. EMOTION_DIRECTED, 32. JUDGMENT, 34. FEELING, 36. DESIRABILITY, 40. AESTHETICS, 92. PREDICAMENT, 94. DESIRING

Table 6: Overtly Emotional frames. Each frame is numbered according to its PMI rank.

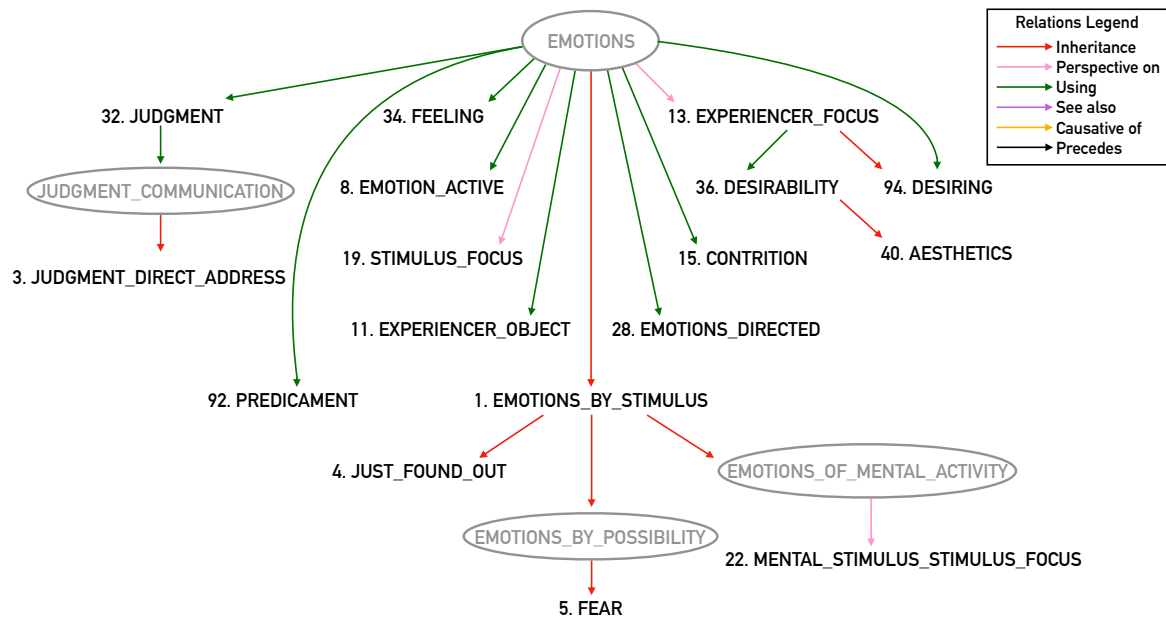


Figure 6: Emotional frames (text in black), which are children of the node EMOTIONS, corresponding to Table 6.

Definition These frames express circumstances that can cause an emotion.

Frames 7. REASSURING, 10. CAUSE_EMOTION, 12. REWARDS_AND_PUNISHMENTS, 16. SENTENCING, 17. CAUSE_TO_START, 21. BUNGLING, 25. PROTEST, 31. ROTTING, 39. KILLING, 41. BEAT_OPPONENT, 42. FIRING, 43. DESTROYING, 45. TERRORISM, 46. DARING, 47. VERDICT, 48. FINISH_COMPETITION, 50. OFFENSES, 55. DEATH, 56. RECOVERY, 57. SUASION, 60. KIDNAPPING, 62. CAUSE_TO_EXPERIENCE, 66. CAUSE_HARM, 67. REVENGE, 69. CATASTROPHE, 70. MISDEED, 71. ARREST, 72. PREVENT_OR_ALLOW_POSSESSION, 75. IMPRISONMENT, 80. ACCOMPLISHMENT, 81. VIOLENCE, 83. SUCCESSFUL_ACTION, 84. RENDER_NONFUNCTIONAL, 87. UNEMPLOYMENT_RATE, 88. WARNING, 89. FORGING, 90. RENUNCIATION, 93. ASSISTANCE, 100. ENTERING_OF_PLEA, 101. REBELLION, 106. ATTACK, 107. REPEL, 108. HOSTILE_ENCOUNTER, 110. ENDANGERING, 111. CAUSE_TO_FRAGMENT, 113. RESCUING, 116. PREVARICATION, 119. SUBVERSION, 121. RESOLVE_PROBLEM, 122. EXPERIENCE_BODILY_HARM, 124. ARSON, 129. MEDICAL_CONDITIONS, 134. EXAMINATION, 138. INFECTING, 143. RUN_RISK, 152. ENDEAVOR_FAILURE, 153. INVADING, 155. THEFT, 158. HOSPITALITY, 159. QUARRELING, 162. MEDICAL_INTERVENTION, 163. BEARING_ARMS, 166. REVEAL_SECRET, 169. ESCAPING, 172. DAMAGING, 173. PRISON, 174. MAKE_COMPROMISE, 177. TRIAL, 178. COMMITTING_CRIME, 180. SURVIVING, 183. SURRENDERING, 186. EXECUTION

Table 7: Emotional frames annotated as Emotion Stimuli. Each frame is numbered according to its PMI rank.

the circumstance and engages in a series of changes – i.e., subjective feelings, neurophysiological, motor and motivational alterations.

Frames concerning evaluations are, e.g., SATISFYING and FAIRNESS_EVALUATION. The latter frame, whose link

to emotions seemed hazy at first, now appears as an emotional exemplar in its own right: the notion of assessment that it brings into play is central to the elicitation of emotions. In this group are also items that qualify events as endangering for the organism (e.g., DIF-

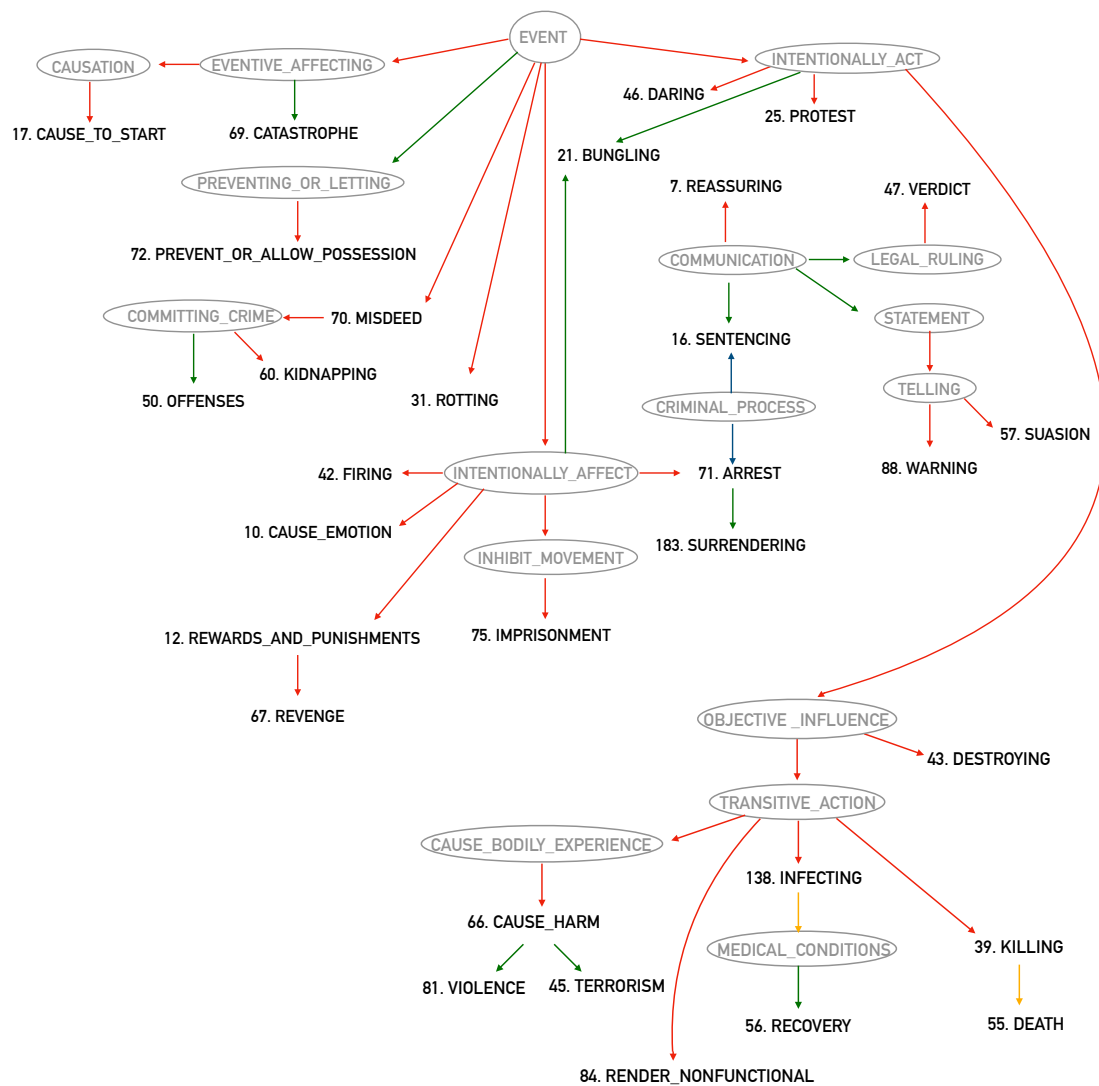


Figure 7: Emotional frames (text in black), deriving from the node EVENTS and expressing factual Emotion Stimuli, extracted from Table 7. The arrow legend is in Figure 6.

FACULTY, RISKY_SITUATION), or as fostering its well-being (e.g., LUCK, WEALTHINESS).

Some of these frames recall the *criteria* that individuals use to evaluate an environment. In the appraisal framework, they are described with a finite number of dimensions (Scherer et al., 2010). One is the coherence of the event with the personal ideals of the experiencer and with societal norms. Frames like FAIRNESS_EVALUATION and MORALITY_EVALUATION convey precisely this type of evaluation. Similarly, GRASP reflects the criterion by which events are appraised in relation to their implications – e.g., Are they relevant to the experiencer’s goals? Can their consequences be estimated? It is indeed evoked by textual chunks that involve a cognizer who acquires knowledge about the significance of a given phenomenon and becomes informed to make predictions about it. Events can also be

evaluated for the degree to which the experiencers are certain about what is going on (e.g., How well does the experiencer understand what is happening in the emotional situation? (Smith and Ellsworth, 1985)), which is echoed by the frame CERTAINTY, and with respect to the urgency of a reaction (REQUIRED_EVENT).

Focusing on such evaluation criteria, appraisal theories claim that specific assessments of events lead to specific emotion experiences. For instance, a lack of certainty likely results in an episode of fear or hope (Smith and Ellsworth, 1985). To an extent, this is accounted for by the relations between frames. CERTAINTY, as an example, is inherited by the node TRUST. Therefore, FrameNet relations seem to explain the affective charge of some of these frames that do not stem from EMOTIONS, but are linked to the EMOTIONS-deriving nodes all the same.

Definition Frames capturing the link between emotions and events, namely, the saliency of the circumstance for the well-being of the experiencer, evaluations, actions, motives and responses that the experiencer takes in reaction to the event.

Frames 2. DISGRACEFUL_SITUATION, 6. MAKING_FACES, 9. FACIAL_EXPRESSION, 14. FAIRNESS_EVALUATION, 18. COMMUNICATION_NOISE, 20. ACCURACY, 23. LUCK, 24. MENTAL_PROPERTY, 26. MAKE_NOISE, 27. BODY_MARK, 29. SATISFYING, 30. COGITATION, 33. SUCCESS_OR_FAILURE, 35. CHEMICAL-SENSE_DESCRIPTION, 37. FRUGALITY, 38. AGREE_OR_REFUSE_TO_ACT, 44. CHAOS, 49. SOCIABILITY, 51. DESERVING, 53. CERTAINTY, 58. OMEN, 59. RISKY_SITUATION, 61. GUILT_OR_INNOCENCE, 63. SUBJECTIVE_INFLUENCE, 64. BEING_QUESTIONABLE, 65. PROMINENCE, 68. VOCALIZATIONS, 73. BIOLOGICAL_URGE, 74. GRASP, 76. DIFFICULTY, 77. MORALITY_EVALUATION, 78. COMING_TO_BELIEVE, 79. STINGINESS, 82. SOCIAL_INTERACTION_EVALUATION, 85. ARTIFICIALITY, 86. FLEEING, 91. HIT_OR_MISS, 95. IMPROVEMENT_OR_DECLINE, 96. WEALTHINESS, 97. CORRECTNESS, 98. COMMITMENT, 102. LEVEL_OF_FORCE_EXERTION, 104. COMPLAINING, 105. REASONING, 109. PEOPLE_BY_MORALITY, 112. SOCIAL_DESIRABILITY, 115. JUSTIFYING, 117. JUDGMENT_COMMUNICATION, 118. WILLINGNESS, 120. SENSATION, 123. INCLINATION, 125. EXPRESSING_PUBLICLY, 130. TRIGGERING, 135. EXPECTATION, 136. EXPEND_RESOURCE, 137. JUDGMENT_OF_INTENSITY, 142. TRUST, 146. OPPORTUNITY, 147. BEING_RELEVANT, 148. DEAD_OR_ALIVE, 150. AWARENESS_STATUS, 151. DYNAMISM, 154. BEING_OPERATIONAL, 157. FAME, 160. BEING_AT_RISK, 161. OPINION, 164. REQUIRED_EVENT, 170. CAUSE_IMPACT, 175. PRECARIOUSNESS, 176. MEET_SPECIFICATIONS, 179. MOTION_NOISE, 181. ATTEMPT, 185. BREATHING, 187. CONFRONTING_PROBLEM, 188. EVENTIVE_AFFECTING, 190. ATTITUDE_DESCRIPTION

Table 8: Emotional frames that capture appraisal-related properties. Each frame is numbered according to its PMI rank.

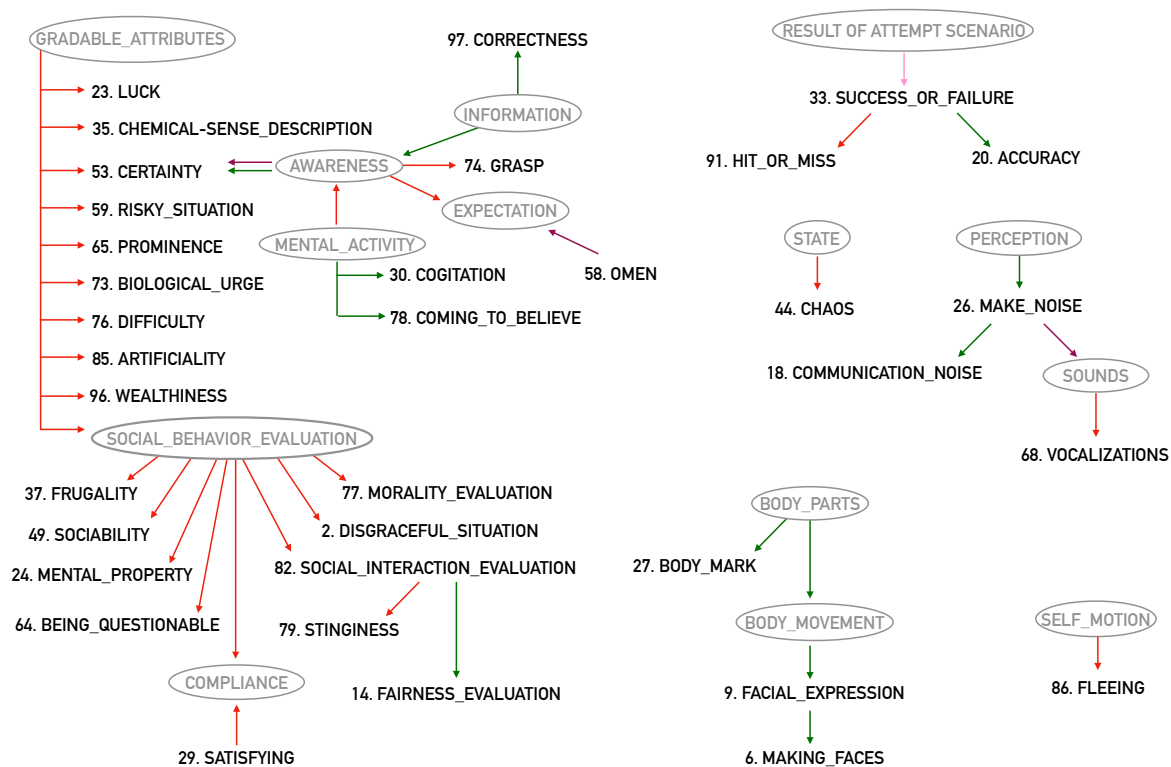


Figure 8: Emotional frames (text in black), expressing appraisal-related concepts (cf. Table 8). The arrow legend is in Figure 6.

We further observe frames that relate to the effects that emotions have on the organism (BIOLOGICAL_URGE exemplifies the involvement of internal, physiological states that can motivate action in response to an event), and frames that correspond to more observable manifestations of the emotion mechanism, such as vocal verbalizations, facial movements, and other diagnostic features that allow people to understand what their interlocutors feel. MAKING_FACES, FACIAL_EXPRESSION, COMMUNICATION_NOISE (evoked by texts like “*For the first week I cried.*”) and MAKE_NOISE seize these components. Other frames, for instance REASSURING and COGITATION (a child node of WORRYING), capture external actions or internal attitudes that can occur in emotional situations.

Additional analyses of frames whose membership to the Appraisal-based cluster is not self-explanatory can be found in Appendix C.

Incidentally Emotional Frames. These 25 frames (see Table 9) rank among the lowest values in the top quartile of the PMI distribution, closer to the cutoff point than the clusters discussed so far. They hardly capture an emotion property or an emotion-inducing event; in fact, they can be argued more affine to the contextually-determined cluster of Section 5.3, to which their PMI values are close. In this analysis, they appear as emotional due to two primary factors. The first one is narrative context, the second is processing errors. We support this analysis by investigating the sentences in which these frames appear.

Regarding narrative context, recall that most COCA sentences contain multiple frames. Therefore, frames can assume emotionality from others in the same sentence, which are often narratively related. BOARD_VEHICLE and RIDE_VEHICLE, for instance, are evoked in texts that have to do with embarking on adventures and journeys: these tend to be emotionally qualified as they often mention personal stances towards such journeys (e.g., if it was pleasant). Instead, REFORMING_A_SYSTEM and CAUSE_TO_RESUME characterise texts that express an idea of personal change, of beginning (e.g., “*We may have reformed, but our enemies have not.*”, “*I felt revived*”). MANIPULATE_INTO_DOING is ascribed to descriptions of bullying episodes; IRREGULAR_COMBATANTS has to do with fighters and hence a notion of brutality (comparable to KILLING and BEARING_ARMS from cluster (2)). MEDICAL_SPECIALTIES is evoked by (potentially stirring) circumstances that are related to healthcare and therapy, and RITE appears in the context of intimate meditations and expressed hopes.

Other cases seem to result directly from mistakes made by the frame identifier. With TEMPERATURE, the automatic role labeller does not understand the

metaphoric use of the word “*cool*”, for which that frame is usually predicted. LINGUISTIC_MEANING is a similar case. It is identified in phrases that are related to meanings and to the “making sense” of a situation, rather than in the context of a discussion about linguistic meaning.

5.2 Nonemotional Frames

Examples of nonemotional frames are in Figure 4 (b), with some corresponding texts in Table 10. We ask the same two questions about nonemotional frames that we asked about emotional frames above.

5.2.1 PMI Values across Lexical Units

To understand the difference between this group and the emotional one, we look at the PMI scores of the frames’ lexical units. Mirroring what we did for the top 35 frames, we focus on the 35 most nonemotional instances at the bottom of the PMI distribution. Here, frames show a much lower internal consistency, and suggest that they act as emotionally coherent units of abstraction only above a certain PMI threshold. Indeed, the scores of lexical units instantiating a nonemotional frame spread away from that of the latter considerably, as exemplified by RELATIONAL_NATURAL_FEATURES (PMI = -.47) whose lexical instantiations encompass a vast range of values, from .01 (for the noun “*summit*”) to -1 (“*shoreline*”), DISTRIBUTED_POSITION (-.48), spanning from the -.07 PMI score of “*envelop*” to -.70 of “*wreathe*”, and BECOMING_SILENT (-.47), where “*quiet*” has a PMI value of -.15 as a noun and -.83 as an adjective.

This outcome is different from what we found for emotional frames, where emotionality is stable for the lexical units within frames (cf. Section 5.1). For nonemotional frames, the picture is not symmetric: the PMI variance of lexical units can be attributed to their presence (mostly) in textual contexts without emotionality, but also in some with an emotion gradation.

5.2.2 Characterising Nonemotional Frames

Compared to the emotional frames, this cluster depends much less on people’s subjective involvement in the state of affairs mentioned in the texts. It includes frames expressing features of objects (e.g., BIOLOGICAL_CLASSIFICATION, ESTIMATED_VALUE, SUBSTANCE_BY_PHASE, MEASURABLE_ATTRIBUTES) or of events which have less relevance for human actors in terms of appraisals (e.g., CHANGE_OF_PHASE, BECOMING_DRY).

5.3 Contextually-determined Frames

Contextually-determined frames are those with PMI values falling in the 2nd or 3rd quartiles of the emotional

Definition	Frames that do not belong to any of the other three groups.
Frames	52. RESPOND_TO_PROPOSAL, 54. INSTITUTIONALIZATION, 99. RITE, 103. LINGUISTIC_MEANING, 114. BOARD_VEHICLE, 126. MANIPULATE_INTO_DOING, 127. MEDICAL_SPECIALTIES, 128. REFORMING_A_SYSTEM, 131. ECONOMY, 132. TEMPERATURE, 133. CO-ASSOCIATION, 139. AFFIRM_OR_DENY, 140. BEHIND_THE_SCENES, 141. APPELLATIONS, 144. RIDE_VEHICLE, 145. EVENT, 149. IRREGULAR_COMBATANTS, 156. CHANGE_OF_LEADERSHIP, 165. PEOPLE_BY_RELIGION, 167. MEDICAL_INTERACTION_SCENARIO, 168. EDUCATION_TEACHING, 171. CAUSE_TO_RESUME, 182. MAKE_AGREEMENT_ON_ACTION, 184. REPRESENTATIVE, 189. TOURING

Table 9: Incidentally Emotional frames. Each frame is numbered according to its PMI rank.

Frame	Text
PATH_TRAVELED	They occur when the orbits of the moons turn edge-on to the Sun and Earth, which happens twice during Jupiter's 12-year circuit of the Sun.
DIRECTIONAL_LOCATIVE_RELATION STORING	It was known he lived across the immense valley below me. Mark your packages with the date they were placed in the freezer so you can keep track of storage times.
MEASURE_AREA	They burned 665,000 acres; roughly 40% of the statewide total of 1.7 million acres.
RELATIONAL_NATURAL_FEATURES BECOMING_SILENT	The shore is crumbling. A silence descends on the tiny room.

Table 10: Examples of nonemotional frames with sentences in which they appear.

Frame	Text
COMMUNICATION_RESPONSE	(N) The answer is, you don't, or at least not with career backups. (E) The answer would be NO!
GIVE_IMPRESSION	(N) Neither candidate seemed to have any awareness of virality . (E) You really seem to be exploding with creativity!
POINT_OF_DISPUTE	(N) The question , crude as it was, hung in the air . (E) The issue is not whether I was a perfect pastor; I was not .

Table 11: Example sentences evoking contextually-determined frames. (E)/(N): emotional/neutral sentences. Words in boldface correspond to predicates.

distribution reported in Figure 3 ($-.16 \leq \text{PMI} \leq .24$). A few examples are provided in Table 11. These items have an ambiguous emotional status, in that they present no clear association with emotionality, nor its absence.

What makes frames contextually-determined? Our hypothesis is that it is possible to set apart these cases from frames that carry an emotional (or nonemotional) load in two, non-mutually exclusive ways. First, by looking at the lexical units internal to frames, once more, to explain the sense in which these frames are different from the most external quartiles in the distribution. Second, by looking at how their emotionality changes as they co-occur with other frames. We explore these two levels separately below.

5.3.1 PMI Values across Lexical Units

The way PMI values distribute across lexical units is more similar to nonemotional frames (Section 5.2) than to emotional frames (Section 5.1.2): Values differ from each other, in such a way that contextually-determined frames, contrary to emotional ones, do not function as emotion-preserving types of units. Cases in point are the verbs “*tell*” and “*assure*”, both evoking the frame TELLING, and whose emotionality association corresponds to the values .07 and .34 (i.e., “*assure*” is most often emotional than not, while “*tell*” is at times emotional); “*disparity*” and “*distinction*” are apart from one another by .51 PMI points (the first of them is the most emotional), despite being units of the same frame (SIMILARITY); likewise, CURE’s lexical units “*rehabilitation*” and “*remedy*” have values falling in different quartiles of the PMI distribution (.25 and -.11, respectively). The distinguishing factor between them and the nonemotional group lies in the fact that lexical units here are more versatile. Those belonging to the nonemotional counterpart are specific to domains without emotionality (cf. MEASUREMENT_AREA, CLOTHING_COMPONENTS), and thus their occurrence in emotional contexts is not only rarer, but an artifact of either the emotion classifier (producing random errors) or of our alignment strategy (which permits nonemotional frames to inherit the emotionality of others, with which they occur). Instead, lexical units of contextually-determined frames lend themselves to assume a wide range of emotional connotations.

5.3.2 Frame Co-occurrence Patterns

Above, when characterising Incidentally Emotional frames, we already alluded to the fact that frames typically co-occur in sentences. We also see this effect for the contextually-determined frames. Figure 9 shows the frequency of these frames in three different scenarios, normalised by the total number of sentences in each of them. The two leftmost columns (Frm_{cont.}) report the

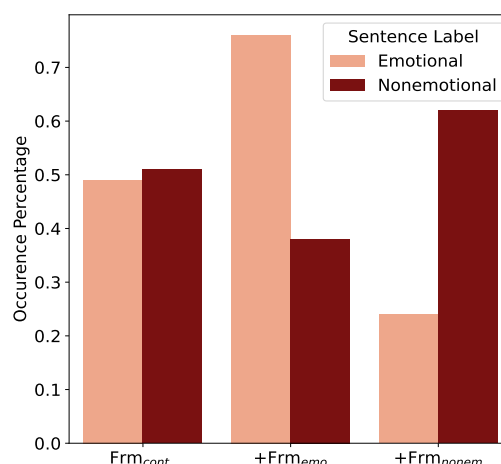


Figure 9: Distribution of emotional and nonemotional sentences evoking contextually-determined frames in isolation (Frm_{cont.}) and accompanied by an emotional frame (+Frm_{emo}) or a nonemotional one (+Frm_{nonemo}).

frequency of frames appearing alone in a text across the two emotion labels, corresponding to >2M emotional and nonemotional sentences. Devoid of frames interactions, these sentences help to clarify what it means for frames to be underspecified with respect to emotionality: based on a manual investigation of such sentences, contextually-determined frames appear to have less to do with properties of things or situations, compared to the nonemotionally-connotated kins. They rather represent such things (FOOD, VEHICLE, BUILDINGS) or processes (CAUSE_EXPANSION, CAUSE_TO_PERCEIVE). We notice that when these frames appear in emotional texts, they do so as side information to the main affective meaning, and do not correspond to the predicate that triggers such emotion content. For instance, CONTINUED_STATE_OF_AFFAIRS in the text “*Glad she’s still on the show.*” is unrelated to the mental state of the subject. The figure also reports the count of sentences with a contextually-determined frame and one that is emotional (+Frm_{emo}), or one that is nonemotional (+Frm_{neu}). From the figure, we see that texts that contain both a contextual frame and one with a positive emotion PMI tend to be emotional; vice versa for the co-presence with a nonemotional frame, found more often in sentences labelled as nonemotional by the classifier.

Overall, the fact that these 378 frames are determined contextually shows an important aspect of the phenomenon under consideration. At times, the relationship that frames hold to their emotion content is underspecified: it is not fixed and bounded to the type of event that they formalise (i.e., it does not necessarily lie at the predicate level), but rather depends on the overall context in which the frame-evoking predicate appears. Emotion meanings make no exception in the lexical semantics panorama, where also other phenom-

ena are to be accounted for *in context* (Cruse, 1986) – e.g., word meanings.

A manual inspection of the data also suggests that compositionality is key in the making of an emotion for those sentences corresponding to +Frm_{emo} and +Frm_{neu} in Figure 9. More precisely, we see two compositional processes. One is a “within frames compositionality”, in which the predicate is (emotionally) underspecified, but its co-presence with certain arguments can turn out emotional or nonemotional. Illustrative in this regard are sentences like “*I remember this point distinctly.*” and “*I remember the magical thinking of my greatest depression.*”, both associated to the frame MEMORY but with different arguments (the first sentence is recognised as nonemotional, the other as emotional). Like in the above examples, many frames are evoked by predicates that serve to introduce topical information, or subordinate sentences. The overall emotionality varies together with the content that they introduce. For instance COMMUNICATION_RESPONSE, TELLING, POINT_OF_DISPUTE, GIVING and GIVE_IMPRESSION have to do with communicative situations that could be loaded with emotionality based on how they are instantiated – what is responded, what is told, what is given (e.g., GIVING in the emotional example “*Cruella gave a gesture of resignation.*”). Similarly, UNDERGO_CHANGES describes a transformation which could be either emotional or nonemotional.

The second compositional process that we notice is an “across frames compositionality”. Frames that appear in combination with a contextually-determined one contribute more to the emotional load of the sentence: the text “[...] *an old girlfriend of mine wrote me this very beautiful letter.*”, which is recognised by the classifier as emotional, evokes MEMORY and the emotional AESTHETICS, while “*The words ‘property value’ are ones I remember.*”, annotated as nonemotional by the classifier, evokes MEMORY and POSSESSION.

6 Discussion

We conducted a PMI-based analysis guided by the research question “are FrameNet frames associated with emotionality?” as well as two leading hypotheses: first, emotional frames constitute a large part of FrameNet, and second, it makes sense to talk about “emotional” frames in the sense that the lexical units within the frame behave coherently. Both assumptions proved correct. Frames that carry emotionality extend beyond the current organization of the database, as many are emotional while having a factual denotation; further, they pass this affective trait on to their lexical units.

Our manual analysis explains what frames have in common from the perspective of emotions, confirming that there are many levels of an emotion mechanism

captured by frame semantics. Some frames depict concepts that seem more descriptive than affective, but it is precisely in this manner that they pick up on some important components of emotions. They correspond to some of the factors that elicit, underlie or manifest an emotion, like events, event evaluations, and emotion effects. The effects components, in particular, not only correspond to phenomena that happen in response to emotion-eliciting events (e.g., FACIAL_EXPRESSION). They can be considered events per se, and consequently, they can evoke specific frames.

We manually group these characteristics in four clusters, motivated by the original structure of FrameNet, and the fact that appraisal models and frames are grounded on a notion of event. Ruppenhofer (2018) already pointed out that appraisal theories can inform an investigation of the emotion vocabulary in FrameNet. We bolster that observation by indicating the frames to which it extends, but one could also identify other emotion properties and other links to theories different from appraisals, and organise the emotional frames accordingly. Take, for instance, the Appraisal-based cluster. In our proposal, it includes both items that contribute to eliciting an emotion (e.g., DIFFICULTY), and items that result from it (e.g., FLEEING). This is a fruitful distinction that can be made to find more fine-grained theoretical coherence in the obtained statistical associations.

Further, we empirically show that there are frames somewhat transparent to emotions: contextually-determined frames reiterate the need to think about emotionality in terms of relations between words, and raise the question of if and how frames influence the emotionality of a text, as well as its automatic classification. To what extent do predicates or arguments contribute to the decisions of an emotion classifier? Is compositionality at play?

We have previously pointed out that emotionality is a continuum, while our study approaches it through categorical lenses. From a practical standpoint, this categorization has the value of generating clear insights. But this choice introduces limitations, not least of which is a certain degree of arbitrariness in such divisions: frames do not necessarily fit into the three nonemotional, contextually-determined, and emotional “boxes” identified with the help of quartiles, and the line between contextually-determined and emotional frames (in particular the Incidentally Emotional ones) is blurred. In fact, the compelling case that emotionality is always a matter of context could be made also for many emotional frames, with the Emotion Stimuli being evident cases (events could stir an emotion or not, depending on who experiences them and how they are rendered in language). Our results prove however that a separation holds, at least

in COCA, between frames for which exhibiting the emotional association is *invariably* contextual, whereas others *maintain a certain level of emotionality* – e.g., Emotion Stimuli have the tendency to denote events with potentially dramatic consequences for their experiencers, see for instance VIOLENCE or CATASTROPHE.

In sum, our analysis reveals that the relationship between the emotionality of a sentence and that of frames is not straightforward. Frames that have a strong positive or negative association to emotionality can be found in texts that express the opposite affective content overall.²¹ Even the frames that FrameNet explicitly associates to the emotion domain are evoked by nonemotional sentences. EMOTIONS_BY_STIMULUS, as an example, is found by the frame identifier in the nonemotional “*I had every right to descend this stair, to walk among the glad company [...]*”, because of the lexical unit “*glad*”. Rather than putting the automatic annotation into question, this outcome sheds light on an important fact. Namely, sentence-level emotionality classifiers can disregard emotional subtleties. A verbal expression might have a predominant connotation to convey (e.g., a nonemotional one, in the example above), and which might be correctly identified by the automatic system; yet, by considering entities besides the subject, different emotion nuances emerge (e.g., the company is glad). Classifiers might fail to account for those, and in such cases the performance of frame identification tools can complement theirs. In line with previous work (Faruqui et al., 2015, i.a.), we thus found that approaches based on embeddings and on human-curated resources help one another also in emotion analysis.

7 Conclusion

The phenomenon of “emotions” is psychological in nature but pervades language. There, the presence of overt markers (the adjectives “*sad*”, “*happy*”, for instance) is not necessary for an emotion to be conveyed. These “untold” emotions spurred much attention in the field of computational emotion analysis (Balahur and

²¹Note that there are signs of domain dependence: frames are more emotional in certain domains of COCA than in others. For example, RUN_RISK has a PMI value of .13 in textual blogs, which raises to .4 in the domain of fiction; and the frame PROTEST turns out considerably more emotional if evoked by fiction- (PMI=.77) than by TV-related texts (.47). Consistent with this observation, a Wilcoxon signed-rank test reveals a significant difference between the general PMI values of the emotional frames reported in Figure 3 and the values of the same frames in the various domains (for all of them, except for TV, p-value < .05). Therefore, emotionality is only partly consistent across genres, and this finding is in line with existing literature on the genre dependence of fine-grained emotions (e.g., Bostan and Klinger, 2018). At the same time, PMI differences are rarely as extreme as to have frames that are emotional in Figure 3 turn into nonemotional in a specific domain (that only happens for ATTITUDE_DESCRIPTION, PRECARIOUSNESS, and TEMPERATURE).

Tanev, 2016; Klinger et al., 2018), which strives to automatise the ability to infer them.

Within such a context, we left traditional, lexical-based approaches of emotion analysis, because interpreting emotions can require a great deal of extralinguistic knowledge. We considered the role that background information plays in emotions understanding, moving our attention to the meeting point between syntax and the U-semantics of Fillmore, which presupposes an acknowledgement of the physical and social world, and therefore accounts for the structural components of real-life events that stimulate emotional responses. This way, our work combined methods for computational linguistics with theories from psychology and linguistics, and it showed how these fields can influence (in fact, fertilize) one another. Below, we summarise the relevance of our findings in this interdisciplinary perspective, and point out promising next steps to take.

Summary of Findings. The observation that frames can be evoked by varied lexical units (thus capturing paradigmatic phenomena) allowed us to disregard the specific terms that instantiate them. We rather asked how frames, as conceptual abstractions that encode world knowledge, are linked to emotionality. We automatically annotated COCA with binary emotion labels and with frames, we investigated the relationship between them, and to answer our research question, we used PMI.

Our results show that there are frames with a prominent emotion import in FrameNet: be they direct children of EMOTION or not, they reflect components of emotions spelled out in the psychological literature. In other words, emotionality is a dimension of meaning that frames possess even though it is not a piece of information directly provided by the database. In addition, our qualitative analysis emphasise that individual predicates do not always carry the same type of emotion load. On the contrary, their import can depend on the context in which the predicate is situated, namely, on syntagmatic facts.

Future Work. We revealed some salient features of frames that open up possible ventures for frame semanticists. Future FrameNet developments could specify what frames carry emotionality with the use of *semantic types* (Fillmore et al., 2004). Semantic types mark general properties of frames and semantic roles, such as variations in the speech use of different lexical units, which could not otherwise be understood from the resource. In FrameNet there already exists a semantic type that is close in spirit to emotions. It indicates the polarity of lexical units like “*compliment*” and “*reprimand*”, both of which instantiate JUDGMENT_DIRECT_ADDRESS and whose valence is indicated

by the semantic types “Positive.judgment” and “Negative.judgment”. It would be possible to adopt the same idea for the semantic clusters proposed in this paper, or for similar partitions. We refrain from modelling this information into FrameNet ourselves – an endeavor that would require careful and lexicographically motivated annotation, which exceeds the scope of our work.

Our insights can also inform computational emotion analysis. Studies in the field could aim at building systems that are simultaneously emotion- and frame-aware. The frames-to-PMI association scores that we make publicly available come handy for that purpose. Upcoming work could deepen the contribution of different parts of texts (e.g., frames, arguments, other words) on automatic emotion predictions – e.g., Do classifiers attend predicates to the same extent when judging a text that evokes an emotional frame and a text that evokes a contextually-determined frame? Lastly, research in the field that follows appraisal theories could concentrate on the intersection between frame semantics, psychology, and emotion analysis: among other events, frames proved able to model the verbal expressions of emotion components, thus capturing the multiple and nuanced realizations through which embodied emotions and the cognitive evaluations underlying them surface in language. In this regard, we have proposed an empirical mapping from frames to appraisals, but it would be important to take the reverse direction as well. Understanding to what extent frames cover the cognitive dimensions documented by appraisal theories could tell us if frame analysis can be applied as an identification strategy of such dimensions, namely, of the criteria that humans use to evaluate events, that lead to an emotion episode, and that also emerge from text: frames could thus be used as input to computational emotion analysis pipelines, making our systems more theoretically grounded.

Acknowledgements

This work has been supported by the German Research Council (DFG), project “Computational Event Analysis based on Appraisal Theories for Emotion Analysis” (CEAT, project number KL 2869/1-2).

References

- Abdul-Mageed, Muhammad and Lyle Ungar. 2017. EmoNet: Fine-grained emotion detection with gated recurrent neural networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 718–728, Vancouver, Canada. Association for Computational Linguistics.
- Alm, Cecilia Ovesdotter, Dan Roth, and Richard Sproat. 2005. Emotions from text: Machine learning for text-based emotion prediction. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 579–586, Vancouver, British Columbia, Canada. Association for Computational Linguistics.
- Baker, Collin F. and Christiane Fellbaum. 2009. WordNet and FrameNet as complementary resources for annotation. In *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*, pages 125–129, Suntec, Singapore. Association for Computational Linguistics.
- Baker, Collin F., Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *COLING-ACL '98: Proceedings of the Conference*, pages 86–90, Montreal, Canada.
- Balahur, Alexandra and Hristo Tanev. 2016. Detecting implicit expressions of affect from text using semantic knowledge on common concept properties. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*. European Language Resources Association (ELRA).
- Banarescu, Laura, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.
- Barrett, Lisa Feldman, Michael Lewis, and Jeannette M Haviland-Jones. 2016. *Handbook of emotions*. Guilford Publications.
- Bostan, Laura Ana Maria, Evgeny Kim, and Roman Klinger. 2020. GoodNewsEveryone: A corpus of news headlines annotated with emotions, semantic roles, and reader perception. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1554–1566, Marseille, France. European Language Resources Association.
- Bostan, Laura-Ana-Maria and Roman Klinger. 2018. An analysis of annotated corpora for emotion classification in text. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2104–2119, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Buckley, Chris. 1993. The importance of proper weighting methods. In *Human Language Technology: Pro-*

- ceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993.
- Buechel, Sven and Udo Hahn. 2017. EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 578–585, Valencia, Spain. Association for Computational Linguistics.
- Buechel, Sven, Susanna Rücker, and Udo Hahn. 2020. Learning and evaluating emotion lexicons for 91 languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1202–1217, Online. Association for Computational Linguistics.
- Canales, Lea and Patricio Martínez-Barco. 2014. Emotion detection from text: A survey. In *Proceedings of the Workshop on Natural Language Processing in the 5th Information Systems Research Working Days (JISIC)*, pages 37–43, Quito, Ecuador. Association for Computational Linguistics.
- Chen, Yanqing and Steven Skiena. 2014. Building sentiment lexicons for all major languages. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 383–389.
- Church, Kenneth and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.
- Clore, Gerald L., Andrew Ortony, and Mark A. Foss. 1987. The psychological foundations of the affective lexicon. *Journal of Personality and Social Psychology*, 53(4):751–766.
- Cohen, Jacob. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.
- Coyne, Bob, Alex Klapheke, Masoud Rouhizadeh, Richard Sproat, and Daniel Bauer. 2012. Annotation tools and knowledge representation for a text-to-scene system. In *Proceedings of COLING 2012*, pages 679–694, Mumbai, India. The COLING 2012 Organizing Committee.
- Cruse, D. Alan. 1986. *Lexical semantics*. Cambridge university press.
- Das, Dipanjan, Nathan Schneider, Desai Chen, and Noah A. Smith. 2010. Probabilistic frame-semantic parsing. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 948–956, Los Angeles, California. Association for Computational Linguistics.
- Davies, Mark. 2015. *Corpus of Contemporary American English (COCA)*.
- Demszky, Dorottya, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A Dataset of Fine-Grained Emotions. In *58th Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Dixon, Thomas. 2012. “Emotion”: The history of a keyword in crisis. *Emotion Review*, 4(4):338–344.
- Eisenberg, Nancy and Paul A. Miller. 1987. The relation of empathy to prosocial and related behaviors. *Psychological bulletin*, 101(1):91.
- Ekman, Paul. 1992. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.
- Etzioni, Oren, Michele Banko, Stephen Soderland, and Daniel S. Weld. 2008. Open information extraction from the web. *Communications of the ACM*, 51(12):68–74.
- Faruqui, Manaal, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. 2015. Retrofitting word vectors to semantic lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1606–1615, Denver, Colorado. Association for Computational Linguistics.
- Felbo, Bjarke, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1615–1625, Copenhagen, Denmark. Association for Computational Linguistics.
- Fellbaum, Christiane, editor. 1998. *WordNet: an electronic lexical database*. MIT Press, Cambridge, Mass.

- Fillmore, Charles J. 1982. Frame semantics. In *Linguistics in the Morning Calm*, pages 111–137. Hanshin Publishing Co., Seoul, South Korea.
- Fillmore, Charles J. 1985. Frames and the semantics of understanding. *Quaderni di semantica*, 6(2):222–254.
- Fillmore, Charles J., Collin F. Baker, and Hiroaki Sato. 2004. FrameNet as a “net”. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC’04)*, Lisbon, Portugal. European Language Resources Association (ELRA).
- Ghazi, Diman, Diana Inkpen, and Stan Szpakowicz. 2015. Detecting emotion stimuli in emotion-bearing sentences. In *Computational Linguistics and Intelligent Text Processing*, pages 152–165. Springer International Publishing.
- Gildea, Daniel and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational linguistics*, 28(3):245–288.
- Grice, Paul. 1975. Logic and conversation. In P. Cole and J. Morgan, editors, *Syntax And Semantics*. Academic Press, New York.
- Hartmann, Silvana, Iliia Kuznetsov, Teresa Martin, and Iryna Gurevych. 2017. Out-of-domain FrameNet semantic role labeling. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 471–482, Valencia, Spain. Association for Computational Linguistics.
- Hobbs, Jerry R. and Andrew S. Gordon. 2011. The deep lexical semantics of emotions. In *Affective Computing and Sentiment Analysis*, pages 27–34. Springer.
- Hofmann, Jan, Enrica Troiano, Kai Sassenberg, and Roman Klinger. 2020. Appraisal theories for emotion classification in text. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 125–138, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Hu, Minqing and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177.
- Ide, Nancy, Collin F. Baker, Christiane Fellbaum, and Rebecca J. Passonneau. 2010. The manually annotated sub-corpus: A community resource for and by the people. In *Proceedings of the ACL 2010 conference short papers*, pages 68–73.
- Jänicke, Stefan, Greta Franzini, Muhammad Faisal Cheema, and Gerik Scheuermann. 2015. On close and distant reading in digital humanities: A survey and future challenges. In *EuroVis (STARS)*, pages 83–103.
- Kiefer, Ferenc. 1988. Linguistic, conceptual and encyclopedic knowledge: Some implications for lexicography. In *Proceedings of EURALEX*, Budapest, Hungary.
- Klinger, Roman, Orphée de Clercq, Saif Mohammad, and Alexandra Balahur. 2018. IEST: WASSA-2018 implicit emotions shared task. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Brussels, Belgium. Association for Computational Linguistics.
- Kong, Lingpeng, Chris Dyer, and Noah A. Smith. 2016. Segmental recurrent neural networks. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.
- Li, Yanran, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.
- Liu, Vicki, Carmen Banea, and Rada Mihalcea. 2007. Grounded emotions. In *Seventh International Conference on Affective Computing and Intelligent Interaction (ACII 2017)*, pages 477–483. IEEE Computer Society.
- Màrquez, Lluís, Xavier Carreras, Kenneth C. Litkowski, and Suzanne Stevenson. 2008. Special issue introduction: Semantic role labeling: An introduction to the special issue. *Computational Linguistics*, 34(2):145–159.
- Martin, James R. and Peter R. White. 2003. *The language of evaluation*, volume 2. Springer.
- Mehrabian, Albert and Norman Epstein. 1972. A measure of emotional empathy. *Journal of personality*, 40(4):525–543.
- Mohammad, Saif. 2012. #emotional tweets. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 246–255, Montréal, Canada. Association for Computational Linguistics.
- Mohammad, Saif and Felipe Bravo-Marquez. 2017. WASSA-2017 shared task on emotion intensity. In

- Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 34–49, Copenhagen, Denmark. Association for Computational Linguistics.
- Mohammad, Saif and Peter D. Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational intelligence*, 29(3):436–465.
- Mohammad, Saif, Xiaodan Zhu, and Joel Martin. 2014. Semantic role labeling of emotions in tweets. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 32–41, Baltimore, Maryland. Association for Computational Linguistics.
- Moors, Agnes, Phoebe C. Ellsworth, Klaus R. Scherer, and Nico H. Frijda. 2013. Appraisal theories of emotion: State of the art and future development. *Emotion Review*, 5(2):119–124.
- Nandwani, Pansy and Rupali Verma. 2021. A review on sentiment analysis and emotion detection from text. *Social Network Analysis and Mining*, 11(1):1–19.
- Oberländer, Laura, Kevin Reich, and Roman Klinger. 2020. Experiencers, stimuli, or targets: Which semantic roles enable machine learning to infer the emotions? In *Proceedings of the Third Workshop on Computational Modeling of People’s Opinions, Personality, and Emotions in Social Media*, Barcelona, Spain. Association for Computational Linguistics.
- Oberländer, Laura Ana Maria and Roman Klinger. 2020. Token sequence labeling vs. clause classification for English emotion stimulus detection. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 58–70, Barcelona, Spain (Online). Association for Computational Linguistics.
- Omdahl, Becky L. 1995. *Cognitive Appraisal, Emotion, and Empathy*. Mahwah, NJ: Lawrence Erlbaum.
- Ortony, Andrew, Gerald L. Clore, and Allan Collins. 1988. *The cognitive structure of emotions*. Cambridge university press.
- Palmer, Martha, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1):71–106.
- Pekar, Viktor, Michael Krkoska, and Steffen Staab. 2004. Feature weighting for co-occurrence-based classification of words. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 799–805, Geneva, Switzerland. COLING.
- Plutchik, Robert. 2001. The nature of emotions. *American Scientist*, 89(4):344–350.
- Preoțiuc-Pietro, Daniel, H. Andrew Schwartz, Gregory Park, Johannes Eichstaedt, Margaret Kern, Lyle Ungar, and Elisabeth Shulman. 2016. Modelling valence and arousal in Facebook posts. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 9–15, San Diego, California. Association for Computational Linguistics.
- Roth, Michael and Mirella Lapata. 2015. Context-aware frame-semantic role labeling. *Transactions of the Association for Computational Linguistics*, 3:449–460.
- Ruppenhofer, Josef. 2018. The treatment of emotion vocabulary in framenet: Past, present and future developments. In Alexander Ziem, Lars Inderelst, and Detmer Wulf, editors, *Frames interdisziplinär: Modelle, Anwendungsfelder, Methoden*, pages 95–122. Düsseldorf University Press.
- Ruppenhofer, Josef, Michael Ellsworth, Myriam Schwarzer-Petruck, Christopher R Johnson, and Jan Scheffczyk. 2016. Framenet ii: Extended theory and practice. Technical report, International Computer Science Institute.
- Russell, James A. and Albert Mehrabian. 1977. Evidence for a three-factor theory of emotions. *Journal of research in Personality*, 11(3):273–294.
- Scarantino, Andrea. 2016. The philosophy of emotions and its impact on affective science. *Handbook of emotions*, 4:3–48.
- Scheff, Thomas J. 1973. Intersubjectivity and emotion. *American Behavioral Scientist*, 16(4):501–511.
- Scherer, Klaus R. 1984. Emotion as a multicomponent process: A model and some cross-cultural data. *Review of personality & social psychology*, pages 37–63.
- Scherer, Klaus R. 1989. Appraisal theory. *Handbook of Cognition and Emotion*.
- Scherer, Klaus R. 2005. What are emotions? And how can they be measured? *Social Science Information*, 44(4):695–729.
- Scherer, Klaus R., Tanja Bänziger, and Etienne Roesch. 2010. *A Blueprint for Affective Computing: A sourcebook and manual*. Oxford University Press.
- Scherer, Klaus R. and Harald G. Wallbott. 1997. The ISEAR questionnaire and codebook. Geneva Emotion Research Group.

- Schuff, Hendrik, Jeremy Barnes, Julian Mohme, Sebastian Padó, and Roman Klinger. 2017. Annotation, modelling and analysis of fine-grained emotions on a stance and sentiment detection corpus. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 13–23, Copenhagen, Denmark. Association for Computational Linguistics.
- Searle, John R. 1976. A classification of illocutionary acts. *Language in society*, 5(1):1–23.
- Shaikh, Mostafa Al Masum, Helmut Prendinger, and Mitsuru Ishizuka. 2009. A linguistic interpretation of the OCC emotion model for affect sensing from text. *Affective Information Processing*, pages 45–73.
- Shen, Dan and Mirella Lapata. 2007. Using semantic roles to improve question answering. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 12–21, Prague, Czech Republic. Association for Computational Linguistics.
- Smith, Craig A. and Phoebe C. Ellsworth. 1985. Patterns of cognitive appraisal in emotion. *Journal of personality and social psychology*, 48(4):186–209.
- Stranisci, Marco Antonio, Simona Frenda, Eleonora Ceccaldi, Valerio Basile, Rossana Damiano, and Viviana Patti. 2022. APPReddit: a corpus of Reddit posts annotated for appraisal. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3809–3818, Marseille, France. European Language Resources Association.
- Strapparava, Carlo and Alessandro Valitutti. 2004. WordNet affect: an affective extension of WordNet. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).
- Swayamdipta, Swabha, Sam Thomson, Chris Dyer, and Noah A. Smith. 2017. Frame-Semantic Parsing with Softmax-Margin Segmental RNNs and a Syntactic Scaffold. *arXiv preprint arXiv:1706.09528*.
- Tafreshi, Shabnam and Mona Diab. 2018. Emotion detection and classification in a multigenre corpus with joint multi-task deep learning. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2905–2913, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Troiano, Enrica, Laura Oberländer, and Roman Klinger. 2023. Dimensional modeling of emotions in text with appraisal theories: Corpus creation, annotation reliability, and prediction. *Computational Linguistics*, 49(1).
- Troiano, Enrica, Sebastian Padó, and Roman Klinger. 2019. Crowdsourcing and validating event-focused emotion corpora for German and English. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4005–4011, Florence, Italy. Association for Computational Linguistics.
- Troiano, Enrica, Sebastian Padó, and Roman Klinger. 2021. Emotion ratings: How intensity, annotation confidence and agreements are entangled. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 40–49, Online. Association for Computational Linguistics.
- Udochukwu, Orizu and Yulan He. 2015. A rule-based approach to implicit emotion detection in text. In *Natural Language Processing and Information Systems*, pages 197–203, Cham. Springer International Publishing.
- Ushio, Asahi, Federico Liberatore, and Jose Camacho-Collados. 2021. Back to the basics: A quantitative analysis of statistical and graph-based term weighting schemes for keyword extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8089–8103, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Wauthier, Fabian L. and Michael Jordan. 2011. Bayesian bias mitigation for crowdsourcing. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc.
- Yu, Liang-Chih, Lung-Hao Lee, Shuai Hao, Jin Wang, Yunchao He, Jun Hu, K. Robert Lai, and Xuejie Zhang. 2016. Building Chinese affective resources in valence-arousal dimensions. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 540–545, San Diego, California. Association for Computational Linguistics.
- Zhuang, Liu, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.

A Limitations

The approach we described in Section 4 is common to data-driven information extraction lines of research which require no human intervention, such as the task of open information extraction (Etzioni et al., 2008), as well as to distant reading, i.e., the application of computational and statistical techniques in the field of digital humanities, aimed at uncovering global patterns in texts (Jänicke et al., 2015). Still, it incurs the risk of mistakes by both the emotion classifier and the frame identifier. The data we study was not collected for the sake of computational emotion analysis nor to study frames, and might differ in tone, topics and linguistic structures from the resources on which our automatic annotators were trained. As a matter of fact, the generalization capabilities of FrameNet-based parsers have been put into question by Hartmann et al. (2017), who found that a state-of-the-art system for SRL loses 16 percentage F1 points when evaluated against out-of-domain data. This issue also applies to emotions. Bostan and Klinger (2018) showed that systems for emotion detection tested out of domain suffer from performance drops as heavy as $\approx .70$ in F1 score. Overall, our findings are limited by the quality of the systems that we employ, but we believe that they provide evidence to learn something about the bond between frames and emotionality.

Some of our design choices could also be instantiated differently. For one thing, our annotation looks at emotions as a binary matter. Follow-up studies could observe if different frames carry specific emotions (anger, joy, etc.). Second, we benefit from word relations in the sense that these give context to identify frames, but we do not leverage roles, leaving this endeavor as our next research step. Third, to measure their association with emotionality, we treat all frames equally and as separate entities. While transparent, this choice does not account for within-sentence frames interactions.

B Corpus Labelling (Emotions)

We associate sentences in COCA to emotions automatically. Using a resource already labelled for emotions by humans could be a safer approach: people’s judgments are arguably more reliable than those of a classifier, and this would have allowed us to only perform the frame-based strand of labelling. Yet, existing resources for affective computing have magnitudes of data points less than we need, and they typically focus on a specific type of texts, such as tweets (Mohammad, 2012), tales (Alm et al., 2005) or news headlines (Bostan et al., 2020). Employing a state-of-the-art classifier specialised in only one domain (i.e., trained on a single re-

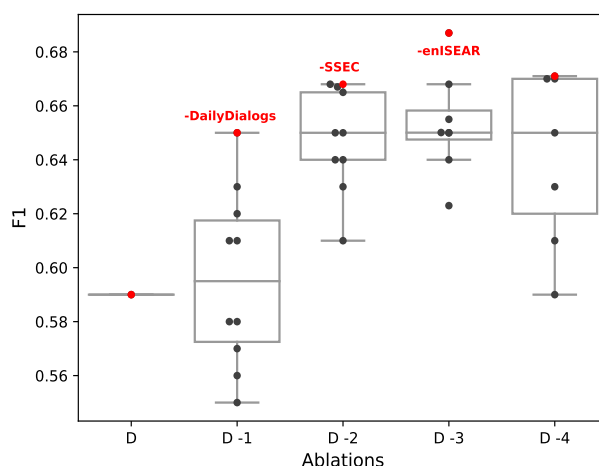


Figure 10: Model Selection: the y axis reports the F1 scores (weighted by the number of examples of each class) of the models evaluated against the annotated COCA sample. We recursively ablate datasets from the training set that yields the best model at the previous step (x axis). Dots are classifiers obtained with an ablation; the red ones indicate the best performing model: from all datasets (D), we remove each separately (“D –1”); from the set on which we obtained the best model (red dot “–DailyDialogs”), we again we remove each dataset, one at a time, thus training the next models on a collection with two datasets less than D (i.e., “D –2”); and so on.

source for emotion analysis) would give no guarantee that the obtained annotations are valid for our data. Moreover, we aim at observing frames as elicited by different emotion expressions, likely to be found in a mixture of textual domains.

Our model selection procedure is shown in Figure 10. Classifiers are plotted as dots in the figure, numbers on the x axis correspond to how many datasets are removed at each successive step. We kept all training parameters constant for the 35 models described in Section 4.1. They were fine-tuned for 10 epochs, setting a learning rate of $2 \cdot 10^{-5}$ a dropout rate of 0.2, and a batch size of 32. We used AdamW as optimizer.

Recursive data elimination proceeds as a backward search. Initially, we train a classifier on all gathered corpora described in Section 4.1 (“D” in the figure, F1=.59); from these resources, we pull out each dataset separately (“D –1”), and observe that the ablation of DailyDialogs is the most beneficial (F1 increases to .65); we move on to the next ablation step and keep using the data that yielded the best performance. From that, we ablate each remaining dataset (i.e., “D –2”): now, the results reached upon removal of SSEC surpass the previously best classifier. We repeat this procedure and reach an upper bound F1 score. From the total of 35 trained models, the most competitive one

	BERT-based	RoBERTa-based
<i>D</i>	F1 = .83	F1 = .86
COCA sample	F1 = .69	F1 = .55

Table 12: BERT- and RoBERTa-based classifiers performance when trained and tested in domain (*D*) vs. trained on *D* and tested on the COCA sample with the majority vote treated as ground truth.

is obtained when removing DailyDialogs, SSEC, and enISEAR (F1=.69 with “D -3”, which outperforms the best model in “D -4”, F1=.67). We use that to annotate COCA. Note that this classifier does not correspond to the one used to select the texts for the test set.

The performances displayed in Figure 10 could be expected. First, it is hard to find classifiers that are seamlessly portable across domains. Bostan and Klinger (2018) conducted multiple experiments showing that classifiers generalise poorly across domains. They report losses as drastic as .82 F1 score when testing on out-of-domain data. For us the loss is less severe (14 points, see Table 12, column BERT-based). Second, our models are learnt on datasets whose original annotation schemata differ from one another.

For a comparison to our BERT-based model selection, we experimented with a RoBERTa-based (Zhuang et al., 2021) emotion annotator trained on the whole concatenation of corpora (*D*). While the latter yielded superior results when evaluated on the in-domain data, it deteriorated on the manually annotated sample of COCA as out-of-domain data. Results are reported in Table 12.

Two viable alternatives for the automatic emotion annotation step could have been: (1) to use two classifiers, having high precision for either of the considered labels – i.e., one dedicated to the labelling of the emotional category and one for nonemotional category, which could arguably be more trustworthy, and (2) to accept texts as emotional or nonemotional if the probability with which the classifier assigns a label exceeds a given threshold. However, the first case would pose the problem of deciding how to treat texts for which the two models are in disagreement with one another. In the other case, we would lose substantial data. Our decision to adopt an individual emotion labeller, with a reasonable F1, bypasses both issues.

Adopting an annotation approach entirely based on human judgments would not be unproblematic either: large data sources compiled via crowdsourcing are noisy since they are labelled by naïve judges (Wauthier and Jordan, 2011); on the other hand, annotations conducted by expert coders are more reliable, but they typically cover smaller data, and this makes empirical observations difficult to draw. We forgo the lat-

ter. Indeed, when it comes to judging emotions, the noisiness problem characterises all human-based annotations, because the task is extremely subjective and therefore can lead to extreme disagreements, irrespective of how trained the coders are. Therefore, should the results of our analysis be due to systematic misclassifications of the automatic annotator, we could assume that similar “errors” are to be found among humans.

C Appraisal-based Frames

While discussing our partition of frames, we have highlighted that many items annotated as Appraisal-based frames tap on evaluations and cognitive processes. They are more than appear at first brush. Some singular examples are:

- REASONING, which often accompanies texts where an evaluation is expressed by means of a dispute described in the text;
- FAME, appearing in sentences with assessments that are either hyperbolic, like “*Believe me it was epic.*”, or that concern one’s reputation and beliefs, like “*To besmirch her reputation is outrageous.*”.

Likewise, the placement of BREATHING, CAUSE_IMPACT and LEVEL_OF_FORCE_EXERTION in this cluster of frames might not be self-explanatory. The first two indicate an emotional reaction, (e.g., sighing and slamming a door). The last usually portrays a property of people or events (e.g., feeling fearless and strong, feeling weak). So do also the following frames:

- DYNAMISM, evoked by texts that express the intensity of an experience;
- MEET_SPECIFICATIONS, coupled in text with mentions of personal achievements, or with expressed sensations of fulfilment.

An Empirical Configuration Study of a Common Document Clustering Pipeline

Anton Eklund, Dept. of Computing Science, Umeå University & Adlede, Sweden antone@cs.umu.se

Mona Forsman, Adlede, Umeå, Sweden mona.forsman@adlede.com

Frank Drewes, Dept. of Computing Science, Umeå University, Sweden drewes@cs.umu.se

Abstract Document clustering is frequently used in applications of natural language processing, e.g. to classify news articles or create topic models. In this paper, we study document clustering with the common clustering pipeline that includes vectorization with BERT or Doc2Vec, dimension reduction with PCA or UMAP, and clustering with K-Means or HDBSCAN. We discuss the interactions of the different components in the pipeline, parameter settings, and how to determine an appropriate number of dimensions. The results suggest that BERT embeddings combined with UMAP dimension reduction to no less than 15 dimensions provides a good basis for clustering, regardless of the specific clustering algorithm used. Moreover, while UMAP performed better than PCA in our experiments, tuning the UMAP settings showed little impact on the overall performance. Hence, we recommend configuring UMAP so as to optimize its time efficiency. According to our topic model evaluation, the combination of BERT and UMAP, also used in BERTopic, performs best. A topic model based on this pipeline typically benefits from a large number of clusters.

1 Introduction

Clustering is an important technique for mining, classifying, and structuring unlabeled text data in an unsupervised manner. Some use cases are the classification of news articles (Iulia-Maria et al., 2020; Radu et al., 2020), social media analysis (Curiskis et al., 2020; Asyaky and Mandala, 2021), and topic modeling (Sia et al., 2020; Churchill and Singh, 2022; Zhang et al., 2022; Zhao et al., 2021). For topic modeling and document classification, practitioners typically use a de-facto standard document clustering pipeline: document vectorization → dimension reduction → clustering; see Figure 1. This pipeline is attractive since it is straightforward to understand and provides flexibility due to its modularity. A popular application of this pipeline is BERTopic (Grootendorst, 2022), which converts the pipeline into a topic model by adding a topic keyword extractor.

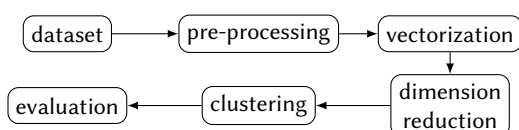


Figure 1: Clustering pipeline overview. The main parts are vectorization, dimension reduction, and clustering.

Since the pipeline components can be chosen from among many algorithms, and those usually depend on multiple parameter settings, it is challenging to analytically determine the best choice of components and their parameters, and the result depends on the concrete application. Further, the effect of the number of dimensions to reduce the vector space to is understudied in research on document clustering. In this paper, we conduct a systematic empirical study of how common embedding techniques, dimension reduction techniques, and clustering algorithms interact. From this, we derive recommendations that, as we hope, can guide practitioners who need to find a suitable configuration for clustering collections of unlabeled documents.

The first component of modern document clustering pipelines usually turns documents into numeric representations, called embeddings. Statistical methods such as Bag-of-Words or TF-IDF (Sammur and Webb, 2010) have been studied as part of topic models created with such a pipeline (Truică et al., 2016), but have nowadays become replaced by neural methods such as Doc2Vec (Le and Mikolov, 2014) and Google’s Transformer-based BERT (Devlin et al., 2019), which outperform the older methods; see Curiskis et al. (2020) and Radu et al. (2020) for the former, and Subakti et al. (2022) for the latter.

The next step, dimension reduction, is added to

avoid degrading performance of clustering algorithms in high-dimensional vector spaces (Steinbach et al., 2004; Zimek, 2014)¹. We study how the reduction to a range of different dimensions affects the quality of the resulting clusterings. There are two major classes of dimension reduction algorithms, those based on matrix factorization, and those based on neighbor graphs. Principle Component Analysis (PCA, by Pearson (1901); Hotelling (1933)) is a well-known and widely used example of the former. The latter, graph-based methods such as UMAP, calculate neighbor relations between points in the vector space, and then project them to a lower dimension, trying to preserve the neighbor relation. UMAP, invented by McInnes et al. (2018), is based on differential geometry and benefits from a solid mathematical foundation. UMAP has many applications, such as bioinformatics (Becht et al., 2019), material sciences (Li et al., 2019) and machine learning (Ordun et al., 2020; Sainburg et al., 2021).

The final step is clustering in the dimension-reduced vector space. In our work, we focus on distance-based clustering algorithms, where the similarity of objects is determined by their distance in the vector space. The clustering literature is extensive (Aggarwal and Zhai, 2012). Among the most popular approaches are algorithms based on determining cluster centroids (K-Means (Lloyd, 1982), K-Medoids (Kaufman and Rousseeuw, 1990)), calculating local density (DBSCAN (Ester et al., 1996), OPTICS (Ankerst et al., 1999), HDBSCAN (Campello et al., 2013; McInnes and Healy, 2017)), computing spectral distributions (SPECTRAL (Ng et al., 2001)), or performing a hierarchical analysis (BIRCH (Zhang et al., 1996), Affinity Propagation (Frey and Dueck, 2007), Mean-Shift (Fukunaga and Hostetler, 1975; Cheng, 1995)). Centroid-based algorithms calculate the distances to cluster centroids to determine which point a cluster should be assigned to. In this work, we use K-means as a representative of this family since it is a widely used algorithm in multivariate data analysis. Density-based algorithms group data points that are in high-concentration areas of the vector space into clusters, with sparser regions in between. DBSCAN is one widely used density-based algorithm, with HDBSCAN being a hierarchical extension. In a comparison of DBSCAN and HDBSCAN for clustering news articles represented by Doc2Vec vectors, Radu et al. (2020) found both to be viable. We use HDBSCAN in this work due to its popularity in text clustering and because it is the default clustering algorithm in BERTopic (Grootendorst, 2022), the popular topic model based on the instance BERT → UMAP → HDBSCAN of the pipeline studied here.

¹The term *curse of dimensionality* was coined by Bellman (2003), originally published as (Bellman, 1957), to refer to the algorithmic disadvantages of a high-dimensional vector space.

Looking at the literature on document clustering, we observe the following in particular:

- (a) Doc2Vec and BERT have been extensively compared with TF-IDF as document representations for clustering, but not with each other;
- (b) the use of dimension reduction in combination with document embeddings and clustering is an understudied method despite its popularity in practice;
- (c) in particular, it is largely unknown how the performance of a clustering system for documents is affected by the number of dimensions of the embedding space.

To shed some light onto these questions, we have studied combinations of Doc2Vec, BERT, PCA, UMAP, K-Means, and HDBSCAN. The choice of these specific methods is motivated in Section 2.

We performed our experiments on collections of news articles because we expect news articles to belong to comparatively distinct topics and be grammatically correct. Moreover, there is a considerable practical demand for systems that can cluster collections of unlabeled news articles because maintaining a consistent tagging of articles even internally in a single publishing house is a significant problem, not to speak of multiple publishers.

We report on an extensive, systematic set of experiments with the task to broadly cluster three different labeled datasets of news articles using combinations of the above-mentioned embeddings and techniques with varying parameter settings, where the datasets are treated as unlabeled datasets and the gold labels are used for performance evaluation. To not only rely on a single quality measure, the quality of a clustering is assessed using both the Adjusted Rand Index (ARI, Hubert and Arabie (1985)) and the Adjusted Mutual Information Index (AMI, Vinh et al. (2010)). Additionally, to assess the intrinsic quality of the resulting topic models independently of the ground truth, a common measure for topic coherence, c_v (Röder et al., 2015), is used. The performance of the pipeline as a topic model is evaluated by adding a topic keyword extractor to each pipeline setup, hence, converting them into BERTopic-style topic models.

The structure of the paper is as follows. Section 2 explains our method. Sections 3 and 4 present and discuss the results of our experiments, respectively. Finally, conclusions are presented in Section 5.

2 Method

To enable experiments with different configurations of the clustering pipeline while keeping the components

separate and individually adjustable, a test suite was designed and implemented. We use the following terminology to separate individual instances of the pipeline architecture of Figure 1 from its parameter settings:

Definition 1 A pipeline with specified vectorization, dimension reduction, and clustering components, but with unspecified parameter settings, is a *setup*. A setup with specific parameter settings is a *configuration*.

The datasets used for training and evaluation are presented in Section 2.1. In Section 2.2, the structure of our test suite is discussed. Section 2.3 explains how we compare and evaluate different configurations.

2.1 Datasets

Three datasets with different characteristics were used as test data. All three consist of news articles written in English. The datasets are fully labeled, meaning that there are no unlabeled articles.

SNACK – *Scraped News Articles Classified with Keywords* consists of publicly available news articles scraped from the Internet in 2021. Topic-related keyword lists were used for classifying the articles, using keywords extracted by a term-based method. Articles classified as TECHNOLOGY (3156), FOOD/DRINK (2246), SPORTS (2836), STOCKS (2208), CONFLICTS (3086) and MOVIES/TV-SERIES (2859) of more than 500 characters were used for our experiments. The classes were chosen because they are largely unrelated. Articles occurring in multiple classes were removed. Unfortunately, the corpus cannot be made available as we do not own the publication rights of the individual articles it consists of. However, the URLs can be provided upon request.

AG News by Zhang et al. (2015) contains 1 000 000 categorized articles. For our study, a subset consisting of 15 000 articles from each of the four categories SPORTS, BUSINESS, SCIENCE/TECHNOLOGY and WORLD was used. This dataset was included to get a perspective on how configurations perform on a dataset consisting of a large number of comparatively short documents.

Reuters is based on the Thomson Reuters Text Research Collection (TRC2)² of 1 800 370 articles from 2008 and 2009. 578 712 of the articles are tagged with keywords. Using the keywords MARKET BONDS (2738), ENVIRONMENT (515), NATURAL DISASTER (777), SOCCER ENGLAND (1974), FILM (844), USA POLITICS (2559) and AUTO (1678) as selectors, a dataset of 11 085 articles was extracted for our experiments.

Name	Articles	Classes	Words
SNACK	16 391	6	7 509 853
AG News	60 000	4	4 520 259
REUTERS	11 085	7	3 148 736

Table 1: The datasets used in this study along with their size statistics.

2.2 The Clustering Pipeline

The version of the clustering pipeline used in our test suite is shown in Figure 1. Recall from Definition 1 that every choice of specific components results in a *setup*, and additionally fixing the parameters of these components yields a *configuration*.

The documents to be clustered are loaded and enter pre-processing. The pre-processing depends on the embedding to be used. For Doc2vec, stopwords, punctuation, and special characters are removed. The WordNet Lemmatizer is used for lemmatizing as it has been shown to be superior to stemming in clustering tasks (Iulia-Maria et al., 2020). For BERT, we follow the cleaning and tokenization steps described by Devlin et al. (2019), using the HuggingFace³ implementation.

Doc2Vec, in the Gensim implementation by Radim Řehůřek⁴, was selected as a typical representative of classical prediction-based neural embeddings. The Doc2Vec training process was run for 15 epochs. BERT, as implemented by HuggingFace³, was chosen as a representative of the Transformer-based class of embeddings. The BERT model was fine-tuned twice on all the sentences of the chosen dataset, using masked language modeling. The texts went through the pre-processing and vectorization only once to save time and to fairly compare the configuration parameters of the other modules.

For the vectorization phase of the pipeline, PCA was chosen to represent the class of matrix factorization techniques since it is widely applied in cases where dimension reduction is needed. UMAP (McInnes et al., 2018) represents the techniques based on neighbor graphs. It was chosen because it outperforms the popular t-SNE with respect to both efficiency and quality (see the original article). The numbers of dimensions tested are shown in Table 2. It is valuable to include reductions to as few as two or three dimensions since those are easy to visualize. As we will see, the effect of an increasing number of dimensions on the scores is not large. Hence, we opted to include comparatively few higher dimensions to reduce the computational resources needed.

²<https://trec.nist.gov/data/reuters/reuters.html>

³<https://huggingface.co/>

⁴<https://radimrehurek.com/gensim/>

Component	Technique	Settings	Value
Vectorization	Doc2Vec	dimensions	300
	BERT	dimensions	768
		layers	12
		attention heads	12
Dim. reduction	both	dimensions	[2, 3, 5, 7, 10, 15, 25, 50]
	PCA	principal components	equal to number of dimensions
	UMAP	$n_neighbors$	[5, 20, 80, 320, 1280, 2560]
Clustering	K-Means	k	[6, 12, 24, 48, 96, 192, 384] ^{SNACK} [4, 8, 16, 32, 64, 128, 256] ^{AG} [7, 14, 28, 56, 112, 224, 448] ^{REUT}
	HDBSCAN	$min_cluster_size$	[5, 10, 20, 40, 80, 160, 320, 640, 1280] ^{SNACK+REUT} [10, 20, 40, 80, 160, 320, 640, 1280, 2560] ^{AG}

Table 2: Major pipeline components and their explored setting configurations.

To investigate the effect of dimension reduction with UMAP, the main setting we manipulate is the variable $n_neighbors$, which determines how many points in the vicinity of a given point should be used to measure local density. A low value makes UMAP focus on the local structure of the vector space whereas a high value emphasizes its global structure. The settings we explored can be found in Table 2.

After dimension reduction, the vectors are L2-normalized, a step which, for simplicity, is not shown in Figure 1. For count-based methods, this normalization is common practice, whereas for neural methods there does not appear to be a clear recommendation as to which approach to use. Initial experiments revealed it to be advantageous for the overall scores to normalize vectors and center them around the origin after dimension reduction so that the clustering algorithm works on normalized vectors. This step could also be performed prior to dimension reduction. Since our initial experiments revealed no significant difference between these options, we chose the former as it will ensure that the norm of all vectors is 1 when the actual clustering algorithm is invoked.

As clustering components, we selected the common K-Means and the more recent HDBSCAN. K-Means is mainly parameterized by the number k of clusters to partition the dataset into. HDBSCAN transforms the vector space based on the local density of the set of points to be clustered and then creates a minimum spanning tree over these points. From that tree, a hierarchy can be created and then converted to a flat structure depending on a parameter $min_cluster_size$. In the respective configurations, we consider a range of settings for the parameters k and $min_cluster_size$ of K-Means and HDBSCAN, respectively, as specified in Table 2. Using a number k different from the number of distinct labels of the dataset allows K-Means to identify

a high-quality clustering with a number of clusters that differs from that of the ground truth. In fact, considering larger values of k is essential when evaluating the system as a topic modeling system.

In contrast to K-Means, which labels all points in the vector space, HDBSCAN detects apparent noise which it then leaves unlabeled. Since our datasets are fully labeled and are thus considered to not contain any noise, this difference makes HDBSCAN suffer in the comparison. To avoid this effect, we use soft clustering for HDBSCAN, meaning that all points get a similarity score with respect to each cluster and are then assigned to the cluster that results in the highest score.

This pipeline is often used to build topic models. To be able to evaluate the pipeline as such, we need to convert each configuration to a topic model. This is done by adapting the c-TF-IDF of BERTopic (Grootendorst, 2022) and assigning top keywords for all clusters. The top 10 keywords are used to represent a cluster as a topic. Further mentions of topic modeling refer to configurations that have the addition of c-TF-IDF, hence, which are topic models in the style of BERTopic.

2.3 Evaluation

In our test suite, we ran the configurations shown in Table 2 to cluster the datasets, and evaluated the quality of the resulting clusterings. For evaluation, we used the gold labels of the datasets together with two different methods of measuring clustering quality: the pair-based measurement Adjusted Rand Index (ARI, Hubert and Arabie (1985)) and the Shannon-based Adjusted Mutual Information Index (AMI, Vinh et al. (2010)). These were chosen because they are widely used in practice and have complementary strengths (Romano et al., 2016): ARI is considered to be advantageous if the ground truth consists of big equal-sized clusters

whereas AMI is preferable when the dataset is unbalanced, containing both large and small clusters.

Scores range from 0 to 1 where 0 marks a random clustering and 1 a clustering that agrees perfectly with the ground truth. We consider a higher score to indicate better clustering even though scores are not comprehensive for all aspects of clustering quality.

In addition to measuring quality by means of comparison with the ground truth, we use the topic coherence measure c_v by Röder et al. (2015) to estimate the intrinsic clustering quality by calculating a score between 0 and 1. The conclusion of Röder et al. (2015) was that c_v is the topic coherence measure most correlated to human judgment. Our coherence calculations employ the window size of 110 also used in Röder et al. (2015). Since c_v is computed by calculating a coherence score for each individual cluster and aggregating the scores, it may favor large numbers of small clusters (one cluster with a low score does not impact the aggregated score as much). However, we found that clusterings with a large number of clusters are not assigned a much higher score than those with a smaller number of clusters. Thus, one can also use c_v to determine an appropriate number of clusters. Hence, we find c_v adequate for comparing the quality of clusterings resulting from different configurations.

2.4 Limitations

While the components whose interactions we study have been chosen to be both typical and representative of a wide range of components that practical clustering pipelines may be composed of, they can only be example instances as there are many other options. We have therefore made our test suite available for download⁵. Some design choices, explained above, were made to keep the project and in particular the number of experiments manageable.

Another limitation of this study lies in the choice of datasets used in the experiments. They all have relatively few categories (at most seven) and are all reasonably balanced. The largest imbalance is found in the Reuters dataset where the largest category, USA POLITICS (2559) is five times larger than the smallest category, ENVIRONMENT (515). There could be many situations where the datasets contain many more categories or have a more unbalanced ground truth. Thus, the results of this work provide approximate parameter values for practitioners to initiate configuring their own system but should not be taken as universal truths.

3 Results

The results are presented as a qualitative analysis with the 2D plots (Figures 2-5) in Section 3.1, and a quantitative analysis presented in Sections 3.2 and 3.3. Details on how the scores were attained and processed are described below.

Combining all the possible configurations that can be constructed from Table 2 yielded 1792 total combinations distributed over 8 setups. The experiments, described in Section 2, were conducted by running all configurations on each dataset. Each configuration was run three times to account for non-deterministic components such as K-Means, UMAP, and the coherence measure c_v . The mean ARI, AMI, and c_v of the three runs on each configuration are considered to be the final scores of the configuration in question. In the text, performance refers to these scores and a better performance is a higher score. Figure 6 shows aggregated results obtained by averaging the performance figures for all configurations of each setup. Since the number of dimensions of the clustering space has a major influence on the results, it is kept as the X-axis, thus giving rise to *trend plots* that describe trends depending on the number of dimensions.

3.1 Visualization in 2D

The 2D plots in Figure 2–5 visualize the vectorized document spaces reduced to two dimensions. While the plots cannot be translated directly into higher dimensions, one can qualitatively compare the vector spaces with the corresponding results in the trend plots. There are clear differences between the 2D vector spaces created with UMAP and PCA. By adding the label color, it is clearly visible that the UMAP reductions keep a more defined geometry of the data corresponding to the original labeling.

3.2 Aggregated Trends per Setup

The trend plots shown in Figure 6 are an attempt to provide a general view of how well the setups work and how this depends on the number of dimensions. In order to obtain a compact comparison of all setups we illustrate their trends for each dataset-metric pair, i.e., there is one figure for each pair. Each figure contains 8 trend lines, one per setup. Each aggregated trend line shows the mean score (vertical axis) over all parameter settings for the given dataset and metric, depending on the dimension setting (horizontal axis).

The aggregated trends in Figure 6 show that the mean score is less in 2D than in 5D and higher. For most setups, the score increases until somewhere between

⁵https://github.com/antoneklund/Systematic_Parameter_Search_News_Article_Clustering

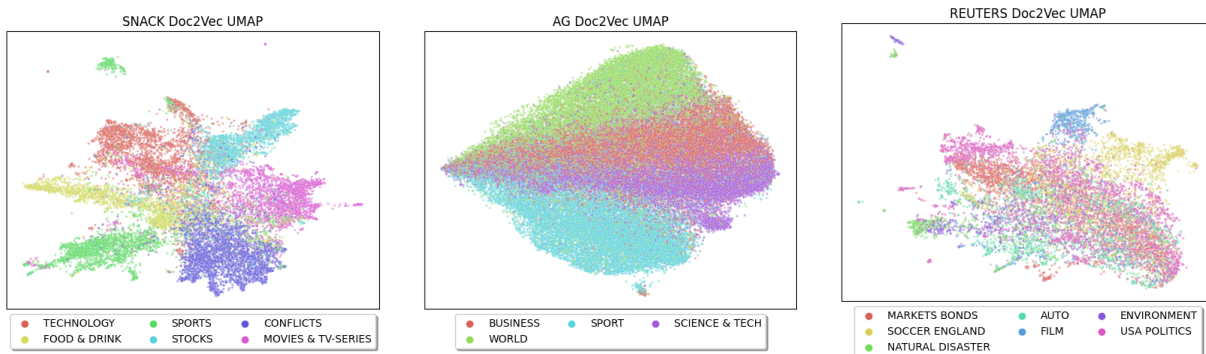


Figure 2: 2D UMAP reductions of Doc2Vec vectors.

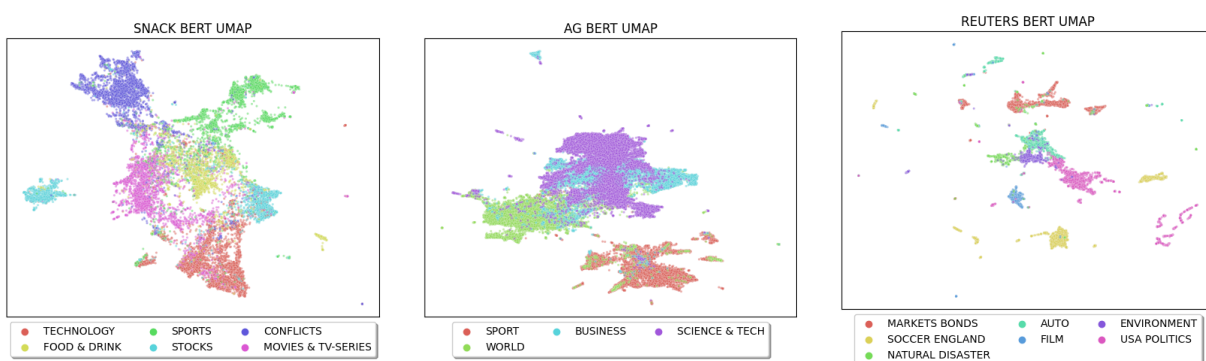


Figure 3: 2D UMAP reductions of BERT vectors.

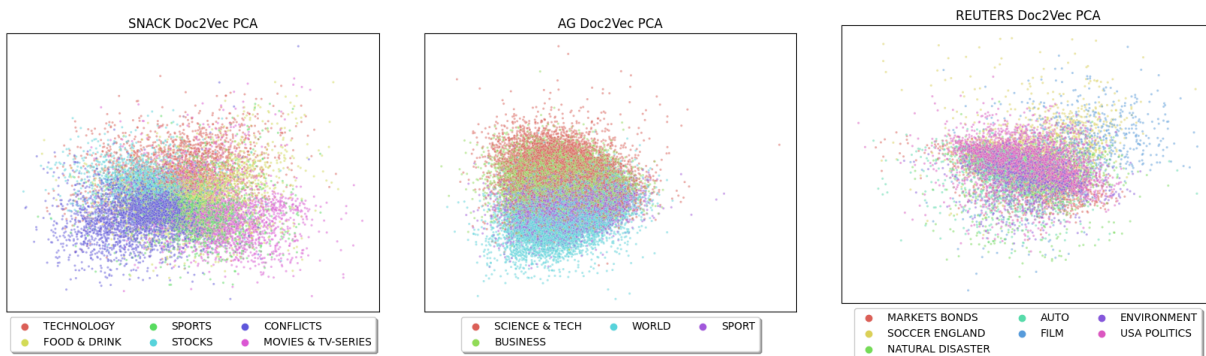


Figure 4: 2D PCA reductions of Doc2Vec vectors.

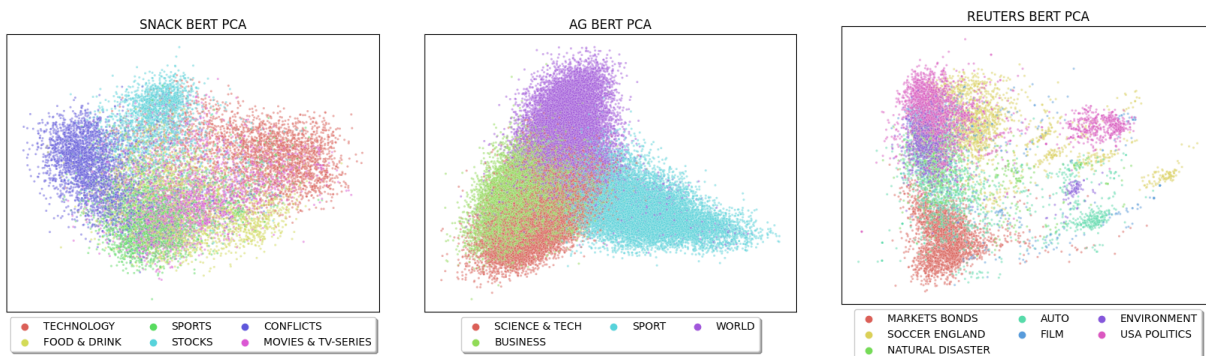


Figure 5: 2D PCA reductions of BERT vectors.

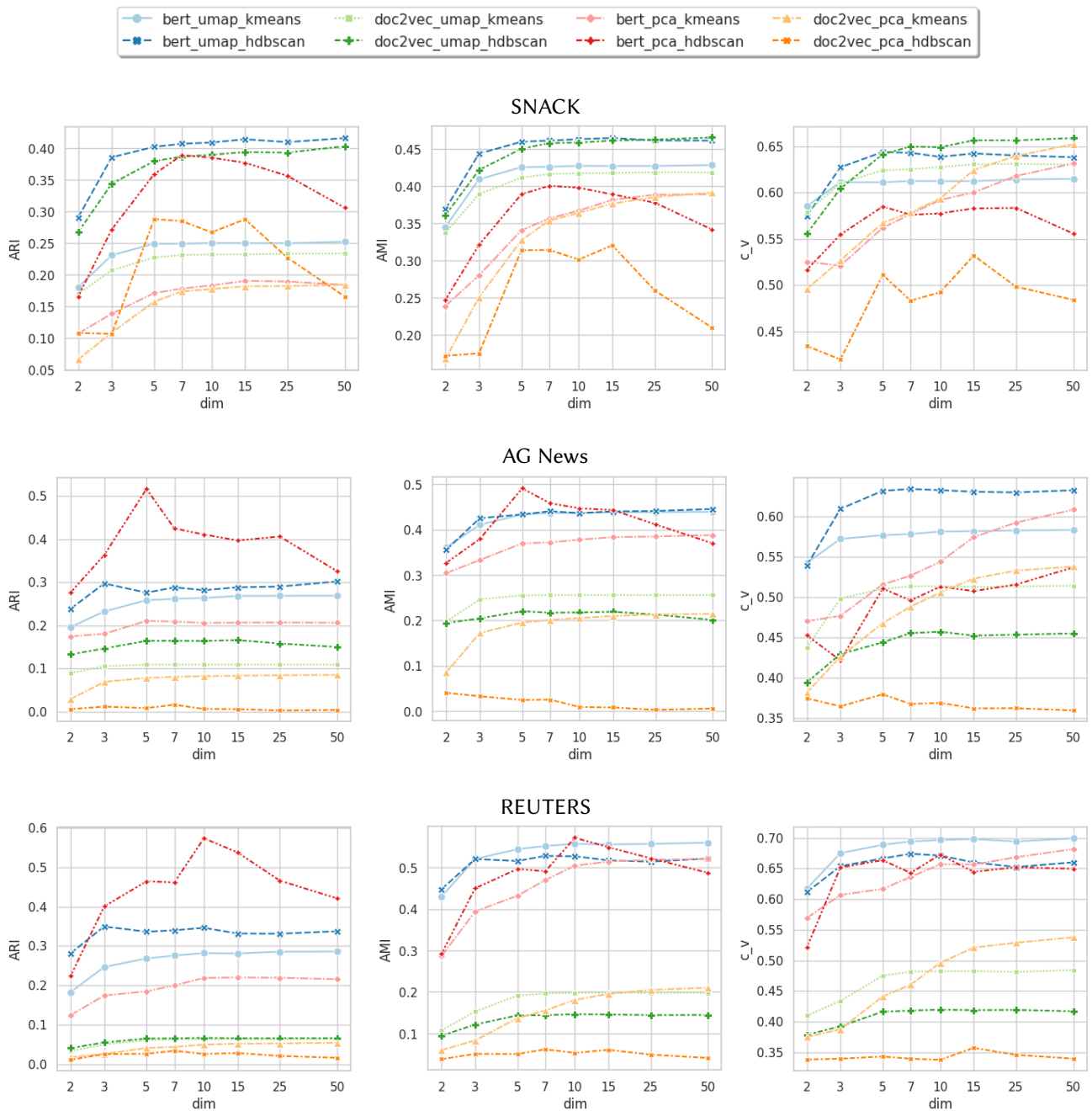


Figure 6: Aggregated trends for the SNACK (top), AG News (middle), and Reuters datasets (bottom). The evaluation metrics are ARI (left), AMI (middle), and c_v (right).

10D and 15D, after which there is no significant change. The exception is the setup `bert_pca_hdbscan`, which decreases after a peak in performance. (On SNACK, the setup `bert_pca_kmeans` is another exception, showing a similar behavior.)

The ARI and AMI scores for the setups does not indicate that more than 50D are needed. However, when looking at the c_v scores in Figure 6, there are setups (`bert_pca_kmeans` and `doc2vec_pca_kmeans`) that are still on a rising trend at 50D.

Some patterns re-occur across the different datasets. Setups that include BERT tend to perform better than those using Doc2Vec. This is true for AG News and Reuters but not for SNACK where the trends look similar for both vectorization methods. Also, more often than not setups that use UMAP seem to give rise to higher scores than the ones using PCA when the other two components are kept unchanged.

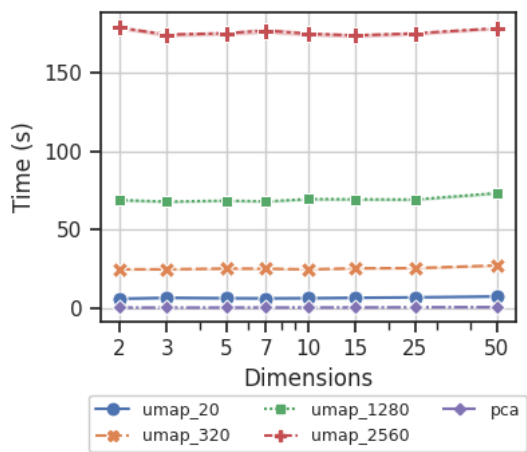


Figure 7: Dimension reduction mean time comparison over dimension for the Reuters dataset. UMAP with $n_neighbors = 20$ is around 6s and PCA is around 0.5s.

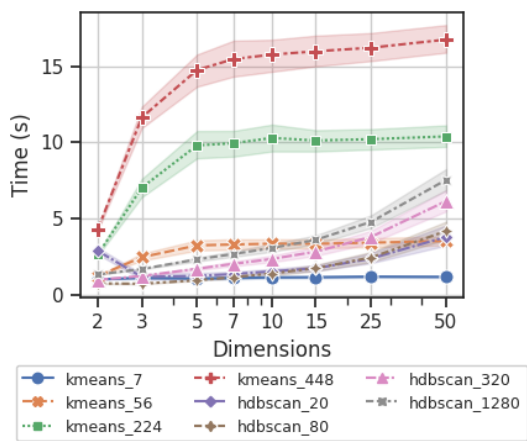


Figure 8: Clustering time comparison over dimension for the Reuters dataset.

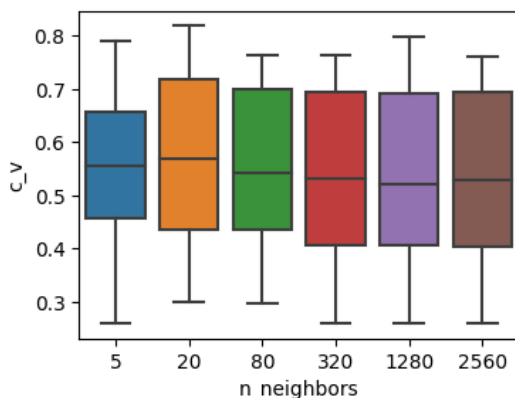


Figure 9: The different scores of c_v on the Reuters dataset depending on the UMAP variable $n_neighbors$.

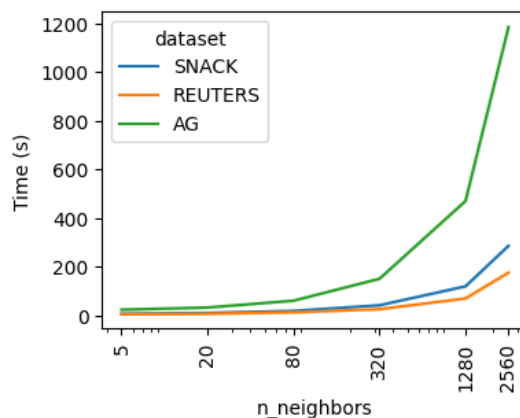


Figure 10: Comparison of time required to perform the UMAP dimension reduction depending on the variable $n_neighbors$.

3.3 Configurations

For the various configurations, there are large quantities of data that mostly tie into the individual datasets. Hence, showing all of them is not meaningful. Therefore, we present a sample of, as we hope, informative configurations in Tables 3–5. We chose the best configuration for each setup. Three tables are presented, one for each of the evaluation metrics ARI, AMI, and c_v . The best-performing configuration for a dataset is highlighted. This is most often bert_umap_hdbscan or bert_umap_kmeans but on the SNACK dataset, doc2vec_umap_hdbscan also achieves a high score.

In addition to these, we chose the Reuters dataset to show the relation between the number of clusters and the final score in Figure 11. The plots report all scores for ARI, AMI, and c_v divided into the different setups. For the Reuters dataset, the ground truth number of clusters is seven, and this is also where we find the highest scores for ARI and AMI. The topic modeling score c_v attains a higher value for a number of clusters larger than the ground truth.

The mean dimension reduction wall times for PCA and UMAP with different settings of the parameter $n_neighbors$ are shown in Figure 7. PCA is faster than UMAP by a large margin. UMAP computation time increases significantly along with $n_neighbors$. However, the computation time is rather unaffected by increasing the number of dimensions from 2D to 50D.

The average c_v score per UMAP $n_neighbors$ setting for the Reuters dataset is plotted in Figure 9. The boxes are similar, which means that the parameter has only a small impact on the score. The best-performing setting is $n_neighbors = 20$ where the mean score is slightly higher than for the other settings. Related to this is the dimension reduction time reported for different $n_neighbors$ that are shown in Figure 10. It can be

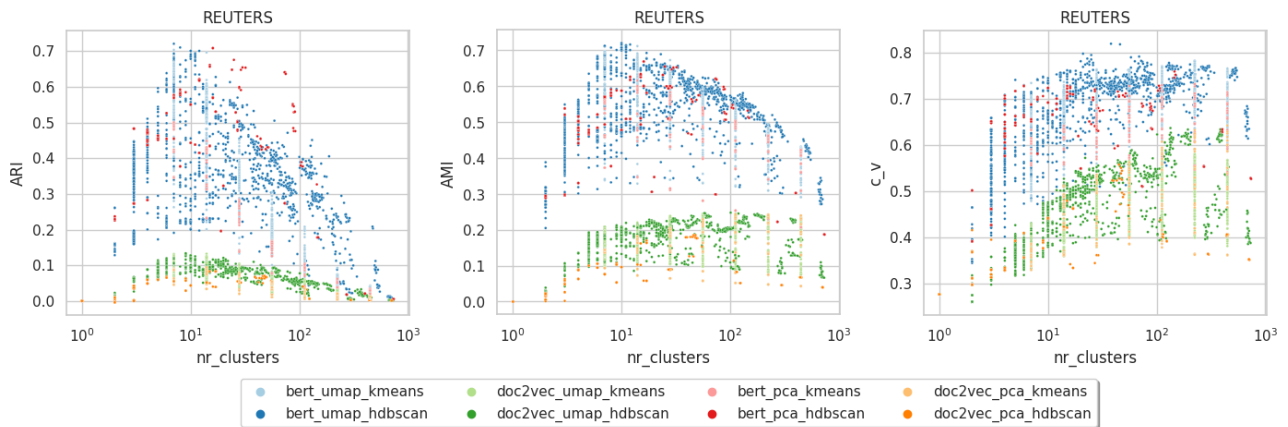


Figure 11: Relation between the number of clusters and the score for different metrics in the Reuters dataset.

Data	Setup	Dim	Alg. setting	Nr of clusters	Time	ARI
SNACK	bert_umap_kmeans	25	6	6	10.82	0.56
	bert_umap_hdbscan	15	320	7	20.27	0.58
	doc2vec_umap_kmeans	15	6	6	12.5	0.55
	doc2vec_umap_hdbscan	50	640	6	21.05	0.58
	bert_pca_kmeans	25	6	6	2.62	0.51
	bert_pca_hdbscan	15	160	6	6.37	0.50
	doc2vec_pca_kmeans	50	6	6	2.02	0.50
	doc2vec_pca_hdbscan	15	160	6	7.26	0.45
AG NEWS	bert_umap_kmeans	10	4	4	31.06	0.67
	bert_umap_hdbscan	50	2560	4	1288.69	0.59
	doc2vec_umap_kmeans	5	4	4	30.19	0.26
	doc2vec_umap_hdbscan	15	160	5	75.11	0.28
	bert_pca_kmeans	50	4	4	5.82	0.64
	bert_pca_hdbscan	5	2560	4	16.79	0.57
	doc2vec_pca_kmeans	50	4	4	3.73	0.17
	doc2vec_pca_hdbscan	3	80	5	4.29	0.09
REUTERS	bert_umap_kmeans	50	7	7	14.62	0.69
	bert_umap_hdbscan	25	160	10	14.47	0.70
	doc2vec_umap_kmeans	15	7	7	5.09	0.12
	doc2vec_umap_hdbscan	25	80	8	6.27	0.13
	bert_pca_kmeans	15	7	7	1.61	0.50
	bert_pca_hdbscan	15	20	16	2.33	0.69
	doc2vec_pca_kmeans	50	14	14	2.68	0.09
	doc2vec_pca_hdbscan	25	10	17	5.66	0.08

Table 3: A table of the best configuration according to ARI for each setup on each dataset. The column ‘Alg. setting’ reports the number k of clusters in K-Means and $min_cluster_size$ in HDBSCAN.

seen that the computation time increases with larger $n_neighbors$ as well as with the size of the dataset.

4 Discussion

The purpose of this study has been to help practitioners limit the time spent on building a clustering system and tuning its hyperparameters. The following discussion is structured according to the three main degrees of freedom, namely the number of dimensions, the choice of components, and the parameter tuning.

Data	Setup	Dim	Alg. setting	Nr of clusters	Time	AMI
SNACK	bert_umap_kmeans	25	6	6	10.81	0.54
	bert_umap_hdbscan	15	320	7	20.27	0.55
	doc2vec_umap_kmeans	15	6	6	11.33	0.54
	doc2vec_umap_hdbscan	25	640	6	16.55	0.55
	bert_pca_kmeans	25	6	6	2.62	0.51
	bert_pca_hdbscan	15	160	6	6.37	0.49
	doc2vec_pca_kmeans	50	6	6	2.02	0.49
	doc2vec_pca_hdbscan	7	160	6	3.23	0.45
AG NEWS	bert_umap_kmeans	10	4	4	31.06	0.64
	bert_umap_hdbscan	25	2560	3	68.6	0.63
	doc2vec_umap_kmeans	3	8	8	30.26	0.31
	doc2vec_umap_hdbscan	15	160	5	75.11	0.31
	bert_pca_kmeans	50	4	4	5.82	0.6
	bert_pca_hdbscan	5	2560	4	16.79	0.54
	doc2vec_pca_kmeans	50	16	16	11.09	0.24
	doc2vec_pca_hdbscan	3	80	5	4.29	0.15
REUTERS	bert_umap_kmeans	50	7	7	14.62	0.69
	bert_umap_hdbscan	25	160	10	14.47	0.71
	doc2vec_umap_kmeans	10	56	56	8.1	0.24
	doc2vec_umap_hdbscan	10	10	62	8.1	0.24
	bert_pca_kmeans	25	14	14	2.18	0.6
	bert_pca_hdbscan	15	20	16	2.33	0.66
	doc2vec_pca_kmeans	50	224	224	10.27	0.24
	doc2vec_pca_hdbscan	25	5	47	6.29	0.18

Table 4: A table of the best configuration according to AMI for each setup on each dataset. The column ‘Alg. setting’ reports the number k of clusters in K-Means and $min_cluster_size$ in HDBSCAN.

4.1 Dimension

The number of dimensions of the clustering vector space is relevant for the clustering result. Too few dimensions will remove relevant information from the vector space, and too many dimensions may make the clustering drop in performance and become computationally inefficient. The difficulty lies in quantifying *too few* and *too many*. The results of this study show that performance typically increases from 2D to somewhere between 10D and 15D, where the increase stagmates. The expected performance drop in higher dimensions due to the curse of dimensionality does not seem

Data	Setup	Dim	Alg. setting	Nr of clusters	Time	c_v
SNACK	bert_umap_kmeans	50	96	96	13.01	0.67
	bert_umap_hdbscan	5	20	41	19.56	0.73
	doc2vec_umap_kmeans	15	12	12	7.00	0.68
	doc2vec_umap_hdbscan	15	40	17	20.98	0.71
	bert_pca_kmeans	50	96	96	11.31	0.65
	bert_pca_hdbscan	50	20	12	16.98	0.64
	doc2vec_pca_kmeans	50	96	96	9.63	0.68
	doc2vec_pca_hdbscan	50	5	105	26.69	0.67
AG NEWS	bert_umap_kmeans	50	246	256	82.13	0.67
	bert_umap_hdbscan	50	20	140	116.43	0.73
	doc2vec_umap_kmeans	25	128	128	87.21	0.57
	doc2vec_umap_hdbscan	5	10	107	52.09	0.60
	bert_pca_kmeans	50	64	64	31.78	0.67
	bert_pca_hdbscan	50	40	14	369.37	0.63
	doc2vec_pca_kmeans	50	128	128	62.03	0.61
	doc2vec_pca_hdbscan	5	10	38	9.59	0.5
REUTERS	bert_umap_kmeans	25	224	224	11.21	0.78
	bert_umap_hdbscan	15	20	63	8.74	0.79
	doc2vec_umap_kmeans	50	112	112	10.81	0.58
	doc2vec_umap_hdbscan	25	5	187	17.53	0.62
	bert_pca_kmeans	50	224	224	11.08	0.72
	bert_pca_hdbscan	50	5	147	12.44	0.75
	doc2vec_pca_kmeans	50	448	448	16.02	0.63
	doc2vec_pca_hdbscan	15	5	47	4.20	0.55

Table 5: A table of the best configuration according to c_v for each setup on each dataset. The column ‘Alg. setting’ reports the number k of clusters in K-Means and *min_cluster_size* in HDBSCAN.

to pose a significant problem for the range of dimensions tested in this article. Hence, for a system that has to perform well on unknown data, a reasonable initial guess would be to use 15D or (moderately) higher.

While a higher-dimensional vector space (within the range in this study) seems to ensure better performance, it has to be weighed against the resulting increase in computation time. As seen in Figure 7, the cost of performing the dimension reduction itself does not significantly depend on the number of dimensions. Instead, the most significant factor affecting the efficiency of the dimension reduction is the size of the dataset and (in the case of UMAP) the $n_neighbors$ parameter as shown in Figure 10. As seen in Figure 8, the clustering times increase very slowly in higher dimensions. Still, the increase in clustering time indicates that the number of dimensions should be kept down if there is no significant performance gain.

We recommend attempting to find a balance between efficiency and desired performance. As previously mentioned, the performance increase tends to stagnate around 15D. Hence, as a rule of thumb, we recommend a reduction to a range from around 15D to 25D. Future work will need to be conducted to study the impact of a number of dimensions higher than 50D, which was the limit in this study.

4.2 Choice of Components

The trend plots in Figures 6 give insights into the component performance, and the 2D plots of the vector spaces in Figures 2–5 add a geometrical view of the results. From this, we draw the following observations.

4.2.1 Vectorization Method

The performance of setups that include BERT is better or similar to that of setups that include Doc2Vec when all other components are the same. We can also see that the highlighted best-performing configurations always include BERT in Tables 3–5. At best, Doc2Vec achieves on-par results with BERT on SNACK. This makes us conclude that BERT as a vectorization method is preferable over Doc2Vec, and we recommend using it in a clustering pipeline.

4.2.2 Dimension Reduction

For dimension reduction, the setups using UMAP yield more stable results, with the scores increasing until they stabilize at around 10D. PCA sometimes shows peaks in scores for configurations around 5D to 15D. However, the performance decreases in higher dimensions as seen in Figure 6. We note in Tables 3–5 that setups with UMAP achieve the top scores. Therefore, UMAP generally seems like a more stable recommendation. However, in this context, it is worth recalling a major advantage of PCA that is not highlighted in the experiments of this article. Namely, that the axes of its coordinate system correspond to the Eigenvectors computed in the course of matrix decomposition. As such, these axes carry a distinct mathematical meaning, which is important for explainability. A common application of this fact is to use the explained variance of the axes for analysis (Raunak et al., 2019).

For visualization purposes, the choice of UMAP is evident when comparing the 2D plots of UMAP in Figures 2 and 3, with PCA in Figures 4 and 5. The vector spaces for the UMAP-reduced datasets form clear clusters without mixing the categories. This result is expected as preserving cluster structure in lower dimensions is something that neighbor graph methods were designed to do.

Nevertheless, if explainability is not a major concern, it seems safe to conclude that UMAP as a component in the document clustering pipeline is preferable over PCA because of both performance stability and visualization properties. This is also supported in the literature by Allaoui et al. (2020). However, PCA performs well in certain configurations and is more efficient. PCA could therefore be preferable in situations where strict time constraints must be obeyed or it is important to be able to interpret the vector space axes.

4.2.3 Clustering Algorithm

Our results show that HDBSCAN generally performs well in combination with UMAP. K-Means also displays good performance but achieves slightly lower scores than HDBSCAN. First and foremost, performance tends to be determined by the other components and particularly the vectorization. It is therefore sensible to leave the choice of clustering algorithm to the practitioner who can visualize the vector space (preferably with UMAP) to obtain information about its shape and make an informed decision (Eklund and Forsman, 2022).

HDBSCAN combined with PCA is the only setup that sometimes exhibits a downward trend after a peak around 5D to 15D. This could signal that HDBSCAN is inferior to K-Means at handling the shrinking variance in distance that occurs in higher dimensions. However, this phenomenon does not occur in all setups involving HDBSCAN, meaning that the behavior cannot be caused by HDBSCAN alone. In fact, the peaks sometimes occur in the best-performing configuration for a dataset. Therefore, we cannot discourage combining PCA with HDBSCAN, but we do advise caution when using this combination.

The rightmost plots (with the metric c_v) in Figure 6 show that a topic model could be successfully created with any combination of UMAP or PCA, and HDBSCAN or K-Means. The performance is again mostly dependent on the vectorization. Furthermore, while the performance increase often seems to stagnate in higher dimensions, setups with PCA and K-Means keep improving. This indicates that increasing the number of dimensions beyond what was done in this study may eventually turn PCA and K-Means into the best-performing combination.

4.3 Parameter Tuning

Parameter tuning is the task most dependent on the dataset. However, being able to trust that the system is well configured is especially important when facing unseen data, and thus when tuning is most difficult. Tables 3–5 contain the best-performing configurations for each setup. These tables give some ideas of what is important when choosing a parameter setting.

One central aspect appears to be the number of clusters. For ARI (Table 3) and AMI (Table 4), it is clear that if the clustering produces a number of clusters closer to the number of gold labels, then the score will be higher. Where this fails, such as setups involving Doc2Vec for the Reuters dataset in Table 4, is when the score is so low that the setup should be discarded no matter the configuration. The recommendation that the number of clusters should stay close to the number of gold labels is also supported by Figure 11, where the highest scores

are obtained by values around seven for $nr_clusters$. In a real-world environment, it could of course be difficult to make practical use of this observation because the “real” number of clusters may not a priori be known. Strategies exist for finding an optimal number of clusters for a dataset that can be used to set the parameter k for K-Means (Kodinariya et al., 2013). In this regard, an advantage of HDBSCAN is that $min_cluster_size$ is related to the dataset size, which is usually known.

The metric c_v used to evaluate topic modeling systems favors pipeline configurations that result in a larger number of clusters than the coarse categorization of the annotated data; see Table 5. This is also indicated by the large values for both c_v and $nr_clusters$ in the rightmost plot in Figure 11. Some benefits of using smaller values of $min_cluster_size$, which yield a larger number of clusters, have been suggested for topic modeling of short social media texts (Asyaky and Mandala, 2021). Our results let us agree with this recommendation for longer news article texts as well. Overall, the clustering algorithms show comparative performances when applied to the same vector space. Hence, there does not seem to be any harm in choosing the algorithm based on domain and application knowledge.

Computational efficiency is an aspect practitioners may have to take into account. UMAP takes considerably longer time to compute than PCA as shown in Figure 7 and supported by the benchmarking comparison found in the UMAP documentation⁶. UMAP complexity is bound by the calculation of $n_neighbors$ and has empirically been shown to be $O(N^{1.14})$ (McInnes et al., 2018). Our results support this by showing a wall time that essentially increases linearly, as shown in Figure 10. From Figure 9 we can also see that $n_neighbors$ in general has a low impact on the overall scores. If there are any patterns, it is that smaller values of $n_neighbors$ are more frequently present in the best-scoring configurations. In UMAP, the parameter $n_neighbors$ is supposed to weigh retaining the local structure against retaining the global structure of the data (smaller vs. larger $n_neighbors$, respectively). Judging from the results in this study, we presume that it is better to focus on preserving the local structure, as also supported in Asyaky and Mandala (2021).

In conclusion, the choice of parameters should be based on how many clusters one expects to find, weighed against any efficiency constraints the system may have. There are indications that the UMAP parameter $n_neighbors$ should be chosen with a lower value to preserve the local structure of the data when working with document embeddings.

⁶<https://umap-learn.readthedocs.io/en/latest/benchmarking.html>

5 Conclusions

After systematically studying different setups of vectorization, dimension reduction, and clustering together with a large number of parameter settings, we conclude that the vectorization component has the most significant impact on the performance of the system and that BERT usually results in a better embedding space for clustering than Doc2Vec. When reducing the vector space, vectors should not be reduced to less than 15D. UMAP most frequently exhibits better performance and visualization capabilities than PCA. However, PCA can be favored if computational efficiency or explainability is required. The clustering algorithms perform roughly on par with each other but with a slight advantage to HDBSCAN over K-Means. The choice of a clustering algorithm ultimately comes down to knowledge about the dataset and application domain. Influencing that choice, and all the parameters of the setup, are mainly the computation time and the number of clusters that the data shall be divided into.

The popularity of the practical pipeline for document clustering and topic modeling studied in this paper is unlikely to decrease in the near future. With this in mind, we think that additional work aiming to evaluate and improve such systems is required. This study used labeled data to assess the performance of different setups and configurations. While we were able to draw a number of general rule-of-thumb conclusions that will hopefully benefit the practitioner, there is no getting around the fact that, ultimately, a lot of domain knowledge is required in concrete practical scenarios. The use of automatic measurements, as done in this study, can be one way of coming up with reasonable settings. However, we believe that such methods have intrinsic limitations in contexts whose end users are humans, e.g., consumers of news articles or readers of online advertisements. In such cases, we believe it to be necessary to complement automatic assessments of the quality of clusterings or topic models by systematic methods based on human judgment. How this can be done in a qualified manner with reasonable budgets appears to be an open question that deserves the focus of future research.

Acknowledgment

We thank the reviewers for their thorough reading of the initial manuscript and their insightful comments which have been useful in revising this paper.

References

Charu C. Aggarwal and ChengXiang Zhai. 2012. A survey of text clustering algorithms. In Charu C. Aggar-

wal and ChengXiang Zhai, editors, *Mining Text Data*, pages 77–128. Springer US, Boston, MA.

Mebarka Allaoui, Mohammed Lamine Kherfi, and Abdelhakim Cheriet. 2020. Considerably improving clustering algorithms using umap dimensionality reduction technique: A comparative study. In Abderrahim El Moataz, Driss Mammass, Alamin Mansouri, and Fathallah Nouboud, editors, *Image and Signal Processing*, pages 317–325. Springer International Publishing, Cham.

Mihael Ankerst, Markus M Breunig, Hans-Peter Kriegel, and Jörg Sander. 1999. Optics: Ordering points to identify the clustering structure. *ACM Sigmod record*, 28(2):49–60.

Muhammad Sidik Asyaky and Rila Mandala. 2021. Improving the performance of hdbscan on short text clustering by using word embedding and umap. In *2021 8th International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*, pages 1–6.

Etienne Becht, Leland McInnes, John Healy, Charles Antoine Dutertre, Immanuel W.H. Kwok, Lai Guan Ng, Florent Ginhoux, and Evan W. Newell. 2019. Dimensionality reduction for visualizing single-cell data using umap. *Nature Biotechnology*, 37:38–44.

Richard Ernest Bellman. 1957. *Dynamic Programming*. Princeton University Press.

Richard Ernest Bellman. 2003. *Dynamic Programming*. Dover Publications.

Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. 2013. Density-based clustering based on hierarchical density estimates. In *Advances in Knowledge Discovery and Data Mining*, pages 160–172, Berlin, Heidelberg. Springer Berlin Heidelberg.

Yizong Cheng. 1995. Mean shift, mode seeking, and clustering. *IEEE transactions on pattern analysis and machine intelligence*, 17(8):790–799.

Rob Churchill and Lisa Singh. 2022. The evolution of topic modeling. *ACM Comput. Surv.*, 54(10s).

Stephan A. Curiskis, Barry Drake, Thomas R. Osborn, and Paul J. Kennedy. 2020. An evaluation of document clustering and topic modelling in two online social networks: Twitter and reddit. *Information Processing & Management*, 57(2):102034.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the*

- North American Chapter of the Association for Computational Linguistics: *Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Anton Eklund and Mona Forsman. 2022. Topic modeling by clustering language model embeddings: Human validation on an industry dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 635–643, Abu Dhabi, UAE. Association for Computational Linguistics.
- Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proc. 2nd Intl. Conf. on Knowledge Discovery and Data Mining (KDD-96)*, pages 226–231.
- Brendan J Frey and Delbert Dueck. 2007. Clustering by passing messages between data points. *science*, 315(5814):972–976.
- K. Fukunaga and L. Hostetler. 1975. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory*, 21(1):32–40.
- Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.
- Harold Hotelling. 1933. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24.
- Lawrence Hubert and Phipps Arabie. 1985. Comparing partitions. *Journal of Classification*, 2:193–218.
- Rădulescu Iulia-Maria, Ciprian-Octavian Truică, Elena Simona Apostol, Alexandru Boicea, Mariana Mocanu, Daniel-Călin Popeangă, and Florin Rădulescu. 2020. Density-based text clustering using document embeddings. In *Proceedings of the 36th International Business Information Management Association Conference (IBIMA)*.
- Leonard Kaufman and Peter J. Rousseeuw. 1990. *Partitioning Around Medoids (Program PAM)*, chapter 2. John Wiley & Sons, Ltd.
- Trupti M Kodinariya, Prashant R Makwana, et al. 2013. Review on determining number of cluster in k-means clustering. *International Journal*, 1(6):90–95.
- Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1188–1196, Beijing, China. PMLR.
- Xin Li, Ondrej E. Dyck, Mark P. Oxley, Andrew R. Lupini, Leland McInnes, John Healy, Stephen Jesse, and Sergei V. Kalinin. 2019. Manifold learning of four-dimensional scanning transmission electron microscopy. *npj Computational Materials*, 5:5.
- Stuart P. Lloyd. 1982. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137.
- Leland McInnes and John Healy. 2017. Accelerated hierarchical density based clustering. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE.
- Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861.
- Andrew Ng, Michael Jordan, and Yair Weiss. 2001. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems*, volume 14. MIT Press.
- Catherine Ordun, Sanjay Purushotham, and Edward Raff. 2020. Exploratory analysis of covid-19 tweets using topic modeling, umap, and digraphs. *arXiv preprint arXiv:2005.03082*.
- Karl Pearson. 1901. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2.
- Robert George Radu, Iulia Maria Rădulescu, Ciprian Octavian Truică, Elena Simona Apostol, and Mariana Mocanu. 2020. Clustering documents using the document to vector model for dimensionality reduction. In *Proceedings of the 22nd IEEE International Conference on Automation, Quality and Testing, Robotics - THETA, AQTR 2020*.
- Vikas Raunak, Vivek Gupta, and Florian Metze. 2019. Effective dimensionality reduction for word embeddings. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 235–243, Florence, Italy. Association for Computational Linguistics.
- Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, WSDM '15*, page 399–408, New York, NY, USA. Association for Computing Machinery.
- Simone Romano, Nguyen Xuan Vinh, James Bailey, and Karin Verspoor. 2016. Adjusting for chance clustering

- comparison measures. *The Journal of Machine Learning Research*, 17(1):4635–4666.
- Tim Sainburg, Leland McInnes, and Timothy Q. Gengner. 2021. Parametric umap embeddings for representation and semisupervised learning. *Neural Computation*, 33(11):2881–2907.
- Claude Sammut and Geoffrey I. Webb, editors. 2010. *TF-IDF*. Springer US, Boston, MA.
- Suzanna Sia, Ayush Dalmaia, and Sabrina J. Mielke. 2020. Tired of topic models? clusters of pretrained word embeddings make for fast and good topics too! In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1728–1736, Online. Association for Computational Linguistics.
- Michael Steinbach, Levent Ertöz, and Vipin Kumar. 2004. The challenges of clustering high dimensional data. In Luc T. Wille, editor, *New Directions in Statistical Physics: Econophysics, Bioinformatics, and Pattern Recognition*, pages 273–309. Springer Berlin Heidelberg, Berlin, Heidelberg.
- Alvin Subakti, Hendri Murfi, and Nora Hariadi. 2022. The performance of bert as data representation of text clustering. *Journal of Big Data*, 9:15.
- Ciprian-Octavian Truică, Florin Rădulescu, and Alexandru Boicea. 2016. Comparing different term weighting schemas for topic modeling. In *2016 18th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*, pages 307–310.
- Nguyen Xuan Vinh, Julien Epps, and James Bailey. 2010. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11(95):2837–2854.
- Tian Zhang, Raghu Ramakrishnan, and Miron Livny. 1996. Birch: An efficient data clustering method for very large databases. In *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data, SIGMOD '96*, page 103–114, New York, NY, USA. Association for Computing Machinery.
- Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015*, volume 28, pages 649–657.
- Zihan Zhang, Meng Fang, Ling Chen, and Mohammad Reza Namazi Rad. 2022. Is neural topic modelling better than clustering? an empirical study on clustering with contextual embeddings for topics. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3886–3893, Seattle, United States. Association for Computational Linguistics.
- He Zhao, Dinh Phung, Viet Huynh, Yuan Jin, Lan Du, and Wray Buntine. 2021. Topic modelling meets deep neural networks: A survey. *arXiv preprint arXiv:2103.00498*.
- Arthur Zimek. 2014. Clustering high-dimensional data. In Charu C Aggarwal and Chandan K Reddy, editors, *Data clustering*, chapter 9, pages 202–229. Citeseer.

Prevention or Promotion? Predicting Author’s Regulatory Focus

Aswathy Velutharambath,

Psychological AI (100 Worte Sprachanalyse GmbH) / Institut für Maschinelle Sprachverarbeitung, University of Stuttgart, Germany
aswathy.velutharambath@100worte.de

Kai Sassenberg, Leibniz-Institut für Wissensmedien, Tübingen, Germany / University of Tübingen, Germany

k.sassenberg@iwm-tuebingen.de

Roman Klinger, Institut für Maschinelle Sprachverarbeitung, University of Stuttgart, Germany

roman.klinger@ims.uni-stuttgart.de

Abstract People differ fundamentally in what motivates them to pursue a goal and how they approach it. For instance, some people seek growth and show eagerness, whereas others prefer security and are vigilant. The concept of regulatory focus is employed in psychology, to explain and predict this goal-directed behavior of humans underpinned by two unique motivational systems – the promotion and the prevention system. Traditionally, text analysis methods using closed-vocabularies are employed to assess the distinctive linguistic patterns associated with the two systems. From an NLP perspective, automatically detecting the regulatory focus of individuals from text provides valuable insights into the behavioral inclinations of the author, finding its applications in areas like marketing or health communication. However, the concept never made an impactful debut in computational linguistics research. To bridge this gap we introduce the novel task of regulatory focus classification from text and present two complementary German datasets – (1) experimentally generated event descriptions and (2) manually annotated short social media texts used for evaluating the generalizability of models on real-world data. First, we conduct a correlation analysis to verify if the linguistic footprints of regulatory focus reported in psychology studies are observable and to what extent in our datasets. For automatic classification, we compare closed-vocabulary-based analyses with a state-of-the-art BERT-based text classification model and observe that the latter outperforms lexicon-based approaches on experimental data and is notably better on out-of-domain Twitter data.

1 Introduction

What motivates a person to pursue a goal and what type of strategies they apply to achieve this goal differs from person to person. For instance, some people brush regularly to maintain healthy teeth and gums, while others do the same to avoid cavities; their goal is the same but the motivation is different. The regulatory focus (RF) theory (Higgins, 1997, 1998) from psychology, explains the goal-directed behavior of humans underpinned by two unique motivational systems – *promotion* and *prevention*. Promotion-focused individuals approach a goal by striving for achievements, taking an eager approach, and are interested in maximizing the gains. On the other hand, prevention-focused ones strive for security, are sensitive to losses, avoid negative outcomes, avert risks, and are vigilant. This framework is predominantly used to explain consumer behavior, organizational behavior, message framing, or information processing (Crowe and Higgins, 1997; Aaker and Lee, 2001; Lanaj et al., 2012; Sassenberg et al., 2014, i.a.).

Automatic detection of regulatory focus helps psychology researchers to bypass the need for manual coding or self-reports which are prone to social desirability. Automatic detection would allow for a more objective and standardized measurement of regulatory focus, removing egocentric biases and subjectivity. In the case of downstream applications, like computer-mediated communication, understanding the behavioral inclination of a person allows one to tailor response messages to fit their motivational orientation, facilitating a more persuasive and effective dialogue between the interlocutors. Such tailoring of messages to match the dominant regulatory focus has proven effective in health communication, promoting positive behavior change in areas like exercise (Latimer et al., 2008a), diet (Latimer et al., 2008b), and dental hygiene (Updegraff et al., 2007). It has also been successfully applied in organizational behavior, marketing, leadership, and many other domains of psychology (Sassenberg and Vliek, 2019, p.51-64). With more than 1,500 journal publications and more than 70,000 citations, the concept is prominent in psychology

Statement	Reg. Focus
(A) I woke up early because I did not want to miss the bus and be late for the class	Prevention
(B) I woke up early because I wanted to be on time for my favorite class	Promotion

Table 1: Examples of prevention and promotion-focused statements

and related disciplines¹. However, this concept has not received any attention in computational linguistics.

Previous studies on the topic of regulatory focus have reported that distinctive linguistic signatures are observed in an individual’s text formulation corresponding to their goal attainment strategies (Semin et al., 2005; Vaughn, 2018). Other studies relied on differences in linguistic features to manipulate regulatory focus or persuade people with a specific regulatory focus. Overall, we conclude that promotion and prevention focus resonate with different linguistic patterns. Inspired by these findings and their practical application in communication, we formulate the novel task of regulatory focus classification as an author profiling task.

Consider the two statements in Table 1, both describing a person’s motivation *to wake up early*. In Statement (A), the person wants to avoid negative outcomes like *missing the bus* or *being late for the class*, which points to a risk-averting motivation or prevention focus. On the contrary, in Statement (B), the person sounds eager and focuses on the positive outcome of *being on time for the class* which warrants promotion focus. As noted earlier, despite the presence of distinctive stylistic variations and linguistic cues, no attempts to automatically classify texts into promotion or prevention-focus have been reported yet. Also, there are no publicly available datasets annotated with regulatory focus categories.

In the course of our study, we create datasets containing event descriptions and social media data, in German, labeled with regulatory focus notions. We use correlation analysis to investigate the linguistic signatures of regulatory focus and ascertain the validity of our datasets. Further, we conduct text classification experiments to explore the possibility of automatically detecting regulatory focus concepts from text. Our experiments show that a BERT-based classifier outperforms lexicon-based approaches popularly used in psychology and the classifier can generalize from experimental data to Twitter data.

The main contributions of the paper are (1) an introduction to the task of regulatory focus classification, (2) experimental and real-world datasets annotated with RF categories, (3) a correlation analysis to

verify to what extent the findings from studies on regulatory focus as observable in the dataset using traditional methods and (4) performance comparison of RF classification using standard measures from psychology vs. state-of-the-art methods from NLP. Our research aims to serve as a starting point for exploring the concept from a computational linguistics perspective and enable future studies. The datasets created as part of this study are freely available for research purposes. They can be accessed together with the corresponding code at <https://www.ims.uni-stuttgart.de/data/author-regulatory-focus-detection>.

2 Background

2.1 Regulatory focus

The regulatory focus theory (RFT) posits that all goal-directed behavior of humans is regulated by two distinct motivational systems, *promotion* and *prevention* (Higgins, 1997, 1998). Promotion-focused individuals are motivated by their growth and development needs, try to attain their *ideal* selves by eagerly approaching a goal and are sensitive to positive outcomes. On the contrary, prevention-focused individuals are motivated by their security needs, try to attain their *ought* selves by vigilantly approaching a goal and are sensitive to negative outcomes (Brockner and Higgins, 2001). RFT has been employed in domains like organizational psychology (Crowe and Higgins, 1997; Lanaj et al., 2012), consumer psychology (Aaker and Lee, 2001; Higgins, 2002) and health communication (Keller, 2006; Kees et al., 2010) to explain phenomena like decision making (Crowe and Higgins, 1997; Higgins, 2002; Sassenberg et al., 2014), social relations (Righetti et al., 2011) and information processing (Aaker and Lee, 2001).

Regulatory focus varies interindividually and situationally. Hence, each individual has a chronic regulatory focus (similar to differences in personality factors). In addition, events can induce a situational regulatory focus. Darkness and strange noises will for instance induce a situational prevention focus. Researchers often employ priming experiments in which they vary (the recall of) events to situationally induce promotion or prevention focus in individuals (Higgins et al., 2001; Hamstra et al., 2014), which is also the main data collection method used in this study. Such approach for text corpus annotation and collection has been shown to work in the NLP context, for instance in emotion classification (Troiano et al., 2019, 2023).

Semin et al. (2005) investigated how an individual’s motivation affects the use of language and reported distinctive linguistic signatures of individuals corresponding to their goal attainment strategies or regulatory focus. They observed that promotion-focused individuals

¹<https://www.webofscience.com/wos/woscc/citation-report/aac080af-4516-427f-bf6f-ae9e89494de9-57fbd01c>

	Promotion	Prevention
Success	Positive activating emotions <i>enthusiasm, happiness, hope, pride, cheerfulness</i>	Positive non-activating emotions <i>contentment, relief, relaxation, calmness</i>
Failure	Negative non-activating emotions <i>disappointment, sadness, dejection, depression</i>	Negative activating emotions <i>anxiety, fear, anger, shame, hate</i>

Table 2: A mapping of emotions related to a regulatory focus category and outcome of a particular situation (success/failure) (drawn following Brockner and Higgins, 2001).

convey intentions and goals in abstract terms characterized by interpretive action verbs (e.g., hurts), state verbs (e.g., hate), and adjectives. On the contrary, individuals with a prevention focus tend to use more concrete terms like descriptive action verbs (e.g., walk, throw). Further in promotion focus, individuals tend to view their goals as hopes and aspirations, while in prevention focus, they tend to view their goals as duties and obligations (Higgins, 1997, 1998). Vaughn (2018) notes differences in language use when describing hopes (focus on positive outcomes) as compared to duties (focus on social relationships). Conley and Higgins (2018) used lexical analysis of essays as an RF measure.

In consumer psychology, the influence of regulatory focus orientation on the persuasiveness of messages has been investigated with reference to “message framing” or the linguistic presentation of information (Aaker and Lee, 2001; Cesario et al., 2013, i.a.). The persuasiveness of a message is enhanced when it is framed to fit the regulatory focus inclination of the recipient or reader (Higgins, 2000). In this study, we focus on the imprints of regulatory focus left behind by the author of a text.

Regulatory focus is a psychological variable that varies inter-individually like a personality trait and varies situationally like emotions, which makes it a manipulable attribute (Higgins et al., 2001; Hamstra et al., 2014). While personality and emotion have been investigated in author profiling studies (Stamatatos et al., 2015; Rangel and Rosso, 2016a, i.a.), regulatory focus has not received any attention there.

Authorship profiling, an application of text analysis relevant for this study, involves assessing the properties of the author like age, gender, personality, and emotion from their linguistic signatures in text (Goswami et al., 2009; Argamon et al., 2003; Nowson and Oberlander, 2006; Pennebaker et al., 2003; Gill et al., 2008; Rangel and Rosso, 2016b). While some of these properties are stable, such as gender and age, others, such as emotion, vary based on the author’s current situation or state of mind. Regulatory focus is a psychological variable that varies inter-individually like a personality trait and varies situationally like emotions (e.g., anxiety Gaudry et al. (1975)), which makes it a manipulable attribute (Higgins et al., 2001; Hamstra et al., 2014). While personality and emotion have been investigated in author

profiling studies (Stamatatos et al., 2015; Rangel and Rosso, 2016a, i.a.), regulatory focus has not received any attention there.

2.2 Emotionality and regulatory focus

The relationship between emotionality and regulatory focus has been explored in a few studies (Higgins et al., 1997; Brockner and Higgins, 2001). Emotions arise from an interaction between a person and a situation. While valence and arousal dimensions help to understand the experienced emotions, regulatory focus aids to explain why an emotion is experienced in a given situation. The nature and magnitude of an emotional reaction when a person succeeds or fails in their attempt to achieve a goal is influenced by their regulatory focus orientation. When a desired end-state (success) is achieved, promotion-focused individuals experience positive activating emotions like cheerfulness and happiness, while prevention-focused individuals experience positive non-activating emotions like relaxation and calm. Similarly, negative non-activating emotions like sadness are related to promotions focus, and negative activating emotions like anger, hate, and fear are linked to prevention focus when an undesired end-state (failure) is encountered. Table 2 shows an approximate mapping of different emotions with respect to regulatory focus and situational outcome (Brockner and Higgins, 2001).

In the current study, we collect data by manipulating situational regulatory focus and present the task of regulatory focus classification from the text. Also, the annotators wield the knowledge of the relationship between emotions and regulatory focus to facilitate better annotation of real-world Twitter data.

3 Data collection

We create three different datasets as part of this study – two containing self-reported event descriptions (EDD-1 and EDD-2) and one manually annotated Twitter dataset (TwD). The event description datasets are created by regulatory focus manipulation experiments and contain self-reported event descriptions provided by participants who were experimentally primed for one of the regu-

Focus	Instructions
Promotion	<p>... a situation in which you felt you made progress towards (<i>being successful in your life / achieving a goal that is important to you</i>).</p> <p>... a situation in which, compared to others, you felt like you were not making any progress towards (<i>being successful in your life / achieving something</i>).</p> <p>... a situation in which you wanted to attain something that was very important to you personally, and in which you were able to do as well as you ideally would like.</p>
Prevention	<p>... a situation in which being careful enough avoided from getting into trouble.</p> <p>... a situation in which lack of caution caused you to get into trouble.</p> <p>... a situation in which you behaved in a way that no one could have found fault with.</p>

Table 3: Instructions used to prime promotion or prevention focus. All instructions started with *Please describe..* The text in italics shows the minor difference in the formulation in the datasets given as (EDD-1/EDD-2).

latory focus conditions. The Twitter dataset contains manually annotated tweets. While the event description datasets, generated using well-established psychological experiments, contain high-quality annotations, they are not naturally produced text. The Twitter dataset, on the other hand, contains real-world data, nevertheless might not be on par with the experimental data in terms of the quality of annotations, given the risk of noise introduced by annotators. However, evaluating the efficiency of models, exposed only to experimental data, on real-world data helps to assess the extent to which such models can be employed in practical applications.

3.1 Self-reported event descriptions

We create two self-reported event description datasets (to which we refer as EDD-1 and EDD-2) using a standard experiment employed in psychology to manipulate regulatory focus (Higgins et al., 2001; Hamstra et al., 2014, i.a.). Participants are asked to recount an event from their past based on a given regulatory focus condition. The instructions are formulated in a way to prime participants to temporarily re-experience a situation in which they held a promotion or a prevention focus. For each condition, three different situations are presented where they succeeded or failed. For example, in the promotion success condition, they are asked, *“Please describe an experience in which you felt you were making progress toward being successful in your life”*, which points to a situation, where they were eagerly seeking a positive end state (being successful) and managed to

achieve it. To induce a prevention focus they are for instance asked, *“Please describe an experience where a lack of caution caused you to get into trouble”*. In this situation, they are primed to recount an event in which they did not exercise caution which resulted in a negative outcome. See Table 3 for an overview of instructions (Appendix A&B shows complete instructions in German).

EDD-1 and EDD-2 are in German and differ only on a few accounts. EDD-1 is a consolidation of data from seven independent studies, both published and unpublished (Sassenrath et al., 2014; Hamstra et al., 2014; Sassenberg et al., 2015). The original purpose of these studies was not to collect data for NLP modeling; instead, for psychology research that required the manipulation of regulatory focus.² The majority of the participants were university students and the number of participants per study ranged between 28 to 172. Every participant contributed to three questions related to one of the regulatory focus conditions. The questions were presented in an open-ended format, so the length and quality of the texts vary substantially. The length of responses ranges from 4 to 308 tokens³ with a mean response length of 38.3 (*median* = 33).

The data collection experiment for the second event description dataset (EDD-2), following the same procedure as EDD-1, is conducted on a crowd-sourcing platform (Clickworker⁴), with the participation restriction as *“working at least 50% of a full employment”* to include a broader demographic. The questionnaire is formulated in terms of goal achievement in a work context. Similar to the previous experiment, we collect three responses per participant corresponding to either promotion or prevention. Participants are mandated to write a minimum of 150 characters for each open-ended question. The questionnaire was completed by 455 participants. The length of responses ranges from 5 to 748 tokens³ with a mean response length of 51.3 (*median* = 41). Table 4 shows examples from EDD-1.

3.2 Annotated tweets

When an individual actively participates in a social media activity, such as posting on Twitter, the action is driven by underlying motivational factors. We build upon this assumption to create a Twitter data dataset (TwD), a social media dataset to investigate the real-world occurrences of the regulatory focus concepts. In order to eliminate noisy content from Twitter and prioritize instances that are more likely to reflect motivational inclinations, we gather a subset of tweets that convey

²We received the data through personal communication and agreed with the original authors to make the data available upon acceptance of this paper.

³We use https://www.nltk.org/api/nltk.tokenize.html#nltk.tokenize.word_tokenize

⁴<https://www.clickworker.com/>

Dataset	RF	Example (German)	Translation (English)
EDD-1	Prom.	(1) Fast immer dann, wenn Durchhaltevermögen über einen längeren Zeitraum gefragt war. (2) Als ich zwei Monate lang nichts tat außer saufen und chillen. (3) Ich konnte am Wochenende zum See fahren, weil ich nicht ganz so viele Klausuren haben wie andere.	(1) Almost always when perseverance was required over a longer period of time. (2) When I didn't do anything for two months except drink and chill. (3) I could go to the lake on weekends, because I did not have quite as many exams as others.
	Prev.	(1) Beim Skifahren habe ich nicht genügend aufgepasst und bin beinahe in einen Baum gefahren. (2) Wir kletterten verbotenerweise als Jugendliche auf ein Dach einer Hütte und wurden erwischt. (3) Zu viel Alkohol auf einer Party hat dazu geführt, dass ich leichtsinnig mein Handy verloren habe.	(1) I didn't pay enough attention when skiing and almost crashed into a tree. (2) As adolescents, we illegally climbed onto the roof of a hut and were caught. (3) Too much alcohol at a party made me recklessly lose my phone.
TwD	Prom.	Wir sind so stolz und erleichtert – unsere Präsentation in Offenburg war ein Erfolg! Danke an alle für die Unterstützung!	We are so proud and relieved – our presentation in Offenburg was a success! Thanks everyone for the support!
	Prev.	Ich hasse es, dass ich nichts aus meinem Leben mache und nur vergammel. Und meine Ernährung ist auch grauenhaft.	I hate that I don't do anything with my life and just rot. And my diet is terrible too.

Table 4: Examples from EDD-1 and TwD with their translations to English.

subjective emotional experiences. We ensure this by selecting tweets that starts with a first-person pronoun (“Ich” or “Wir”) and at least one emotion word (See Appendix C.1 for the list of emotion words). The messages to be annotated are sub-sampled from the period 2016 to 2019. Further, they are required to contain less than 20 % hashtag tokens, no images, no URLs, and not the word “corona”. We sampled 1,500 instances to be annotated.

Annotating tweets with regulatory focus categories is quite challenging for non-expert annotators. Preparatory to the actual annotation, training sessions are conducted to make sure the concepts are understood well. We update the annotation guidelines accordingly (see Appendix C) to support the quality of the annotation. We instructed the annotators to label each instance with one of the four labels – (1) *prevention*, (2) *promotion*, (3) *neither promotion nor prevention* or (4) *not sure*. The instances labeled as *neither promotion nor prevention* or *not sure* were disregarded to retain only the most confident instances.⁵ From the 1,500 annotated tweets, we retained instances in which both annotators agreed on the two labels *promotion* (Cohen’s $\kappa=.42$) or *prevention* ($\kappa=.39$), which amounts to 923 (61.5%) tweet instances. Table 4 shows some examples.

To understand the characteristics and differences between the datasets we look into some descriptive statistics on the distribution of labels and tokens in the collected datasets as shown in Table 5. The label distribution is relatively balanced in the event description

⁵In Appendix F we include a discussion on the occurrence of *neutral* instances in real-world data and address regulatory focus classification as a tertiary classification task.

Dataset	Label stats			Token stats		
	Prom.	Prev.	Tot.	Max	Min	Median
EDD-1	776	799	1575	309	4	33.0
EDD-2	678	582	1260	746	5	41.0
TwD	655	268	923	61	11	22.0

Table 5: Descriptive statistics of labels and tokens distribution in the collected datasets.

datasets which can be attributed to the collection procedure. However, in the Twitter dataset, the imbalance is prominent as the data is representative of real-world occurrences wherein out of the filtered 923 annotated tweets around 70% are labeled as promotion. Regarding the text length, we note that the TwD dataset containing only tweets maintains a minimum word count of 11 words, while the event description datasets contain very long as well as very short texts. So, in the real-world scenario that we are considering in this study, the texts are relatively short and prevention scenarios are scarce compared to promotion.

4 Linguistic correlation analysis

As discussed in Section 2, previous research has reported that authors’ regulatory focus orientation leaves distinctive markers in their language use. Semin et al. (2005) studied abstractness or concreteness of words used, while Brockner and Higgins (2001) looked into expressed emotionality and Vaughn (2018) investigated

the differences in the description of hopes vs. duties. Conley and Higgins (2018) used lexical analysis of essays as a measure to regulatory focus.

To investigate these linguistic features, they use the psychological categories defined in dictionary-based methods like *Linguistic Inquiry and Word Count* (LIWC, Pennebaker et al., 2015). Our analysis aims at confirming that these findings on the linguistic intricacies of regulatory focus are observable in our datasets as well. This serves on one side as a replication study of previous work and on the other side as a preliminary study for developing automatic RFT classifiers based on these lexical resources. We consider the datasets EDD-1, EDD-2, and TwD for this analysis.

4.1 Method

A commonly used dictionary-based method employed to analyze text samples automatically is to count words corresponding to psychologically relevant categories, which is also referred to as the word-count approach (Stone and Hunt, 1963; Gottschalk and Gleser, 1979; Berry et al., 1997, i.a.). We use this closed-vocabulary approach to understand the relationship between a set of predefined psychological categories and regulatory focus. To our knowledge, there are no publicly available dictionaries that encapsulate different psychological concepts, let alone in German. For this reason, we resort to two commercially available text analysis systems with support for German, namely LIWC (Pennebaker et al., 2015) and 100W⁶ (Spitzer, 2019).

LIWC is one of the most popularly used tools in psychology for automated text analysis. It relies on exact matches with words, word stems, and selected emoticons. 100W employs various NLP disambiguation techniques on top of the lexicons. For instance, it disambiguates word senses named entity recognition and word embeddings to count only specific senses of a word. Both tools do not provide access to their raw dictionaries but return the relative frequency of terms in each category per text.

We use the measure of *point-biserial correlation* (Glass and Hopkins, 1996) to explain the correlation between the regulatory focus label of instances (a discrete value) and the relative frequency of any given psychological variable (a continuous value). If n is the total number of instances in the dataset, then the point-biserial correlation ρ_{pb} is calculated as

$$\rho_{pb} = \frac{\mu_{prev} - \mu_{prom}}{\sigma_n} \sqrt{\frac{n_{prev}n_{prom}}{n(n-1)}}, \quad (1)$$

where μ_{prom} and μ_{prev} are the mean values of the continuous variable for promotion and prevention labeled instances respectively, σ_n the standard deviation of the

continuous variable, and n_{prom} and n_{prev} the frequencies of the promotion and prevention labels, respectively, within the dataset. The point-biserial correlation coefficient ranges from -1 to $+1$ indicating perfect negative and perfect positive correlation, respectively. A high positive correlation coefficient suggests that the relative frequency of the psychological variable tends to be higher when the instance label is prevention. Conversely, a high negative correlation coefficient indicates that the relative frequency of the variable is higher when the label is promotion. The magnitude and sign of the correlation coefficient provide insights into the strength and direction of the relationship between the regulatory focus label and the psychological variable.

To account for the potential issue of multiple comparisons and control the family-wise error rate, we apply Bonferroni correction (Bonferroni, 1936), a method to adjust the significance level when conducting multiple statistical tests simultaneously. It divides the desired overall significance level (α) by the number of comparisons to derive the adjusted significance level for each individual test. In our study, since we perform multiple point-biserial correlation tests between the regulatory focus label and various psychological variables, we divide the α level by the number of correlations to obtain the adjusted α level.

4.2 Experimental setup

We use the German version of the LIWC 2015 dictionary (DE-LIWC2015) and the 100W API to analyze all instances in our datasets, to obtain the relative frequencies corresponding to each of the included psychological categories. For the analysis, we take into account 80 LIWC categories and all 49 categories from 100W⁷. To calculate the point-biserial correlation, we use the implementation from `scipy`.⁸ For Bonferroni correction, we set the desired overall significance level (α) to 0.05 and the adjusted significance level is calculated by dividing it by the number of psychological categories in each lexicon. We also calculate 95% confidence intervals for each point biserial correlation coefficient, considering the adjusted alpha level and the degrees of freedom.⁹

4.3 Results

We look into those categories which show a high correlation with promotion and prevention focus labels and compare the observed trends of different psychological categories in our datasets with previously reported relationships between these concepts and the regulatory focus orientation of the author. We report the point

⁷See Appendix D for the complete list of categories.

⁸<https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.pointbiserialr.html>

⁹We use the percent point function (`t.ppf`) from `scipy.stats`

⁶<https://www.100worte.de/en/science>

categories	LIWC						100W					
	EDD-1		EDD-2		TwD		EDD-1		EDD-2		TwD	
	corr.	CI	corr.	CI	corr.	CI	corr.	CI	corr.	CI	corr.	CI
achievement	-.38*	[-.47, -.29]	-.36*	[-.46, -.26]	-.11	[-.24, .03]	-.18*	[-.28, -.08]	-.30*	[-.41, -.19]	-.04	[-.17, .09]
adjective	-.19*	[-.29, -.08]	-.21*	[-.32, -.10]	-.12	[-.26, .01]	-.11*	[-.22, -.01]	-.09	[-.21, .02]	.04	[-.09, .18]
affect	-.17*	[-.27, -.07]	-.17*	[-.28, -.06]	-.08	[-.21, .05]	—	—	—	—	—	—
affiliation	.12*	[.02, .22]	.02	[-.10, .13]	-.03	[-.16, .11]	.05	[-.06, .15]	-.07	[-.19, .04]	.03	[-.10, .16]
anger	.10	[-.01, .20]	.03	[-.08, .15]	.62*	[.54, .70]	.15*	[.04, .25]	.08	[-.04, .19]	.60*	[.51, .68]
anxiety	.08	[-.03, .18]	.06	[-.05, .18]	.12	[-.01, .25]	.20*	[.10, .30]	.20*	[.09, .32]	.10	[-.04, .23]
article	-.11*	[-.21, -.00]	.07	[-.05, .19]	-.04	[-.18, .09]	-.11*	[-.21, -.00]	.05	[-.06, .17]	-.01	[-.14, .13]
auxverb	.14*	[.04, .25]	.10	[-.02, .21]	-.22*	[-.34, -.09]	.08	[-.02, .19]	.04	[-.08, .15]	-.23*	[-.35, -.10]
clout	.11*	[.00, .21]	.11	[.00, .23]	-.17*	[-.30, -.04]	—	—	—	—	—	—
compare	-.16*	[-.26, -.06]	-.18*	[-.30, -.07]	.04	[-.09, .18]	—	—	—	—	—	—
differ	-.04	[-.15, .06]	-.13*	[-.24, -.02]	.17*	[.04, .30]	—	—	—	—	—	—
discrep	.06	[-.05, .16]	.02	[-.09, .14]	.10	[-.03, .23]	.18*	[.08, .28]	.14*	[.02, .25]	.00	[-.13, .14]
drives	-.14*	[-.24, -.04]	-.22*	[-.33, -.10]	-.11	[-.24, .02]	—	—	—	—	—	—
feel	-.13*	[-.24, -.03]	-.19*	[-.30, -.08]	.01	[-.13, .14]	—	—	—	—	—	—
feminine	.11*	[.00, .21]	.01	[-.11, .12]	.07	[-.06, .20]	.16*	[.06, .26]	.08	[-.04, .20]	-.10	[-.23, .03]
focuspresent	.11*	[.01, .21]	-.04	[-.15, .08]	-.22*	[-.34, -.09]	—	—	—	—	—	—
i	-.07	[-.18, .03]	-.07	[-.18, .05]	.14*	[.01, .27]	-.07	[-.18, .03]	-.06	[-.18, .05]	.17*	[.04, .30]
insight	-.15*	[-.25, -.05]	-.14*	[-.26, -.03]	-.03	[-.16, .10]	—	—	—	—	—	—
negemo	.09	[-.01, .20]	.12*	[.00, .23]	.36*	[.25, .48]	.19*	[.09, .29]	.11	[-.01, .22]	.35*	[.23, .47]
negpower	—	—	—	—	—	—	.16*	[.05, .26]	.18*	[.07, .29]	.08	[-.05, .22]
posachieve	—	—	—	—	—	—	-.27*	[-.36, -.17]	-.33*	[-.43, -.22]	-.02	[-.16, .11]
posemo	-.26*	[-.36, -.16]	-.26*	[-.37, -.16]	-.39*	[-.51, -.28]	-.01	[-.12, .10]	-.09	[-.20, .03]	-.26*	[-.38, -.13]
reward	-.30*	[-.40, -.21]	-.37*	[-.47, -.27]	-.11	[-.24, .02]	-.26*	[-.36, -.16]	-.38*	[-.48, -.28]	-.07	[-.20, .06]
risk	.25*	[.15, .35]	.27*	[.16, .38]	.01	[-.13, .14]	.25*	[.15, .35]	.29*	[.18, .40]	.08	[-.06, .21]
sadness	.01	[-.10, .11]	.01	[-.11, .12]	-.23*	[-.36, -.11]	.01	[-.10, .11]	-.02	[-.14, .10]	-.23*	[-.36, -.10]
social	.14*	[.04, .25]	.06	[-.06, .18]	.03	[-.10, .17]	—	—	—	—	—	—
sv	—	—	—	—	—	—	-.04	[-.15, .06]	-.06	[-.18, .06]	.39*	[.27, .50]
tone	-.27*	[-.36, -.17]	-.26*	[-.37, -.15]	-.47*	[-.57, -.37]	—	—	—	—	—	—
we	.16*	[.06, .26]	.06	[-.06, .18]	-.05	[-.18, .08]	.16*	[.06, .26]	.06	[-.06, .18]	-.04	[-.17, .10]
work	-.40*	[-.49, -.31]	-.29*	[-.39, -.18]	.00	[-.13, .14]	—	—	—	—	—	—

Table 6: Point-biserial correlation between regulatory focus labels (prevention–promotion) and relative frequencies for the categories in LIWC and 100W discussed in Section 4.3. n takes the value of 1575, 1260 and 923 for EDD-1, EDD-2 and TwD respectively. The correlations considered significant (p -value < 0.05) are marked with a * symbol.

biserial correlation coefficient, and the lower bound and upper bound of the 95% confidence interval, of the categories for which the correlation was statistically significant after Bonferroni correction. Table 6 shows the point-biserial correlation between the regulatory focus labels and the relative frequencies of categories in LIWC and 100W, mentioned in the following discussion. In order to ensure meaningful and reliable conclusions, we exclude categories that appear in only one of the lexicons and exhibit statistically significant correlations in only one of the datasets. This decision was based on the understanding that drawing substantial conclusions from such observations would be challenging and could potentially lead to unreliable findings.

Risk and reward: LIWC and 100W approximate the prevention and promotion concepts with their categories *risk* and *reward* respectively (Meier et al., 2019; Spitzer, 2019). In the event description datasets, the *risk* category of LIWC significantly correlated with prevention (.25

for EDD-1, .27 for EDD-2) and *reward* category with promotion (.30 for EDD-1, .37 for EDD-2) categories. For 100W, the values are slightly lower yet significant for *risk* (.25 for EDD-1, .29 for EDD-2) and *reward* (.26 for EDD-1, .38 for EDD-2). In the Twitter dataset, they show only a very weak correlation to the same categories, but also statistically significant. We conclude that the *risk* and *reward* categories represent an approximation of the regulatory focus concepts in EDD-1/2.

Emotionality: In promotion focus, individuals experience positive-activating emotions like cheerfulness and happiness on successfully achieving the desired goal. While in prevention focus they experience positive non-activating emotions like relaxation and relief. Similarly, on failing to attain the desired goal, in promotion focus, people experience negative non-activating emotions like sadness. At the same time, in prevention focus they experience negative activating emotions like anger and hate (Higgins, 1997; Brockner and Higgins, 2001).

LIWC and 100W represent affective states in the categories *positive emotion*, *negative emotion*, *tone*, *anxiety*, *anger*, and *sadness*. They do not include categories corresponding to all different magnitudes of emotional activation (e.g., calmness, fear, hope), which proves to be a drawback of these lexicon-based methods in capturing the characteristics of regulatory focus.

We make following observation to be aligned with previous studies. In both 100W and LIWC, *anger*, a negative activating emotion, correlates with prevention focus (.6 and .62 resp.) and *sadness*, a negative non-activating emotion, correlates with promotion focus (.23) in the Twitter data. In 100W, the *anxiety* category is positively correlated to prevention in the event description datasets (.2 for EDD-1/2). The *tone* category in LIWC, representing overall the positive tone of a text, highly correlates with promotion focus in all datasets (.27 for EDD-1, .26 for EDD-2 and .47 for TwD). The Twitter dataset reflects findings on emotionality more reliably than the event description datasets.

Abstractness vs. concreteness: Semin et al. (2005) argued that markers of abstractness and concreteness in language are associated with the promotion and prevention focus, respectively. They attributed state verbs (e.g., *love*, *hate*), interpretive action verbs (i.e., *hurt*, *console*) and adjectives to abstractness and descriptive action verbs (e.g., *walk*, *throw*) to the concreteness of language. The category *adjective* in both LIWC and 100W shows a significant correlation to promotion focus in event description datasets (.19 for EDD-1, .2 for EDD-2, and .11 for EDD-1), reinforcing the claim made in Semin et al. (2005). The mentioned verb classes are, however, not included as psychological categories in LIWC. In 100W only *descriptive action verbs* and *state verbs* are defined, but they do not show any consistent pattern across datasets. We conclude that not all aspects of language abstraction are represented.

Hopes and duties: Goals are viewed as duties and obligations in prevention focus, and as hopes and aspirations in promotion focus. Vaughn (2018) observed that people talk more about positive outcomes when describing hopes which are reflected in the categories *positive emotion*, *reward*, and *achievement*. While describing duties the focus is on maintaining social relationships which is represented in the categories *social processes* and *affiliation*.

Corroborating with these findings, a significant correlation with the promotion label is observed for the LIWC categories *positive emotion* (.26 for EDD-1, .26 for EDD-2), *reward* (.30 for EDD-1, .37 for EDD-2) and *achievement* (.38 for EDD-1, .36 for EDD-2) for event description datasets. For Twitter data and 100W lexicon, significant correlation patterns are not observed. The

social processes and *affiliation* categories do not show any consistent pattern across lexicons or datasets.

We construe that some, but not all linguistic markers from studies on regulatory focus are discernible in our datasets. Existing dictionary-based methods have the drawback that they capture emotionality only in terms of a few psychological categories (e.g., anger, sadness, anxiety) and do not include activating and non-activating emotions discussed in Section 2.2. Additionally, there are significantly correlated categories not being investigated in previous studies (*drives*, *feel*).

5 Classification experiments

The linguistic correlation analysis sheds some light on the strengths and limitations of traditional automated text analyses. We go one step further to assess how well we can automatically predict the regulatory focus of the author from the text. To this end, we explore open and closed vocabulary text classification methods.

5.1 Methods

5.1.1 Closed vocabulary approach

We use the LIWC and the 100W analyses used earlier as candidates for the closed vocabulary approach. We consider all psychological categories defined in both tools and as noted earlier, these tools do not provide access to the raw dictionaries, instead, return the relative frequency of terms in each category per text. We use these relative frequency values and reweight them with logistic regression on the training data.

5.1.2 Open vocabulary approach

We use two machine-learning-based approaches. The first is a tf-idf-bag-of-words logistic regression classifier with unigrams and bigrams. The second is BERT-based (Devlin et al., 2019), a bidirectional transformer-based language model pre-trained with masked token prediction and next-sentence prediction objectives. We use the `deepset/gbert-large`¹⁰ (Chan et al., 2020) model which is trained on a large dataset sourced from Common Crawl, German Wikipedia, legal data, movie subtitles, parliament speeches, and books. We fine-tune BERT on our regulatory focus-annotated data for a sequence classification task.

5.2 Experimental setup

We conduct our classification experiments on the three regulatory focus labeled datasets. We conduct 10-fold cross-validation on the event description datasets EDD-1, EDD-2, and EDD-1+2, and identify the best event

¹⁰<https://huggingface.co/deepset/gbert-large>

description dataset suited for the task. We then evaluate all our models trained on the best dataset on the out-of-domain TwD dataset.

For the LIWC and 100W experiments, we weight their output with logistic regression models from *sklearn*¹¹ with default parameters. The details of each model are as follows:

LR_LIWC: We use the German version of the LIWC 2015 dictionary (DE-LIWC2015). The output file from the software contains relative frequencies of words from all psychological categories per document. We use these relative frequency values as features.

LR_100W: The API from 100W accepts a text document to be analyzed and returns a JSON response containing relative frequencies of all psychological categories. We use these relative frequency values as features.

LR_TFIDF: We use the *TfidfVectorizer* from *sklearn*¹² to vectorize the documents and use NLTK¹³ to remove the German stopwords.

GBERT: We use the pre-trained German BERT model *deepset/GBERT-large* with 24-layer, 1024-hidden, 16-heads and 335M parameters. For fine-tuning, we use the *BertForSequenceClassification*¹⁴ implementation from Hugging Face (Wolf et al., 2020). During fine-tuning, we set the number of epochs to 8, the learning rate to 10^{-5} , and the batch size to 16. Additionally, to prevent over-fitting, we monitor the validation loss and stop training if it does not improve for 5 steps. For other hyper-parameters, we used the default values from the implementation.¹⁵

We split the datasets into training, validation, and test sets by allocating 80% for training, 10% for validation, and 10% for testing. We perform 10-fold cross-validation and in each fold, we assess the performance of the model trained on 80% of the data, on two separate test datasets: the 10% reserved as the test set and the annotated Twitter data. The validation split of data was utilized only for GBERT fine-tuning.

5.3 Results

Table 7 shows the performance of comparison between all models and a random baseline, of 10-fold cross-validation on EDD-1, EDD-2, and EDD-1+2 datasets. Table 7 also presents the results of the models trained on the best dataset evaluated in the TwD dataset.

We see that the performance of both closed vocabulary-based approaches LR_100W and LR_LIWC are almost similar on all of the event description dataset

settings. The LR_TFIDF model on the other hand outperforms closed-vocabulary models by a good margin on all event description datasets (average accuracy of 0.86 on EDD-1, 0.81 on EDD-2 and 0.87 on EDD-1+2). Also, as hypothesized, the GBERT model outperforms all other models in the majority of the experiments, with an average accuracy above 0.91 when trained on any of the event description datasets.

Evaluations on the event description datasets show that models perform better when trained on a combination of both event description datasets, possibly because it adds more diversity to the topics and helps the models to learn better generalizations. So we conclude EDD-1+2 to be the best dataset and use it as training data to test generalizations on out-of-domain datasets.

For LR_100W and LR_LIWC, despite their reasonably good performance on event description datasets, the accuracy on the out-of-domain TwD dataset is in most cases only slightly better than the random baseline and sometimes worse. Additionally, the number of instances labeled as prevention (268/923) is quite low compared to the promotion label. The closed-vocabulary approaches have a high recall compared to other methods, with 100W giving the best recall for prevention. Overall, LR_LIWC performs better than LR_100W on the out-of-domain dataset.

The LR_TFIDF model outperforms LR_100W and LR_LIWC on TwD datasets with an accuracy of 0.58. This observation supports the argument that there are possibly more lexical features that capture the regulatory focus concepts, however, cannot essentially be represented as dictionaries of words.

Finally, the GBERT model outperforms all other models when trained on the EDD-1+2 dataset.¹⁶

5.4 Error analysis

We conduct an error analysis in order to get a comprehensive understanding of the best model's (GBERT+EDD-1+2) behaviour and its generalization capability to real-world Twitter instances. The analysis involved comparing the representations generated by the pre-trained language model before and after fine-tuning for the regulatory focus classification task. Additionally, we manually examine instances misclassified by the model to identify common error patterns.

Representation comparison: To understand how fine-tuning for the task affects representations generated by the model, we employ a t-SNE visualization which reduces the high-dimensional representations

¹¹https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

¹²https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

¹³[nltk.corpus.stopwords.words\('german'\)](https://www.nltk.org/stopwords.html)

¹⁴https://huggingface.co/transformers/model_doc/bert.html

¹⁵See Appendix E for more information on fine-tuning.

¹⁶To ensure a fair comparison, we conducted experiments using non-linear models, such as SVM, Random Forest, and Gradient Boosting. However, logistic regression was observed to produce more stable results across LIWC, 100W, and tf-idf features and across datasets. See Appendix F for comparison of these results.

Cross validated on corresponding dataset								
Train Dataset	Model	Promotion			Prevention			Acc
		P	R	F ₁	P	R	F ₁	
EDD-1	random	.49 ± .06	.50 ± .06	.49 ± .05	.51 ± .07	.50 ± .06	.50 ± .06	.50 ± .04
	LR_LIWC	.78 ± .03	.76 ± .05	.77 ± .03	.77 ± .05	.79 ± .05	.78 ± .04	.77 ± .02
	LR_100W	.77 ± .06	.76 ± .04	.76 ± .03	.77 ± .05	.77 ± .06	.77 ± .03	.77 ± .03
	LR_TFIDF	.89 ± .02	.83 ± .04	.85 ± .02	.84 ± .05	.89 ± .02	.86 ± .03	.86 ± .02
	GBERT	.91 ± .05	.93 ± .03	.92 ± .03	.94 ± .03	.91 ± .05	.92 ± .03	.92 ± .03
EDD-2	random	.53 ± .06	.48 ± .06	.50 ± .05	.45 ± .07	.49 ± .07	.47 ± .06	.49 ± .05
	LR_LIWC	.77 ± .04	.77 ± .05	.77 ± .04	.73 ± .07	.73 ± .05	.73 ± .05	.75 ± .04
	LR_100W	.78 ± .03	.77 ± .06	.77 ± .03	.73 ± .07	.74 ± .06	.74 ± .05	.76 ± .04
	LR_TFIDF	.77 ± .05	.94 ± .03	.84 ± .03	.90 ± .06	.67 ± .07	.77 ± .06	.81 ± .04
	GBERT	.90 ± .04	.94 ± .05	.92 ± .03	.93 ± .05	.87 ± .05	.90 ± .02	.91 ± .02
EDD-1+2	random	.52 ± .04	.50 ± .05	.51 ± .04	.50 ± .04	.52 ± .03	.51 ± .03	.51 ± .03
	LR_LIWC	.78 ± .02	.77 ± .03	.77 ± .02	.76 ± .03	.78 ± .02	.77 ± .02	.77 ± .01
	LR_100W	.78 ± .04	.77 ± .03	.77 ± .04	.76 ± .03	.77 ± .04	.76 ± .03	.77 ± .03
	LR_TFIDF	.86 ± .02	.88 ± .03	.87 ± .02	.88 ± .03	.85 ± .02	.86 ± .02	.87 ± .02
	GBERT	.94 ± .02	.91 ± .04	.93 ± .02	.92 ± .04	.94 ± .03	.93 ± .01	.93 ± .02
EDD-1+2	Best Model (trained on EDD-1+2) tested on TwD							
	random	.71 ± .02	.50 ± .02	.58 ± .02	.29 ± .02	.50 ± .03	.37 ± .02	.50 ± .02
	LR_LIWC	.79 ± .01	.46 ± .04	.58 ± .03	.34 ± .01	.70 ± .03	.46 ± .01	.53 ± .02
	LR_100W	.76 ± .03	.33 ± .03	.46 ± .03	.31 ± .01	.75 ± .04	.44 ± .02	.45 ± .01
	LR_TFIDF	.77 ± .02	.57 ± .05	.66 ± .04	.37 ± .03	.59 ± .02	.45 ± .02	.58 ± .04
	GBERT	.82 ± .04	.61 ± .10	.70 ± .05	.41 ± .03	.66 ± .14	.50 ± .05	.63 ± .04

Table 7: Cross-validation results (summarized as mean ± standard deviation) for all models trained on different event description datasets

Example (German)	Translation (English)	Gold Label
1. Ich hasse mein Leben langsam, ich hab einfach kein Glück... Ich finde keine Arbeit und werde deswegen ange- meckert	1. I'm starting to hate my life, I just don't have any luck.... I can't find a job and I get bitched at for it	prevention
2. Ich hasse den "Sommer" ... Ich kann da nie einschlafen, weil es zu warm ist ._.	2. I hate the "summer" ... I can never fall asleep there because it's too warm ._.	prevention
3. Ich habe Angst.Angst dich zu verlieren oder Angst wie ich damit klar kommen werde wenn du nicht mehr da bist.	3. I am afraid of losing you or afraid of how I will cope when you are gone.	prevention
4. Ich hoffe nur Sie lesen nicht allzu viele von den Kom- mentaren hier unter Ihrem Beitrag! So viel Hass und Hetze würde ich selbst mit Ihrem Gehalt nicht lange durch- stehen! Bleiben Sie stark für eine tolerante, weltoffene Gesellschaft.	4. I just hope you don't read too many of the comments here under your post! I wouldn't last long with that much hate and agitation even on your salary! Stay strong for a tolerant, open-minded society.	promotion
5. Ich bin so froh das Chingy nichts passiert ist. Ich wäre wortwörtlich fast vor Sorge gestorben.Zum Glück ist es nochmal "gut" ausgegangen..	5. I am so glad that nothing happened to Chingy. I would have literally almost died of worry.fortunately it is once again "well" ended.	promotion
6. Die leute waren traurig und wütend.Ich bin froh dass sie friedlich geblieben sind nach diesem Tag.	6. People were sad and angry. I'm glad they stayed peaceful after that day.	promotion

Table 8: Instances from TwD dataset misclassified by the GBERT+EDD-1+2 model.

into a two-dimensional space (van der Maaten and Hinton, 2008). Figure 1 shows this visualization on the test splits generated using the deepset/gbert-large model, before and after fine-tuning.

We see distinct clusters after fine-tuning. However, in the TwD data it lacks clear separability compared to the event description dataset. This questions the extent of the model’s ability to generalize to real-world instances and emphasizes the need to investigate and understand the types of errors made by the model.

Common error patterns: We extend the error analysis by manually going through misclassified instances to understand the pattern and characteristics of the model’s most frequent errors. We take into account the tweets that have been classified incorrectly in every fold in the 10-fold cross-validation setting. Table 8 shows examples corresponding to the two main types of errors discussed in this section.

We observe that the emotion *hate* is completely absent in the event description dataset, despite being one of the most frequently occurring emotions in the Twitter data, accounting for about 25% of the instances in the annotated data. The emotion *hate* is a negative activating emotion often associated with a prevention motive. Interestingly, when examining misclassified instances related to prevention, we find that 87% of them contain the emotion word *hate* (Examples 1, 2). However, due to the absence of this emotion word in the training data, the model was unable to capture this particular nuance accurately. Other false negatives in the promotion class also point to the fact that the model fails to capture the relationship between emotions and regulatory focus category accurately (Example 3).

In misclassified instances of promotion, a common error arises when the text mentions a negative event and is followed by an expression of optimism or anticipation for something positive (Examples 4, 5, 6). This occurrence refers to the emotion of *hope*, which is associated with promotion. Many incorrectly classified promotion tweets exhibit this pattern. These instances express elements of both promotion and prevention, hence the model encounters challenges in accurately classifying them.

6 Conclusion

In this study, we bring attention to regulatory focus, a construct used in psychology to explain the goal-oriented behavior of humans. To encourage NLP research into this topic, we introduce the novel task of regulatory focus classification (*promotion vs. prevention*) and datasets of experimental and real-world data annotated with the concept. Our correlation analysis with

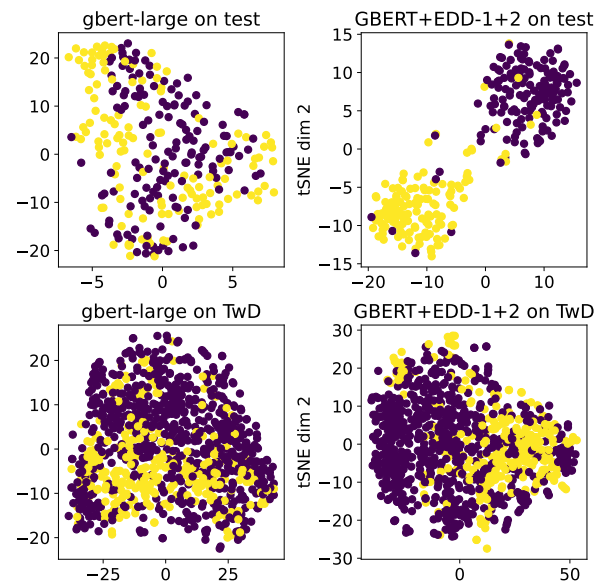


Figure 1: t-SNE visualization of representations generated by the pre-trained deepset/gbert-large model before (left) and after fine-tuning (right) on event-description data for regulatory focus classification.

lexicons uncovers corroborating evidence from previous research and also highlights some limitations of dictionary-based approaches. Further, we apply automatic text classification methods for regulatory focus detection. The results show that a language-model-based classifier outperforms models which rely on lexical-level features. Our best model identifies the regulatory focus inclination of a person from text with high accuracy and can be considered a strong baseline for future research. Further, by evaluating the best model on manually annotated Twitter data, we confirm the generalization capability of the model on unseen domains. We achieve good results by disregarding the preconceived relationship between an *a priori* list of words and psychological categories. Instead, relying only on the language model’s capability to learn them shows the best performance on the task. Nevertheless, a model that can combine these two aspects would be worth investigating further.

We also acknowledge that tweets might be too short or sometimes too vague in terms of context for the model to make a reliable prediction. As RFT is a concept in between stable traits and variable states, consolidating multiple texts from the same author could be one possible way to produce a more accurate prediction on the author’s regulatory focus.

Regulatory focus detection can find practical applications in general computer-mediated communication and human-computer interaction, where automatically identifying the needs, motivations, traits, etc., of the collocator, ensures more efficient communication. For

instance, a message that addresses the needs and motivation of the collocutor could be more persuasive or be received more positively. In future research, we would like to investigate paraphrasing of a given text to fit the regulatory focus of the counterpart and to what degree it influences the persuasiveness of a text.

7 Limitations

In this study we consider regulatory focus as a binary classification problem as supported by the framework of regulatory focus theory. While it was deemed appropriate for the current study, it may not be adequate for the real-world applications like Twitter. This is because there could be *neutral* instances which do not reflect the motivational orientation of the author owing to the limited context. Even though we heuristically subset tweets expressing emotional experience, by reducing it to a binary classification task, our classifier could potentially be misrepresenting the regulatory focus landscape in real-world scenarios. Additionally, a truly neutral motivational orientation is not well supported in the current theoretical framework.

In order to ensure practical applicability, future work could explore establishing predetermined conditions or criteria for selecting potential texts that can be used to identify the regulatory focus of authors. By defining specific guidelines or requirements for text inclusion, a focused analysis can be conducted on the relevant texts that provide valuable insights into individuals' regulatory focus orientations.

8 Ethical considerations

The regulatory focus manipulation experiment collects personal experiences from participants which can be classified as sensitive data. However, the study was conducted online and we do not store any personally identifiable information of the participants, to ensure that the original author cannot be traced back from the data. Before the start of the regulatory focus manipulation experiment, informed consent was read and explicitly acknowledged by the participants. Instructions to the participants detailed the purpose and procedure of the study, the remuneration, and data handling (see Appendix B for full instructions). Participation in the study was voluntary and participants were compensated as agreed, after completing the task. They were also informed that they could quit the experiment at any point or revoke the consent before submission.

We acknowledge that a system which can predict the regulatory focus accurately can not only be used to promote positive behavior change in areas like health care. It can also raise serious ethical concerns. Auto-

mated assessment of psychological constructs from text can potentially be employed to profile people based on their regulatory focus orientation, manipulate or persuade them in targeted marketing, political campaigns, or other persuasive endeavors. Also, employing inaccurate systems in downstream applications may result in unintended consequences as the system can make incorrect assessment about the behavioral inclinations of the person.

If automatic detection of regulatory focus is implemented in any application, the end-users should be explicitly notified that the system assumes knowledge of an individual's personality and behavioral patterns, and might entail biases. To prevent any kind of misuse, it is crucial to establish ethical guidelines and ensure transparency in the usage and obtain informed consent from the users. Responsible use, strict data governance, and clear communication about the limitations and potential risks of the system are essential to safeguard individuals' rights.

The study we presented in this paper is a novel attempt to automatically label regulatory focus which could be lacking in many aspects. We acknowledge that the bias contained in the data and the chosen method may inadvertently perpetuated or amplified. We do not advocate the use of our methods in any fully automated downstream applications.

9 Acknowledgment

Roman Klinger's research is partially funded by the German Research Council (DFG), projects KL 2869/1-2 and KL 2869/5-1. We thank the reviewers whose feedback contributed to improve the manuscript considerably.

References

- Aaker, Jennifer L. and Angela Y. Lee. 2001. "I" seek pleasures and "we" avoid pains: The role of self-regulatory goals in information processing and persuasion. *Journal of Consumer Research*, 28(1):33–49.
- Argamon, Shlomo, Moshe Koppel, Jonathan Fine, and Anat Rachel Shimoni. 2003. Gender, genre, and writing style in formal written texts. *Text & talk*, 23(3):321–346.
- Berry, Diane S., James W. Pennebaker, Jennifer S. Mueller, and Wendy S. Hiller. 1997. Linguistic bases of social perception. *Personality and Social Psychology Bulletin*, 23(5):526–537.
- Bonferroni, Carlo. 1936. Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 8:3–62.

- Brockner, Joel and Tory E. Higgins. 2001. Regulatory focus theory: Implications for the study of emotions at work. *Organizational Behavior and Human Decision Processes*, 86(1):35–66.
- Cesario, Joseph, Katherine S. Corker, and Sara Jelinek. 2013. A self-regulatory framework for message framing. *Journal of Experimental Social Psychology*, 49(2):238–249.
- Chan, Branden, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Conley, Mark A. and E. Tory Higgins. 2018. Value from fit with distinct motivational field environments. *Basic and Applied Social Psychology*, 40(2):61–72.
- Crowe, Ellen and Tory E. Higgins. 1997. Regulatory Focus and Strategic Inclinations: Promotion and Prevention in Decision-Making. *Organizational Behavior and Human Decision Processes*, 69(2):117–132.
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Gaudry, Eric, Peter Vagg, and Charles D Spielberger. 1975. Validation of the state-trait distinction in anxiety research. *Multivariate behavioral research*, 10(3):331–341.
- Gill, Alastair J., Robert M. French, Darren Gergle, and Jon Oberlander. 2008. Identifying emotional characteristics from short blog texts. In *30th Annual Conference of the Cognitive Science Society*, pages 2237–2242. Cognitive Science Society Washington, DC.
- Glass, G.V. and K.D. Hopkins. 1996. *Statistical Methods in Education and Psychology*. Allyn and Bacon.
- Goswami, Sumit, Sudeshna Sarkar, and Mayur Rustagi. 2009. Stylometric analysis of bloggers' age and gender. In *Proceedings of the Third International Conference on Weblogs and Social Media, ICWSM 2009, San Jose, California, USA, May 17-20, 2009*. The AAAI Press.
- Gottschalk, Louis A. and Goldine C. Gleser. 1979. *The Measurement of Psychological States Through the Content Analysis of Verbal Behavior*. Berkley: University of California Press.
- Hamstra, Melvyn R. W., Kai Sassenberg, Nico W. Van Yperen, and Barbara Wisse. 2014. Followers feel valued—When leaders' regulatory focus makes leaders exhibit behavior that fits followers' regulatory focus. *Journal of Experimental Social Psychology*, 51:34–40.
- Higgins, Tory E. 1997. Beyond pleasure and pain. *American Psychologist*, 52(12):1280–1300.
- Higgins, Tory E. 1998. Promotion and prevention: Regulatory focus as a motivational principle. In Mark P. Zanna, editor, *Advances in experimental social psychology*, volume 30, pages 1–46. Academic Press.
- Higgins, Tory E. 2000. Making a good decision: value from fit. *American psychologist*, 55(11):1217.
- Higgins, Tory E. 2002. How Self-Regulation Creates Distinct Values: The Case of Promotion and Prevention Decision Making. *Journal of Consumer Psychology*, 12(3):177–191.
- Higgins, Tory E., Ronald S. Friedman, Robert E. Harlow, Lorraine Chen Idson, Ozlem N. Ayduk, and Amy Taylor. 2001. Achievement orientations from subjective histories of success: Promotion pride versus prevention pride. *European Journal of Social Psychology*, 31(1):3–23.
- Higgins, Tory E., James Y. Shah, and Ronald S. Friedman. 1997. Emotional responses to goal attainment: strength of regulatory focus as moderator. *Journal of personality and social psychology*, 72 3:515–25.
- Kees, Jeremy, Scot Burton, and Andrea Heintz Tangari. 2010. The impact of regulatory focus, temporal orientation, and fit on consumer responses to health-related advertising. *Journal of Advertising*, 39(1):19–34.
- Keller, Punam A. 2006. Regulatory Focus and Efficacy of Health Messages. *Journal of Consumer Research*, 33(1):109–114.
- Lanaj, Klodiana, Chu-Hsiang Chang, Russell E. Johnson, et al. 2012. Regulatory focus and work-related outcomes: a review and meta-analysis. *Psychological bulletin*, 138(5):998.
- Latimer, Amy E., Susan E. Rivers, Tara A. Rench, Nicole A. Katulak, Althea Hicks, Julie Keany Hodorowski, Edward Tory Higgins, and Peter Salovey. 2008a. A field experiment testing the utility of regulatory fit messages for promoting physical activity. *Journal of experimental social psychology*, 44 (3):826–832.
- Latimer, Amy E, Pamela Williams-Piehot, Nicole A Katulak, Ashley Cox, Linda Mowad, E Tory Higgins, and Peter Salovey. 2008b. Promoting fruit and vegetable

- intake through messages tailored to individual differences in regulatory focus. *Annals of Behavioral Medicine*, 35(3):363–369.
- van der Maaten, Laurens and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605.
- Meier, Tabea, Ryan L. Boyd, James W. Pennebaker, Matthias R. Mehl, Mike Martin, Markus Wolf, and Andrea B. Horn. 2019. “LIWC auf Deutsch”: The development, psychometrics, and introduction of DE-LIWC2015. *PsyArXiv*.
- Nowson, Scott and Jon Oberlander. 2006. The Identity of Bloggers: Openness and Gender in Personal Weblogs. In *AAAI spring symposium: Computational approaches to analyzing weblogs*, pages 163–167. Palo Alto, CA.
- Pennebaker, James W., Ryan Boyd, Kayla Jordan, and Kate Blackburn. 2015. *The development and psychometric properties of LIWC2015*. University of Texas at Austin.
- Pennebaker, James W., Matthias R. Mehl, and Kate G. Niederhoffer. 2003. Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology*, 54(1):547–577.
- Plutchik, Robert. 2001. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist*, 89(4):344–350.
- Rangel, Francisco and Paolo Rosso. 2016a. On the impact of emotions on author profiling. *Information Processing & Management*, 52(1):73–92. Emotion and Sentiment in Social and Expressive Media.
- Rangel, Francisco and Paolo Rosso. 2016b. On the impact of emotions on author profiling. *Information processing & management*, 52(1):73–92.
- Righetti, Francesca, Catrin Finkenauer, and Caryl Rusbult. 2011. The benefits of interpersonal regulatory fit for individual goal pursuit. *Journal of Personality and Social Psychology*, 101(4):720.
- Sassenberg, Kai, Florian Landkammer, and Johann Jacoby. 2014. The influence of regulatory focus and group vs. individual goals on the evaluation bias in the context of group decision making. *Journal of Experimental Social Psychology*, 54:153–164.
- Sassenberg, Kai, Claudia Sassenrath, and Adam K. Fetterman. 2015. Threat \neq prevention, challenge \neq promotion: The impact of threat, challenge and regulatory focus on attention to negative stimuli. *Cognition and Emotion*, 29(1):188–195.
- Sassenberg, Kai and Michael LW Vliek. 2019. Self-regulation strategies and regulatory fit. *Social Psychology In Action: Evidence-Based Interventions From Theory To Practice*, pages 51–64.
- Sassenrath, Claudia, Kai Sassenberg, Devin G. Ray, Katharina Scheiter, and Halszka Jarodzka. 2014. A motivational determinant of facial emotion recognition: Regulatory focus affects recognition of emotions in faces. *PLOS ONE*, 9(11):1–9.
- Semin, Gün R., Tory E. Higgins, Lorena Gil de Montes, Yvette Estourget, and Valencia J. 2005. Linguistic signatures of regulatory focus: how abstraction fits promotion more than prevention. *Journal of personality and social psychology*, 89(1):36–45.
- Spitzer, Daniel. 2019. Von der Wissenschaft zur Anwendung-Entwicklung einer KI-unterstützten Textanalyse. <https://tinyurl.com/100W-textanalysis>.
- Stamatatos, Efstathios, Martin Potthast, Francisco Rangel, Paolo Rosso, and Benno Stein. 2015. Overview of the pan/clef 2015 evaluation lab. In *Proceedings of the 6th International Conference on Experimental IR Meets Multilinguality, Multimodality, and Interaction - Volume 9283, CLEF’15*, page 518–538, Berlin, Heidelberg, Springer-Verlag.
- Stone, Philip J. and Earl B. Hunt. 1963. A computer approach to content analysis: Studies using the general inquirer system. In *Proceedings of the May 21-23, 1963, Spring Joint Computer Conference, AFIPS ’63 (Spring)*, page 241–256, New York, NY, USA. Association for Computing Machinery.
- Troiano, Enrica, Laura Oberländer, and Roman Klinger. 2023. Dimensional modeling of emotions in text with appraisal theories: Corpus creation, annotation reliability, and prediction. *Computational Linguistics*, 49(1):1–72.
- Troiano, Enrica, Sebastian Padó, and Roman Klinger. 2019. Crowdsourcing and validating event-focused emotion corpora for German and English. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4005–4011, Florence, Italy. Association for Computational Linguistics.
- Updegraff, John A., David K. Sherman, Faith S. Luyster, and Traci L. Mann. 2007. The effects of message quality and congruency on perceptions of tailored health communications. *Journal of Experimental Social Psychology*, 43(2):249–257.
- Vaughn, Leigh A. 2018. Contents of Hopes and Duties: A Linguistic Analysis. *Frontiers in Psychology*, 9:757.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A EDD-1 data creation

A.1 Experiment questionnaire

Bedingung 1: In diesem letzten Teil möchten wir Sie bitten, sich an einige persönliche Erlebnisse aus Ihrer Vergangenheit zu erinnern. Dabei kann es sich beispielsweise um Erfahrungen handeln, die Sie im Laufe Ihrer Schulzeit bzw. Ihres Studiums oder in Ihrem Privatleben gemacht haben. Bitte beschreiben Sie in einigen Sätzen drei verschiedene Erlebnisse Ihrer Vergangenheit:

1. Bitte beschreiben Sie ein Erlebnis, bei dem Sie das Gefühl hatten, Sie machen Fortschritte dahingehend, in Ihrem Leben erfolgreich zu sein.
2. Bitte beschreiben Sie ein Erlebnis, bei dem Sie das Gefühl hatten, Sie machen keine Fortschritte dahingehend, in Ihrem Leben erfolgreich zu sein.
3. Bitte beschreiben Sie ein Erlebnis, bei dem Sie im Vergleich zu anderen Personen dazu fähig waren, das zu bekommen, was Sie wollen.

Bedingung 2: In diesem letzten Teil möchten wir Sie bitten, sich an einige persönliche Erlebnisse aus Ihrer Vergangenheit zu erinnern. Dabei kann es sich beispielsweise um Erfahrungen handeln, die Sie im Laufe Ihrer Schulzeit bzw. Ihres Studiums oder in Ihrem Privatleben gemacht haben. Bitte beschreiben Sie in einigen Sätzen drei verschiedene Erlebnisse Ihrer Vergangenheit:

1. Bitte beschreiben Sie ein Erlebnis, bei dem eine ausreichende Vorsicht Sie davor bewahrt hat, in Schwierigkeiten zu geraten.
2. Bitte beschreiben Sie ein Erlebnis, bei dem eine mangelnde Vorsicht dazu geführt hat, dass Sie in Schwierigkeiten geraten sind.
3. Bitte beschreiben Sie sowie ein Erlebnis, bei dem Sie sich so verhalten haben, dass niemand etwas daran hätte aussetzen können.

A.2 Experiment questionnaire (translation)

Condition 1: In this last part, we would like you to recall some personal experiences from your past. These can be, for example, experiences you had during your school years or studies or in your private life. Please describe in a few sentences three different experiences from your past:

1. Please describe an experience where you felt you were making progress toward being successful in your life.

2. Please describe an experience in which you felt you were not making progress toward being successful in your life.

3. Please describe an experience where you were able to get what you want compared to other people.

Condition 2: In this last part, we would like you to recall some personal experiences from your past. These can be, for example, experiences you had during your school or university years or in your private life. Please describe in few sentences three different experiences from your past:

1. Please describe an experience in which being sufficiently careful kept you from getting into trouble.
2. Please describe an experience where a lack of caution caused you to get into trouble.
3. Please describe an experience in which you behaved in a way that no one could have found fault with.

B EDD-2 data creation

B.1 Introduction

Liebe Untersuchungsteilnehmerin, lieber Untersuchungsteilnehmer, vielen Dank für die Bereitschaft, an der Studie teilzunehmen! Bitte lesen Sie sich die folgenden Informationen sorgfältig durch und entscheiden dann über Teilnahme oder Nichtteilnahme an dieser Studie.

Inhalt: In dieser Studie untersuchen wir, wie unterschiedliche Zielverfolgungsstrategien zusammenhängen. Dazu werden wir Sie bitten, offene Fragen zu Situationen aus der Vergangenheit zu beantworten, in denen Sie (un)erfolgreich Ziele verfolgt haben. Danach folgen einige Fragen zu Ihrem Verhalten am Arbeitsplatz und zu der Verfolgung von Leistungszielen.

Studienablauf und Bezahlung Insgesamt dauert die Studie in etwa 8-10 Minuten. Alle Teilnehmenden erhalten dafür eine Entlohnung von 1.50 €. Die Studie sollte zusammenhängend am Computer, Laptop oder Tablet (nicht auf dem Handy) bearbeitet werden. Voraussetzung für Ihre Teilnahme ist, dass Sie mindestens 18 Jahre alt sind und fließend Deutsch sprechen.

Vertraulichkeit und Handhabung der Daten Alle personenbezogenen Daten werden streng vertraulich behandelt und nur für Forschungszwecke verwendet. Durch Ihre Bestätigung unten erlauben Sie uns, Ihre Antworten für wissenschaftlichen Zwecken auszuwerten und in vollständig anonymisierter Form anderen Wissenschaftlern öffentlich zur Verfügung zu stellen. Am Ende der Umfrage haben Sie nochmals die Möglichkeit, diese Einwilligung zu widerrufen. Danach ist ein Rückzug der Daten nicht mehr möglich, da die Daten anonym gespeichert werden und wir nicht in der Lage sind, Ihre Daten zu identifizieren. Sollten Sie Fragen bezüglich Ihrer Daten oder Datenspeicherung haben, können Sie unsere Datenschutzbeauftragten kontaktieren: XXXX

- Ich bin mindestens 18 Jahre alt und habe die Informationen gelesen und verstanden. Ich erkläre mich damit einverstanden, an der Studie teilzunehmen.
- Ich möchte nicht an der Studie teilnehmen.

B.2 Experiment questionnaire

In diesem ersten Teil möchten wir Sie bitten, sich an einige persönliche Erlebnisse aus Ihrer Vergangenheit zu erinnern. Wir interessieren uns für Erfahrungen, die Sie gemacht haben, während Sie ein Ziel verfolgt haben - beispielsweise während der Arbeit oder im privaten

Kontext. Bitte beschreiben Sie in einigen Sätzen drei verschiedene Erlebnisse Ihrer Vergangenheit (jeweils mindestens 150 Zeichen):

1. Bitte beschreiben Sie ein Erlebnis, bei dem Sie das Gefühl hatten, Sie machen Fortschritte dahingehend, in Bezug auf ein Ihnen wichtiges Ziel erfolgreich zu sein.
2. Bitte beschreiben Sie ein Erlebnis, bei dem Sie das Gefühl hatten, Sie machen keine Fortschritte dahingehend, etwas zu erreichen.
3. Bitte beschreiben Sie ein Erlebnis, bei dem Sie im Vergleich zu anderen Personen dazu fähig waren, das zu bekommen, was Sie wollten.

In diesem ersten Teil möchten wir Sie bitten, sich an einige persönliche Erlebnisse aus Ihrer Vergangenheit zu erinnern. Wir interessieren uns für Erfahrungen, die Sie gemacht haben, als Sie ein Ziel verfolgt haben - beispielsweise während der Arbeit oder im privaten Kontext. Bitte beschreiben Sie in einigen Sätzen drei verschiedene Erlebnisse Ihrer Vergangenheit (jeweils mindestens 150 Zeichen):

1. Bitte beschreiben Sie ein Erlebnis, bei dem ausreichende Vorsicht Sie davor bewahrt hat, in Schwierigkeiten zu geraten.
2. Bitte beschreiben Sie ein Erlebnis, bei dem eine mangelnde Vorsicht dazu geführt hat, dass Sie in Schwierigkeiten geraten sind.
3. Bitte beschreiben Sie ein Erlebnis, bei dem Sie sich so verhalten haben, dass niemand etwas daran hätte aussetzen können.

B.3 Introduction (translation)

Dear participant, thank you for your willingness to participate in the study! Please read the following information carefully and then decide whether to participate or not in this study.

Content: In this study, we will investigate how different goal pursuit strategies are related. For this purpose, we will ask you to answer open-ended questions about situations from the past in which you have (un)successfully pursued goals. This will be followed by some questions about your behavior at work and about the pursuit of performance goals.

Study procedure and payment: In total, the study will take about 8-10 minutes. All participants will receive a payment of 1.50 €. The study should be completed contiguously on a computer, laptop or tablet (not on a cell phone). To participate, you must be at least 18 years old and fluent in German.

Confidentiality and data handling: All personal data will be kept strictly confidential and will only be used for research purposes. By confirming below, you allow us to evaluate your answers for scientific purposes and make them publicly available to other researchers in a completely anonymized form. At the end of the survey, you will again have the opportunity to revoke this consent. After that, it is no longer possible to retrace the data, as the data is stored anonymously and we are not able to identify your data. If you have any questions regarding your data or data storage, you can contact our data protection officers: XXXX

- I am at least 18 years old and have read and understood the information. I agree to participate in the study.
- I do not wish to participate in the study.

B.4 Experiment questionnaire (translation)

Condition 1: In this first part, we would like you to recall some personal experiences from your past. We are interested in experiences you had while pursuing a goal - for example, during work or in a private context. Please describe in a few sentences three different experiences from your past (at least 150 characters each):

1. Please describe an experience in which you felt you were making progress toward being successful in a goal that was important to you.

2. Please describe an experience in which you felt you were not making progress toward achieving something.
3. Please describe an experience where you were able to get what you wanted compared to other people.

Condition 2: In this first part, we would like you to recall some personal experiences from your past. We are interested in experiences you had when pursuing a goal - for example, during work or in a private context. Please describe in a few sentences three different experiences from your past (at least 150 characters each):

1. Please describe an experience where sufficient caution kept you from getting into trouble.
2. Please describe an experience where a lack of caution caused you to get into trouble.
3. Please describe an experience in which you behaved in a way that no one could have found fault with.

C Twitter data creation

C.1 List of emotion words

For detecting emotion words we created a list of words that are represented in Plutchik's emotion wheel (Plutchik, 2001) and two additional items representing shame and pride.

Emotion words: klar, wüt, angewider, betrüb, erstaun, erschrock, bewunder, begeister, froh, bereit, verärger, ablehn, traurig, überrasch, ängst, vertrau, akzeptier, gelass, neugierig, gereiz, gelangweil, nachdenk, verwirr, besorg, stolz, aufmerksam, klar, optimist, verlieb, streitlust, hass, bereund, enttäusch, ehrfürchtig, fügsam, scham

C.2 Annotation Guidelines

C.2.1 Definition & Examples

According to Regulatory focus theory human behavior or thoughts are motivated by a need for achievement (promotion focus) or a need for security (prevention focus). Promotion-focused individuals are motivated by achievement, are more risk seeking and approach tasks eagerly. Prevention focused individuals take a risk-averting approach, are more vigilant and value security. The examples below demonstrate how variation in regulatory focus is captured in formulation of text. In the annotation task that follows only tweets in German are included and for adding diversity, examples cover different domains and not only tweets.

C.2.2 Regulatory Focus and emotion

Prevention and promotion are related to distinct sets of emotions. Emotions triggered in the context of success (i.e., a positive situation) or failure i.e., in a negative situation) can clearly be connected to promotion or prevention focus. Positive activating emotions like cheerfulness and happiness (success situation), and negative non-activating emotions like sad and depressed (failure situation) are indicators of promotions focus. While positive non-activating emotions like relaxed, unstressed, calm, calming down etc.,(success situation) and negative activating emotions like anger, hate, fear etc.,(failure situation) are prevention focus indicators. Below Figure 2 shows the emotions related to a regulatory focus category and outcome of a particular situation (success/failure).

1. Prevention Focus

- (a) Die Forschung hat gezeigt, dass Vitamin C vor Krankheiten wie z. B. Erkältungen schützt.

Explanation: This example emphasises on protection or avoiding sickness, hence it is prevention focus

- (b) Der 100% Grapefruit-Saft sichert den Tagesbedarf an Vitamin C.
Explanation: This formulation instils a sense of security, hence is prevention focus.
- (c) Habe meine praktische Fahrprüfung bestanden, war doch einfacher als gedacht. Die Straße muss nicht mehr auf mich warten.
Explanation: The expression "einfacher als gedacht" shows the person was prepared for the difficult task, pointing to prevention focus
- (d) Wir konnten uns endlich den Traum vom eigenen Haus erfüllen. Wir sind sooo dankbar! *Explanation: "endlich" refers to a feeling of relief which is a prevention emotion*
- (e) Die Welt fühlt sich manchmal so abweisend an. Früher hatte ich noch ein Gefühl von Sicherheit.
- (f) Können wir drauf vertrauen, dass sich unsere Politiker genug ernsthafte Gedanken gemacht haben über die Risiken des Klimawandels?
- (g) Von Reisen rät doch jeder im Moment ab, richtig so, ist doch viel zu gefährlich!
- (h) Ich bin kein Impfgegner, Impfungen retten Leben, bestes Beispiel Polio oder Tetanus. Aber einen mRNA Impfstoff zu bekommen, der weniger als 6 Monate getestet wurde... sorry, da kann ich auch Russisch Roulette spielen. Ich hatte Covid übrigens bereits und nix bis auf Husten.

2. Promotion Focus

- (a) Forschung hat gezeigt, dass Vitamin C Ihre Gesundheit stärkt.
Explanation: Compared to the prevention formulation, you can see that this statement emphasises on positive outcome, hence this is promotion focus.
- (b) Unser 100% Grapefruit-Saft hat drei Mal mehr Vitamin C als andere Fruchtsäfte. Richtig gut, oder?
Explanation: Here the statement focuses on advantage rather than security.
- (c) Die Früchte werden nur zur besten Erntezeit verarbeitet und schmecken daher so gut.
Explanation: The emphasis here again is on the plus points or advantages, hence promotion focused

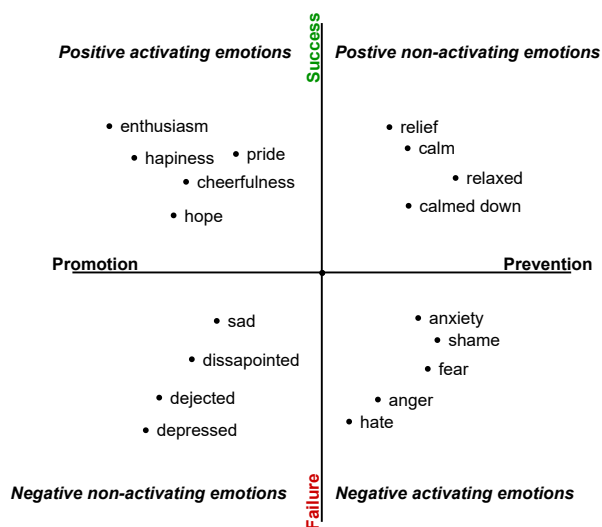


Figure 2: An approximate representation of emotions related to a regulatory focus category and outcome of a particular situation (success/failure) (drawn following Brockner and Higgins, 2001).

- (d) Heute nochmal fünf Kilo mehr geschafft. Habe mein Monatsziel fast erreicht, so kann es weitergehen.
- (e) Ich bin heute Morgen früh aufgestanden, weil ich zum Beginn meines Psychologieunterrichts um 8:30 Uhr in der Schule sein wollte, der normalerweise hervorragend ist.
- (f) Ich freue mich auf meinen neuen Job bei amnesty. Dort kann ich nicht nur Geld verdienen sondern mich auch für meine Werte einsetzen.
- (g) Ich habe mir ein neues Fahrrad gekauft. Ich wusste gar nicht wieviel Spass es machen kann in der Freizeit die nähere Umgebung zu erkunden.
- (h) In nur 6 Monaten wurden 50% der Deutschen einmal geimpft. Seid doch mal ehrlich, dass sowas geht hätte vor Corona auch niemand gedacht.

C.2.3 Task Description

Familiarize yourself with the concepts mentioned in the previous section. Note the difference in text formulation for prevention and promotion focus. Ask for more examples, if the concept is not clear. The annotation task requires you to annotate each given tweet with the one of the following labels.

1. *prevention*
2. *promotion*

3. *neither promotion not prevention*

4. *not sure*

Take into consideration the emotion expressed in the context of success or failure. Even though it is more common to see positive emotion in promotion focus text, it is not always the case.

C.2.4 Annotation Environment

The annotation task will be carried out in google sheets. You have to read the text in the column *tweet*, decide which regulatory focus category the tweet belongs and choose a label from the drop-down in the column *label*. If you have any feedback about the instance, please use the *comments* column.

D Psychological categories

For the linguistic correlation analysis we included all 49 categories from the 100W api and 80 categories from DE-LIWC2025. We excluded only those categories referring to punctuations and the categories *fillers*, *other* and *Dic* as they are not relevant in the context of current study. Table 9 shows the categories from both lexicon that were used in this study.

LIWC categories

Analytic (Analytic Thinking), Authentic (Authentic), Clout (Clout), Sixltr (Words > 6 letters), Tone (Emotional tone), WPS (Words/sentence), achiev (Achievement), adj (Common adjectives), adverb (Common Adverbs), affect (Affective processes), affiliation (Affiliation), anger (Anger), anx (Anxiety), article (Articles), assent (Assent), auxverb (Auxiliary verbs), bio (Biological processes), body (Body), cause (Causation), certain (Certainty), cogproc (Cognitive processes), compare (Comparisons), conj (Conjunctions), death (Death), differ (Differentiation), discrep (Discrepancy), drives (Drives), family (Family), feel (Feel), female (Female references), focusfuture (Future focus), focuspast (Past focus), focuspresent (Present focus), friend (Friends), function (Total function words), health (Health), hear (Hear), home (Home), i (1st pers singular), informal (Informal language), ingest (Ingestion), insight (Insight), interrog (Interrogatives), ipron (Impersonal pronouns), leisure (Leisure), male (Male references), money (Money), motion (Motion), negate (Negations), negemo (Negative emotion), netspeak (Netspeak), nonflu (Nonfluencies), percept (Perceptual processes), posemo (Positive emotion), power (Power), ppron (Personal pronouns), prep (Prepositions), pronoun (Total pronouns), quant (Quantifiers), relativ (Relativity), relig (Religion), reward (Reward), risk (Risk), sad (Sadness), see (See), sexual (Sexual), shehe (3rd person singular), social (Social processes), space (Space), swear (Swear words), tentat (Tentative), they (3rd person plural), time (Time), verb (Common verbs), we (1st pers plural), work (Work), you_formal (2nd pers formal), you_plur (2nd person plural), you_sing (2nd person singular), you_total (2nd person)

100W categories

DAV (Descriptive Action Verb), achieve (Achievement), adjective (Adjective), adverb (Adverb), affil (Affiliation), agent (Active voice), anger (Anger), anxiety (Anxiety), article (Article), auxverb (Auxiliary Verb), booster (Intensifiers), conj (Conjunctions), discrep (Discrepancy), feminine (Feminine), future (Future focus), ich (First Person singular), impersonalPronouns (Impersonal Pronouns), masculine (Masculine), money (Money), motion (Motion), negAchieve (Negative Achievement), negAffil (Negative Affiliation), negEmo (Negative Emotion), negPower (Negative Power), negation (Negation), numbers (Numbers), past (Past focus), patient (Passive voice), personalPronouns (Personal Pronouns), posAchieve (Positive Achievement), posAffil (Positive Affiliation), posEmo (Positive Emotion), posPower (Positive Power), power (Power), preposition (Preposition), quant (Quantity), relativ (Absolutness), reward (Reward), risk (Risk), sadness (Sadness), shehe (Third Person plural), space (Space), speak (Speak), strictNegationPrepositions (Strict Negation Prepositions), sv (State Verb), swear (Swear Words), time (Time), we (First Person plural), you (Second Person singular)

Table 9: List of psychological variables and their corresponding categories in both LIWC and 100W lexicons used in the current study

E Training details for GBERT

We fine-tuned the pre-trained German BERT model deepset/GBERT-large for the regulatory focus classification task. Figure 3 shows the training and validation loss averaged across folds for each epoch. The number of epochs are varying in some cases because we set an early stopping criteria to stop training if the validation loss does not improve for 5 steps. We use the setting `load_best_model_at_end` to save the model with best performance on the validation set, rather than the model from the last training epoch.

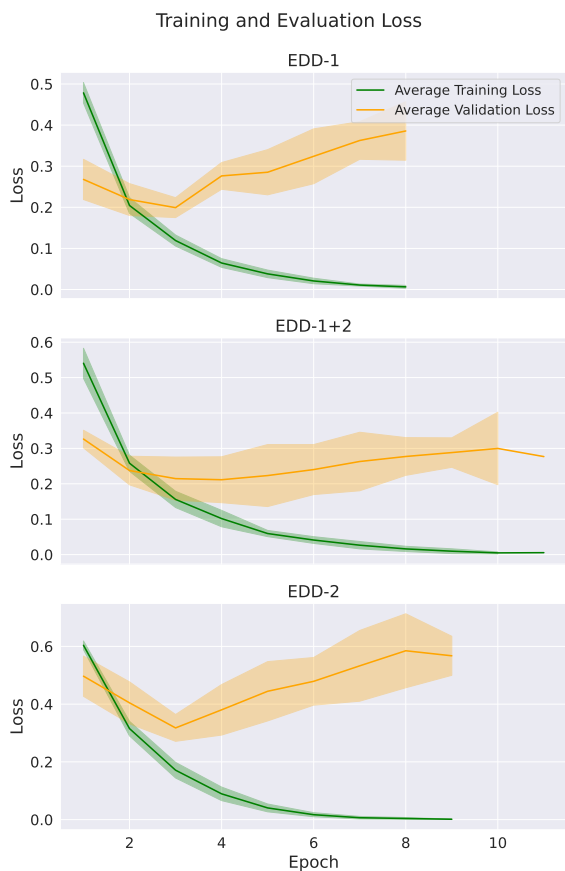


Figure 3: Training and validation loss for each dataset averaged over folds for each epoch.

F Additional experiment results

In closed-vocabulary methods, in addition to the linear models discussed in the paper, we conducted regulatory focus classification using three non-linear models: support vector machines (SVM), random forest, and gradient boosting and the three feature sets: LIWC, 100W, and TF-IDF vectors. We use the default hyper-parameters for the model in the `scikit-learn` python

package. The experiments are conducted with the same setup as discussion in Section 5.2 for the linear models. Figure 4 displays the results of 10-fold cross-validation on both the event description dataset and the Twitter dataset for all non-linear models and the logistic regression model. On comparing the results, we observe that the SVM_TFIDF model outperforms other non-linear models. However, the logistic regression model (Logreg_TFIDF) achieves almost similar results and the standard deviation suggests that logistic regression model might be more stable in comparison. Furthermore, the performance of both SVM_TFIDF and Logreg_TFIDF on Twitter data is comparable.

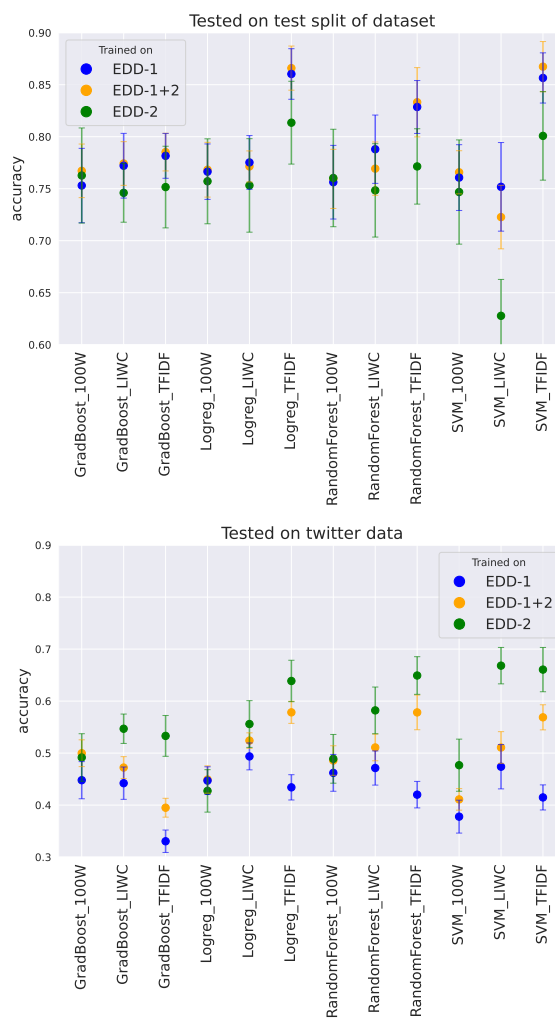


Figure 4

G Binary vs. tertiary classification

Individuals exhibit varying degrees of regulatory focus based on the given situation and context. The notion of a completely neutral regulatory focus, where an individual lacks any inclination towards promotion or prevention, is quite rare. When an individual is engaging in a social media activity like posting in Twitter, they have an active motivational orientation. However, it is possible that it is hard to identify the regulatory focus of the author when there is no sufficient contextual information to make an accurate prediction. This is reflected in the annotation task as well, where the annotators did not choose either of the two labels. As shown in Figure 5, the distribution of labels is skewed with only 3.65% instances labeled as *neutral*.

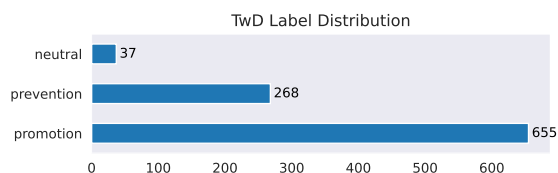


Figure 5: Distribution of labels in Twitter data.

To understand whether the state-of-the-art model used in the study is also able to handle regulatory focus classification as a three class problem, we trained and tested the model using the annotated Twitter data. We fine-tuned the pre-trained German BERT model `deepset/GBERT-large` on the Twitter data with instances labelled as *promotion*, *prevention* and *neutral*. In the *neutral* label we consolidated instances labeled as *neither promotion nor prevention* or *not sure* by both annotators. We split each of the datasets into training, validation, and test sets using an 80-10-10 split.

class	precision	recall	F ₁
promotion	0.949	0.962	0.955
prevention	0.889	0.896	0.889
neutral	0.400	0.300	0.311

Table 10: Results of GBERT model trained on Twitter data labeled with *promotion*, *prevention* and *neutral* labels

Table 10 shows the results for regulatory focus classification as a 3-class problem. Considering the limited number of instances for the neutral label (3% of the dataset), the model’s relatively poor performance on that label is expected. However, it demonstrates good performance on both the promotion and prevention labels. The results could be suggesting that the

distinction between promotion-focused and prevention-focused content is more evident and discernible compared to instances exhibiting a neutral regulatory focus.

Unsupervised Text Embedding Space Generation Using Generative Adversarial Networks for Text Synthesis

Jun-Min Lee, Korea Advanced Institute of Science and Technology, I-BRICKS, ljm56897@gmail.com

Tae-Bin Ha, I-BRICKS, taebinalive@gmail.com

Abstract Generative Adversarial Network (GAN) is a data synthesis model that creates plausible data through the competition between a generator and a discriminator. Although GAN has been extensively studied for image synthesis, it has inherent limitations when applied to natural language generation. This is because natural language is composed of discrete tokens, and the generator faces challenges in updating its gradient through backpropagation. Therefore, most text-GAN studies generate sentences starting with a random token (or prompt) based on a reward system. Thus, the generators of previous studies are pre-trained in an autoregressive manner before adversarial training, resulting in data memorization where synthesized sentences reproduce the training data. In this paper, we synthesize sentences using a framework similar to the original GAN. More specifically, we propose Text Embedding Space Generative Adversarial Networks (TESGAN), which generate continuous text embedding spaces instead of discrete tokens to address the gradient backpropagation problem. Furthermore, TEGAN conducts unsupervised learning that does not directly refer to the text of the training data to overcome the data memorization issue. Also, TEGAN enables unconditional text synthesis during the inference phase by using random noise instead of tokens or prompts for text synthesis. By adopting this novel method, TEGAN can synthesize new sentences, demonstrating the potential of unsupervised learning for text synthesis. We look forward to extended research that combines large-scale language models with a new perspective on viewing text as continuous spaces.

1 Introduction

Generative Adversarial Network (GAN), as proposed by Goodfellow et al. (2014), is a popular model for data synthesis. GAN is an unconditional data generation algorithm that aims to generate plausible data in an unsupervised manner by fostering competition between a generator and a discriminator to capture the real data distribution. When GAN was initially introduced, it primarily focused on image synthesis, and extensive research was conducted to achieve high-quality synthetic data results (Arjovsky et al., 2017; Radford et al., 2015; Karras et al., 2018, 2021). Furthermore, GAN is commonly employed in the field of computer vision for data augmentation through image synthesis (Sandfort et al., 2019; Bowles et al., 2018; Antoniou et al., 2018; Tran et al., 2021). The GAN generator learns implicit density based on the discriminator’s loss without direct reference to the training data. Consequently, GAN can prevent data memorization, where the model reproduces the training data. Additionally, GAN can synthesize various data by using random noise instead of a specific starting point, such as a designated start token.

Similar to images, unconditional text generation

can function as a data augmentation technique by generating new text that resembles a given dataset. It also has practical applications, such as creating new documents by generating fictitious text information suitable for direct use. Consequently, several studies have attempted to apply GAN to natural language, but they have encountered limitations in natural language generation. The challenge arises from the fact that natural language is composed of discrete tokens, making it challenging for the GAN generator to directly update gradients through backpropagation. The gradient backpropagation issue in text-based GANs was first discussed by Yu et al. (2017), and numerous subsequent text-GAN research efforts aimed to address this problem using gradient policy-based reinforcement learning with a reward system. Furthermore, the previous text-GAN approaches necessitated pre-training the generator with supervised learning (autoregressive) before adversarial training due to convergence issue with the generator (Yu et al., 2017). Accordingly, we discovered that the generators of previous text-GAN approaches reproduce the training data (leading to data memorization) during text synthesis due to the autoregressive-based pre-training process, which becomes a significant

issue in generative models.

This paper introduces a novel framework known as Text Embedding Space Generative Adversarial Networks (TESGAN)¹, which enables backpropagation and prevents data memorization. TESGAN does not rely on a supervised, pre-trained autoregressive-based generator that generates discrete tokens for text synthesis. Our generator generates continuous text embedding spaces for text synthesis instead of discrete tokens, allowing training with gradient backpropagation. Furthermore, the fact that TESGAN deals with continuous spaces makes it possible for TESGAN’s generator to be trained within the original GAN framework to mimic the real text embedding space. Moreover, TESGAN enables unconditional text generation, as it does not require the selection of a starting token (or prompt) for text synthesis. Our seed interpretation model then synthesizes sentences by interpreting the imitated continuous text embedding space created by the generator. During sentence synthesis, data memorization does not occur because TESGAN does not directly refer to the training text data but only learns from the continuous text embedding space. We use two datasets to conduct performance evaluations and general applicability experiments based on synthetic text generated by TESGAN. To assess the quality and diversity of synthesized text, we employ evaluation metrics such as Fréchet BERT Distance (FBD), Multi-sets-Jaccard (MSJ) (Alihosseini et al., 2019), Language Model score (LM) (de Masson d’Autume et al., 2019; Caccia et al., 2020), and Self-BLEU (SBL) (Zhu et al., 2018). In addition, we conducted human evaluations, and TESGAN achieved the highest average score. Lastly, we calculate the data memorization ratio and present the synthesized sentences to assess the potential of unsupervised learning and continuous embedding spaces for text synthesis.

2 Related Works

The most common method of text generation is to use an autoregressive-based language model via teacher forcing (Williams and Zipser, 1989). For example, extensive studies have been conducted on models using recurrent neural network (RNN) with Long Short-Term Memory (LSTM) cells (Hochreiter and Schmidhuber, 1997). Using LSTM, Graves (2013) successfully generated handwriting by predicting sequences, and Wen et al. (2015) synthesized sentences under specific conditions. Bowman et al. (2016) generated text after learning text embedding spaces with an autoregressive-based LSTM model and a variational autoencoder (VAE) architecture (Kingma and Welling, 2014). Policy Gradient

with BLEU (PG-BLEU) calculates the BLEU (Papineni et al., 2002) score of synthesized sentences and takes them as a reward when updating the generator using policy gradient.

Numerous investigations have been conducted to utilize GANs for text synthesis. Sequence GAN (SeqGAN) (Yu et al., 2017) attempted to address the backpropagation problem by employing gradient policy-based reinforcement learning with a reward system. However, SeqGAN faced a reward sparsity issue, leading Lin et al. (2017) to introduce RankGAN, which replaced the previous regression-based discriminator with a novel ranker. RankGAN trains the discriminator to assign higher scores to more realistic sentences. MaskGAN (Fedus et al., 2018) utilized an LSTM-based generator to fill in masked parts of sentences with tokens during training. Since MaskGAN uses discrete tokens, gradient backpropagation is not possible for the generator. To overcome this challenge, the authors employed the actor-critic method, using the probabilities of candidate tokens from the discriminator as rewards during training. Che et al. (2017) proposed Maximum Likelihood Augmented Discrete GAN (MaliGAN), which synthesizes text by minimizing Kullback-Leibler divergence (Kullback and Leibler, 1951). LeakGAN (Guo et al., 2018) alleviated issues related to sparseness and the lack of intermediate information by providing leaked information from the discriminator.

Several studies have aimed to address the gradient backpropagation problem without relying on reward-based reinforcement learning. TextGAN (Zhang et al., 2017) introduced kernel-based moment-matching, which enforces empirical distributions of real and synthetic text by using LSTM and Convolutional Neural Networks (CNN) for the generator and the discriminator, respectively. Feature Mover GAN (FM-GAN) (Chen et al., 2018) defined the feature-mover’s distance (FMD) and learned it by minimizing the FMD between real and fake sentences. Both TextGAN and FM-GAN utilized LSTM generators that generate discrete tokens using the *soft-argmax* trick instead of relying on reinforcement learning. Relational GAN (RelGAN) (Nie et al., 2019) applied relational recurrent neural networks (Santoro et al., 2018) and attempted to address the gradient backpropagation issue using Gumbel-softmax (Jang et al., 2017). However, since these approaches employed autoregressive (e.g., LSTM) generators, they explicitly referenced the training text data during model training. Consequently, previous studies faced challenges in avoiding complete data memorization while synthesizing sentences due to an autoregressive generator. Lastly, Transformer-based Implicit Latent GAN (TILGAN) (Diao et al., 2021) adopted a similar approach to TESGAN for addressing the gradient backpropagation issue based on the

¹<https://github.com/ljm565/TESGAN>

embedding space. However, TILGAN differs from TEGAN in that it was trained on a latent space compressed by the encoder, configured as an autoencoder transformer, and did not utilize embeddings learned from real language models.

Most of the aforementioned text-GAN models require the first token or prompt to synthesize text due to their autoregressive generators. TEGAN stands apart from these models as it generates text embedding space directly from random noise, eliminating the need for selecting tokens. Our TEGAN is the first text-GAN model that learns the real text embedding space without relying on an autoregressive generator.

3 Text Embedding Space GAN

TEGAN aims to generate the seeds required for synthesizing plausible text. These generated seeds (fake seeds) from the generator, along with the real seeds from the real text, are passed to the discriminators for training within the GAN framework. Once the training of TEGAN is complete, the pre-trained seed interpretation model synthesizes text using the fake seed created by the generator.

3.1 Seed for Text Synthesis

We denote a text sequence as $S = w_1, \dots, w_T$ (T is the sequence length). An autoregressive-based language model calculates the probability of the text sequence S as a product of conditional probabilities. If we assume that S is a complete sentence, then the sequence $D = S_1, \dots, S_N$ (N is the dialogue length) can be viewed as multi-turn sentences. Let $S_1 = w_1, \dots, w_m$ and $S_2, \dots, S_N = w_{m+1}, \dots, w_M$ denote the first sentence and the subsequent sentences, respectively (M is the total length of the multi-turn sentences D). The subsequent sentences after the first sentence can be predicted from a product of conditional probabilities:

$$p(S_2, \dots, S_N | S_1) = \prod_{i=m+1}^M p(w_i | w_1, \dots, w_{i-1}) \quad (1)$$

In other words, the first sentence can generate subsequent text using an autoregressive-based language model trained with multi-turn sentences. Therefore, the first sentence can serve as a seed. Meanwhile, the generator of TEGAN generates the first sentence as a continuous embedding space instead of discrete tokens to enable gradient backpropagation. Consequently, a continuous embedding space of the first sentence is defined as a seed.

3.2 Seed Interpretation Model

We define a seed in Section 3.1, and the seed interpretation model $f_\theta(\cdot)$ is used to synthesize text based on the seed. To synthesize text, the seed interpretation model must first be trained with multi-turn sentences in an autoregressive manner, similar to general language modeling, before adversarial training, as shown in Figure 1 (left), with the following loss function:

$$\mathcal{L}_{LM} = -\frac{1}{N} \sum_{n=1}^N \log \frac{\exp(x_n, y_n)}{\sum_{c=1}^C \exp(x_n, c)} \quad (2)$$

This enables the generator to synthesize appropriate text by utilizing the fake embedding space it creates during the inference phase. More detailed explanations will be provided in the following section. As a result, the model has to be trained on data consisting of multi-turn sentences $D = S_1, \dots, S_N$, where each sentence has a maximum length of L , meaning the total number of tokens in D is $N \times L$. When constructing the multi-turn sentence data, the special token $[CLS]$ is inserted only at the beginning of the first sentence, and each sentence is distinguished by adding the special token $[SEP]$ at the end. If the length of the tokenized sentence is less than L , the sentence is padded with the special token $[PAD]$:

$$S_1 = w_1^1, \dots, w_{|S_1|}^1, \dots, w_L^1 \quad (3)$$

$$(w_1^1 = [CLS], w_{|S_1|}^1 = [SEP], w_{l>|S_1|}^1 = [PAD])$$

$$S_{i(i>1)} = w_1^i, \dots, w_{|S_i|}^i, \dots, w_L^i \quad (4)$$

$$(w_{|S_i|}^i = [SEP], w_{l>|S_i|}^i = [PAD])$$

where S_1 and S_i represent a seed sentence and subsequent text. Let H_{real} denote the real seeds from TEGAN. As shown in Figure 1, the real seed is an embedding space of a sentence obtained by applying the sum of the token embedding and the positional embedding to the sigmoid function. Since most sentences can exist before others as long as the seed interpretation model is trained with multi-turn sentences, a significant portion of them can be used as seeds for text generation. Therefore, most of the sentences can be used as real seeds:

$$H_{real} = \sigma(W_{emb}(S_1) + W_{pos}(S_1)) \quad (5)$$

$$\approx \sigma(W_{emb}(S_n) + W_{pos}(S_n)) \in \mathbb{R}^{L \times d}$$

where L and d represent sequence length and embedding dimensions. H_{real} from the real text can be viewed as continuous spaces, similar to images, and the well-pretrained seed interpretation model can predict the next sentence S_{n+1} properly as illustrated in Figure 1 (right):

$$S_{n+1} = f_\theta(\sigma(W_{emb}(S_n) + W_{pos}(S_n))) = f_\theta(H_{real}) \in \mathbb{Z}^L \quad (6)$$

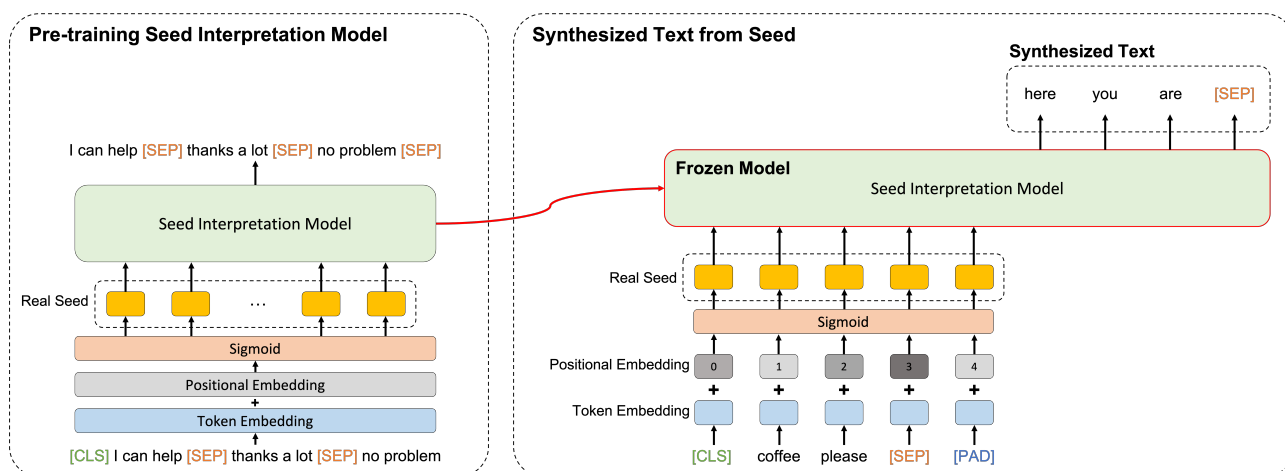


Figure 1: Illustration of the seed interpretation model. The seed interpretation model is pre-trained with multi-turn sentences before adversarial training (left). After pre-training, the model’s parameters are frozen, allowing it to synthesize text from the seed. The right figure implies that text can be synthesized from the seed. The [PAD] tokens following the [SEP] tokens are omitted in the left part for clarity.

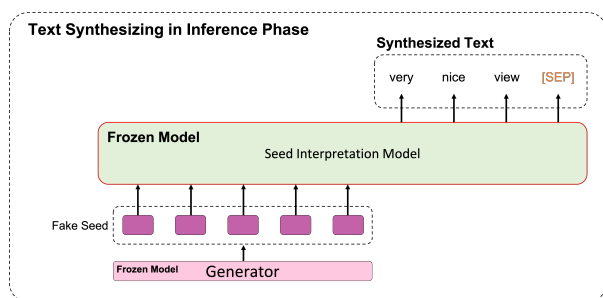


Figure 2: Illustration of text synthesizing method using the seed interpretation model in the inference phase.

As a result, text synthesis is carried out as the seed passes through the seed interpretation model to predict the subsequent sentence.

Applying to Unconditional Text Synthesis

Here, we assume that the training of the TEGAN framework, including adversarial training, is fully completed and describe how the seed interpretation model synthesizes text during the inference phase. Let $f_{\theta}^*(\cdot)$ and $g_{\phi}^*(\cdot)$ denote the frozen seed interpretation model and frozen generator, respectively. We can now synthesize text during the inference stage using the well-trained $f_{\theta}^*(\cdot)$ and $g_{\phi}^*(\cdot)$. At this point, $g_{\phi}^*(\cdot)$ will generate a fake seed H_{fake} with the same dimensions as H_{real} , as shown in Equation 7. Then, $f_{\theta}^*(\cdot)$ can synthesize text using the fake seed, as shown in Figure 2. In other words, if the generator can skillfully create fake seeds H_{fake} that imitate the distributions of H_{real} , then H_{fake} can also generate appropriate subsequent sentences (a.k.a synthetic text). However, no matter how

excellently H_{fake} is generated by the generator, it is useless if it cannot be interpreted; therefore, training the seed interpretation model is crucial. We use the pre-trained GPT-2 (Radford et al., 2018)² model and fine-tune it with multi-turn text data to serve as the seed interpretation model. In addition, this model is only used to provide H_{real} from the real text with frozen parameters during adversarial training. Thus, the seed interpretation model never affects the training of the generator and the discriminator during adversarial training, and vice versa. More detailed specifications of the seed interpretation model are explained in Appendix A.

Synthesizing text based on the generated fake seeds H_{fake} by the generator is entirely different from autoregressive prompting. This is because prompting methods (Wei et al., 2021; Ouyang et al., 2022; Chung et al., 2022) function by providing discrete tokens as input to a model that generates the next tokens based on the previous one. On the other hand, the generator of TEGAN creates continuous spaces H_{fake} for synthesizing text from random noise, enabling unconditional text synthesis without explicit human instruction. Furthermore, research is actively being conducted to leverage continuous spaces (learnable query) for flexible model training (Lester et al., 2021; Alayrac et al., 2022; Li et al., 2023; Dai et al., 2023). Most research explores various methods, including using a fixed learned query after model training or memorizing multiple learned queries and selecting them selectively as needed in different situations. However, the TEGAN framework differs from the mentioned studies in that its primary objective is to generate appropriate queries through the generator to

²https://huggingface.co/docs/transformers/model_doc/gpt2

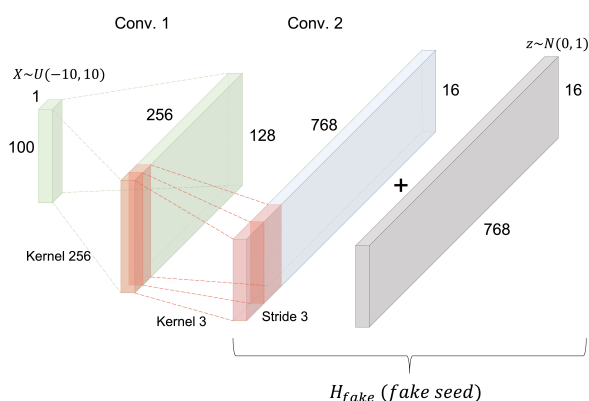


Figure 3: Illustration of the generator. P-TESGAN makes perturbed seeds by adding zero-centered normal distribution noise z (gray) to the output (blue) from the generator.

produce appropriate sentences.

3.3 Generator

Both real and fake seeds (H_{real} and H_{fake}) are essential for adversarial training. Real seeds can be obtained from the real text, as described in the seed interpretation model, and fake seeds are generated by the generator. In most text-GAN models reported so far, fake sentences are obtained from a text-based pre-trained autoregressive generator. Consequently, data memorization occurs, where several synthetic sentences reproduce the training data. To prevent data memorization, our generator does not use a pre-trained autoregressive-based model and does not explicitly reference the text in the training data during adversarial training. Our generator aims to create suitable fake text embedding spaces H_{fake} in an unsupervised manner (GAN framework) by referencing real text continuous spaces H_{real} .

As shown in Figure 3, the generator $g_\phi(\cdot)$ is composed of two convolutional layers and generates seeds from the uniform distribution noise X within an interval of $[-10, 10]$ to create diverse forms of the seeds. Additionally, using random noise has the advantage of not having to select the first token in the text synthesis process after model learning. The final H_{fake} can be obtained by Equation 7:

$$H_{fake} = g_\phi(X \sim U(-10, 10)) \in \mathbb{R}^{L \times d} \quad (7)$$

where L and d represent sequence length and embedding dimensions. As a result, the embedding space created by the generator has the same dimension as the real seed. Moreover, we also compare an additional model, perturbed TEGAN (P-TESGAN). P-TESGAN creates perturbed seeds by adding zero-centered nor-

mal distribution noise z to the generator output. P-TESGAN is expected to learn more robustly by perturbing the generator output. Please refer to Appendix A for detailed model information.

3.4 Objective Functions

Since the generator does not refer to text during adversarial training, its performance is determined by the loss of the discriminator. Thus, we propose four types of loss to update the parameters of the generator and the discriminator.

3.4.1 Discriminators

Sentence structure is important for constructing a complete sentence. Since the real seeds are made from perfect sentences, they maintain structural representations of sentences. Therefore, the fake seeds should capture the structural features of the real seeds. As shown in Figure 4a, we use Bidirectional Encoder Representations from Transformer (BERT) (Devlin et al., 2019) $d_\alpha(\cdot)$ called Seed Structure Discriminator (SSD) to capture the structural features of sentences, and the first hidden state is used to predict whether the seed is real or fake.

The order of tokens is also important for constructing sentences. It is possible to predict whether a sentence’s order representation of a seed is correct because both real and fake seeds have a dimension of (*sequence length * embedding dimensions*). To do this, as shown in Figure 4b, we use Bidirectional LSTM $d_\beta(\cdot)$ called Seed Order Discriminator (SOD) to consider both forward and backward directions of sentences. The concatenation of the first and the last hidden states is used to predict whether the seed is real or fake.

During adversarial training, both discriminators are trained to predict whether the seeds are real (label 1) or fake (label 0), while the generator is trained to fool the discriminators by predicting fake seeds as 1. The loss function of the discriminators is defined by the following equation, which updates both the discriminators and the generator:

$$\mathcal{L}_D = -\frac{1}{N} \sum_{i=1}^N [y_i \log x_i + (1 - y_i) \log (1 - x_i)] \quad (8)$$

$x = \text{predicted}, y = \text{target}$

Additional information regarding the size and descriptions of the discriminators can be found in Appendix A.

3.4.2 Generator Helpers

During adversarial training, it is challenging for the generator to learn solely from the discriminators introduced in Section 3.4.1. Therefore, in this section, two

Text Embedding Space GAN for Text Synthesis

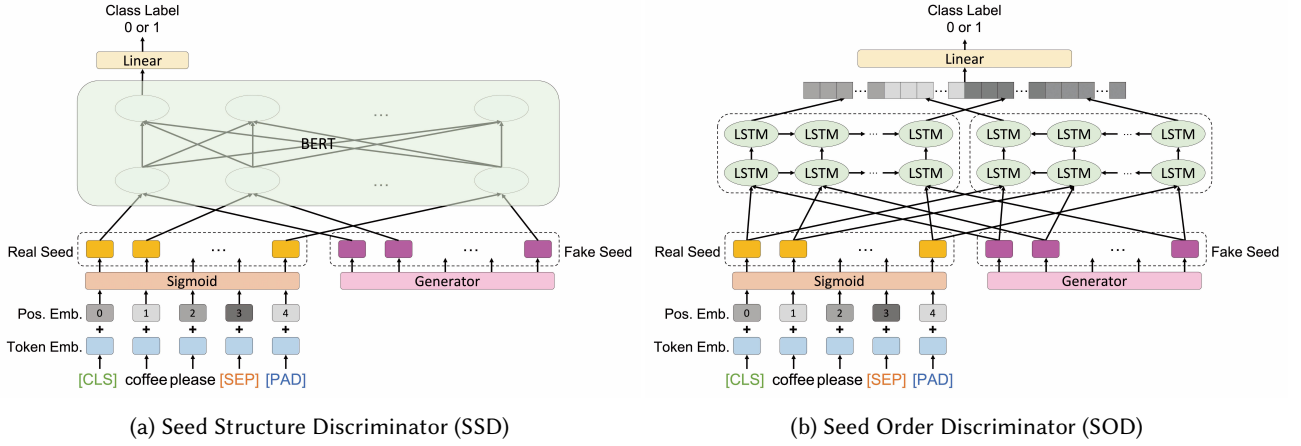


Figure 4: Illustrations of the two discriminators. SSD predicts whether the seed is real or fake using the $[CLS]$ special token’s feature. SOD considers both forward and backward contexts of the seed.

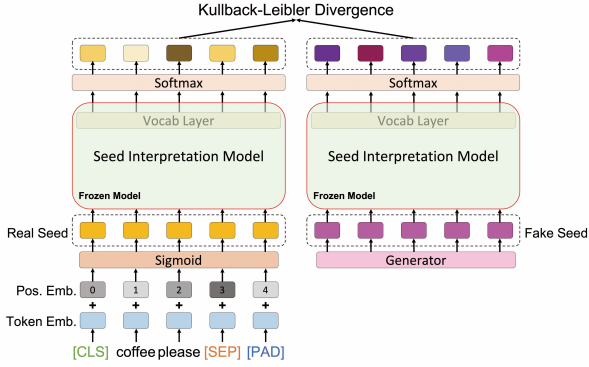


Figure 5: Illustration of Seed Distribution Prediction (SDP). SDP is used to enhance the fake seeds of the generator during adversarial training by minimizing the distance between real and fake seed distributions.

auxiliary tasks are introduced to aid the training of the generator.

Capturing the distribution of the text embedding space is important, and for this purpose, we employ Seed Distribution Prediction (SDP). However, since the text embedding space cannot be directly used as a probability distribution, the output of the seed interpretation model is utilized. Specifically, when a seed passes through the frozen seed interpretation model, the output dimension of $(sequence\ length * vocabulary\ size)$ is obtained through the softmax function, which can be used as a probability distribution. The loss is calculated using the Kullback-Leibler divergence between the distributions of the real and the fake seeds. SDP is used solely for updating the generator during adversarial training:

$$\mathcal{L}_{SDP} = \sigma(f_{\theta}^*(H_{real})) \log \frac{\sigma(f_{\theta}^*(H_{real}))}{\sigma(f_{\theta}^*(H_{fake}))} \quad (9)$$

where the σ and $f_{\theta}^*(\cdot)$ mean *softmax* function and the frozen seed interpretation model. More detailed figure of SDP is illustrated in Figure 5.

The sentences used as the seeds are composed of tokens explained in Equation 3. Additionally, we apply Seed Frame Prediction (SFP) since the structures of seeds are somewhat formalized. Therefore, we calculate the Mean Absolute Error (MAE) and Mean Squared Error (MSE) to make the form of a fake seed similar to a real one. If we train the fake seeds using MAE and MSE, the fake seeds from the generator can become blurred. However, the loss of SFP is relatively small compared to that of SSD, SOD, and SDP; therefore, SFP does not adversely affect the generator. SFP is used only for updating the generator during adversarial training:

$$\mathcal{L}_{SFP} = \|\mu_r - \mu_f\|_2^2 + \|H_{real} - H_{fake}\|_1 \quad (10)$$

$$\mu_r = avg(H_{real}), \mu_f = avg(H_{fake})$$

The full loss function, including the seed interpretation model’s loss, is described in Appendix B.

4 Text Synthesis Experiments

4.1 Dataset

In this experiment, we use two datasets consisting of multi-turn sentences to train the seed interpretation model and perform the text synthesis experiment.

DailyDialog³ (Li et al., 2017) is multi-turn conversation data used for training open-domain dialogue generation models. It consists of chit-chat-style multi-turn English conversations, and we select this data for domain-independent text synthesis. This dataset is used to evaluate the performance of TEGAN and other baselines. **IMDb**⁴ (Maas et al., 2011) contains highly polar movie

³<http://yanran.li/dailydialog>

⁴<https://huggingface.co/datasets/imdb>

Algorithm 1 Text Embedding Space GAN

Require: Seed interpretation model f_θ ; Generator g_ϕ ; BERT discriminator d_α ; LSTM discriminator d_β ; Multi-turn data $D = \{S_{1:N}\}$; Sentence data $S_i = \{w_{1:L}^i\}$.

- 1: Pre-train f_θ using D .
- 2: Initialize g_ϕ , d_α , d_β with random weights
 $\phi, \alpha, \beta \sim N(0, 0.08)$.
- 3: Freeze the f_θ .
- 4: **while** TESGAN converges **do**
- 5: **for** d-steps (during odd epoch) **do**
- 6: Get real data from f_θ using S with positive label 1.
- 7: Make fake data from g_ϕ with negative label 0.
- 8: Update α and β via results of d_α and d_β .
- 9: **end for**
- 10: **for** g-steps **do**
- 11: Make fake data from g_ϕ with positive label 1.
- 12: Calculate SDP and SFP.
- 13: Update ϕ via results of d_α , d_β , SDP and SFP.
- 14: **end for**
- 15: **end while**

reviews and is widely used for sentiment classification tasks. Each human-written movie review consists of several sentences, and we used this data as multi-turn data. This dataset is rougher and has a larger volume compared to DailyDialog. We evaluate the general applicability by synthesizing sentences based on IMDb-trained TEGAN. Statistics of the two datasets are shown in Appendix C.

4.2 Training Steps

TEGAN training has two steps. First, the seed interpretation model must be pre-trained with multi-turn data to interpret the seeds. In the performance and general applicability experiments, we train the model on the 11k and 25k multi-turn sets of DailyDialog and IMDb, respectively, as shown in Table 7. Then, the model that achieves the highest BLEU-4 score in the validation set is selected in each experiment.

The second step is adversarial training. After pre-training the seed interpretation model, the generator and the discriminator learn through adversarial training. For adversarial training, real and fake seeds are created by the embedding part of the frozen seed interpretation model and the generator, respectively. Since real seeds can be generated from a significant number of sentences, all 87k and 300k sentences in each training set used in the experiment mentioned above are used to create the real seeds via Equation 3. We also generate the same number of fake seeds as real seeds for adversarial training, and the following equation represents

what the discriminator and generator aim to optimize during adversarial training:

$$\begin{aligned} \mathcal{D}_{\mathcal{L}} &= \max_{\alpha, \beta} \mathbb{E}_{x \sim H_{real}} \left[\log d_{\alpha, \beta}(x) \right] \\ \mathcal{G}_{\mathcal{L}} &= \max_{\phi} \mathbb{E}_z \left[\log d_{\alpha, \beta}(g_\phi(z)) \right] + \mathcal{L}_{SDP} + \mathcal{L}_{SFP} \end{aligned} \quad (11)$$

where $d_{\alpha, \beta}$ means SSD, SOD respectively. $\mathcal{D}_{\mathcal{L}}$ implies updating the parameters of the discriminator to accurately predict real seeds as 1 from the perspective of real seeds. $\mathcal{G}_{\mathcal{L}}$ also means updating the generator so that the discriminator predicts the fake seeds created by the generator as 1. This approach helps partially resolve the learning imbalance problem between the generator and the discriminator (Goodfellow et al., 2014). Further discussion of the above pseudocode and optimization methods is covered in Section 6.1. Lastly, hyperparameters and experiment setup are described in Appendix D.

4.3 Evaluation Metric

Target-oriented evaluation metrics, such as BLEU and ROUGE (Lin, 2004), are not suitable for evaluating synthetic text. This is because each synthesized sentence from random noise has no corresponding target, and the generative models aim to synthesize plausible data based on real data distribution without copying the training data. Therefore, we employ several metrics that can evaluate unconditional text generation.

4.3.1 Fréchet BERT Distance (FBD)

de Masson d’Autume et al. (2019) proposed Fréchet Embedding Distance (FED) to evaluate the quality of synthetic text, inspired by Fréchet Inception Distance (FID) (Heusel et al., 2017). Alihosseini et al. (2019) proposed FBD, an improved version of FED, to measure the quality and diversity of synthesized text using a pre-trained BERT. The features of real and synthesized text obtained by the pre-trained BERT are assumed to have Gaussian distributions, and FBD is the distance between them:

$$FBD = \sqrt{\|\mu_r - \mu_f\|_2^2 + \text{tr}(\Sigma_r + \Sigma_f - 2(\Sigma_r \Sigma_f)^{0.5})} \quad (12)$$

where μ and Σ show the mean vectors and the covariance matrices of the real and fake seed features.

4.3.2 Multi-Sets-Jaccard (MSJ)

Each synthesized sentence has no corresponding target; thus, we select MSJ (Alihosseini et al., 2019), which calculates the score between real and synthesized text sets. The Jaccard Index determines the similarity of two sets, calculating the ratio of the cardinality of their intersection to that of their union. Inspired by the Jaccard

Index, MSJ) focuses on the similarity of the n-gram frequencies of text in the two sets, s_r and s_f , which are the real and synthesized text sets, respectively:

$$MSJ_n = \frac{\sum_{g \in G_n} \min(C_n(g, s_r), C_n(g, s_f))}{\sum_{g \in G_n} \max(C_n(g, s_r), C_n(g, s_f))} \quad (13)$$

where G_n and $C_n(g, s)$ mean the n-gram in $s_r \cup s_f$ and the normalized counts of the n-gram in set s . Additionally, this n-gram-based synthesized sentence evaluation method is a common approach in the field of unconditional text generation (Yu et al., 2017; Press et al., 2017; Fedus et al., 2018).

4.3.3 Language Model score (LM)

de Masson d’Autume et al. (2019); Caccia et al. (2020) proposed LM, which can evaluate the quality of generated samples using a well-trained language model. LM measures the quality of generated samples, meaning that scores of the bad samples are poor under a well-trained language model. We select the pre-trained GPT-2² as a well-trained language model. LM is calculated as the cross-entropy results between the output and input of GPT-2.

4.3.4 Data Synthesis Ratio (DSR)

DSR considers not only the data memorization ratio between the training and synthesized data but also synthetic diversity itself. Short sentences identical to training data, such as “I’m fine”, can be synthesized by coincidence. Therefore, sentences longer than two-thirds of the maximum sentence length that perfectly reproduce the training data are considered memorized data. Considering these conditions, we can calculate DSR using the following equation:

$$R_{syn} = \frac{|S_{syn} - S_{train}|}{|S_{syn}|}, R_{unq} = \frac{|S_{unq}|}{|S_{syn}|} \quad (14)$$

$$DSR = \frac{2 * R_{syn} * R_{unq}}{R_{syn} + R_{unq}}$$

where S_{syn} and S_{train} indicate synthesized and training text set respectively. S_{unq} means the set of unique sentences of synthesized text results. If the synthesized sentences in S_{syn} do not reproduce any of the sentences in S_{train} , R_{syn} would be 1. Similarly, if the synthesized text in S_{syn} is all unique, R_{unq} will be 1. The final DSR is calculated as the harmonic mean of R_{syn} and R_{unq} ratios.

4.3.5 Self-BLEU (SBL)

Zhu et al. (2018) first proposed SBL to measure diversity of token combination. The original BLEU evaluates the degree of n-gram overlap (similarity) between one hypothesis sentence and multiple reference sentences.

However, unconditionally generated text does not have specific targets, so it is not suitable for BLEU evaluation. SBL is widely used to solve this problem. SBL can evaluate n-gram-level similarity by regarding one sentence as a hypothesis and the rest as references in a synthetic text set. Since SBL evaluates based on the generated text set itself, it is not able to evaluate the quality of the synthetic text, but it is possible to evaluate the diversity of token combinations based on n-gram. Additionally, the difference between SBL and DSR lies in their evaluation criteria. DSR assesses data memorization by comparing the generated text set with the training dataset, while also considering the diversity of not n-gram-based but generated complete sentences themselves.

4.4 Baselines

In this paper, we compare our two models with the following approaches: LSTM-based Maximum Likelihood Estimation (MLE-L), PG-BLEU, SeqGAN, RankGAN, and MaliGAN. MLE-L represents the pre-training result of the generator, which all the other models undergo before adversarial training. The pre-trained generators with the lowest loss in the validation set were chosen for each method, including MLE-L. We compare these models with our original TEGAN and P-TEGAN, which is trained by adding zero-centered normal distribution noise z to the generator’s output. We also evaluate the GPT-2-based pre-trained seed interpretation model (MLE-G) used in the TEGAN framework. Since MLE-based models are trained without adversarial training, they are shown as baselines in Figure 6. Finally, to demonstrate that the results of the TEGAN-based models are not solely dependent on the seed interpretation model but rather on seeds created by the generator, we present the outcomes when using Gaussian random noise as input for the seed interpretation model.

5 Results

5.1 Metric-based Evaluation

In this section, we compare the results of the synthesized text at every epoch of adversarial training using the metrics mentioned in Section 4.3. This experiment was performed with models trained on the DailyDialog dataset. Since Fréchet BERT Distance (FBD) and Multi-Sets-Jaccard (MSJ) require a real text corpus, the test set is used as the real text corpus. Data Synthesis Ratio (DSR) is calculated with the training set as the data memorization ratio needs to be computed.

The FBDs of the TEGAN-based models are lower than the MLE-based autoregressive results, while the

Text Embedding Space GAN for Text Synthesis

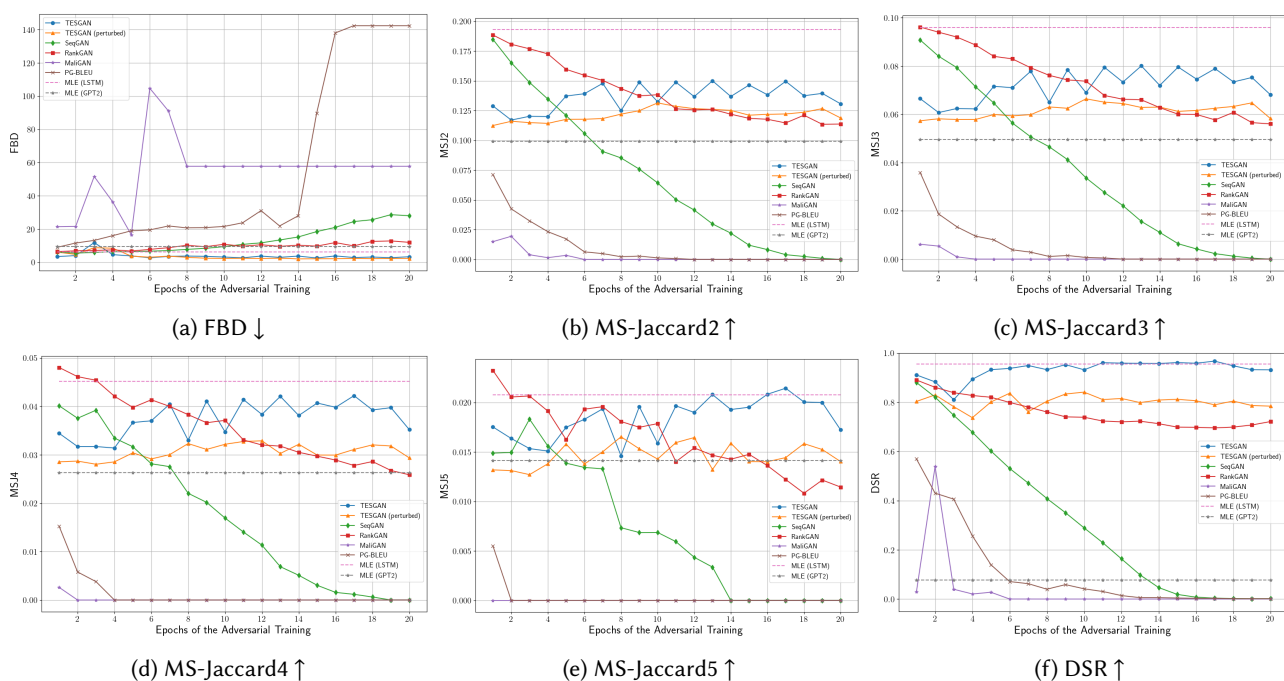


Figure 6: Illustration showing the results of the text-GAN models. In previous research, adversarial training is conducted after the generator pre-training. MLE is represented as a baseline because it is a supervised pre-trained generator without adversarial training.

Method	FBD ^{Q,D} ↓	MSJ4 ^{Q,D} ↑	MSJ5 ^{Q,D} ↑	DSR ^M (R_{syn}, R_{unq}) ↑	LM* ^Q ↓	SBL3* ^D ↓	SBL4* ^D ↓
TESGAN (ours)	2.899	0.042	0.021	0.967 (1, 0.936)	4.236	0.743	0.623
P-TESGAN (ours)	2.274	0.032	0.014	0.841 (0.997, 0.727)	3.642	0.790	0.702
SeqGAN (Yu et al., 2017)	6.153	0.040	0.015	0.880 (0.883, 0.877)	5.094	0.420	0.266
RankGAN (Lin et al., 2017)	6.409	0.048	0.023	0.890 (0.895, 0.886)	5.123	0.446	0.290
MaliGAN (Che et al., 2017)	21.436	0.003	0	0.030 (1, 0.015)	-	-	-
PG-BLEU (Yu et al., 2017)	9.002	0.015	0.006	0.569 (0.555, 0.584)	4.584	0.628	0.484
MLE-L (Yu et al., 2017)	6.284	0.045	0.021	0.955 (0.925, 0.987)	5.168	0.403	0.242
MLE-G	9.592	0.026	0.014	0.078 (1, 0.040)	3.543	0.948	0.944
Random Noise †	14.142	0.016	0.006	0.930 (1, 0.869)	4.562	0.516	0.404

Table 1: Performance of the models. P-TESGAN denotes the perturbed TEGAN. † is the result of directly entering Gaussian random noise as an input to the seed interpretation model. The second group of models consists of autoregressive models. * denotes a metric not considered when selecting the best model. - denotes that the confidence of the result is low because the quality of the synthesized sentence is poor. The superscript of each metric represents what each metric can measure (Q: quality, D: diversity, M: data memorization).

baselines increase after having the lowest value at the end of the first epoch of adversarial training, as shown in Figure 6a. In terms of MSJ, as shown in Figure 6b-6e, the previous studies report lower values than the TEGAN-based models at the end of adversarial training, despite having higher results in the beginning. On the other hand, MSJ results of the TEGAN-based models slightly increase during adversarial training. Moreover, some MSJ5 results of the original TEGAN are higher than MLE-L during adversarial training, as shown in Figure 6e. In the case of DSR, as shown in Figure 6f, the original TEGAN also increases during

adversarial training, and some results are higher than MLE-L. On the other hand, the results of the previous studies decrease during adversarial training, resulting in lower values than the TEGAN-based models in the end. As adversarial learning progresses, the results of the baselines deteriorate because the LSTM generator tends to generate only a few unique sentences. Furthermore, we will explain the reasons why the MLE-G results of the GPT-2 base are relatively poor in the following section.

We chose the best model of each method considering the FBD, MSJ, and DSR results because these met-

TESGAN (17-epoch, DailyDialog)	P-TESGAN (10-epoch, DailyDialog)	Random Noise
I'm so glad you finally got on the train. I just lost my job. Yeah. You mean the network connection? What happened? So you have to wait for a while.	Hello, Mr. Smith. I'm Mary. I just want to tell you the truth. It's the end of the world. What do you want to do in this company? He just broke up with Ann.	Anything I have called three weeks Is Is Is Is Is Is Left and go to go to go to go Mr Moon, Mr Moon...Mr Moon... are you have finished 6 items?
TESGAN (18-epoch, IMDb)		
This is probably one of the best of the best of the series. I was bored to think about how stupid this movie was. "The Deadly Loved One" is the story of a rebellious college basketball I have to say, this is the worst film I have ever seen. I was very excited to see it, anticipating Christmas eve. This movie was one of the best of the year for me.		

Table 2: Example of unconditionally synthesized sentences. P-TESGAN denotes the perturbed TESGAN.

rics evaluate quality, diversity, and data memorization. We evaluated the text generated by each model per epoch using the metrics mentioned above and compared the best-performing models⁵. According to Table 8 in Appendix E, we compare the baselines at 1-epoch with TESGAN and P-TESGAN at 17 and 10-epochs, respectively. After selecting the best models, we calculated the Language Model score (LM) and Self-BLEU (SBL) based on the text generated by each model. As shown in Table 1, the TESGAN-based models show the highest results in FBD and DSR. Also, the TESGAN-based models show comparable results in MSJ compared to the baselines and display the highest results among the adversarial-based methods in terms of LM score. However, in terms of SBL, TESGAN-based models perform worse than the baselines. In addition, the results of Gaussian random noise demonstrate that the TESGAN results are attributed to the seeds from the generator. The first group of Table 2 shows the synthetic text by TESGAN and Gaussian random noise. In conclusion, MLE-L is a supervised pre-trained generator applied before adversarial training, but most of the result curves of the prior methods showed lower performance than MLE-L during adversarial training. On the other hand, our TESGAN-based models showed better results than MLE-L or improved performance during adversarial training. Finally, the results according to the epoch of the LM and SBL of each model are shown in Appendix E.

5.2 Analysis of Autoregressive Models

In this section, we will analyze the results of the MLE-based autoregressive models. Other baseline models pretrain an LSTM-based generator before starting adversarial training, while the TESGAN framework employs a GPT-2-based pre-trained seed interpretation model. The results of the MLE-based models in Sec-

tion 5.1 are based on the evaluation of corpora generated in an autoregressive manner using the two pre-trained models. MLE-based models generate sentences in an autoregressive manner, starting from a specific [*start_token*] and predicting the next token. If the model predicts the next token in a greedy manner, all generated sentences would be identical, exhibiting deterministic behavior. To prevent this, MLE-based models sample the next token based on the probability logits (Yu et al., 2017). This way, MLE-L results in diverse token choices since the logit probability differences are not large. On the other hand, MLE-G training fits the data better than LSTM-based models, resulting in significantly larger differences in the logits of the next token. As a consequence, MLE-G is relatively deterministic compared to MLE-L. Therefore, when generating sentences without the use of the softmax temperature technique (Hinton et al., 2015), MLE-G delivers high quality, but it struggles to produce a variety of sentences. In practice, sentences generated by MLE-G lack diversity, which led to relatively lower results in Section 5.1. However, it is worth noting that while diversity may be lacking, the quality of the generated sentences is high, and this aspect will be demonstrated in the next section.

5.3 Human Evaluation

We conducted human evaluations based on the corpora generated by each model. The corpora, comprising 50 randomly selected unique sentences that do not duplicate those from the training set, were assessed by 10 annotators. We asked annotators to give higher scores to corpora that contained more natural sentences on a scale from 1 to 5. The scores presented in Table 3 represent the average scores assessed by each person. According to Table 3, TESGAN received the highest score, with MLE-G achieving the second-highest result. As mentioned in Section 5.2, MLE-G, despite facing challenges in generating diverse sentences, was able to pro-

⁵Detailed results are shown in Appendix E

Method	Avg. Score
TESGAN (ours)	4.2
P-TESGAN (ours)	3.4
SeqGAN (Yu et al., 2017)	2.4
RankGAN (Lin et al., 2017)	2.0
MaliGAN (Che et al., 2017)	1.0
PG-BLEU (Yu et al., 2017)	1.6
MLE-L (Yu et al., 2017)	3.0
MLE-G	3.8

Table 3: Human evaluation scores (1 ~ 5).

duce high-quality sentences.

5.4 General Applicability

In this section, we trained TESGAN with IMDB using a larger volume than DailyDialog to assess the general applicability of TESGAN. The IMDB-trained TESGAN is evaluated with both the DailyDialog test set (zero-shot) and the IMDB test set (non-zero-shot). Figures 7a and 7b display the zero-shot and non-zero-shot test results of the IMDB-trained TESGAN, respectively. Additionally, the zero-shot results generally exhibit a similar trend to the non-zero-shot test, suggesting that the model is being trained without bias toward the training data. Furthermore, the second group in Table 2 presents the synthetic text results of the IMDB-trained TESGAN. Both the zero-shot and text synthesis results indicate that TESGAN’s outcomes do not vary significantly depending on the dataset, implying that TESGAN generalizes well and can be trained on diverse datasets. Figure 7c also illustrates the DSR, LM, and SBL results of the IMDB-trained TESGAN. Since these metrics are evaluated not on the test set but on generated text data, they consistently yield results regardless of the zero-shot test.

5.5 Error Analysis

We also conducted three additional TESGAN trainings without setting a manual seed in the code to confirm reproducibility. To assess whether the sentences generated by the model for each trial converge as adversarial training progresses, we calculated the Standard Error of the Mean (SEM) based on the average results. SEM is equivalent to the standard deviation of a sample mean taken from a population and represents the standard deviation that indicates the extent of variability in sample means. SEM is calculated by $\frac{\sigma}{\sqrt{n}}$ (σ and n denote average results and the number of trials). As a result, the overall tendency of training outcomes during adversarial training is similar. Furthermore, the SEM of each epoch decreases during adversarial training, indicating that each TESGAN converges. Figure 8 displays

the results of the four experiments, including the average results and SEM.

6 Discussion

6.1 Generator and Training Strategy

We found that the performance of the TESGAN framework depends on the generator’s architecture. When ReLU was used, dying ReLU (Lu et al., 2019) occurred, where the negative values became zero, making it unsuitable for text synthesis where diversity is important. Additionally, the hyperbolic tangent (tanh) was not adequate due to the problem of gradient vanishing (Wang et al., 2019). Consequently, we adopted Leaky ReLU (Maas et al., 2013) as the activation function between two convolutional layers of the generator. Furthermore, deep structures and batch normalization tended to result in monotonous text synthesis. Therefore, we designed the generator’s layers to be wide rather than deep without batch normalization.

We also observed that the convergence of TESGAN depends on the parameter update rate of the discriminators and the generator. As in Algorithm 1, the discriminators update their parameters only during odd training epochs to allow the generator to catch up with the discriminator’s learning because the convergence of the generator is commonly slower than that of the discriminator. When the discriminators updated their parameters at every training epoch, the same as the generator, adversarial training became unbalanced. Additionally, we conducted further experiments by changing the update frequency of the generator from once to three times per mini-batch step. When the generator updated only once per step, the same as the discriminators, it could not keep up with the learning of the discriminators. On the other hand, when the generator updated three times per step, the discriminators could not keep up with the learning of the generator. Therefore, we chose to update the generator twice per step, resulting in the generator being updated four times more frequently per two epochs than the discriminators, as explained in Algorithm 1.

6.2 Ablation Study

In this section, we confirm the effect of the four objective functions in Section 3.4, and the results are shown in Table 4. When Seed Order Discriminator (SOD) and Seed Distribution Prediction (SDP) were not used, there was a significant difference in the results, indicating that SOD and SDP are important for high-quality text synthesis. Since MSJ evaluates text based on the n-gram of tokens, the order of the synthesized text is important. Accordingly, the MSJ results of the ”w/o SOD”

Text Embedding Space GAN for Text Synthesis

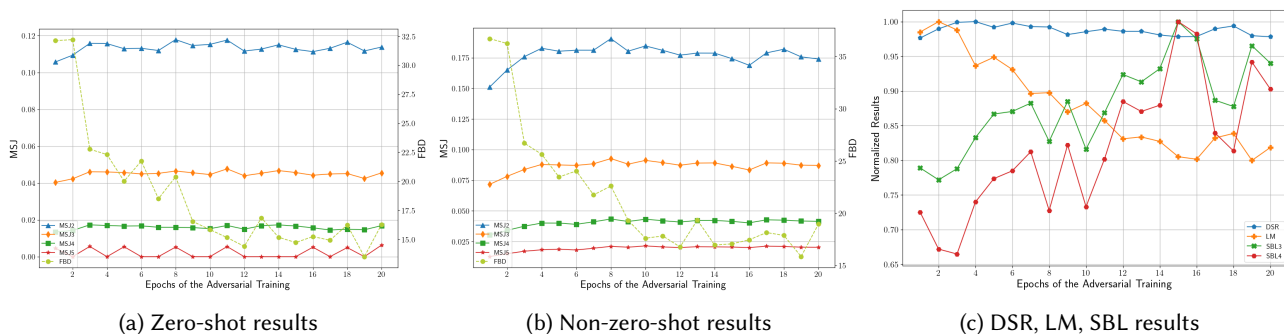


Figure 7: Zero-shot, non-zero-shot results of IMDb-trained TEGAN. DSR, LM, and SBL results of IMDb-trained TEGAN are normalized for ease of viewing.

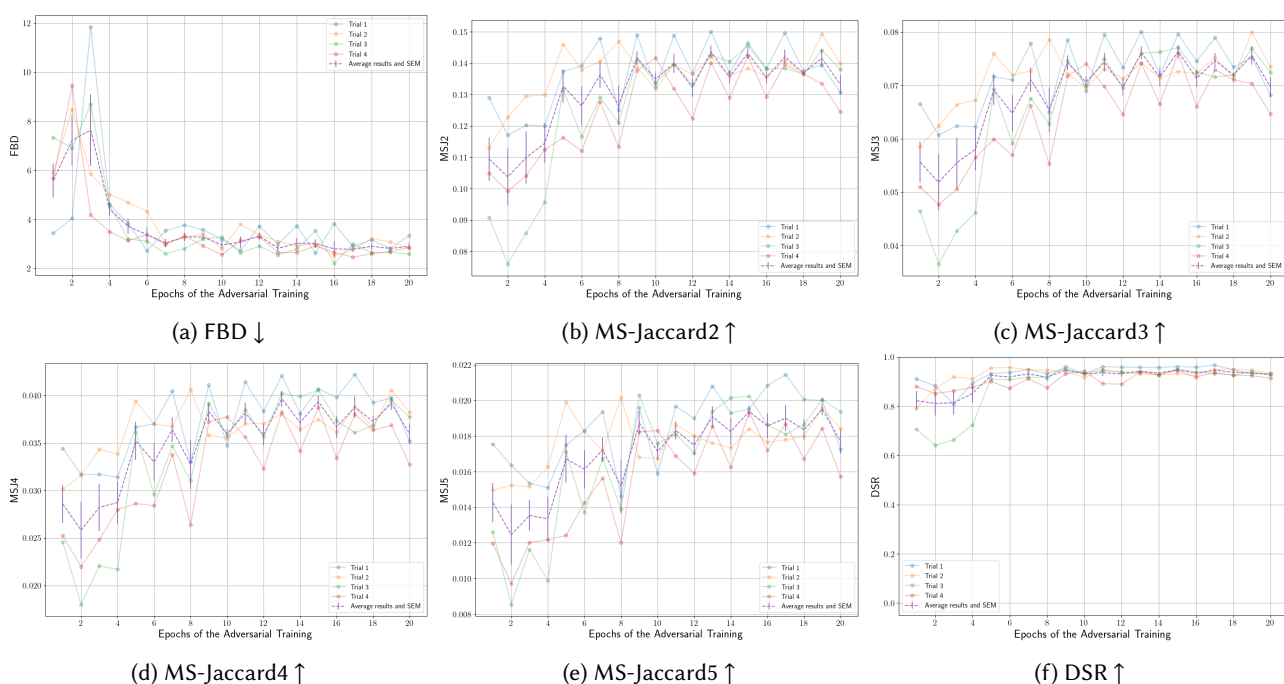


Figure 8: Illustration of TEGAN training results. TEGANs show similar trends for every trial, and SEMs decrease during adversarial training.

in Table 4 are worse than those of the "w/o Seed Structure Discriminator (SSD)", which proves that SOD can capture the token order representations. The four objective functions were used to achieve a good overall result, demonstrating that each of the four objective functions is playing a unique role.

6.3 Activation Function Study

The results varied depending on the activation functions used at the end of the seed-making process. We conducted experiments on sigmoid, tanh, and non-use cases during the seed-making process, and their results are shown in Table 5. Table 6 shows the quality of the synthetic text for tanh and non-use cases, and the results are worse than those using sigmoid in Table 2.

However, according to Table 5, the DSR results of the non-use case are higher than the sigmoid case. Thus, we can see that a higher DSR does not always mean good quality because DSR only considers data memorization. Therefore, we select the model using the sigmoid activation function, which has better results for FBD and MSJ, and moderately high DSR.

6.4 Data Memorization Study

The pre-trained GPT-2, which has 124M parameters and is used as the seed interpretation model, has been trained on relatively large corpora. Therefore, we need to confirm whether the low data memorization comes from transfer learning or the TEGAN framework. We trained three smaller seed interpretation models from

Method	FBD ↓	MSJ2 ↑	MSJ3 ↑	MSJ4 ↑	MSJ5 ↑	DSR	LM* ↓
TESGAN w/o SSD (13)	2.8346	0.1377	0.0744	0.0394	0.0204	0.9497	4.1833
TESGAN w/o SOD (2)	4.7724	0.1125	0.0596	0.03115	0.0164	0.8656	4.7011
TESGAN w/o SDP (13)	38.1301	0.0580	0.0252	0.0099	0.0041	0.7390	-
TESGAN w/o SFP (16)	2.9202	0.1477	0.0790	0.0422	0.0209	0.9463	4.2339
TESGAN (17)	2.8994	0.1496	0.0789	0.0422	0.0214	0.9669	4.2361

Table 4: Results of the ablation study. Numbers in parentheses indicate the training epoch of the selected model. * denotes a metric not considered when selecting the best model. - denotes that the confidence of the result is low because the quality of the synthesized sentence is poor.

Activation	FBD ↓	MSJ2 ↑	MSJ3 ↑	MSJ4 ↑	MSJ5 ↑	DSR ↑	LM* ↓
TESGAN	None	47.261	0.108	0.060	0.032	0.016	0.982
	Tanh	9.780	0.110	0.057	0.030	0.015	0.871
	Sigmoid	2.899	0.150	0.079	0.042	0.021	0.967
P-tesGAN	None	54.002	0.111	0.059	0.030	0.015	0.958
	Tanh	20.158	0.118	0.061	0.031	0.015	0.937
	Sigmoid	2.274	0.131	0.066	0.032	0.014	0.841

Table 5: Performance according to the activation functions of the generator. * denotes a metric not considered when selecting the best model. - denotes that the confidence of the result is low because the quality of the synthesized sentence is poor.

TESGAN with Tanh	P-tesGAN with Tanh
You are a little I ' m sorry to see you off. You ' Ve come I ' m sorry. You ' d like a tour to see the dentist. You are late.	You ' re a book? You are late. I ' m doing ' t " all day ' s I don ' t know what time it is? I ' m sorry to hear this!
TESGAN without activation	P-tesGAN without activation
I ' d like to say it! I ' d like to Yes, do you want to buy? I ' s right over there? What ' s the matter? I got a bite the food?	I ' s a big, that ' s right. I like the back ones. They look like a shop. I ' , this ' , this ' ! be real, I have a problem with my English textbooks. I ' s faster, George. I ' d like to go

Table 6: Synthesized sentences by tanh and non-use cases in Table 5. P-tesGAN denotes the perturbed TESGAN.

scratch to measure the data memorization and they have 54M, 75M, and 96M parameters each. As shown in Figure 9, DSR is high regardless of the number of model parameters during adversarial training, indicating that the low data memorization comes from the TESGAN framework.

7 Conclusion

In this work, we proposed a novel unsupervised text synthesis framework, TESGAN. TESGAN facilitated the gradient backpropagation of natural language discrete tokens by creating a continuous text embedding space called a seed. In most text-GAN studies, data memorization had been inevitable because the generator had to be pre-trained with an autoregressive approach be-

fore adversarial training. Therefore, we introduced TESGAN, which mitigated the data memorization issue by applying an unsupervised GAN framework that does not directly refer to the training data. TESGAN improved text synthesis performance during adversarial training and resulted in the best or comparable results in terms of evaluation metrics. Additionally, TESGAN exhibited the lowest data memorization ratio, and the data memorization study confirmed that these results were attributable to the TESGAN framework. Furthermore, TESGAN achieved the highest scores in human evaluations. The ablation study highlighted the importance of the four objective functions, and the synthetic text results from a large dataset-trained TESGAN demonstrated its general applicability. This paper underscores the potential of continuous embedding spaces in conjunction with discrete tokens for text

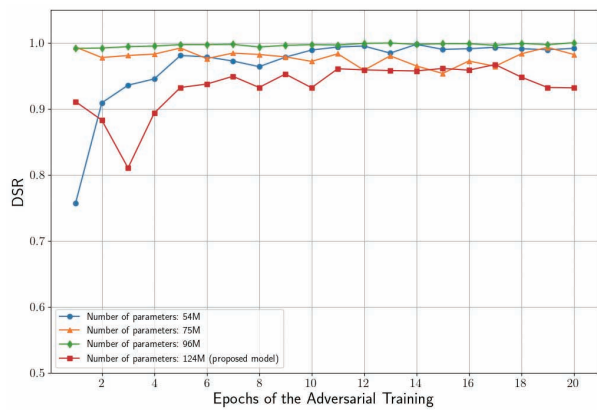


Figure 9: DSR results according to scales of seed interpretation model.

synthesis through unsupervised learning. By integrating the concept of viewing text as a continuous space with publicly available Large Language Models (Touvron et al., 2023), models can synthesize more expressive sentences, and we anticipate that many follow-up studies will emerge.

References

- Alayrac, Jean-Baptiste, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: a visual language model for few-shot learning.
- Alihosseini, Danial, Ehsan Montahaei, and Mahdieh Soleymani Baghshah. 2019. Jointly measuring diversity and quality in text generation models. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 90–98, Minneapolis, Minnesota. Association for Computational Linguistics.
- Antoniou, Antreas, Amos Storkey, and Harrison Edwards. 2018. Data augmentation generative adversarial networks.
- Arjovsky, Martin, Soumith Chintala, and Léon Bottou. 2017. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223. PMLR.
- Bowles, Christopher, Liang Chen, Ricardo Guerrero, Paul Bentley, Roger Gunn, Alexander Hammers, David Alexander Dickie, Maria Valdés Hernández, Joanna Wardlaw, and Daniel Rueckert. 2018. Gan augmentation: Augmenting training data using generative adversarial networks.
- Bowman, Samuel R., Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21, Berlin, Germany. Association for Computational Linguistics.
- Caccia, Massimo, Lucas Caccia, William Fedus, Hugo Larochelle, Joelle Pineau, and Laurent Charlin. 2020. Language gans falling short. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26–30, 2020*. OpenReview.net.
- Che, Tong, Yanran Li, Ruixiang Zhang, R Devon Hjelm, Wenjie Li, Yangqiu Song, and Yoshua Bengio. 2017. Maximum-likelihood augmented discrete generative adversarial networks.
- Chen, Liqun, Shuyang Dai, Chenyang Tao, Dinghan Shen, Zhe Gan, Haichao Zhang, Yizhe Zhang, Ruiyi Zhang, Guoyin Wang, and Lawrence Carin. 2018. Adversarial text generation via feature-mover’s distance. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems, NIPS’18*, page 4671–4682, Red Hook, NY, USA. Curran Associates Inc.
- Chung, Hyung Won, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.
- Dai, Wenliang, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning.
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Diao, Shizhe, Xinwei Shen, Kashun Shum, Yan Song, and Tong Zhang. 2021. TILGAN: Transformer-based implicit latent GAN for diverse and coherent text generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4844–4858, Online. Association for Computational Linguistics.
- Fedus, William, Ian J. Goodfellow, and Andrew M. Dai. 2018. Maskgan: Better text generation via filling in the _____. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.
- Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.
- Graves, Alex. 2013. Generating sequences with recurrent neural networks. *CoRR*, abs/1308.0850.

- Guo, Jiaxian, Sidi Lu, Han Cai, Weinan Zhang, Yong Yu, and Jun Wang. 2018. Long text generation via adversarial training with leaked information. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.
- Heusel, Martin, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Hinton, Geoffrey, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*.
- Hochreiter, Sepp and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.
- Jang, Eric, Shixiang Gu, and Ben Poole. 2017. Categorical reparameterization with gumbel-softmax. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.
- Karras, Tero, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2021. Alias-free generative adversarial networks.
- Karras, Tero, Samuli Laine, and Timo Aila. 2018. A style-based generator architecture for generative adversarial networks.
- Kingma, Diederik P. and Jimmy Ba. 2014. Adam: A method for stochastic optimization. Cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.
- Kingma, Diederik P. and Max Welling. 2014. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.
- Kullback, S. and R. A. Leibler. 1951. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79 – 86.
- Lester, Brian, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Li, Junnan, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models.
- Li, Yanran, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.
- Lin, Chin-Yew. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Lin, Kevin, Dianqi Li, Xiaodong He, Zhengyou Zhang, and Ming-Ting Sun. 2017. Adversarial ranking for language generation. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS’17*, page 3158–3168, Red Hook, NY, USA. Curran Associates Inc.
- Lu, Lu, Yeonjong Shin, Yanhui Su, and George Em Karniadakis. 2019. Dying relu and initialization: Theory and numerical examples. *ArXiv*, abs/1903.06733.
- Maas, Andrew L., Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.
- Maas, Andrew L., Awni Y. Hannun, and Andrew Y. Ng. 2013. Rectifier nonlinearities improve neural network acoustic models. In *in ICML Workshop on Deep Learning for Audio, Speech and Language Processing*.
- de Masson d’Autume, Cyprien, Mihaela Rosca, Jack W. Rae, and Shakir Mohamed. 2019. Training language gans from scratch. In *Neural Information Processing Systems*.
- Nie, Weili, Nina Narodytska, and Ankit Patel. 2019. Relgan: Relational generative adversarial networks for text generation. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.
- Ouyang, Long, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training

- language models to follow instructions with human feedback.
- Papineni, Kishore, Salim Roukos, Todd Ward, and Weijing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Press, Ofir, Amir Bar, Ben Bogin, Jonathan Berant, and Lior Wolf. 2017. Language generation with recurrent generative adversarial networks without pre-training. *CoRR*, abs/1706.01399.
- Radford, Alec, Luke Metz, and Soumith Chintala. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks.
- Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2018. Language models are unsupervised multitask learners.
- Sandfort, Veit, Ke Yan, Perry Pickhardt, and Ronald Summers. 2019. Data augmentation using generative adversarial networks (cycleGAN) to improve generalizability in ct segmentation tasks. *Scientific Reports*, 9.
- Santoro, Adam, Ryan Faulkner, David Raposo, Jack Rae, Mike Chrzanowski, Théophane Weber, Daan Wierstra, Oriol Vinyals, Razvan Pascanu, and Timothy Lillicrap. 2018. Relational recurrent neural networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems, NIPS’18*, page 7310–7321, Red Hook, NY, USA. Curran Associates Inc.
- Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.
- Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.
- Tran, Ngoc-Trung, Viet-Hung Tran, Ngoc-Bao Nguyen, Trung-Kien Nguyen, and Ngai-Man Cheung. 2021. On data augmentation for gan training. *Trans. Img. Proc.*, 30:1882–1897.
- Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Wang, Xin, Yi Qin, Yi Wang, Sheng Xiang, and Haizhou Chen. 2019. Reltanh: An activation function with vanishing gradient resistance for sae-based dnns and its application to rotating machinery fault diagnosis. *Neurocomputing*, 363:88–98.
- Wei, Jason, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2021. Finetuned language models are zero-shot learners. *CoRR*, abs/2109.01652.
- Wen, Tsung-Hsien, Milica Gašić, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Semantically conditioned LSTM-based natural language generation for spoken dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1711–1721, Lisbon, Portugal. Association for Computational Linguistics.
- Williams, Ronald J. and David Zipser. 1989. A learning algorithm for continually running fully recurrent neural networks. *Neural Computation*, 1(2):270–280.
- Yu, Lantao, Weinan Zhang, Jun Wang, and Yong Yu. 2017. Seqgan: Sequence generative adversarial nets with policy gradient. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1).
- Zhang, Yizhe, Zhe Gan, Kai Fan, Zhi Chen, Ricardo Henao, Dinghan Shen, and Lawrence Carin. 2017. Adversarial feature matching for text generation. In *International Conference on Machine Learning*, pages 4006–4015. PMLR.
- Zhu, Yaoming, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texus: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR ’18*, page 1097–1100, New York, NY, USA. Association for Computing Machinery.

A Details of Models

A.1 Seed Interpretation Model

The seed interpretation model $f_{\theta}(\cdot)$ is necessary to predict subsequent sentences a seed makes. Therefore, the seed interpretation model must be trained with multi-turn sentences in an autoregressive way. Our interpretation model inherits the 12-layer GPT-2, derived from the decoder of the transformer language model (Vaswani et al., 2017), and has 124M parameters. We used the model achieving the highest NLTK BLEU-4⁶ in the validation set of each dataset in Table 7 as the seed interpretation model.

A.2 Generator

The generator $g_{\phi}(\cdot)$ consists of two 1D transposed convolutional layers and has 3.3M parameters. The first and the second layers conduct convolution with 128 and 16 filters, respectively. Since sentences vary according to types of tokens and their order, forms of the real seed H_{real} are also varied. The generator generates seeds from the uniform distribution noise X with an interval of $[-10, 10]$ to make diverse forms of the seeds. Furthermore, the fake seeds generated by the deep convolutional layers and batch normalization results tend to synthesize only monotonous sentences. Thus, layers of the generator are constructed not deeply but widely and the generator does not have batch normalization layers. Leaky ReLU is used as the activation function between the two convolutional layers.

A.3 Seed Structure Discriminator (SSD)

Sentence structure is important for constructing a complete sentence. For example, “*I love you so much*” is structurally error-free, but “*I love like so much*” and “*I love*” are not. Because real seeds are created from perfect sentences, they retain the structural representation of sentences. Therefore it is important that the fake seeds should capture the structural features of the real seeds. We assume that every sentence can be the first sentence in multi-turn cases. Thus, the real seeds are obtained from sentences where the [CLS] token is inserted at the beginning like Equation 3. We use the 2-layer BERT $d_{\alpha}(\cdot)$ to capture the structural features of sentences, and the [CLS] token’s feature is used to predict whether the seed is real (label 1) or fake (label 0). In addition, real and fake seeds do not pass through the embedding part of the BERT because they are already embedding spaces. Finally, the BERT used in SSD has 54M parameters.

⁶https://www.nltk.org/_modules/nltk/translate/bleu_score.html

A.4 Seed Order Discriminator (SOD)

The order of tokens is important for constructing sentences. For example, “*I love you so much*” is syntactically correct, but “*I you love so much*” and “*I love you much so*” are not. We use a 2-layer Bidirectional LSTM to consider both forward and backward directions of sentences and the model has 24M parameters. The concatenated hidden states of the last token ([SEP] or [PAD]) and the first token ([CLS]) are used to predict whether the seed is real (label 1) or fake (label 0).

B Loss Function

Here, we show whole loss functions of TEGAN:

Seed Interpretation Model :

$$\mathcal{L}_{LM} = -\frac{1}{N} \sum_{n=1}^N \log \frac{\exp(x_n, y_n)}{\sum_{c=1}^C \exp(x_n, c)}$$

Adversarial Training Training

d - step :

$$\mathcal{L}_D = -\frac{1}{N} \sum_{i=1}^N \left\{ [y_i \log x_i + (1 - y_i) \log (1 - x_i)]_{SSD}^{real/fake} + [y_i \log x_i + (1 - y_i) \log (1 - x_i)]_{SOD}^{real/fake} \right\}$$

g - step :

$$\mathcal{L}_G = -\frac{1}{N} \sum_{i=1}^N \left\{ [y_i \log x_i + (1 - y_i) \log (1 - x_i)]_{SSD}^{fake} + [y_i \log x_i + (1 - y_i) \log (1 - x_i)]_{SOD}^{fake} + \sigma(f_{\theta}(H_{real})) \log \frac{\sigma(f_{\theta}(H_{real}))}{\sigma(f_{\theta}(H_{fake}))} + \|\mu_r - \mu_f\|_2^2 + \|H_{real} - H_{fake}\|_1 \right\}$$

(15)

The first loss function in Equation 15 is cross-entropy and is used to train the seed interpretation model. The loss functions used in adversarial training operate differently in the discriminator and generator steps. In the discriminator step (d-step), the loss function is designed to train the discriminator to distinguish between real and fake seeds, predicting them as 1 and 0, respectively. On the other hand, in the generator step (g-step), the loss function aims to train the generator to predict fake seeds as 1. Additionally, SDP and SFP losses are added to assist the generator learning during the g-step.

C Statistics of Datasets

Table 7 shows the statistics of the two datasets used in this paper. We excluded single-turn reviews when constructing the IMDb multi-turn dataset. For the baseline performance experiments, we generated fake seeds equal to the number of sentences in the DailyDialog dataset to conduct TEGAN training (adversarial

training). Similarly, for the general applicability experiments, we generated nearly 300k fake seeds to conduct the experiments with IMDb datasets.

D Hyperparameters

The TESGAN framework has two training steps. The first step is seed interpretation model training. The multi-turn data for seed interpretation model training were limited to a maximum of four and eight turns in performance (DailyDialog-trained) and general applicability (IMDb-trained) experiments, respectively. Also, the maximum length of the sentence was set to 16 and 32 for each experiment (total sequence length of each experiment was 64 and 256). The 12-layer GPT-2 is used as the seed interpretation model, and both hidden and embedding dimensions are 768. We adopted the byte-pair-encodings (BPE) (Sennrich et al., 2016) tokenizer with 50,260 vocabularies in the seed interpretation model. We used the Adam optimizer (Kingma and Ba, 2014) with $1e^{-3}$ learning rate to train the seed interpretation model and set the mini-batch size to 100.

In the adversarial training phase, the sentences used as the seeds are composed of tokens explained in Equation 3, and the length of each sentence is set to 16 including special tokens. The discriminators are updated by the loss of SSD and SOD during adversarial training. Also, the generator is updated not only by the loss of SSD and SOD but also that of SDP and SFP.

The fake seeds are generated from the uniform distribution noise X with an interval of $[-10, 10)$ by the generator, which has two convolutional processes. In addition, the Leaky ReLU with slope 0.5 and the sigmoid are used in the middle and the end of the generator, respectively. We used the Adam optimizer with $2e^{-4}$ learning rate when training DailyDialog because the generator has difficulty converging when the learning rate exceeds $4e^{-4}$. However, when training on IMDb, a larger dataset than DailyDialog, we set the learning rate to $5e^{-4}$. The BERT and the LSTM models, used as SSD and SOD respectively, consist of two layers and 768 hidden dimensions. Both discriminators used the Adam optimizer with $5e^{-4}$ and $1e^{-3}$ learning rate, respectively. When the learning rates of the discriminators were larger than the proposed values, adversarial learning was imbalanced. Also, the mini-batch size was set to 128 during adversarial training. Lastly, all the above experiments took place on a machine with Ubuntu 18.04.5 and an NVIDIA RTX 3090 GPU.

E Additional Results

We provide evaluation results of the text generated by each model per epoch. Table 8 shows the results of 1,

5, 10, 15, 17, and 20-epoch results of each model. Also, Figure 10 shows LM and SBL results of the TESGAN-based models and the baselines. In Figure 10, the SBL results of the baselines tend to increase.

Text Embedding Space GAN for Text Synthesis

Statistics	DailyDialog			IMDb		
	Train	Validation	Test	Train	Validation	Test
# of multi-turn set	11,118	1,000	1,000	24,890	12,500	12,390
Total sentences	87,170	8,069	7,740	299,137	150,369	148,768
Avg. # of turns per set	7.84	8.07	7.74	12.02	12.03	12.01
Avg. # of words per sentence	11.30	11.21	11.44	19.34	19.40	19.28
Avg. # of tokens per sentence	14.51	14.39	14.69	24.25	24.31	24.20

Table 7: Statistics of datasets

Method	Epoch	FBD ↓	MSJ2 ↑	MSJ3 ↑	MSJ4 ↑	MSJ5 ↑	DSR (R_{syn}, R_{unq}) ↑
TESGAN	1	3.441	0.129	0.067	0.034	0.018	0.911 (1, 0.836)
	5	3.826	0.137	0.072	0.037	0.018	0.932 (1, 0.873)
	10	3.185	0.132	0.069	0.035	0.016	0.932 (0.999, 0.873)
	15	2.624	0.146	0.080	0.041	0.020	0.961 (1, 0.925)
	17	2.899	0.150	0.079	0.042	0.021	0.967 (1, 0.936)
	20	3.339	0.131	0.068	0.035	0.017	0.932 (1, 0.872)
P-TESGAN	1	6.146	0.112	0.057	0.029	0.013	0.803 (1, 0.671)
	5	3.746	0.118	0.060	0.030	0.016	0.801 (0.998, 0.669)
	10	2.274	0.131	0.066	0.032	0.014	0.841 (0.997, 0.727)
	15	2.132	0.121	0.061	0.030	0.014	0.812 (1, 0.683)
	17	2.274	0.122	0.063	0.031	0.014	0.789 (0.998, 0.653)
	20	2.309	0.119	0.058	0.029	0.014	0.784 (0.997, 0.646)
SeqGAN	1	6.153	0.185	0.091	0.040	0.015	0.880 (0.883, 0.877)
	5	6.373	0.121	0.065	0.032	0.014	0.602 (0.639, 0.568)
	10	9.432	0.064	0.034	0.017	0.007	0.289 (0.34, 0.251)
	15	18.471	0.012	0.006	0.003	0	0.019 (0.025, 0.015)
	17	24.504	0.004	0.002	0.001	0	0.004 (0.006, 0.003)
	20	28.048	0	0	0	0	0.001 (0.013, 0.001)
RankGAN	1	6.409	0.189	0.096	0.048	0.023	0.890 (0.895, 0.886)
	5	6.778	0.160	0.084	0.04	0.016	0.82 (0.851, 0.791)
	10	10.862	0.138	0.074	0.037	0.018	0.739 (0.799, 0.687)
	15	9.732	0.118	0.060	0.030	0.015	0.699 (0.791, 0.625)
	17	9.893	0.115	0.058	0.028	0.012	0.696 (0.792, 0.62)
	20	12	0.114	0.056	0.026	0.011	0.721 (0.836, 0.634)
MaliGAN	1	21.436	0.015	0.006	0.003	0	0.030 (1, 0.015)
	5	16.589	0.003	0	0	0	0.027 (1, 0.014)
	10	57.769	0	0	0	0	0 (1, 0)
	15	57.769	0	0	0	0	0 (1, 0)
	17	57.769	0	0	0	0	0 (1, 0)
	20	57.769	0	0	0	0	0 (1, 0)
PG-BLEU	1	9.002	0.071	0.036	0.015	0.006	0.569 (0.555, 0.584)
	5	18.974	0.017	0.008	0	0	0.139 (0.472, 0.082)
	10	21.568	0.001	0.001	0	0	0.041 (0.792, 0.021)
	15	89.764	0	0	0	0	0.004 (0.773, 0.002)
	17	142.384	0	0	0	0	0.003 (1, 0.001)
	20	142.384	0	0	0	0	0.001 (1, 0.001)

Table 8: Performance of each model per epoch.

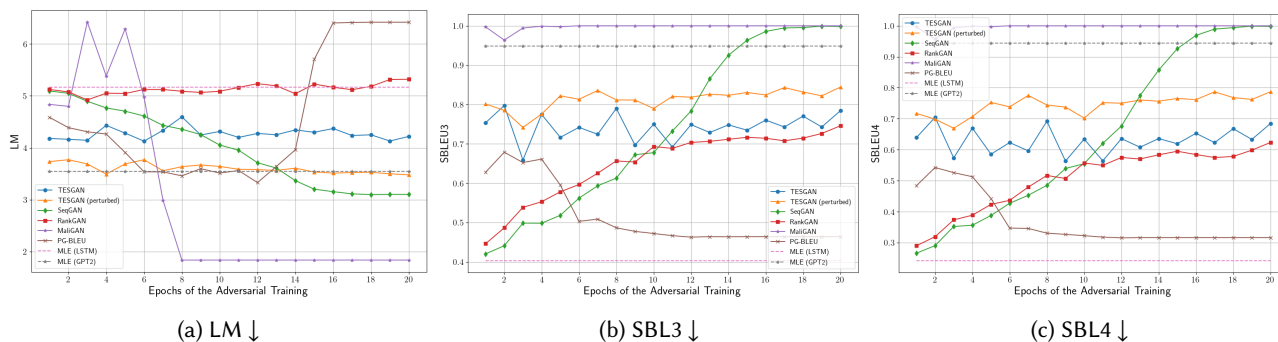


Figure 10: LM, SBL results of TEGSAN-based models and baselines trained with DailyDialog.

QUA-RC: the semi-synthetic dataset of multiple choice questions for assessing reading comprehension in Ukrainian

Mariia Zyrianova, KTH Royal Institute of Technology, Stockholm, Sweden mariiaz@kth.se

Dmytro Kalpakchi, KTH Royal Institute of Technology, Stockholm, Sweden dmytroka@kth.se

Abstract In this article we present the first dataset of multiple choice questions (MCQs) for assessing reading comprehension in Ukrainian. The dataset is based on the texts from the Ukrainian national tests for reading comprehension, and the MCQs themselves are created semi-automatically in three stages. The first stage was to use GPT-3 to generate the MCQs zero-shot, the second stage was to select MCQs of sufficient quality and revise the ones with minor errors, whereas the final stage was to expand the dataset with the MCQs written manually. The dataset is created by the Ukrainian language native speakers, one of whom is also a language teacher. The resulting corpus has slightly more than 900 MCQs, of which only 43 MCQs could be kept as they were generated by GPT-3.

1 Introduction

Assessing reading comprehension is of interest both for the native speakers of any language (for instance, through PISA (OECD, 2019) assessments), and for the foreigners learning the language (e.g., through IELTS¹ for English, DELE² for Spanish, or DELF³ for French). In both cases the skills are frequently assessed on the same scale, namely the one proposed by the Common European Frame of Reference (CEFR; Council of Europe (2001)). One of the assessment formats recommended on any CEFR-level is multiple choice questions (MCQs), which consist of the following components:

- *stem*, typically a question inquiring about some information from the text;
- *key*, the correct answer for the stem;
- *distractors*, wrong but plausible options.

The key and the distractors together are called *alternatives*. Note that reading comprehension MCQs require carefully selected texts, which are absolutely crucial, since reading comprehension MCQs are not designed to stand on their own.

In practice, the assessment with MCQs is rather popular because it enables fast, automatic, and thus objective grading. On the other hand, creating MCQs is

comparatively slow and requires a lot of manual efforts, which motivated the research on NLP methods for generating MCQs automatically. As Ch and Saha (2018) report, researchers have tried different techniques for MCQ generation, ranging from the manually created pipelines to more recent methods based on learning from data. Indeed, the introduction of large language models (LLMs), such as BERT (Devlin et al., 2019), or GPT-3 (Brown et al., 2020), resulted in new approaches being tested for many NLP tasks, not least for MCQ generation, especially for English (Vachev et al., 2022; Raina and Gales, 2022; Dijkstra et al., 2022). By comparison, MCQ generation problem (particularly for reading comprehension) received much less attention in other languages, and especially in Ukrainian. In this work we aim to bridge the gap for Ukrainian by making the following contributions:

- We present the first (to the best of our knowledge) dataset of Ukrainian MCQs for reading comprehension called QUA-RC. The dataset contains more than 900 MCQs (for example, the English translation of one such MCQ is provided in Figure 1), and is designed with the Ukrainian-first mindset (instead of being a translation of another dataset). The texts are taken from the real-world Ukrainian reading comprehension tests, and the MCQs themselves are created semi-automatically using GPT-3 (zero-shot), followed by manual curation and then manual expansion of the dataset.

¹<https://www.ielts.org/>

²<https://www.dele.org/>

³<https://fiaf.org/exams/delf-dalf/>

- At the same time, we evaluate GPT-3 on the task of generating MCQs for reading comprehension in Ukrainian in a zero-shot manner. Our evaluation reveals extensive shortcomings of this approach with less than 10% of MCQs judged to be of sufficient quality.

Both the dataset, and the accompanying source code are available on GitHub: <https://github.com/dkalpakchi/QUA-RC>.

Text:
 [...] Is there at least one city in Ukraine that can be viewed as an example in these terms? "It is Lviv, which is a pioneer city and a role model for the whole country in the attitude towards animals. There is an excellent communal enterprise that registers pets, keeps a clear electronic account of homeless four-legged friends and tracks their number," says Oleksandra Mezinova, head of the Kyiv animal shelter.

Stem:
Which Ukrainian city is seen as exemplary in its attitude to animals?

Alternatives:
 (A) Kyiv
(B) Lviv
 (C) Kharkiv
 (D) Zaporizhzhia

Figure 1: An example MCQ with an accompanying text from the collected QUA-RC dataset (translated from Ukrainian into English). The alternative in **bold** denotes the key, whereas all the other alternatives (in this case (A), (C), and (D)) denote the distractors.

2 Related work

To the best of our knowledge there has been no prior work on creating datasets of MCQs specifically for Ukrainian first, let alone semi-automatically.

In parallel with this work [Bandarkar et al. \(2023\)](#) have developed Belebele benchmark where they have created a parallel reading comprehension dataset in 122 languages, with Ukrainian being among them. The texts and MCQs in the dataset have been manually translated from English with reportedly rigorous curation process. The texts for this dataset were taken from three sources: WikiNews, WikiVoyage and WikiBooks. By their nature, such texts contain mostly facts, lacking, for instance, literary devices or dialogues, and often

appear additionally structured (compared to narrative texts) for ease of reading. Moreover, the translations for the dataset were produced to maximize the alignment between 122 languages, which could lead to the increased use of Translationese ([Gellerstam, 1986](#)), as the authors themselves note. By contrast, the texts used in our dataset are taken directly from the Ukrainian national tests for reading comprehension, meaning they are guaranteed to not contain Translationese, and are considered to be of suitable quality by the experts.

The translated datasets in Ukrainian are scarce even when looking at the broader field of Question Answering. The only work that we are aware of is an attempt at translating the SQuAD dataset ([Rajpurkar et al., 2016](#)) to Ukrainian⁴. However, it is unclear to what extent the translations have been curated, and the dataset contains no distractors (similar to the original SQuAD).

In general, the idea of creating synthetic QA datasets is not new, and has been rejuvenated by the advent of Large Language Models (LLMs). For instance, [Alberti et al. \(2019\)](#) produced synthetic question-answer pairs by using three different BERT ([Devlin et al., 2019](#)) models fine-tuned on SQuAD2 ([Rajpurkar et al., 2018](#)) to perform three different tasks: (1) extract the potential answer, (2) generate the question for that answer, and (3) answer this new question to check for the roundtrip consistency and filter-out the inconsistent questions.

The idea of creating synthetic MCQ datasets is not new either. For instance, [Kalpakchi and Boye \(2023\)](#) generated MCQs using OpenAI’s GPT-3 ([Brown et al., 2020](#)) in a zero-shot manner. After curating the output, 44% of MCQs turned out to be of acceptable quality. In this work we build on the work of [Kalpakchi and Boye \(2023\)](#) and expand it in the following ways:

- we perform our experiment in Ukrainian, which differs from English much more than Swedish, in multiple ways: (1) it uses a different script, (2) it is characterised by a relaxed word order, and (3) it is more morphologically complex;
- our prompt attempts for a fine-grained control by requesting MCQs with a different number of alternatives (e.g., one MCQ with two alternatives, three MCQs with three alternatives, and two MCQs with four alternatives) to get an indication of the extent to which such format control is possible;
- we removed the request for MCQs of varying complexity since GPT-3 could not arrange the MCQs in the order of increasing complexity, as reported by [Kalpakchi and Boye \(2023\)](#).

⁴<https://huggingface.co/datasets/FIDO-AI/ua-squad>

Additionally, in contrast to Kalpakchi and Boye (2023), we also attempt to revise the generated MCQs that did not meet the quality standards. Furthermore, we expand the dataset with manually written MCQs, instead of relying entirely on the synthetically generated MCQs, thus taking a semi-automatic approach. We also conduct a pilot investigation and check to what extent the synthesised MCQs could inspire the creation of the new ones.

3 Data

Any MCQ dataset for reading comprehension consists of the texts and MCQs based on these texts. The choice of texts is crucial in this endeavour as it partly defines what kinds of MCQs would appear in the dataset (e.g., those testing simple text scanning skills, or more advanced, asking the reader to compare or contrast). In this paper we took the texts from the Ukrainian national tests in the Ukrainian language and literature, which are part of the university admission exams in Ukraine, called External independent evaluation, EIE (Ukr. “Зовнішнє незалежне оцінювання, ЗНО”). Specifically, we took the texts from the “Reading” section of the tests administered between 2007 and 2021 (the last year before the radical change of format). We have cleaned the texts by removing titles and/or subtitles of the original texts, numeration of the text parts, and other notes (e.g., names of the authors, number of the words included to the text). Additionally, we have filtered out texts that included non-continuous elements (following the definition of OECD (2019), e.g., lists) or relied on images for the narration.

Furthermore, we were forced to split the vast majority of the texts into parts, which resulted in 62 excerpts from the 32 original texts. The reason behind the aforementioned splitting is illustrated by Figure 3, which shows that one word in Ukrainian corresponded to between slightly less than 7 and 8.5 GPT-3 tokens, (in stark contrast to roughly 1.33⁵ tokens per word for English).

While the aforementioned problem is often solved using the sliding window approach, we would like to argue that it is not sufficient for this particular problem. The reason behind this is that the generated MCQs need to go beyond “local” factual questions about the information that is presented in a couple of sentences. Indeed, we are also interested in the MCQs that test higher-order reading skills (e.g. making high-level inferences or drawing conclusions from a text), for instance, MCQs with such stems as “What is the main idea of the text?”, “Why did X do Y and not Z?”, or

⁵Based on the information here: <https://help.openai.com/en/articles/4936856-what-are-tokens-and-how-to-count-them>

“What is the relationship between X and Y, according to the text?”, which are prevalent in real-world reading comprehension tests. If we require a model to generate such stems in a reliable way, the whole text must be provided to a generation model.

Bearing in mind that we asked GPT-3 to generate N_q MCQs per text, we have empirically identified that the excerpts should be at most 250 words long to allow enough space for the MCQs themselves. Additionally, we took only those excerpts which discuss a particular topic and/or convey a certain idea, so that each of them can be perceived as a standalone text. The extracted 62 excerpts are divided into the following three types:

- *Narrative* texts mainly convey facts or tell a story informing the reader about something or somebody. The texts can be of an *encyclopedic* nature (providing summarised knowledge on a certain object or phenomenon), or *biographical* (narrating life of famous people). Contrary to the Wikipedia-style factual texts, narrative texts in our dataset include literary devices (e.g., metaphors).
- *Descriptive* texts portray something or somebody by giving detailed characteristics of their appearance or features. These texts can include elements of narrative texts.
- *Argumentative* texts convey a certain opinion or a set of opinions (of one or several people) aiming to persuade the reader and/or encourage them to take a certain action. These texts can include elements of narrative and/or descriptive texts.

Later in the article we will refer to these 62 excerpts as simply *texts*.

4 Method

In this work we have investigated the three-stage semi-automatic approach to creating the MCQ dataset. **At the first stage**, we have seeded GPT-3 with the following prompt **in Ukrainian** in an attempt to synthesise N_q^T MCQs:

Напиши N_q^T різних завдань до даного тексту для перевірки розуміння прочитаного. У кожному завданні має бути одне запитання, пронумероване арабськими цифрами (1, 2, 3 ...). З цих N_q^T завдань S_2^T містити два варіанти відповіді, S_3^T містити три варіанти відповіді, S_4^T містити чотири варіанти відповіді. Варіанти відповіді повинні мати вигляд переліку, позначеного буквами (а, б, в, г). З усіх варіантів другий варіант (б) завжди має бути правильною відповіддю.

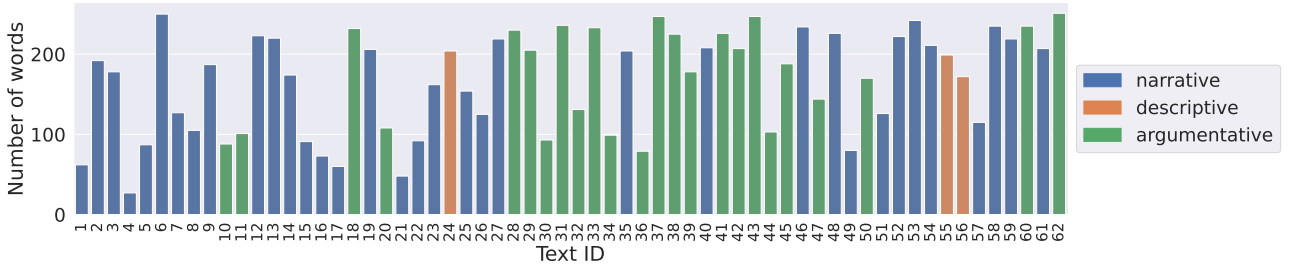


Figure 2: The distribution of the text length in words (defined as space-separated tokens).

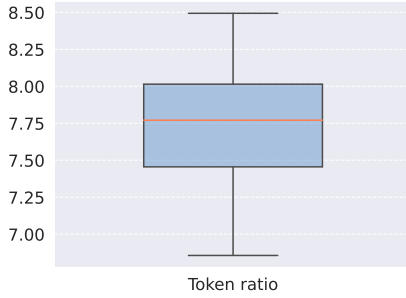


Figure 3: The boxplot showing the distribution of the token ratio $\frac{N_{GPT}}{W_T}$ for the 62 texts from Figure 2, whiskers denote the minimum and maximum values.

У кожному завданні правильною має бути лише одна відповідь.

To aid the reader, we supply the English translation of the prompt, although we stress again that the prompt was fed to GPT-3 in **Ukrainian**.

Write N_q^T different reading comprehension tasks for this text. In each task there should be one question, enumerated with arabic numbers (1, 2, 3 ...). From these N_q^T tasks, S_2^T contain two answer alternatives, S_3^T contain three answer alternatives, S_4^T contain four answer alternatives. Answer alternatives should be in the form of a list, marked by letters (а, б, в, г). From all these alternatives, the second alternative (б) must always be the correct answer. In each task there must be only one correct answer.

The number N_q^T was calculated as follows:

$$N_q^T = \max \left(3, \left\lceil \frac{W_T}{\bar{W}} \right\rceil \right) \quad (1)$$

In Equation 1 W_T denotes the number of space-separated tokens in the text T , and \bar{W} denotes the average number of space-separated tokens per text in the corpus. In this article we have empirically calculated $\bar{W} = 14$ based on the collected 62 texts.

Each S_x^T is a string of the form “ N_x^T <should>”, where N_x^T is the requested number of MCQs with x alternatives, and <should> is the correct form of the Ukrainian verb “мати” (equivalent to the Eng. *should* in this context) grammatically aligned with the number N_x^T , which is calculated as follows:

$$N_x^T = \left\lfloor \frac{N_q^T}{3} \right\rfloor + \mathbb{1}_{N_q^T \% 3 > 4-x} \quad (2)$$

In Equation 2, $\mathbb{1}_{N_q^T \% 3 > 4-x}$ is an indicator function taking the value of 1 if $N_q^T \% 3 > 4-x$ holds, and 0 otherwise. Since $N_q^T \% 3 \leq 2$, the aforementioned condition enables distributing the remainder $N_q^T \% 3$ roughly equally between N_x^T , by first incrementing N_4^T , and then N_3^T .

At the second stage we went through all synthesised MCQs and divided them into the following three types:

- *Kept* denote MCQs of sufficient quality that did not require any corrections. We use N_k^T to denote the number of such MCQs for the text T .
- *Revised* denote MCQs that were manually corrected keeping the stem, the key, and at least one distractor semantically equivalent to (or even the same with) the original ones. Such correction is possible if the original MCQ meets the following three conditions: (1) it is possible to understand the meaning of the original stem and correct its deficiencies, (2) the key answers the new stem correctly, and (3) at least one distractor is still plausible but wrong for the new stem. If the key was not present in the original MCQ, the condition (2) is ignored, and introducing the key counts as correction. We use N_r^T to denote the number of such MCQs for the text T .
- *Discarded* denote MCQs failing to meet at least one condition for being revised. We use N_d^T to denote the number of such MCQs for the text T .

For the sake of simplicity, we will refer to the discarded MCQs and the original MCQs behind the revised ones

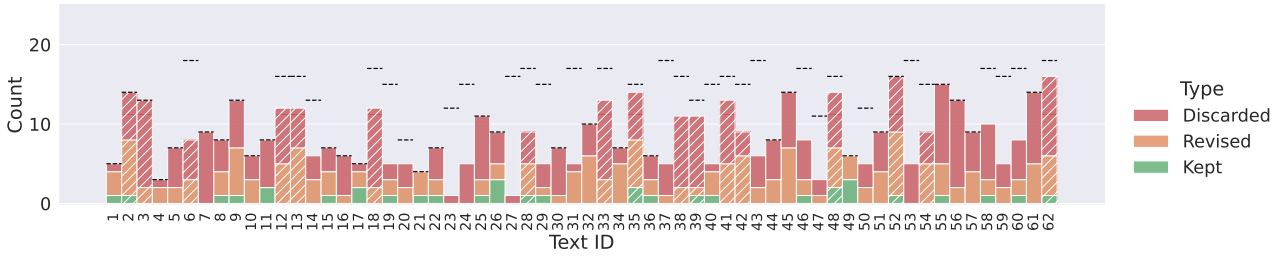


Figure 4: Histogram showing the number of MCQs per text generated by GPT-3. The MCQs are divided into types defined for the second stage. The black dashed lines indicate the *requested* number of MCQs for each text, whereas the height of each bar indicates the *actual* number of generated MCQs. The bars with diagonal hatching indicate the texts for which GPT-3 stopped generating due to reaching the maximum size of its context window (4096 tokens).

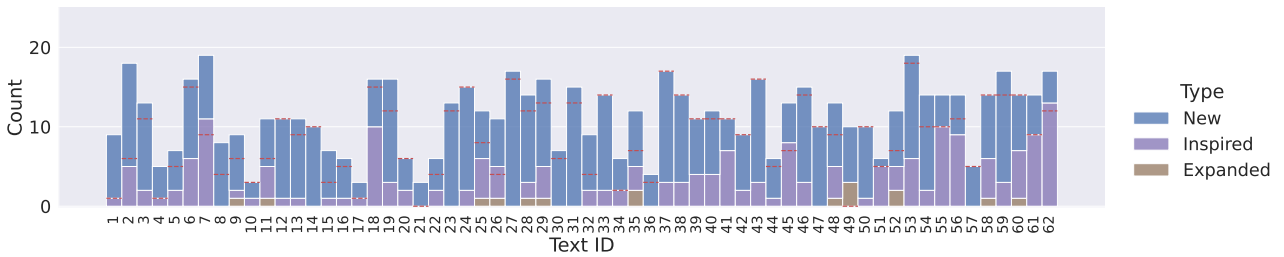


Figure 5: Histogram showing the number of manually added MCQs per text after the third stage. The red dashed lines indicate the minimum required number of MCQs for each text to reach the black dashed lines in Figure 4.

as *MCQs of insufficient quality*. For these MCQs we have identified and categorised the problems causing their poor quality, as described in Section 4.1.

To re-iterate, the introduced revisions are meant to keep the original meaning of the stem and alternatives if it can be derived. If such revisions are impossible, we proceed to the next stage and create a new MCQ.

At the third stage we attempted to complete the dataset with manually written MCQs so that there are *at least* N_q^T MCQs of sufficient quality for *each* text. To be more specific, this means that we needed to write at least $\max(0, N_q^T - N_k^T - N_r^T)$ for each text T . Here we differentiate between three types of MCQs:

- *Expanded* denote MCQs that keep both the original stem and *all* alternatives (consisting of at least the key and one distractor) but introduce more distractors. We use N_e^T to denote the number of such MCQs for the text T .
- *Inspired* denote MCQs which meet at least one of the two conditions: (1) the stem is changed by adding/removing parts compared to the original stem, and (2) none of the distractors are semantically equivalent to any of the original ones. We use N_i^T to denote the number of such MCQs for the text T .
- *New* denote MCQs with entirely new stems, although the alternatives could be taken from the

original MCQ(s). We use N_n^T to denote the number of such MCQs for the text T .

Formally, the goal of this stage is that for every text T the following inequality holds:

$$N_n^T + N_i^T + N_e^T + N_r^T + N_k^T \geq N_q^T \quad (3)$$

4.1 Problem categorisation

We have categorised the problems found in MCQs of insufficient quality based on the impact of these problems on the further revision. Some problems required simple fixes of grammatical errors, whereas others forced us to re-write the entire stem. The rule of thumb is that the larger the re-written part is, the more severe the problem is considered. More specifically, we have grouped the problems into the following four categories:

1. *Formatting errors* – the stem or the alternatives do not follow the formatting requested in the prompt. Such errors can be easily edited.
2. *Language errors* – problems related to inaccurate use of language in terms of its syntax, punctuation, grammatical or lexical norms, while the meaning of the stem and the alternatives is clear. Such problems can be fixed by referring to and following a particular language rule or dictionary.

3. *Semantic errors* – problems which (might) lead to misinterpretation of the stem or the alternatives, or completely prevent a reader from understanding the meaning of these. Depending on the type of fault such errors may be fixed by editing of the stem or the alternative(s). Stems which are *not in the interrogative form* are included to this category since it is not always possible to keep the original (often clear) meaning of the stem after transforming it into a question.
4. *Content-related errors* – problems which keep MCQs incomplete (e.g., abruptly cut stem, lack of alternatives) or affect the stem or the alternative(s) so that their meaning does not correspond to that conveyed by the related text. Such problems usually cannot be fixed by editing, so the MCQ is to be completely re-written (though certain elements of it can still be used as a source of inspiration for a new MCQ).

For explanations and examples of individual errors belonging to each category we refer to Appendix A.

5 Evaluation

To categorise the MCQs as outlined in Section 4, we have manually annotated all MCQs generated by GPT-3. The annotations of the generated MCQs were performed by the first author of this paper who has background in teaching. However, we followed an iterative annotation process (annotating – discussing issues – re-annotating) with both authors (native speakers of Ukrainian) contributing to the discussion and re-annotation. The manually added MCQs, created by the first author, were mostly annotated by the second author (although even here we followed the very same iterative annotation process). Both kinds of annotations were performed using the Textinator annotation tool (Kalpakchi and Boye, 2022).

The results of the annotations for the second stage described in Section 4 are presented in Figure 4. As can be seen from the figure, GPT-3 has produced the required N_q^T MCQs only for slightly less than half of the texts (30 out of 62 texts). Interestingly, although in line with findings of Kalpakchi and Boye (2023), for slightly more than half of cases where GPT-3 did not reach N_q^T (18 out of 32) the generation was stopped because of reaching the stop token (hatched bars in Figure 4), and **not** the maximum number of tokens, meaning more MCQs could potentially be generated for 18 texts.

The number of MCQs that could be kept as they are (green in Figure 4) is very low, only 43 of 525 MCQs, and is distributed unequally among the texts. The number of MCQs that could be revised (orange in Figure 4) also differs substantially between the texts. In total for 36

texts (58% of texts) the number of discarded MCQs is larger or equal to the number of kept and revised ones *together*. This observation reveals a substantial problem with using GPT-3 for generating MCQs in Ukrainian, since discarded MCQs are those that could not be revised without re-writing the major parts of the MCQ.

Recall that we have also requested different number of MCQs with two, three, and four alternatives, attempting to keep each number roughly equal to one third of N_q^T . Figure 6 shows the distribution of the number of alternatives for the generated MCQs per text. As can be seen, GPT-3 failed to meet the aforementioned request for *all* texts. For some texts GPT-3 has also generated MCQs with only one alternative, or even no alternatives at all. Most frequently, GPT-3 generated MCQs with either two or four alternatives, with three alternatives being very rare. This suggests that GPT-3 might have an inductive bias towards generating two or four alternatives (as such cases might have been much more frequent in its training data). Additionally, we note that most of the kept MCQs (green in Figure 6) had only two alternatives (which often were of *yes/no* type), of which only *one* was a distractor.

In an attempt to reach N_q^T MCQs per text we have proceeded to the third stage described in Section 4, which is summarised in Figure 5. Note that for all texts where GPT-3 stopped generating MCQs of its own accord we could manually add the required number of MCQs (and beyond that). For 9 texts most of the newly added MCQs were in fact inspired by the deficient ones produced by the GPT-3. This indicates that MCQs produced by the GPT-3 could potentially be used as an inspiration for the MCQs rather than blindly relied upon. At the same time, we note that for 10 texts none of the MCQs produced by GPT-3 provided the inspiration for the new MCQs (entirely blue bars in Figure 5).

In total, our efforts on correcting the generated MCQs and adding the new ones resulted in expanding the dataset from 43 automatically generated MCQs that could be kept as they are, to 926 MCQs. Observe that MCQs with the same stem but with the alternatives of different types are counted as *different* MCQs. To exemplify, consider the stem “Who wrote the stories about Hercule Poirot?”, and the following three sets of alternatives: (1) Agatha Christie, Arthur Conan Doyle; (2) An English, A French; (3) A woman, A man. While the stem is the same, the first set of alternatives inquires about full names, the second – about nationalities, and the third – about gender. Depending on the text, some of these things might be stated verbatim, while others would need to be inferred, resulting in MCQs of various difficulty. This is why we count the aforementioned example as three different MCQs with two alternatives each, rather than one MCQ with six alternatives.

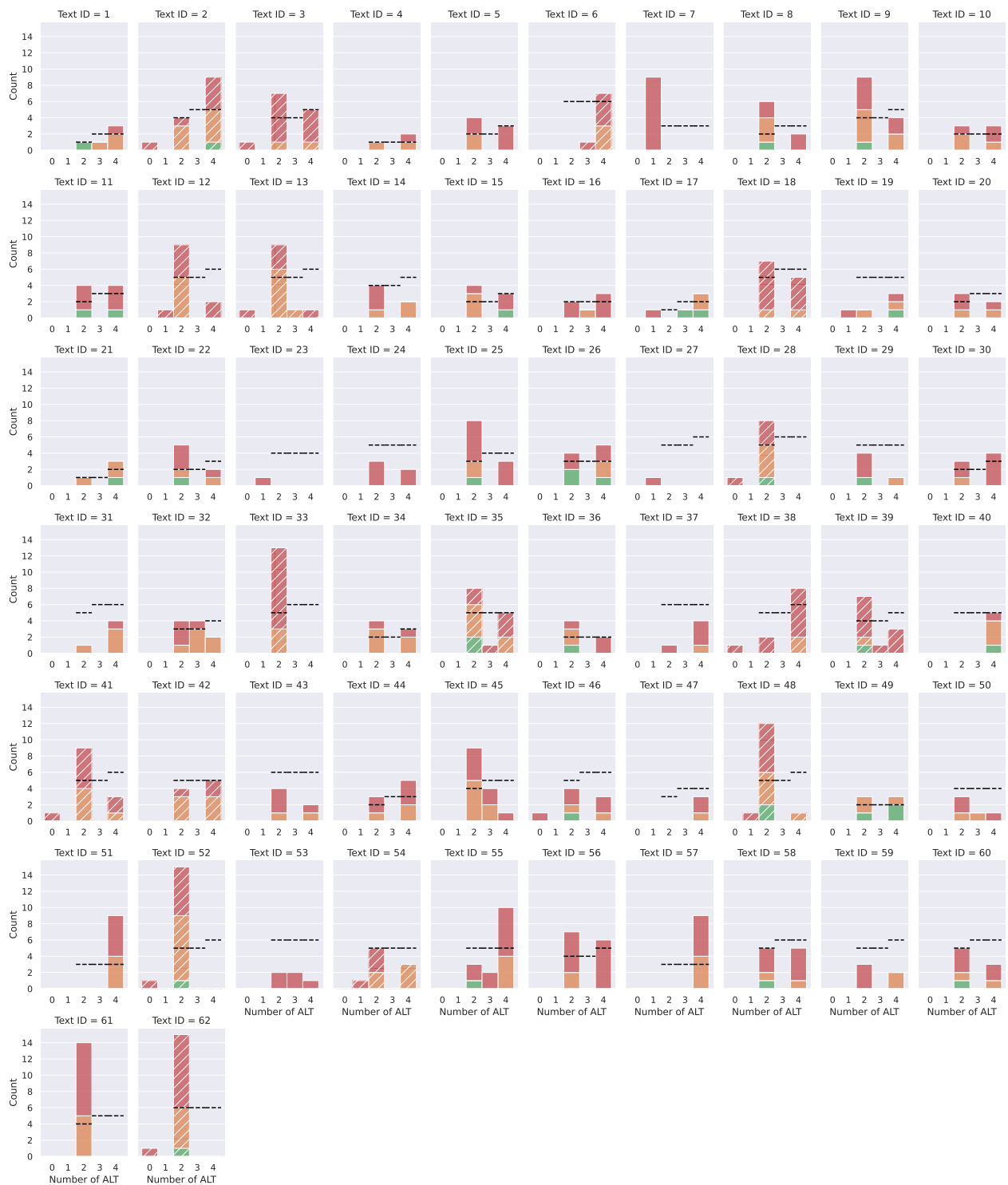


Figure 6: Histogram showing the number of alternatives for MCQs per text generated by GPT-3. The MCQs are divided into types defined for the second stage (the color legend is the same as in Figure 4). The black dashed lines indicate the *requested* number of MCQs with the specified number of alternatives for each text, whereas the height of each bar indicates the *actual* number of generated MCQs with this number of alternatives. Similarly to Figure 4, the bars with diagonal hatching indicate the texts for which GPT-3 stopped generating due to reaching the maximum size of its context window (4096 tokens).

For further analysis of the discarded MCQs and the original MCQs behind the revised ones, the distribution of the identified errors is presented in Figure 7. As can be seen, two the most frequent kinds of errors are associated with the formatting errors (yellow bars in Figure 7), **the least severe category of errors** from Section 4.1. These errors signify MCQs that did not follow the formatting requested in the prompt. For instance, consider the following two MCQs:

1. Скільки мільйонів статей має Вікіпедія?
а) 280 б) 2 в) Більше двох г) Менше двох
Відповідь: Більше двох.
2. В якому році з'явилася книга «Скаутинг для хлопців»? Відповідь: (а) 1906 (б) 1908.

For this example the exact translations do not matter but note the word “Відповідь” (Eng. *Answer*) that is present in both MCQs. In the first MCQ it comes *after* the four alternatives and provides the correct answer (not following the request in the prompt of simply making the correct answer the second one). In the second MCQ, this word comes *before* the alternatives and is absolutely redundant. While such formatting inaccuracies might seem minor, they impede fully automatic processing of MCQs, i.e. getting the stem, the key and each distractor as separate strings.

The second category of problems by severity is language errors (light orange bars in Figure 7). Observe that fixable grammatical errors in the stem and the alternatives belong to the top three most frequent errors, accompanied by the lexical errors in the stem. One interesting kind of grammatical errors made by GPT-3 is introduced by the use of anglicisms, where words or phrases are translated word-by-word from English, as in the stem below:

У якій мові вона робить записи українських пісень?
(*In what language does she record Ukrainian songs?*)

Here the beginning of the stem “У якій мові” is a word-by-word mapping of the English *In what language*, whereas the correct phrase in Ukrainian contains only two words, namely “Якою мовою”. Another example which is likely an anglicism concerns capitalisation of nationalities, as in the stem below:

Що пропонував Французький літературознавець Жюль Ренар?
(*What did the French literary critic Jules Renard promote?*)

Here the capitalisation of the nationality *French* is transferred to the stem in Ukrainian as “Французький”, although nationalities must not be capitalised in Ukrainian. These two small examples suggest that

GPT-3 might be prone to Translationese (Gellerstam, 1986) and use the direct translations of phrases from English. This hypothesis seems plausible given that texts in English constituted 92.64% of the training data of GPT-3, whereas texts in Ukrainian constituted only 0.00763%⁶. However, further investigations on the matter are required.

An interesting example of lexical errors are rusanisms, for instance, as in the stem below:

Як ван Гог відносився до своєї праці?
(*How did van Gogh relate to his work?*)

Here the Ukrainian word “відносився” (*vidnosyvjsja*, Eng. *related*) is likely taken from the Russian “относился” (*otnosilsja*, Eng. *treated*), whereas the correct verb in Ukrainian is “ставився” (*stavivsja*, Eng. *treated*). This suggests that the Ukrainian texts that were included in the training data of GPT-3 have not necessarily been lexically correct to the fullest extent, something that should be investigated further.

The third category of problems by severity is semantic errors (dark orange bars in Figure 7). Here the three most frequent errors are all stem-related, namely ambiguous formulation, misleading grammatical errors and too literal text interpretation. The last one is especially interesting, since one of the motivations behind the use of large language models is exactly to avoid such cases. To exemplify, consider the following MCQ generated by GPT-3:

Text: Галина Бабій, радіожурналіст: Буваючи на відпочинку чи у відрядженнях за кордоном, зауважила, що вільний обмін книжками там дуже поширений. [...]
(*Halyna Babiy, radio journalist: While on vacation or on a business trip abroad, I noticed that the free exchange of books is very common there. [...]*)

В якому місці Галина Бабій зауважила поширення обміну книжками?
(*In which place did Halyna Babiy notice the spread of book exchange?*)
а) у відпочинку (*in vacation*)
б) за кордоном (*abroad*)

Observe that the MCQ is based on the single provided sentence which itself does not point to a specific place but rather to a situation (on vacation or on a business trip). Hence asking “in what place” is inappropriate in these circumstances, let alone the fact that this detail is very minor and is unlikely to be asked in a real-world reading comprehension test.

⁶As reported here: https://github.com/openai/gpt-3/blob/master/dataset_statistics/languages_by_word_count.csv

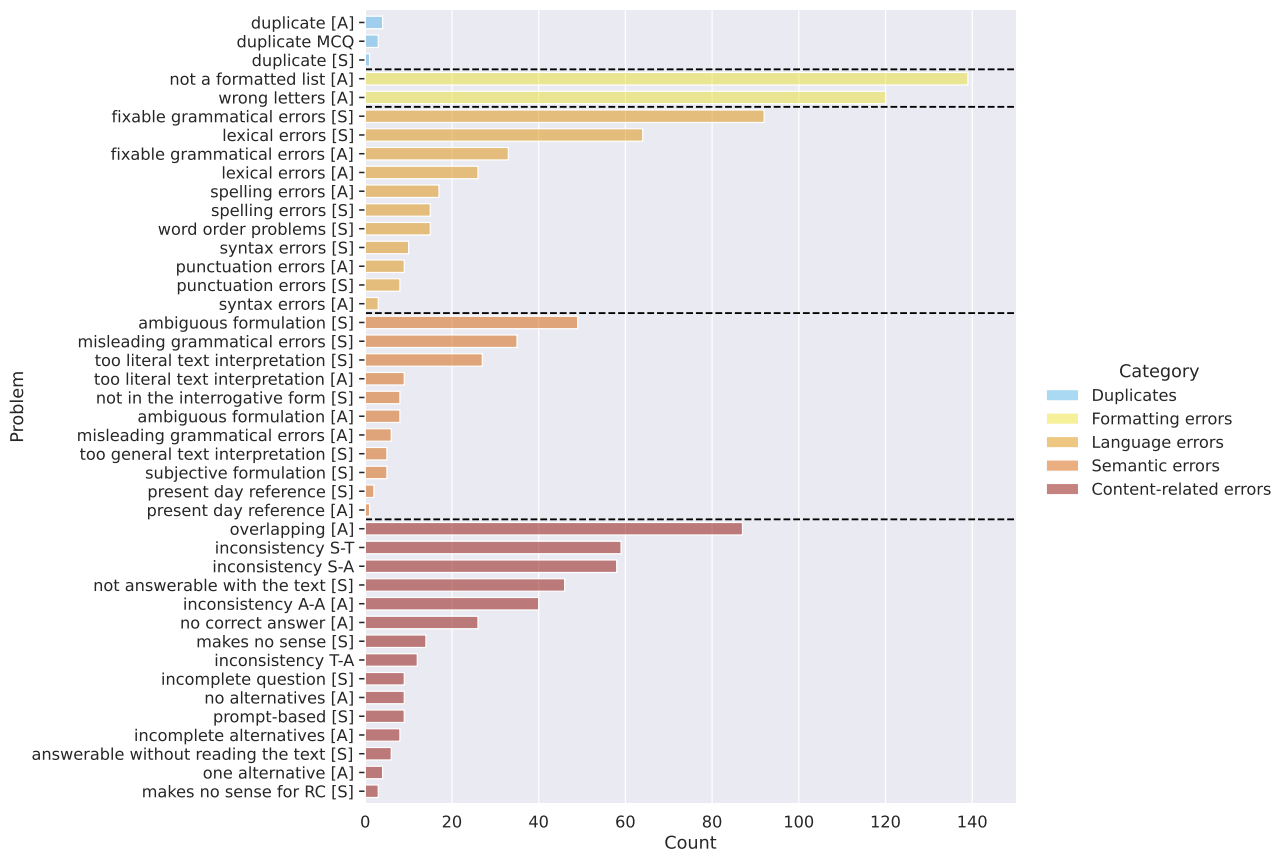


Figure 7: Histogram showing the distribution of problems in the discarded MCQs and the original MCQs behind the revised ones. The problems are categorised as described in Section 4.1, the categories in the legend are ordered from the least severe (at the top) to the most severe (at the bottom). The problems ending with [S] are stem-related, whereas problems ending with [A] are related to at least one of the alternatives within. RC stands for “reading comprehension”.

The final and the most severe problem category contains complex problems with the three most frequent being overlapping alternatives, inconsistency between the stem and the text or the stem and the alternatives. Note that the fourth problem which is very close to the TOP-3 signifies stems that are unanswerable by the text. One particularly interesting problem in this category concerns prompt-based MCQs, such as the one below:

Чи має перелік варіантів відповіді бути позначений буквами (а, б, в, г)?
(Should the list of the answer alternatives be marked by the letters (a, b, c, d)?)
 а) Так (Yes)
 б) Ні (No)

Clearly, the MCQ above asks about the prompt (provided in Section 4), and not about the content of an actual text. This phenomenon was not observed by Kalpakchi and Boye (2023) when applying GPT-3 for generating MCQs in Swedish. Such discrepancy between our and their findings calls for an empirical in-

vestigation across languages and across LLMs aimed at defining the cases when the prompt and the text are not separated by LLMs.

One category that we have not discussed previously are duplicate MCQs (blue in Figure 7). These MCQs constitute cases when the whole MCQ or its part (a stem or a set of alternatives) completely repeats another one or has semantically the same meaning with it. We have not encountered any instances of fully duplicated MCQs, word by word. While some MCQs were semantically equivalent to each other, we have kept them in the dataset.

6 Discussion

The presented dataset of MCQs is created semi-automatically and has its own limitations regarding both the automatic and the manual parts. One limitation that concerns both parts is that the dataset follows no particular principles for ordering either the MCQs for each text, or the alternatives within one

MCQ. While any or both of these could play a role in a real-life testing scenario, we are unaware of any systematic investigation on this matter. Furthermore, any such investigation would be constrained to the particular groups of students, something that is beyond our control in this work. Hence, all alternatives in our dataset are presented in a random order.

Regarding the automatic part, as we have previously discussed, GPT-3 seems to have a number of problems related to Translationese, i.e. applying the phrases or grammar rules of other languages (most notably, English and Russian) to Ukrainian. Bearing in mind that Ukrainian texts constituted only a tiny part of the training data of GPT-3 (0.00763%), such finding is to be expected. Avoiding such problems is one of the strongest arguments either for language models trained specifically for Ukrainian language, or on multilingual language models, where texts in all languages are represented equally.

That said, we do not believe that *fine-tuning* GPT-3 on Ukrainian texts is a feasible way forward due to multiple reasons. First and foremost, to the best of our knowledge, it is currently impossible to estimate what quantity of texts would be enough to reach increase in the model's performance. Secondly, we believe that fine-tuning for this task would require high quality texts in Ukrainian, which are not readily available copyright-free. Lastly, it is very likely that the amount of texts in English in the training data of GPT-3 is higher than all (copyright-free) texts in Ukrainian we will be able to find. All of these arguments together with the cost of fine-tuning GPT-3, and potentially maintaining access to the fine-tuned version for the general public, make such approach practically infeasible for this research.

Another discovery worth further investigation concerns the cases where GPT-3 failed to identify the boundary between the prompt and the supplied text. In our investigation, this manifested itself as MCQs asking about the details of the prompt rather than about the content of the supplied text. Our suggestion is to conduct a systematic investigation on whether such problem occurs across languages and language models (and, ideally, also across NLP tasks).

We have also noticed that GPT-3 did not succeed in “decoding” literary devices (e.g., metaphors, rhetorical figures) and phraseological units, as in the MCQ below:

Text: Але ретельні рентгенівські дослідження засвідчили, що всі роботи митця написані зі «швидкістю виконання й без вагань», «на одному подихові». [...]
(*But careful x-ray studies proved that all the works of the artist were written with “speed of execution and without hesitation”, “in one breath”. [...]*)

На якому подихові були написані всі роботи В. ван Гога?

(*In what breath were all the works of V. van Gogh written?*)

- а) Довгому (*Long*)
- б) Одному (*One*)
- в) Короткому (*Short*)
- г) Завеликому (*Too large*)

Here *in one breath* is a phraseological unit with a stable meaning of “very quickly, without difficulties”, and its component parts cannot be separated from each other. Below there is an example of the MCQ which contains a metaphor:

Text: Квітка вступила до нью-йоркської консерваторії. Оперне майбутнє не склалося, а її американською «дійсністю» стає... рекламний конвеєр, і ось вона — цей янгол — співає дивним тембром сто мільйонів разів якісь «трелі»-заставки для кока-коли.

Було в її кар'єрі й залучення до «великого» кіно. Але це так — мимохідь — так і не розквітла для «Оскара». Але родичі чітко усвідомили: призначення цього херувима не кока-кола, а щось неземне. [...]

(*Kvitka entered the New York Conservatory. The future in the Opera did not materialise, and her American “reality” becomes... an advertising conveyor belt, and here she - this angel - is for a hundred million times singing some “trills” - screensavers for Coca-Cola - in a strange timbre. Her career also involved “big” movies. It was, though, very circumstantial, and she never blossomed for “Oscar”. However, her relatives clearly understood: the destination of this cherub is not Coca-Cola ads, but something otherworldly. [...]*)

Що було призначенням херувима Квітки?
(*What was the destination of the Kvitka's cherub? or What was the destination of Kvitka, the cherub?*)

- а) Кока-кола (*Coca-Cola*)
- б) Щось неземне (*Something otherworldly*)

Here *the cherub* is a metaphor to describe Kvitka herself, and not her property; neither was Kvitka a real cherub - something that GPT-3 did not manage to catch.

We note that the aforementioned performance problems were documented when we tested GPT-3 in a zero-shot way (e.g., just a prompt with a task specification, without any examples). It is theoretically possible that giving some examples of texts and MCQs for these texts (i.e. formulating the problem as few-shot) could bring more MCQs of sufficient quality. However, in practice, given that one word in Ukrainian amounts

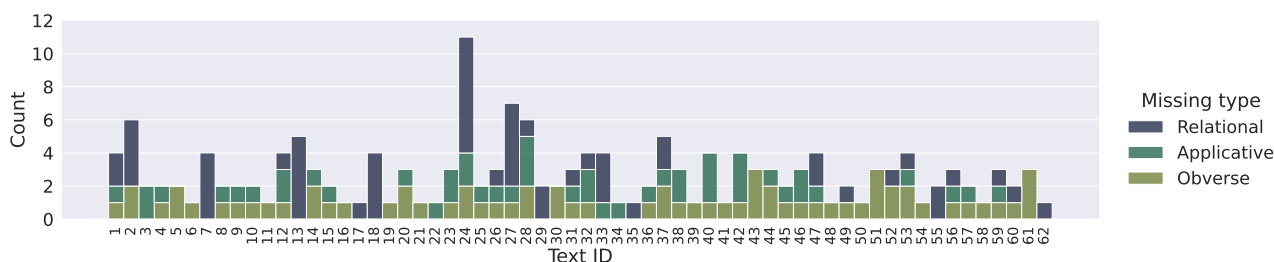


Figure 8: Histogram showing the distribution of new manually written MCQs of three missing MCQ types per text.

on average to about eight tokens, the example texts and their MCQs would take a considerable chunk of the tokens available for GPT-3, leaving little to no space for the actual text and its generated MCQs.

Regarding the manual part, both new and inspired MCQs were written aiming to diversify the MCQs structurally as well as content-wise. For instance, we noticed that certain types of MCQs frequently used in the real EIE tests were absolutely absent from the MCQs generated by GPT-3. Hence, aiming to both increase the diversity and create MCQs resembling the EIE examinations, we attempted to add the missing types of MCQs for each text. More specifically, we limited ourselves to the following three *missing MCQ types*:

- *Relational MCQs* are those asking to establish the relations between two or more elements. Such MCQs are often based on comparison of objects, defining similarities/differences between them or their advantages/disadvantages. Note that the aforementioned relationships should NOT be stated verbatim in the given text for an MCQ to count as relational. For instance, the stem “Яку перевагу бачить автор у театрі перед соціальними мережами?” (“*What advantage of the theater compared to social networks does the author see?*”) would give a rise to a relational MCQ, if such advantage is not written verbatim.
- *Obverse MCQs* are those requiring to detect the opposite from that directly stated in the text. In such MCQs, the stem often includes a clause with a negation which is absolutely necessary to find the key correctly. The negation is often expressed by the particle “не” (*not*), or words such as “відсутній” (*absent*), or “заперечувати” (*to deny*). Together with that, if the text itself focuses on describing what is not happening (e.g., factors which do not cause a certain disease) and the stem requires to name the opposite (e.g., what can cause the named disease), such MCQ is also considered obverse. However, if a stem retains the negation which is already stated in the text (e.g., still asking which factors cannot cause the named disease, while the factors are mentioned

in the text as those not leading to the disease), such MCQ is not considered obverse.

- *Applicative MCQs* are those asking to apply the knowledge from the text to a hypothetical real-life situation introduced in the stem. Typically, a reader is required to first locate the relevant piece(s) of information in the text, extract the knowledge from there, and then correctly apply this knowledge to the given situation. Note that these MCQs require to extract the established knowledge, and NOT someone’s opinion. For instance, “Нобеліантом якої країни стане громадянин України вірменського походження, який на момент присудження премії мешкає у Франції?” (“*A citizen of Ukraine of Armenian origin who lives in France at the time of awarding the prize will become a Nobel laureate of which country?*”) is the stem of an applicative MCQ requiring the reader to understand the formal rules for awarding the Nobel Prize.

Our goal was to investigate whether it was possible to manually create at least one MCQ for each of the aforementioned missing MCQ types.

The results of the aforementioned endeavour are presented in Figure 8. As can be seen, we could create an MCQ of at least one missing type for *each text*. At the same time, only 13 out of 62 texts received at least one MCQ of *all* missing types. This shows that not every kind of text could provide grounds for every missing MCQ type. For instance, the obverse MCQs could be created for the vast majority of the texts, since it is usually enough to have a single fact (which are usually abundant in the texts of various genres) for such MCQ. On the contrary, applicative MCQs require the text to include some kind of knowledge that can be applied, which, for instance, immediately excludes vast majority of the biographies and descriptive texts. Similarly, relational MCQs can not be written for each and every text, since they require at least two objects or concepts that could be compared/contrasted. Additionally, shorter texts (especially those consisting of only a couple of sentences) tend to give less opportunities for these kinds of MCQs.

In addition to the MCQ types mentioned above, there are also so-called *tabular* questions. These MCQs are associated with the texts that contain information that could have been arranged in a table. For instance, one of the texts in our dataset describes the history of comic books and includes information about the names of the comic books in different countries. Such information could be represented as a table with the name of the country in one column and the corresponding name of the comic book in another. A tabular MCQ from our dataset for this text is:

У якому рядку правильно визначено відповідність між країною походження та терміном, що використовується?

(In which line the correspondence between the country and the used term is correctly specified?)

США - «комікс», Франція - «мальовані історії», Японія - «манга», Україна - «стрічка малюнків»

(The US - "comics", France - "drawn stories", Japan - "manga", Ukraine - "picture tape")

США - «комікс», Франція - «стрічка малюнків», Японія - «манга», Україна - «мальовані історії»

(The US - "comics", France - "picture tape", Japan - "manga", Ukraine - "drawn stories")

США - «стрічка малюнків», Франція - «комікс», Японія - «манга», Україна - «мальовані історії»

(The US - "picture tape", France - "comics", Japan - "manga", Ukraine - "drawn stories")

США - «мальовані історії», Франція - «комікс», Японія - «манга», Україна - «стрічка малюнків»

(The US - "drawn stories", France - "comics", Japan - "manga", Ukraine - "picture tape")

The notable feature of tabular MCQs is that one could create many MCQs by simply varying the number of items in each alternative (the number of countries in this example), and matching the values of different columns in various ways (grouping the country name with its name of the comic books in this example). In our dataset, we tend to keep only a few examples of tabular MCQs, without providing its all possible variations.

Similarly, MCQs which include names and digits are mainly represented in one variation only, while the names can often be altered from the full to the shortened ones (or vice versa), including those with only first letters of the name left, and the numbers can be represented in digits or in words.

An opposite kind of MCQs with no variation in alternatives, are the yes/no MCQs, which most frequently contain only two alternatives ("Yes" and "No"), sometimes more (e.g., including "Maybe"). Among

MCQs generated by GPT-3, only 21 were of such type, of which only one was kept as it was and nine were revised. For such MCQs, the set of alternatives is always fixed. The same concerns the stems like "What is the theme of the text?" where the stem is fixed, while the alternatives change with each new text. Writing such MCQs equates to constructing only a stem or only a set of alternatives which makes the process faster but of the lower priority for automation. Taking that into account, we did not add such MCQs manually.

Another transformation that could expand the dataset is the use of synonymous reformulations or paraphrases of the given stems, which might potentially allow to manipulate difficulty of the MCQs but where an extensive coverage is hardly reachable. Mainly focused on covering content-related aspects, we leave vocabulary alterations and difficulty evaluation for the future work.

However, already from the MCQs included into this dataset (from those created both automatically and manually), we have noticed that their difficulty might vary depending on a personality-based factor (OECD, 2019) of previous knowledge – the knowledge a reader already had before beginning to read the given text. Since such factor can hardly be controlled, it can be tricky to judge whether an MCQ is suitable for testing the reader's reading skills rather than their previous knowledge. For instance, MCQs which are to some extent based on the so-called "common knowledge" may require making complex inferences with respect to the text but become absolutely trivial for people with the relevant previous knowledge. We noticed that the particular kinds of previous knowledge for which it is true are stable facts (those that could be verified from multiple credible sources and are not based on opinions), meanings of idioms, and definitions of terms. To exemplify further, consider the following text and an MCQ:

Text: Микола Леонтович збирав народні пісні й адаптував їх для хорового співу. [...]

Композитор, як різьбяр, зробив навколо основної послівочки витончену оправу. Поєднавши прийоми народного багатоголосся з досягненням класичної поліфонії, він домігся того, що кожен голос почав відігравати самостійну роль, відтворюючи найтонші зміни настрою. Леонтович кілька разів переробляв твір, аж поки 1916 року не створив досконалий хорал.

(Mykola Leontovych collected folk songs and adapted them for choral singing. [...] The composer, like a carver, made an elegant frame around the main song. By combining the techniques of folk polyphony with the achievements of classical polyphony, he made each voice play an independent role, reproducing the most subtle mood

changes. *Leontovych revised the work several times until the year 1916, when he managed to create a perfect chorale.*)

Ким був Микола Леонтович за фахом?
(*Who was Mykola Leontovych by profession?*)

- а) Композитором (*A composer*)
- б) Різьбярем (*A carver*)
- в) Хористом (*A chorister*)
- г) Етнологом (*An ethnologist*)

Here M. Leontovych's profession might be a completely unknown fact for some people while they are able to infer the information from the text. However, students of a music school or students with broad knowledge in arts and/or Ukrainian culture are likely to answer this MCQ without even reading the text.

Another example of an MCQ which is potentially answerable without reading the text is presented below:

Text: Практика надання допомоги безпритульним тваринам сягає XVII ст. Саме 1695 р. в Японії, у місті Едо (нині Токію), з'явився перший (з відомих нам) притулок для собак. [...]

(*The practice of helping homeless animals dates back to the 17th century. It was in 1695 in Japan, in the city of Edo (now Tokyo), that the first (we know about) shelter for dogs appeared. [...]*)

Яке місто мало назву Едо?
(*Which city used to be named Edo?*)

- а) Токію (*Tokyo*)
- б) Львів (*Lviv*)
- в) Київ (*Kyiv*)
- г) Кіото (*Kyoto*)

Here students might be completely unaware of the first name of the city of Tokyo; however, it is a stable real-life fact which a student can know from school subjects (e.g., geography, arts) or other sources not related to the reading material used for a reading comprehension test.

In practice it means that no pre-generated set of MCQs can be blindly taken as it is for real-life learning and is still to be verified by a person (likely, a teacher) who knows their target audience, peculiarities of the learning process of this audience, exact objectives of a test, and so on.

7 Conclusions

Despite the format of MCQs being widely used in the Ukrainian educational system, specifically for reading comprehension tests as part of the university admission exams, automatic generation of these questions in Ukrainian has not been introduced yet. Inspired by

what has been achieved in the NLP field for other languages, we created a semi-synthetic MCQ dataset for reading comprehension in Ukrainian which can be used as training or evaluation data for models specialising in MCQ generation and answering.

As expected, to achieve the sufficient quality of the dataset, manual editing was necessary to fix the errors made by GPT-3 and diversify the whole dataset by creating additional MCQs. However, the extent to which human assistance appeared necessary is surprisingly high - more than 90 per cent of the generated MCQs. The faults by the model are named and additionally categorised according to their impact on further revision.

We also note that prompted to generate MCQs with a different number of alternatives, GPT-3 failed to meet the request, which means that this aspect of generating tasks in the multiple-choice format appears to be hardly controllable in the zero-shot scenario, which is a likely way real-world teachers would interact with such models.

Additionally, we found that the effective context window size for Ukrainian is much smaller than $\frac{4096}{1.33}$ words, since one word in Ukrainian is roughly equal to 8 tokens of GPT-3. Such limitation prevented us from generating MCQs on real-length reading comprehension texts, and calls for development of models with tokenisers keeping the token-to-word ratio closer to 1.

That given, this particular model, GPT-3 (as of July 2023), does not seem to be appropriate for reading comprehension MCQs generation in Ukrainian. However, more tests with other models and languages are needed to determine the extent to which LLMs can be used as a helpful generative tool for the stated task.

Acknowledgements

We gratefully acknowledge the financial support for this research provided by the Knut and Alice Wallenberg Foundation to the first author.

References

- Alberti, Chris, Daniel Andor, Emily Pitler, Jacob Devlin, and Michael Collins. 2019. Synthetic QA corpora generation with roundtrip consistency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6168–6173, Florence, Italy. Association for Computational Linguistics.
- Bandarkar, Lucas, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2023. The belebele benchmark: a parallel reading comprehension

- dataset in 122 language variants. *arXiv preprint arXiv:2308.16884*.
- Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Ch, Dhawaleswar Rao and Sujan Kumar Saha. 2018. Automatic multiple choice question generation from text: A survey. *IEEE Transactions on Learning Technologies*, 13(1):14–25.
- Council of Europe. 2001. *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Dijkstra, Ramon, Zülküf Genç, Subhradeep Kayal, Jaap Kamps, et al. 2022. Reading comprehension quiz generation using generative pre-trained transformers.
- Gellerstam, Martin. 1986. Translationese in Swedish novels translated from English. *Translation studies in Scandinavia*, 1:88–95.
- Kalpakchi, Dmytro and Johan Boye. 2022. Textinator: an internationalized tool for annotation and human evaluation in natural language processing and generation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 856–866, Marseille, France. European Language Resources Association.
- Kalpakchi, Dmytro and Johan Boye. 2023. Quasi: a synthetic question-answering dataset in Swedish using GPT-3 and zero-shot learning. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 477–491, Tórshavn, Faroe Islands. University of Tartu Library.
- OECD. 2019. *PISA 2018 Assessment and Analytical Framework*. OECD Publishing, Paris.
- Raina, Vatsal and Mark Gales. 2022. Multiple-choice question generation: Towards an automated assessment framework. *arXiv preprint arXiv:2209.11830*.
- Rajpurkar, Pranav, Robin Jia, and Percy Liang. 2018. Know what you don’t know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.
- Rajpurkar, Pranav, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.
- Vachev, Kristiyan, Momchil Hardalov, Georgi Karadzhov, Georgi Georgiev, Ivan Koychev, and Preslav Nakov. 2022. Leaf: Multiple-choice question generation. In *European Conference on Information Retrieval*, pages 321–328. Springer.

A Error typology

In this section we present a more detailed account of the problem categories in Section 4.1. The errors within each category are listed in the alphabetic order.

Formatting errors

1. *Not a formatted list* – the generated alternatives are arranged in a line, which does not comply with the prompt and thus is considered to be an error made by the model.
2. *Wrong letters* – the generated alternatives are marked with symbols other than prompted (а, б, в, г), which also includes upper case or a different script (e.g., latin) used by the model.

Language errors

1. *Fixable grammatical errors* – a phrase/sentence contains a faulty, uncommon, or controversial usage of the Ukrainian language (e.g., misuse of grammar cases, verb tenses or voices, breaking grammatical alternation rules, etc.), which can be clearly identified and fixed by applying corresponding rules.
2. *Lexical errors* – a word/phrase is used in an inappropriate meaning, repeats another root word or phrase from the same sentence, is missing or redundant to express the intended idea.
3. *Punctuation errors* – punctuation marks are missing, redundant, or incorrectly used in the sentence, according to the Ukrainian grammar.
4. *Spelling errors* – a word/phrase is formed incorrectly in terms of choice of letters, order of letters, capitalisation or usage of special characters/symbols according to the rules of Ukrainian.
5. *Syntax errors* – a phrase/sentence is incorrectly built in terms of agreement between its parts or choice of the parts of speech (mainly function words), which might partially or completely prevent the reader from understanding the meaning.
6. *Word order problems* – an incorrect or awkward placing of the words in a sentence which makes it more difficult to understand the meaning or breaks sentence structure rules fixed in the grammar.

Semantic errors

1. *Ambiguous formulation* – a phrase/sentence is formulated in an unclear way which allows several possible interpretations in the given context.

2. *Misleading grammatical errors* – a faulty, uncommon, or controversial usage of the Ukrainian language (e.g., misuse of grammar cases, verb tenses or voices, breaking grammatical alternation rules, etc.) or a combination of these which prevents from understanding the meaning. The only possible fix to the problem is re-writing the major part(s) of the stem or alternative(s).
3. *Too literal text interpretation* – a phrase or sentence is taken verbatim from the text to the stem or alternative so that a corresponding part of the MCQ sounds incomplete, unclear, or weird.
4. *Too general text interpretation* – a phrase or sentence in the stem is extracted from the text without all necessary details for the stem to be evident, context-related, and answerable with one of the given alternatives.
5. *Not in the interrogative form* – the stem is given in the form of a fill-in-the-gap or continue-the-sentence tasks which does not comply with the prompt and thus is considered as an error made by the model.
6. *Present-day reference* – the requested information is related to the period of time defined by the words “currently”, “recently”, “today” (or similar) in the text, while the same formulation in the stem tends to become irrelevant with the pass of time and might then confuse a reader. Moreover, sometimes the key changes as time passes by, for instance, if the stem inquires about the number of months between “today” and some event, which means that the key will require adjustment with the pass of time.
7. *Subjective formulation* – the stem or alternative requires evaluation of an object or phenomenon based on a reader’s personal opinion, feelings or experience, where the reader’s answer will likely lack grounds for support or disapproval, and consequently, for objective evaluation.

Content-related errors

1. *Answerable without reading the text* – it is possible to answer the question by analysing the stem and alternatives, without reading the given passage.
2. *Incomplete alternatives* – the process of generating alternatives was started but then stopped for some reason, so the alternatives are cut.
3. *Incomplete question* – the process of generating the stem was started but then stopped for some reason, so both the stem and the set of alternatives are cut.

4. *Inconsistency A-A* – alternatives within one and the same MCQ do not correspond in represented type of content. For instance, the stem asks about the kind of the objects, while the alternatives are “long and round” (naming the form), “white and blue” (naming the colour).
5. *Inconsistency S-A* – information requested by the stem does not match with the type of information provided by one or more of the alternatives within one and the same MCQ. For instance, the stem asks about the shape of the objects, whereas at least one of the alternatives provides colors.
6. *Inconsistency S-T* – information requested by the stem is not provided or cannot be inferred from the text.
7. *Inconsistency T-A* – information provided in the text does not correspond semantically to that in the alternative.
8. *Makes no sense* – the combination of words/phrases in the stem makes its meaning either incomprehensible or hardly plausible for a real life context-related situation.
9. *Makes no sense for RC* – the stem is grammatically and semantically correct (or can be easily edited to be correct) but focuses on the details from the given text which are not strictly important to make relevant inferences and understand the meaning.
10. *No alternatives* – not a single alternative was generated by the model; it is, though, possible, that the model still presented the correct answer for the corresponding stem.
11. *No correct answer* – among the generated alternatives, not a single one can be considered as a key to the stem.
12. *Not answerable with the text* – the given text provides no information for a reader to be able to answer the generated stem.
13. *Prompt-based* – the generated MCQ or its part is based on information from the prompt rather than that from the given text.
14. *One alternative* – from the requested number of alternatives (two, three, or four), only one alternative was generated by the model, which does not comply with the prompt and thus is considered as an error made by the model.
15. *Overlapping alternatives* – more than one alternative satisfy the conditions stated in the stem and thus result in more than one key for the MCQ, not complying with the prompt.