Can Humans Identify Domains?

Maria Barrett ⊜ **Max Müller-Eberstein** ⊜ **Elisa Bassignana** ⊜
Amalie Brogaard Pauli ⊜ **Mike Zhang** ⊜ **Rob van der Goot** ⊜

IT University of Copenhagen, Aarhus University, Aalborg University, Pioneer Centre for AI
{mamy, elba, robv}@itu.dk
ampa@cs.au.dk, jjz@cs.aau.dk

Abstract

Textual *domain* is a crucial property within the Natural Language Processing (NLP) community due to its effects on downstream model performance. The concept itself is, however, loosely defined and, in practice, refers to any non-typological property, such as genre, topic, medium or style of a document. We investigate the core notion of domains via human proficiency in identifying related intrinsic textual properties, specifically the concepts of genre (communicative purpose) and topic (subject matter). We publish our annotations in **TGeGUM**: A collection of 9.1k sentences from the GUM dataset (Zeldes, 2017) with single sentence and larger context (i.e., prose) annotations for one of 11 genres (source type), and its topic/subtopic as per the Dewey Decimal library classification system (Dewey, 1979), consisting of 10/100 hierarchical topics of increased granularity. Each instance is annotated by three annotators, for a total of 32.7k annotations, allowing us to examine the level of human disagreement and the relative difficulty of each annotation task. With a Fleiss' kappa of at most 0.53 on the sentence level and 0.66 at the prose level, it is evident that despite the ubiquity of domains in NLP, there is little human consensus on how to define them. By training classifiers to perform the same task, we find that this uncertainty also extends to NLP models.

Keywords: domain, genre, topic, multi-annotation

1. Introduction

The concept of "domain" is ubiquitous in Natural Language Processing (NLP), as differences between "sublanguages" have strong effects on model transferability (Kittredge and Grisham, 1986). This issue of domain divergence has prompted comprehensive surveys on how to best adapt language models (LMs) trained on one or more source domains to more specific targets (Ramponi and Plank, 2020; Ramesh Kashyap et al., 2021; Saunders, 2022), and remains an open issue, even with LMs of increasing size (Ling et al., 2023; Singhal et al., 2023; Wu et al., 2023). Despite its importance, what constitutes a domain remains loosely defined, typically referring to any non-typological property that degrades model transferability. In practice, textual properties with the largest domain effects relate to a document's genre/medium/style (McClosky, 2010; Plank, 2011; Müller-Eberstein et al., 2021b), topic (Lee, 2001; Karouzos et al., 2021), or mixtures thereof (Aharoni and Goldberg, 2020). More broadly, domains can be viewed as a high-dimensional space with variation across the aforementioned properties, plus factors such as author personality, age, or gender (Plank, 2011, 2016).

We attempt to gain a better understanding of the foundational concept of domain, by taking a step back from modeling this phenomenon, and instead investigating whether humans themselves can distinguish between different instantiations of domain-related properties of textual data. In linguistics literature, these properties are separated into register, style and genre (Biber, 1988; Biber and Conrad, 2009, 2019), of which we choose to focus on *genre*, as it distinguishes itself from register and style by remaining consistent across complete texts. In ad-

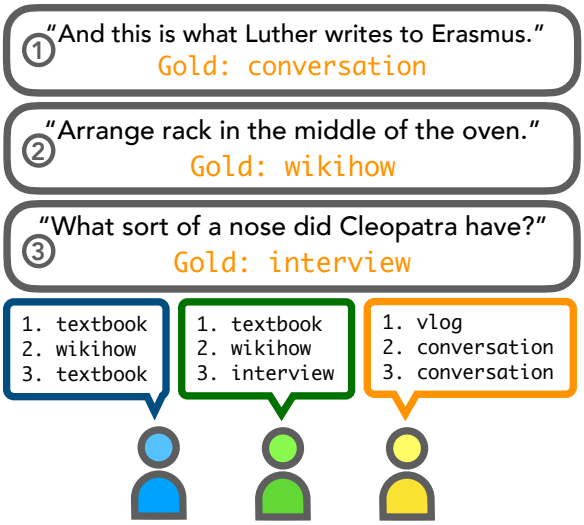


Figure 1: Graphical illustration of our triple-annotation setup with gold genre labels.

⊜ All authors contributed equally.

LREC-COLING 2024, pages 2745–2765

dition, we examine the orthogonal factor of *topic*, i.e., the subject matter of a text, which can be expressed independently of genre (Kessler et al., 1997; Lee and Myaeng, 2002; Stein and Zu Eissen, 2006; Webber, 2009). We operationalize these two factors analogously to van der Wees et al. (2015) as genre stemming from different source types with distinct communicative styles, and topic being the principal subject matter of a given text.

More formally, our main research question is: *To what extent can humans detect genres and topics from text alone, and how does this align with machines?* We investigate the human proficiency in detecting these intrinsic properties by turning our attention to the Georgetown University Multilayer Corpus (GUM; Zeldes, 2017),[1] a large-scale multi-layer corpus consisting of texts from 11 different source types (henceforth *genre*). These act as gold annotations against which we compare the manual genre labels provided by 12 human annotators for the entirety of the corpus (Figure 1). In addition, the annotators supply a new annotation layer regarding the texts' subject matter (henceforth *topic*). As no gold labels are available for topic, they are annotated according to Dewey Decimal Classification (DDC; Dewey, 1979), a library classification system that allows new books to be added to a collection based on the subject matter. The DDC consists of 10 topics, 100 fine-grained topics, and 1,000 even finer-grained topics, of which we investigate the former two in detail and provide a preliminary study on the latter.

To understand the importance of context, we have annotators label genre and topic at both the sentence and prose level (defined as sequences of five sentences), and compare annotator agreement. Due to the subjective uncertainty associated with these types of characteristics, we gather three annotations per instance, measure their agreement, and release them in their unaggregated form as multi-annotations for future research.

Finally, we investigate the ability of machines to identify the same characteristics by training multiple ablations of genre and topic classifiers. Concretely, these experiments examine the difficulty of discerning each property, whether metadata or human notions of genre are more easily recoverable, as well as which level of context is most appropriate for the different ways in which the genre and topic label distributions can be represented.

Overall, this work is the first to explore the discernability of domain by both humans *and* machines. In Section 5, we further discuss the implications of our findings, both with respect to domain-sensitive downstream applications, as well as for the NLP community's more general definition of domain. Our contributions thus include:

- **TGeGUM** (Topic-Genre GUM), a multi-layer extension of GUM, covering 9.1k sentences triple-annotated for a diverse set of 11 genres and 10/100 topics (Section 3).[2]
- An in-depth exploratory data analysis of the human annotations concerning annotator disagreement, uncertainty, and overall trends for domain characteristics across different context sizes (Section 4).
- A case study on the capability of NLP models to discern the human notions of genre and topic, as well as an analysis of which factors affect classification performance (Section 5).

2. Related Work

Domains Initially coined as "sublanguages" (Kittredge and Lehrberger, 1982; Kittredge and Grisham, 1986), domains have long been a topic of study in traditional linguistics and NLP (Lee, 2002; Lee and Myaeng, 2002; Stein and Zu Eissen, 2006; Eisenstein et al., 2014; van der Wees et al., 2015; Plank, 2016). Some of the early work mentioning domains as textual categories include Sekine (1997); Ratnaparkhi (1999), which categorize texts into, e.g., "general fiction", "romance & love", and "press:reportage". However, as also mentioned by Lee (2002); Lee and Myaeng (2002); Plank (2011); van der Wees et al. (2015), the concept of domain is under-defined. Plank (2011) considers domains as a multi-dimensional space, spanning all kinds of variability between texts, such as genre, topic, style, medium, etc. In this work, we follow a definition of domains similar to van der Wees et al. (2015), focusing on two of the largest dimensions of variability: i.e., *genres* (the communicative purpose and style) as well as *topics* (the subject matter). The former is closely tied to the source of a text, such as academic papers versus fiction books, while the latter may include subjects such as sports, politics, and philosophy, which can occur in multiple genres.

Automatic Domain Detection In NLP, automatic domain detection is essential for ensuring robust downstream performance, as it degrades with increasing levels of domain shift (Ramponi and Plank, 2020). Since this issue occurs independently of the application, domain classification has been explored in many contexts. Generally, the problem is either phrased in terms of a binary task, i.e., whether a target text matches the domain of the training data or not (e.g., Tan et al., 2019; Pokharel and Agrawal, 2023), or a multi-label classification task, in which the exact domain is to be

[1] https://gucorpling.org/gum/

[2] Data and code can be found at bitbucket.org/robvanderg/humans-and-domains.

determined (e.g., Müller-Eberstein et al., 2021a). Here, we use the latter approach as it requires a more formalized operationalization of domain.

At a broader level, genre is frequently used as a proxy for domain, as it has lower internal variability than many more specific dimensions, including topic (Kessler et al., 1997; Webber, 2009). Its automatic detection has been leveraged for selecting training data for transfer learning across a broad range of applications, such as classification (Ruder and Plank, 2017; van der Goot et al., 2021a; Gururangan et al., 2020) and generative tasks (Aharoni and Goldberg, 2020). Beyond English, genre has further been shown to provide a cross-lingually consistent signal for enabling more robust transfer in syntactic parsing (Müller-Eberstein et al., 2021a).

Topics provide a more granular differentiation between texts, also with close ties to domain. Automatically detecting topics has more immediate practical implications, as knowledge of the subject matter is critical for many downstream information extraction systems (Liu et al., 2021; Bassignana and Plank, 2022) and more datasets with topic annotations are available (Sandhaus, 2008; Maas et al., 2011; Wang and Manning, 2012; Zhang et al., 2015); however, these works typically contain source data from only a single corpus.

Going beyond prior work with limited sets of post-hoc topic labels for single-genre corpora, we build on the general-purpose DDC system (Dewey, 1979) for libraries and apply its hierarchical set of 10/100 topics to a corpus containing data from 11 genres. By building on the existing annotations of the GUM dataset (Zeldes, 2017), we further enable research not only ascertaining to domain classification for its own sake, but also with applications to other downstream NLP tasks.

Multi-annotations Given the subjective nature of domains and their associated properties of genre and topic, each text in our dataset is annotated multiple times and retains individual labels without aggregating them. This approach of *multi-annotations* (Plank, 2022) avoids obscuring human uncertainty in the annotation process and has benefits both for tasks with high variability, such as ours, as well as tasks for which a ground truth is typically assumed.

E.g., Plank et al. (2014) map part-of-speech (POS) tags from Gimpel et al. (2011) to the universal 12-tag set by Petrov et al. (2012), retaining five *crowdsourced* POS labels per token.

For Relation Classification (RC), Dumitrache et al. (2018) obtained annotations for 975 sentences for medical RC, where each sentence is annotated by at least 15 annotators on average.

For Natural Language Inference (NLI), Nie et al. (2020) released ChaosNLI: A dataset with 4,645 examples and 100 annotations per example for some existing data points in the development set of SNLI (Bowman et al., 2015), MNLI (Williams et al., 2018), and Abductive NLI (Bhagavatula et al., 2020). For a more in-depth overview of multi-annotation datasets, we refer to Uma et al. (2021).

3. The Dataset

3.1. Source Data

The source dataset on top of which we build our domain-related annotations is the GUM corpus which in turn incorporates data from a wide variety of sources. We use the portion of the GUM corpus released as part of the Universal Dependencies project (UD; Nivre et al., 2017), i.e., excluding Reddit. Since a text's source is closely tied to its communicative purpose, we consider GUM's *data source* metadata field of each instance as the gold genre label. For the topic, no equivalent gold label is discernible from the metadata.

The entire dataset is annotated both at sentence and prose level to investigate the importance of context for genre and topic annotation. For this purpose, we follow the gold sentence segmentation provided by GUM. We opted for these blocks instead of paragraphs, as the latter are not natural dividers for all text types and can have a high variety of conventions and functions across genres. To avoid the same annotator observing the same sentence individually as well as in prose, we shuffle the dataset such that annotations of a sentence with and without context are distributed across different annotators, while maintaining coverage of the full dataset.

3.2. Annotation Procedure

Since there are no official descriptions of the genres in GUM, our annotation guidelines refer to the descriptions from the homepages of the websites of the source or the corresponding abstracts from Wikipedia. For topic annotation, we follow the Dewey Decimal library classification system (Dewey, 1979) consisting of 10/100/1,000 hierarchical topics of increased granularity. We consider the 10 high-level and the 100 mid-level classes for the coarse- and fine-grained topic annotations. We constrain our guidelines such that topic-100 should always be a sub-type of topic-10. For example, if topic-100 is "520 Astronomy", then topic-10 should be "500 Science". When none of the topic-100 labels fit the fine-grained topic of the instance, the annotators were allowed to leave the more specific topic blank, i.e., annotating topic-100 with the same label as topic-10. In addition, we include the *no-topic* label for when it is not possible

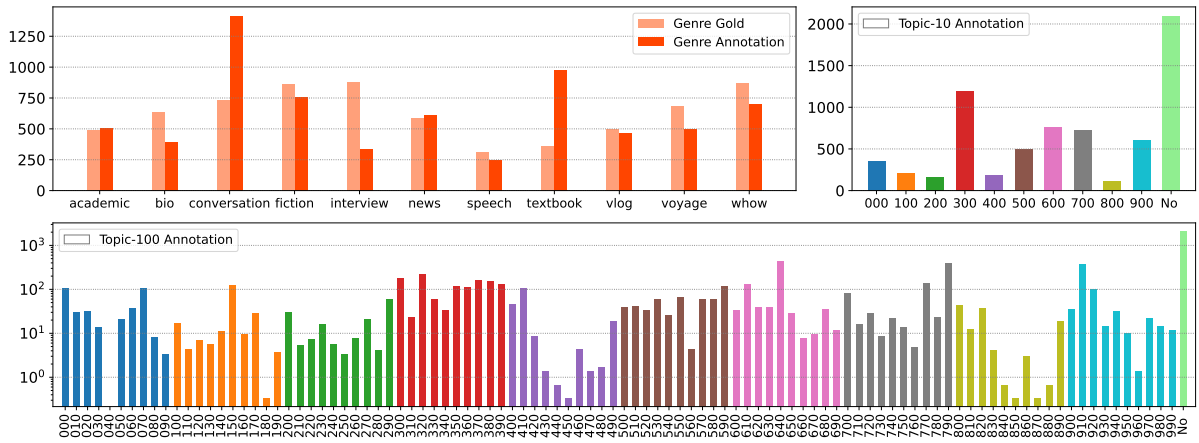


Figure 2: Frequency distributions of the labels in gold genre labels, annotations of genres, annotations of topic-10, and annotations of topic-100 (log scale) on sentence level. For the human annotations, the number is divided by three in order to align with the (unique) gold label. The mapping of topic-10 and topic-100 labels can be found in Appendix F. The tag "No" in the topic annotations refers to *no-topic*.

to identify a specific topic from the provided text., such as for very short sentences, like "Ok" or "I agree with that."

We completed an initial annotation round of 20 instances with all annotators and authors of this paper to evaluate the guidelines and annotation setup. None of this data is included in the final dataset. We continued with groups of three annotators annotating different subsets of the data. After an introductory meeting, further unclarities were discussed asynchronously throughout the process. Annotators were asked to pose their questions in general terms and to not use direct examples as to not bias the other annotators on specific instances. We did not conduct inter-annotator studies over the course of annotation and only had minor guideline revisions during the annotation process since we are mostly interested in human intuitions of genre and topic, and there are no gold labels for the topic task.

Annotators could indicate whether they were unsure about the annotation of a specific instance, and were also asked to provide notes/comments, if applicable. The annotation rate started at approximately 80–150 instances per hour. To ensure a similar amount of effort across annotators, we asked them to aim for approximately 150 instances per hour (also considering that annotation speed increases over time).

In total, we hired 12 annotators, who were paid 34,21 EUR per hour (before tax) for a total of 32 hours per person over a period of 4 weeks. The mean age was 27 ($\pm$2), and their highest completed education was equally split between a bachelor's and a master's degree. All rated their English skills as either C2/proficient or native. Seven annotators were reported to be female, three male, and two other/non-binary.

	Instances		Annotations	
	Sentence	Prose	Sentence	Prose
Train	6,911	1,358	20,733	4,074
Dev.	1,117	217	3,351	651
Test	1,096	221	3,288	663
Total	9,124	1,796	27,372	5,388

Table 1: Dataset Statistics: Note that each instance has three associated annotations.

3.3. Dataset Statistics

Table 1 shows the final dataset statistics of **TGeGUM**. The dataset includes around 9.1K sentences, and 1.8K prose, each of them annotated by three individual annotators for genre, coarse-grained topic, and fine-grained topic.

In Figure 2, we report the sentence-level distribution of gold labels and human annotations, reporting the average number of annotations per label (total number of annotations divided by three annotators) to align with the singular gold genre metadata. For topic-10 and topic-100 we only report the human annotations as no gold labels exist.

Comparing gold and annotated genre labels, we observe a skew towards *conversation* and *textbook*. We hypothesize that this is due to the small amount of context an annotator receives. For example, the sentence "Is that all that's left?" with the gold genre label *fiction* is annotated by all annotators as *conversation*. Another example is the

	Kappa			Maj. Acc.
	Genre	Topic-10	Topic-100	Genre
Sentences	0.5260	0.5213	0.4239	67.68
Prose	0.6582	0.5238	0.3838	81.11

Table 2: Agreement scores across annotators, and accuracy of majority vote among annotators compared to gold genre labels.

sentence “Some of the greatest poetry has been born out of failure and the depths of adversity in the human experience.” with gold label *interview*. All annotators annotated this example as *textbook*.

For topic, we note that despite skewness, almost all 100 topics are used. The *300 Social sciences* including, e.g., *320 Political science* and *370 Education*, stand out as being the most prevalent topics. The most frequent label, however, is *no-topic*, indicating that it is challenging to identify a specific topic given only one sentence and that individual sentences can be associated with different topics, depending on the surrounding context.

The genre distribution at the prose level (Appendix D) reveals a more accurate distribution for *conversation*-like utterances; however, the general skew towards *textbook* remains. Concerning topic, the main contrast to the sentence-level distributions is the reduction of the *no-topic* label, confirming that more context is crucial for this task.

4. Exploratory Data Analysis

In addition to the previous aggregated overview, we are interested in exploring whether domain characteristics are recoverable by humans in a consistent manner. While we can compare human annotations to the original gold labels for genre, no equivalent is available for topic. Therefore, we place more emphasis on inter-annotator agreement, in the form of Fleiss’ Kappa (Fleiss, 1971), to measure intuitive alignment and ease of identification. Table 2 and Figure 3 shows this agreement across the different genres, topics and levels of available context.

4.1. Human Genre Detection

Accuracy and Agreement Considering that annotation guidelines were phrased to avoid any intentional alignment to an existing ground truth (i.e., annotators were unaware of the existence of gold genre labels), an accuracy of 67.68% at the sentence level shows that genre is recoverable to a far higher degree than by random chance or by a majority baseline. This further increases to 81.11% given more context at the prose level and is also reflected in the increase from moderate inter-annotator agreement (0.53) to substantial agreement (0.66).

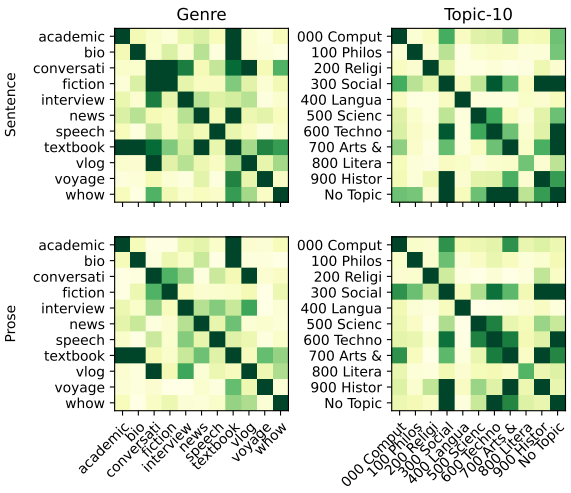


Figure 3: Confusion matrix with all annotated pairs of labels for Genre and Topic-10 (across all annotators) in our training data: The darker the color, the higher the number of annotations for that label pair. The diagonal can be seen as agreement, whereas off-diagonal is a proxy for disagreement.

The additional context appears to help differentiate genres that have more similarities to each other. This phenomenon is especially pronounced for spoken-language data, such as *conversation*, *interview* and *vlog*, which differ with respect to genre-specific conventions such as who the speech is directed towards (i.e., bi-directional, interviewee, video viewer), or how formal the register is. Both properties are more easily discernible across multiple turns.

Nonetheless, even given more context, high amounts of confusion remain between certain genres such as non-fiction texts of the type *academic*, *biography*, and *textbook*. These are intuitively similar to each other and may require even more context to distinguish. Generally, genres appear to lie on a more continuous spectrum that is difficult to discretize in conceptually similar cases.

Human Uncertainty In case of uncertainty, annotators were encouraged to select a “best guess” label and to indicate uncertainty by ticking a checkbox. In addition to overall uncertainty, we also hypothesize that sentence length affects accuracy due to the amount of information available. To evaluate these two effects for genre detection, we measure the Pearson correlation between human accuracy concerning the gold label, with 1) sentence length, 2) the number of uncertainty flags (Table 3). As expected, longer sentences are annotated correctly more often. Figure 4 further highlights how spoken-language genres have a strong

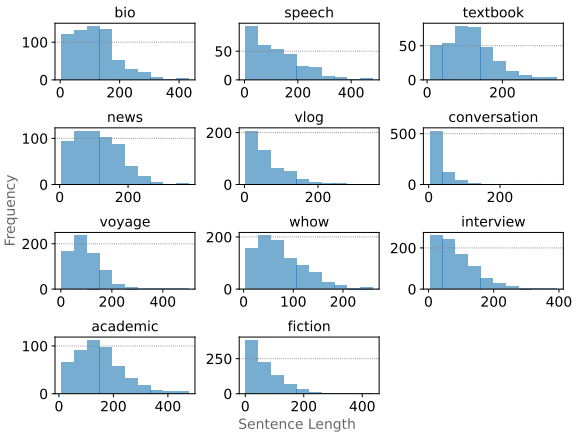


Figure 4: Frequency of sentence lengths, measured by the number of characters, per gold genre.

skew towards shorter sentences, and for which annotators have the lowest agreements. Additionally, sentences marked as "unsure" align with gold labels less often, showing that annotators appear to have well-calibrated judgments of their own uncertainty, even for this relatively difficult task.

4.2. Human Topic Detection

Agreement In the absence of gold labels, inter-annotator agreement allows us to estimate the difficulty of discerning broader vs granular topics. For the 10 broader topics, Table 2 shows a moderate agreement of 0.52 for both the sentence and prose levels. As expected with an order of magnitude more labels, Topic-100 sees a drop in agreement to 0.42 and an additional drop to 0.38 at the prose level. While this may seem counter-intuitive due to topic's higher specificity compared to genre, Figure 3 sheds some light on this peculiarity: In contrast to genre, topic has a *no-topic* label (Section 3.2), which, in turn, is used frequently by all annotators at the sentence level, due to the absence of any subject matter in many shorter utterances—especially in speech. Given the additional context, topic becomes more apparent, and agreement spreads toward more topics along the diagonal. As such, sentence-level agreement mainly hinges on *no-topic*, while prose-level annotations agree more with respect to actual topics. This is less apparent for 10-topic kappa, for which this effect cancels out, but is more prevalent with 100 topics, where the shift away from *no-topic* at the prose level comes with a much wider spread of topics, thereby reducing overall agreement, despite having a higher level of true topic annotations.

Overall, topics which were most consistently identified include *social sciences*, *arts & recreation*, *technology*, *science* and *history & geography*. On the other hand, *literature* was least consistently annotated and most frequently confused with the aforementioned topics, potentially due to its broader scope compared to the others.

	Sent	Prose
length vs unsure	-0.1126*	-0.0474
length vs correct	0.1267*	-0.0385
unsure vs correct	-0.2948*	-0.3411

Table 3: Correlations across utterance length, correct predictions of human majority vote, and the number of unsure annotations. * indicates statistical significance for $p < 0.05$.

1,000 Topics After completing the full set of genre and topic-10/100 annotations with three annotators per instance, the remaining time of the annotators was spent on a preliminary study to label the most fine-grained categories of DDC. With 1,000 labels, this task is substantially more difficult. We obtained a total of 904 sentences and 172 prose sequences with three annotations each.[3] Measuring inter-annotator agreement at this level of granularity, we find a Fleiss' Kappa of 0.32 for sentences and 0.26 for prose. Although substantially lower than for coarser topic granularities as well as genre, this score still indicates above-random agreement among annotators. Similarly to the previous topic results, prose-level context allows humans to detect more actual topics than *no-topic*, leading to lower overall agreement but a broader coverage of actual topics.

In general, despite the importance of topic to downstream applications (i.e., topic classification as a task in itself), there is no clear human consensus regarding discrete topic classification. Similarly to genre, topic appears to be a concept for which human intuition shares some agreement at a broader level, but is also spread along a continuum—especially as granularity increases.

5. Modeling Domain

Following our examination of human notions of genre and topic, we investigate automatic methods' ability to model the same properties. Ablating across different setups for representing the multiple annotations per instance (Section 5.1), we train models to classify genre and topic at different levels of granularity (Section 5.2) and evaluate their ability to learn the underlying distribution (Section 5.3). While pre-neural work typically performed document-level classification (Webber, 2009; Petrenz and Webber, 2011), contemporary

[3]From 3,918 total annotations, we discarded instances with less than three completed annotations.

trends have shifted towards the sentence-level (Aharoni and Goldberg, 2020; Müller-Eberstein et al., 2021b). Leveraging our multi-level annotations, we investigate genre and topic classification at both the sentence and prose-level, mirroring our human annotation setup.

5.1. Setup

Most work on modeling multiple annotators is based on tasks consisting of only two or three labels, e.g., hate speech detection, or RTE (Uma et al., 2021). An exception is Kennedy et al. (2020), who use multiple classification heads to predict a score for a variety of aspects of hate speech, which are then used to predict a final floating point score for hate speech detection. Other related work predicts multiple task labels simultaneously (e.g., Demszky et al., 2020; Kiesel et al., 2023; Piskorski et al., 2023), however these are typically discrete and do not model annotator certainty. We propose a variety of methods to model the distribution of the annotations (overview in Figure 5):

Majority Discretizes the labels using a majority vote, and uses a single classification head to predict it. For the distribution similarity metric (see below), we assign a score of 1.0 to the chosen label.

PerLabel-Regression Converts the human annotations to scores per label and then predicts these as a regression task. Each label has its own decoder head, trained using an MSE loss, and mapped to the [0;1] range afterwards.

PerLabel-Classification Converts the human annotations into score bins and predicts them as four possible labels: “0.0”, “0.33”, “0.66”, “1.0”.

PerAnnotator One decoder head modeling each annotator, that predicts their annotation as a discrete label. Afterwards, the three predictions are converted to a distribution.

We evaluate these models using the standard accuracy over each singular predicted label (i.e., highest score or majority). In addition, we conduct a finer-grained evaluation that takes the multi-annotations into account. For this purpose, we propose a similarity metric for comparing the predicted and annotated label distribution per instance. Let n be the number of label types, and X and Y are label distributions that sum to 1, with a score for each label. Then, the distributional similarity per instance can be computed as:

$$distr\_sim = 1 - \frac{\sum_{l=0}^{n} |X_n - Y_n|}{2} \quad .$$

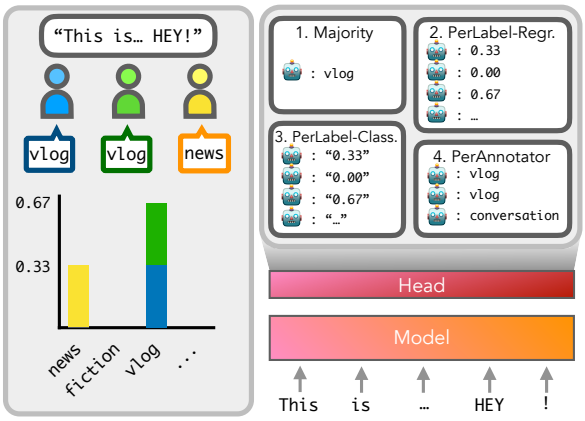


Figure 5: The target value each model variant is trained to predict: 1) Majority vote. 2) PerLabel-Regr(ession) on label distributions. 3) PerLabel-Class(ification), on score bins per label. 4) PerAnnotator, three different annotations.

The resulting score between 0 and 1 represents the distributions’ similarity. Note that we compare model predictions to the human annotations, which are not a gold standard; here, we aim to determine whether the human ability to discern these concepts is easy to model.

We implement all our model variants in the MaChAmp (van der Goot et al., 2021b) toolkit v0.4 using default parameters. MaChAmp is a toolkit focused on multi-task learning for NLP, and allowed us to implement all varieties of the tasks described earlier. Each way of phrasing the task is implemented on top of a single language model for fair comparison. From an initial evaluation of the bert-large-cased (Devlin et al., 2019), luke-large-lite (Yamada et al., 2020), deberta-v3-large (He et al., 2021), xlm-roberta-large (Conneau et al., 2020) LMs on the gold genre labels, we identify that DeBERTa has the highest accuracy; hence we use it in the following experiments.

5.2. Classification Results

We examine which notion of domain is more learnable and distinguishable for a model; genre or topic? Since genre has associated ground truth labels, we additionally examine whether the human annotators’ perception of genre or the ground truth genre is easier to learn.

We establish a majority vote based on the human annotations; in case of a tie, the first element in the annotation list is chosen as the label, both for sentences and prose. This happens in ∼10% of cases for genre and topic-10 (sentence and prose), and ∼20% cases for topic-100.

Table 4 shows accuracy and macro-F1 scores of the annotators’ majority vote evaluated against

	Accuracy	Macro-F1	$\|N\|$
Sentence	67.68	59.92	1,117
Prose	81.11	74.75	217

Table 4: Performance of annotators' majority vote compared with the gold genre (development set).

the gold genre. As noted previously, more context (prose level) helps disambiguate the genre.

To evaluate how well a model can align with the human intuition of genres and topics, we fine-tune an LM on the majority labels of the annotators. We compare the performance on the gold genre labels (the only task for which we have gold labels) and compare the accuracy and macro-F1 scores (Table 5). We notice the following:

Sentences 1) Unsurprisingly, DeBERTa fine-tuned on the gold genre labels (gold_genre) is better aligned with the ground truth genre than the human majority vote, i.e., 73.20 (Table 5) versus 67.68 (Table 4) accuracy at the sentence level (note that other LMs performed worse). 2) In contrast, the fine-tuned DeBERTa model has higher accuracy when trained and tested on the human majority vote (maj_gerne) than when using gold genre labels (gold_genre), i.e., 75.88 versus 73.20, although macro-F1 is lower. This indicates that less common genre labels are easier to learn from gold labels, while more frequent genres are easier to learn based on human intuitions. 3) Despite topic-10 having fewer classes than genre, the notion of topic appears to be more difficult for a model to learn (lower F1). 4) The skew of the fine-grained topics (maj_topic-100) and the difficulty of the long tail become apparent in the large divergence across the accuracy and macro-F1 score.

Prose 5) In contrast to the sentence level, our fine-tuned DeBERTa model generalizes better to the gold genre labels (gold_genre) than the human majority vote (maj_genre). At this level of context, the majority vote topic is also harder for a model to learn than the majority vote genre.

5.3. Distributional Results

In Figure 6, we report the results of the models trained on all instances (sentences and prose) with DeBERTaV3-large.[4] The main trends show that the model performs better on the genre task. Unsurprisingly, for topics, the granularity of the labels impacts performance.

[4]Training on sentences and prose separately leads to similar trends (Appendix B).

		Accuracy	Macro-F1
Sent.	gold_genre	73.20± 0.02	70.74± 0.02
	maj_genre	75.88± 0.01	67.04± 0.01
	maj_topic-10	75.56± 0.02	60.54± 0.07
	maj_topic-100	64.55± 0.00	18.43± 0.02
Prose	gold_genre	89.49± 0.02	88.02± 0.03
	maj_genre	80.83± 0.01	74.97± 0.03
	maj_topic-10	67.74± 0.01	50.35± 0.03
	maj_topic-100	52.35± 0.01	16.04± 0.02

Table 5: Accuracy and Macro-F1 on test split, for DeBERTa models fine-tuned and evaluated on gold genre, human majority vote for genre, and human majority vote for topic-10/100 (standard deviations across five seeds).

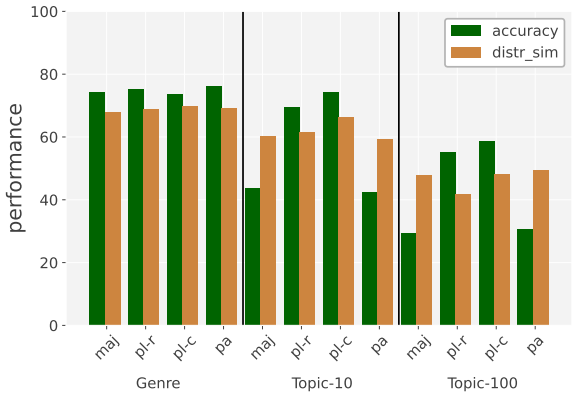


Figure 6: Accuracy and distributional similarity on test split, for DeBERTa models trained on target labels based on Majority vote (maj), PerLabel-Regression/Classification (pl-r/c), PerAnnotator labels (pa); standard deviations across five seeds.

By modeling the annotation distributions (i.e., PerLabel-Regression/Classification), we can outperform the majority vote model. However, distributional similarity decreases with increased label granularity (i.e., from topic-10 to topic-100), showing that it is difficult for models to calibrate to diverging human judgments. Interestingly, the per-label models achieve comparable or higher scores on the $distr\_sim$ metric, showing that the examined LMs model label distributions more easily than annotator behavior.

6. Conclusion

To examine the widely used but scarcely defined notion of *domain*, this work provides the first investigation of human intuitions of this property in the form of **TGeGUM**: a collection of 9.1k sentences annotated with 11 genres and 10/100 topics by three annotators per instance, using an annotation

procedure designed to capture human variability instead of forcing alignment (Section 3).

Our exploratory analysis (Section 4) shows that despite the subjective nature of this task, as reflected in a Fleiss' Kappa of 0.53–0.66, humans can identify certain domain characteristics consistently from one sentence alone. Nonetheless, genres with a high similarity benefit substantially from added context. This is even more crucial for identifying topics, where we observe a shift from annotators not being able to discern any topic at all to being able to reach an above-random agreement, even when presented with 100 or 1,000 topics.

Finally, our experiments of modeling these domain characteristics automatically (Section 5) show that genre is easier to model than topic. For both the agreements between human annotators, and the performance from the automatic model, we see that context is crucial for the genre classification task, but not for topic classification, where adding context even leads to decrease in scores if the label space is large.

Overall, this work highlights that despite the importance of "domain", there is little consensus regarding its definition, both in the NLP community as well as in our human annotations. Taking a closer look at what intuition predicted, further reveals that genres and topics are difficult to discretize completely, and that a continuous space of domain variability may be more suited for characterizing these phenomena.

7. Ethics Statement

Our approach to modeling human label variation is intrinsically linked to the larger issue of human social bias. As highlighted by Plank (2022), significant social implications are tied to the study of label variation. In the context of our research, it is essential to acknowledge that variations in labeling might stem from societal biases and disparities. To address this, we recognize the necessity of addressing bias mitigation techniques as we aim to create more equitable and just models. However, we also contend that our focus on modeling generic subjects, such as genre and topic, may carry less severe implications compared to more subjective tasks like hate speech detection (Akhtar et al., 2021; Davani et al., 2022). The differences in annotations within our work may primarily relate to two categories: "Missing Information" and "Ambiguity" (Sandri et al., 2023).

Another ethical facet we must address is the potential biases present in the classification system we use. In particular, the Dewey Decimal Classification System, which is the de-facto standard for libraries worldwide, has been found to exhibit prejudice (Gooding-Call, 2021). For example, the classification of information related to religion, specifically within class 200, demonstrates a clear skew, with a majority of subjects (six out of ten) reserved for Christianity-related topics. The remaining four slots are designated for other dominant religions, with an *other* section meant to encompass all other belief systems. This reveals an inherent bias toward Christianity, which can affect the accessibility of non-dominant religions and belief systems. There are alternatives to knowledge organization systems like the Dewey Decimal Classification, as suggested by Franzen (2022), to promote a more inclusive and equitable information landscape.

8. Acknowledgments

Many thanks to our annotators: Nina Sand Horup, Leonie Brockhaus, Birk Staanum, Constantin-Bogdan Craciun, Sofie Bengaard Pedersen, Yiping Duan, Axel Sorensen, Henriette Granhøj Dam, Trine Naja Eriksen, Cathrine Damgaard, and the other two anonymous annotators. Maria Barrett is supported by a research grant (34437) from VILLUM FONDEN. Mike Zhang is supported by the Independent Research Fund Denmark (DFF) grant 9131-00019B. Elisa Bassignana and Max Müller-Eberstein are supported by the Independent Research Fund Denmark (DFF) Sapere Aude grant 9063-00077B. Amalie Pauli is supported by the Danish Data Science Academy, which is funded by the Novo Nordisk Foundation (NNF21SA0069429) and VILLUM FONDEN (40516).

9. Bibliographical References

Roee Aharoni and Yoav Goldberg. 2020. Unsupervised domain clusters in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7747–7763, Online. Association for Computational Linguistics.

Sohail Akhtar, Valerio Basile, and Viviana Patti. 2021. Whose opinions matter? perspective-aware models to identify opinions of hate speech victims in abusive language detection. *ArXiv preprint*, abs/2106.15896.

Elisa Bassignana and Barbara Plank. 2022. CrossRE: A cross-domain dataset for relation extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3592–3604, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman,

Hannah Rashkin, Doug Downey, Wen-tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Douglas Biber. 1988. *Variation across speech and writing*. Cambridge University Press, Cambridge.

Douglas Biber and Susan Conrad. 2009. *Register, Genre, and Style*, 1 edition. Cambridge Textbooks in Linguistics. Cambridge University Press, Cambridge.

Douglas Biber and Susan Conrad. 2019. *Register, genre, and style*. Cambridge University Press.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, pages 92–110.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

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.

Melvil Dewey. 1979. Dewey decimal classification and relative index.

Anca Dumitrache, Lora Aroyo, and Chris Welty. 2018. Crowdsourcing semantic label propagation in relation classification. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 16–21, Brussels, Belgium. Association for Computational Linguistics.

Jacob Eisenstein, Brendan O'Connor, Noah A Smith, and Eric P Xing. 2014. Diffusion of lexical change in social media. *PloS one*, 9(11):e113114.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5).

Mara Franzen. 2022. Alternatives to the dewey decimal system.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-speech tagging for Twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 42–47, Portland, Oregon, USA. Association for Computational Linguistics.

Anna Gooding-Call. 2021. Racism in the dewey decimal system.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing.

Constantinos Karouzos, Georgios Paraskevopoulos, and Alexandros Potamianos. 2021. UDALM: Unsupervised domain adaptation through language modeling. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2579–2590, Online. Association for Computational Linguistics.

Chris J Kennedy, Geoff Bacon, Alexander Sahn, and Claudia von Vacano. 2020. Constructing interval variables via faceted rasch measurement and multitask deep learning: a hate speech application. *ArXiv preprint*, abs/2009.10277.

Brett Kessler, Geoffrey Nunberg, and Hinrich Schutze. 1997. Automatic detection of text genre. In *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 32–38, Madrid, Spain. Association for Computational Linguistics.

Johannes Kiesel, Milad Alshomary, Nailia Mirzakhmedova, Maximilian Heinrich, Nicolas Handke, Henning Wachsmuth, and Benno Stein. 2023. SemEval-2023 task 4: ValueEval: Identification of human values behind arguments. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2287–2303, Toronto, Canada. Association for Computational Linguistics.

Richard Kittredge and Ralph Grisham. 1986. *Analyzing Language in Restricted Domains: Sublanguage Description and Processing*. Lawrence Erlbaum Associates.

Richard Kittredge and John Lehrberger. 1982. *Sublanguage: Studies of language in restricted semantic domains*. Walter de Gruyter.

David Lee. 2002. Genres, registers, text types, domains and styles: clarifying the concepts and navigating a path through the bnc jungle. In *Teaching and Learning by Doing Corpus Analysis*, pages 245–292. Brill.

David YW Lee. 2001. Genres, registers, text types, domain, and styles: Clarifying the concepts and navigating a path through the bnc jungle. *Language Learning & Technology*, 5(3):37–72.

Yong-Bae Lee and Sung Hyon Myaeng. 2002. Text genre classification with genre-revealing and subject-revealing features. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 145–150.

Chen Ling, Xujiang Zhao, Jiaying Lu, Chengyuan Deng, Can Zheng, Junxiang Wang, Tanmoy Chowdhury, Yun Li, Hejie Cui, Xuchao Zhang, Tianjiao Zhao, Amit Panalkar, Wei Cheng, Haoyu Wang, Yanchi Liu, Zhengzhang Chen, Haifeng Chen, Chris White, Quanquan Gu, Jian Pei, and Liang Zhao. 2023. Domain specialization as the key to make large language models disruptive: A comprehensive survey. *ArXiv*.

Zihan Liu, Yan Xu, Tiezheng Yu, Wenliang Dai, Ziwei Ji, Samuel Cahyawijaya, Andrea Madotto, and Pascale Fung. 2021. Crossner: Evaluating cross-domain named entity recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13452–13460.

Andrew L. Maas, 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.

David McClosky. 2010. *Any domain parsing: automatic domain adaptation for natural language parsing*. Ph.D. thesis, Brown University.

Max Müller-Eberstein, Rob van der Goot, and Barbara Plank. 2021a. Genre as weak supervision for cross-lingual dependency parsing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4786–4802, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Max Müller-Eberstein, Rob van der Goot, and Barbara Plank. 2021b. How universal is genre in Universal Dependencies? In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, pages 69–85, Sofia, Bulgaria. Association for Computational Linguistics.

Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020. What can we learn from collective human opinions on natural language inference data? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9131–9143, Online. Association for Computational Linguistics.

Joakim Nivre, Željko Agić, Lars Ahrenberg, Lene Antonsen, Maria Jesus Aranzabe, Masayuki Asahara, Luma Ateyah, Mohammed Attia, Aitziber Atutxa, Elena Badmaeva, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, John Bauer, Kepa Bengoetxea, Riyaz Ahmad Bhat, Eckhard Bick, Cristina Bosco, Gosse Bouma, Sam Bowman, Aljoscha Burchardt, Marie Candito, Gauthier Caron, Gülşen Cebiroğlu Eryiğit, Giuseppe G. A. Celano, Savas Cetin, Fabricio Chalub, Jinho Choi, Yongseok Cho, Silvie Cinková, Çağrı Çöltekin, Miriam Connor, Marie-Catherine de Marneffe, Valeria de Paiva, Arantza Diaz de Ilarraza, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Marhaba Eli, Ali

Elkahky, Tomaž Erjavec, Richárd Farkas, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta Gonzáles Saavedra, Matias Grioni, Normunds Grūzītis, Bruno Guillaume, Nizar Habash, Jan Hajič, Jan Hajič jr., Linh Hà Mỹ, Kim Harris, Dag Haug, Barbora Hladká, Jaroslava Hlaváčová, Petter Hohle, Radu Ion, Elena Irimia, Anders Johannsen, Fredrik Jørgensen, Hüner Kaşıkara, Hiroshi Kanayama, Jenna Kanerva, Tolga Kayadelen, Václava Kettnerová, Jesse Kirchner, Natalia Kotsyba, Simon Krek, Sookyoung Kwak, Veronika Laippala, Lorenzo Lambertino, Tatiana Lando, Phuong Lê Hồng, Alessandro Lenci, Saran Lertpradit, Herman Leung, Cheuk Ying Li, Josie Li, Nikola Ljubešić, Olga Loginova, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Héctor Martínez Alonso, André Martins, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Gustavo Mendonça, Anna Missilä, Verginica Mititelu, Yusuke Miyao, Simonetta Montemagni, Amir More, Laura Moreno Romero, Shunsuke Mori, Bohdan Moskalevskyi, Kadri Muischnek, Nina Mustafina, Kaili Müürisep, Pinkey Nainwani, Anna Nedoluzhko, Luong Nguyễn Thị, Huyền Nguyễn Thị Minh, Vitaly Nikolaev, Rattima Nitisaroj, Hanna Nurmi, Stina Ojala, Petya Osenova, Lilja Øvrelid, Elena Pascual, Marco Passarotti, Cenel-Augusto Perez, Guy Perrier, Slav Petrov, Jussi Piitulainen, Emily Pitler, Barbara Plank, Martin Popel, Lauma Pretkalniņa, Prokopis Prokopidis, Tiina Puolakainen, Sampo Pyysalo, Alexandre Rademaker, Livy Real, Siva Reddy, Georg Rehm, Larissa Rinaldi, Laura Rituma, Rudolf Rosa, Davide Rovati, Shadi Saleh, Manuela Sanguinetti, Baiba Saulīte, Yanin Sawanakunanon, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Lena Shakurova, Mo Shen, Atsuko Shimada, Muh Shohibussirri, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Aaron Smith, Antonio Stella, Jana Strnadová, Alane Suhr, Umut Sulubacak, Zsolt Szántó, Dima Taji, Takaaki Tanaka, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Francis Tyers, Sumire Uematsu, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Gertjan van Noord, Viktor Varga, Veronika Vincze, Jonathan North Washington, Zhuoran Yu, Zdeněk Žabokrtský, Daniel Zeman, and Hanzhi Zhu. 2017. Universal dependencies 2.0 – CoNLL 2017 shared task development and test data. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Philipp Petrenz and Bonnie Webber. 2011. Squibs: Stable classification of text genres. *Computational Linguistics*, 37(2):385–393.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2089–2096, Istanbul, Turkey. European Language Resources Association (ELRA).

Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023. SemEval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multi-lingual setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2343–2361, Toronto, Canada. Association for Computational Linguistics.

Barbara Plank. 2011. *Domain adaptation for parsing*. Ph.D. thesis, University of Groningen.

Barbara Plank. 2016. What to do about non-standard (or non-canonical) language in nlp. In *KONVENS 2016, Ruhr-University Bochum*.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 507–511, Baltimore, Maryland. Association for Computational Linguistics.

Rhitabrat Pokharel and Ameeta Agrawal. 2023. Estimating semantic similarity between in-domain and out-of-domain samples. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 409–416, Toronto, Canada. Association for Computational Linguistics.

Abhinav Ramesh Kashyap, Devamanyu Hazarika, Min-Yen Kan, and Roger Zimmermann. 2021. Domain divergences: A survey and empirical

analysis. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1830–1849, Online. Association for Computational Linguistics.

Alan Ramponi and Barbara Plank. 2020. Neural unsupervised domain adaptation in NLP—A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6838–6855, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Adwait Ratnaparkhi. 1999. Learning to parse natural language with maximum entropy models. *Machine learning*, 34:151–175.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Sebastian Ruder and Barbara Plank. 2017. Learning to select data for transfer learning with Bayesian optimization. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 372–382, Copenhagen, Denmark. Association for Computational Linguistics.

Evan Sandhaus. 2008. The new york times annotated corpus.

Marta Sandri, Elisa Leonardelli, Sara Tonelli, and Elisabetta Jezek. 2023. Why don't you do it right? analysing annotators' disagreement in subjective tasks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2428–2441, Dubrovnik, Croatia. Association for Computational Linguistics.

Danielle Saunders. 2022. Domain adaptation and multi-domain adaptation for neural machine translation: A survey. *Journal of Artificial Intelligence Research*, 75:351–424.

Satoshi Sekine. 1997. The domain dependence of parsing. In *Fifth Conference on Applied Natural Language Processing*, pages 96–102, Washington, DC, USA. Association for Computational Linguistics.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, pages 1–9.

Benno Stein and Sven Meyer Zu Eissen. 2006. Distinguishing topic from genre. In *Proceedings of the 6th International Conference on Knowledge Management (I-KNOW 06)*, pages 449–456. Citeseer.

Ming Tan, Yang Yu, Haoyu Wang, Dakuo Wang, Saloni Potdar, Shiyu Chang, and Mo Yu. 2019. Out-of-domain detection for low-resource 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)*, pages 3566–3572, Hong Kong, China. Association for Computational Linguistics.

Alexandra N Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Rob van der Goot, Ahmet Üstün, and Barbara Plank. 2021a. On the effectiveness of dataset embeddings in mono-lingual,multi-lingual and zero-shot conditions. In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 183–194, Kyiv, Ukraine. Association for Computational Linguistics.

Rob van der Goot, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf, and Barbara Plank. 2021b. Massive choice, ample tasks (MaChAmp): A toolkit for multi-task learning in NLP. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 176–197, Online. Association for Computational Linguistics.

Marlies van der Wees, Arianna Bisazza, Wouter Weerkamp, and Christof Monz. 2015. What's in a domain? analyzing genre and topic differences in statistical machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 560–566, Beijing, China. Association for Computational Linguistics.

Sida Wang and Christopher Manning. 2012. Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th*

Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 90–94, Jeju Island, Korea. Association for Computational Linguistics.

Bonnie Webber. 2009. Genre distinctions for discourse in the Penn TreeBank. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 674–682, Suntec, Singapore. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *ArXiv*.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. LUKE: Deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454, Online. Association for Computational Linguistics.

Amir Zeldes. 2017. The gum corpus: creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

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, December 7-12, 2015, Montreal, Quebec, Canada*, pages 649–657.

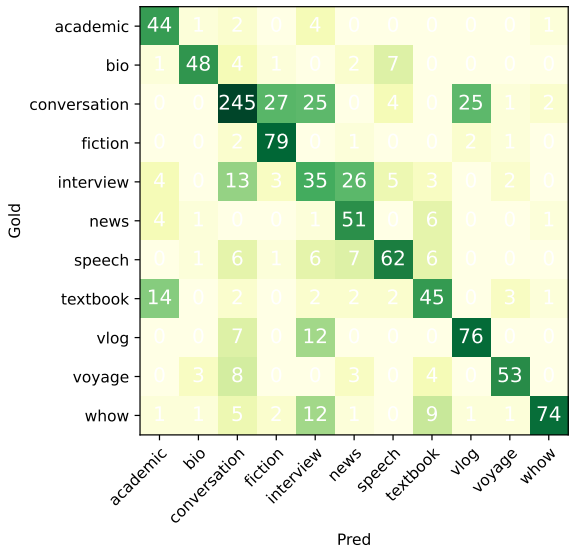


Figure 7: Confusion matrix on the sentence level, numbers are summed over all five random seeds.

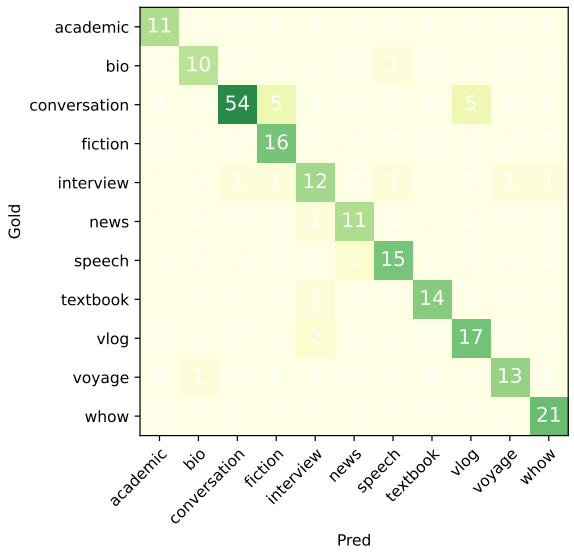


Figure 8: Confusion matrix on the prose level, numbers are summed over all five random seeds.

Appendix

A. Confusion Matrices Genre

In Figure 7-Figure 9 we plot the confusion matrices of our DeBERTa model trained on the gold genre labels. The conversation genre shows to be the most difficult label; it is commonly confused with fiction, interview and vlog; which also overlap in length (Section 4).

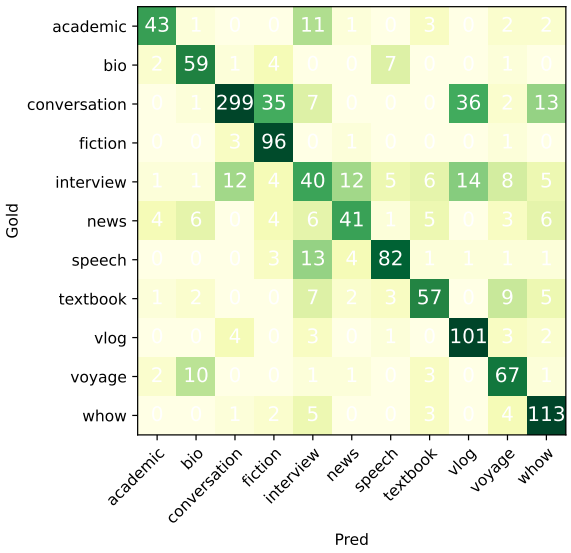


Figure 9: Confusion matrix on all data, numbers are summed over all five random seeds.

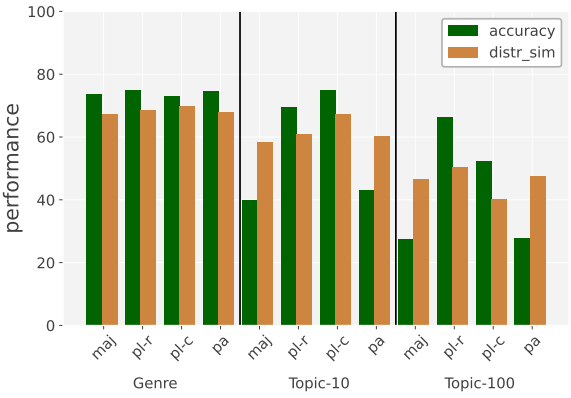


Figure 10: Results of our proposed models on the sentence level data.

B. Sentence and Prose Results

In Figure 10 we show the results of our proposed models trained and evaluated only on the sentence level data. Figure 11 has the same evaluation on the prose level data.

C. Visualization of Embeddings

We encode sentences using Sentence-BERT (Reimers and Gurevych, 2019), apply a PCA-downprojection, and color each sentence according to gold genres, our majority-vote genre annotations, as well as majority-vote topic-10 annotations. The results are shown in Figures 12–14.

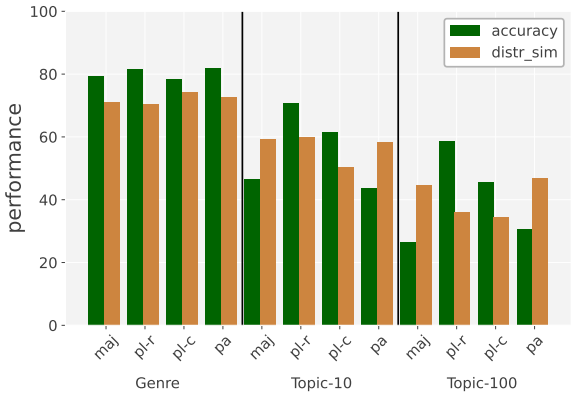


Figure 11: Results of our proposed models on the prose level data.

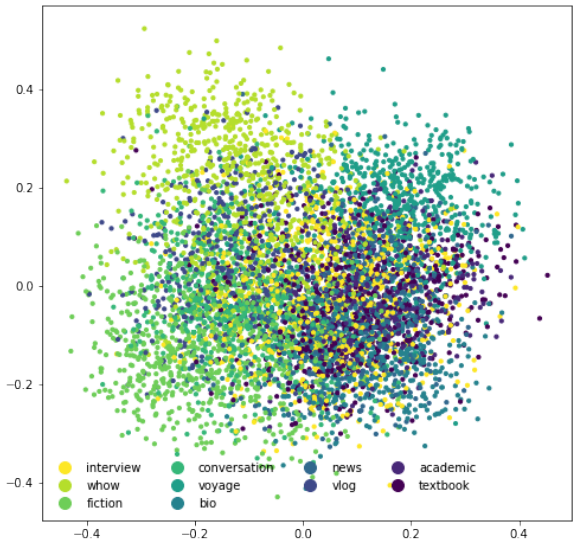


Figure 12: PCA plot of sentence embeddings with the gold genres.

D. Prose-level Statistics

Label statistics on the prose level are shown in Figure 15. While general trends, such as the majority genres and topics remain the same as on the sentence level, additional context spreads annotations more evenly, and allows for disambiguations such as for spoken data genres. This is also reflected in the higher alignment between gold and annotated genre labels—both in terms of number, but also in terms of accuracy (Table 2). For topic, we further observe almost an order of magnitude fewer no-topic annotations, which are consequently distributed across the spectrum of actual topics.

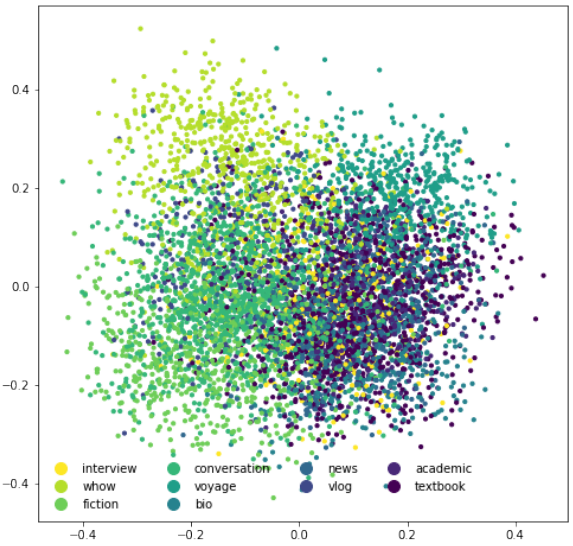


Figure 13: PCA plot of sentence embeddings with our annotation for genres, majority vote is used for each instance.

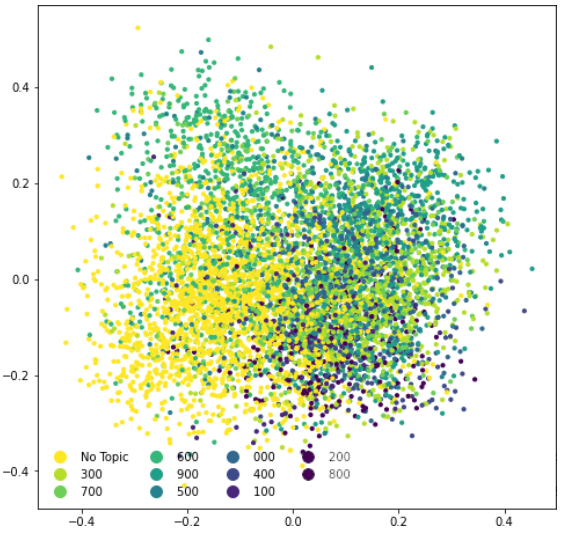


Figure 14: PCA plot of sentence embeddings with our annotation for coarse topics, majority vote is used for each instance.

E. Annotator Comments

Annotators were provided with a free-form field to provide optional comments regarding each annotation. Of the final dataset, 3.9% of annotations have an annotator comment attached, with a median length of 38 characters. They primarily contain explanations of annotations which were marked with high annotator uncertainty.

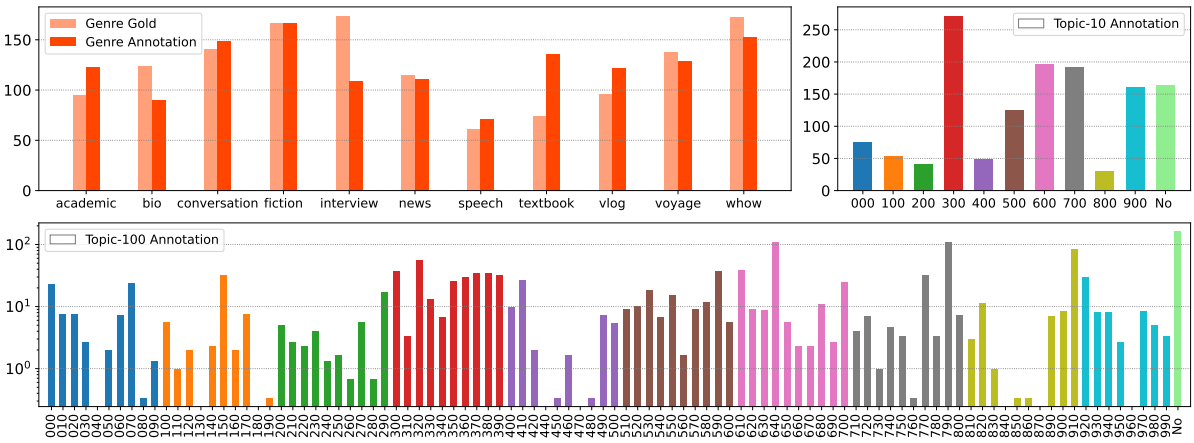


Figure 15: **Distribution of Labels (Prose).** Frequency distributions of the labels in gold genre labels, annotations of genres, annotations of topic-10, and annotations of topic-100 (log scale). For the annotations, the number is divided by three to get an average distribution. The mapping of topic-10 and topic-100 labels can be found in Appendix F. The tag "No" in the topic annotations means "No topic".

F. Guidelines

Goal/Task

In this annotation project, we are interested in knowing what the topic and genre is of a sentence and whether we humans can identify these. For Topics, we make use of the Dewey Decimal Classification (DDC) system. For genres, we make use of the genres provided in the Georgetown University Multilayer Corpus (GUM) corpus. The goal is to put the sentence/paragraph at hand into the most probable class (determined by you).

Genre has a *one-layer* annotation scheme, while **Topic** has a *two-layer* annotation scheme, which we will refer to as L1 and L2. We want to annotate for all three. There is an option for "Not Sure" (abbreviated to "NS"). This is when you feel that the label for the sentence is not present in the options. In addition, feel free to add any notes for clarification (e.g., clarify your choice or something else).

Preliminaries

Below we give an introduction to the topics and genre labels of this annotation project. It takes around 15-20 minutes to read. Note that you don't have to remember the label numbers. This introduction is to make you aware of the definition of the classes. All the labels are present in the annotation spreadsheet

Introduction Genres

We make use of the text types (genres) in the GUM corpus. These genres do not have a specific number like the topics above. Therefore we simply enumerate them. The genres are the following:

- Academic
- Bio
- Conversation
- Fiction
- Interview
- News
- Speech
- Textbook
- Vlog
- Voyage
- Whow

Brief explanation of the genre classes

- **Academic** (writing) is nonfiction writing adhering to academic standards and disciplines. It includes research reports, monographs, and undergraduate versions. It uses a formal style, references other academic work, and employs consistent rhetorical techniques to define scope, situate in research, and make new contributions.

- A **bio**graphy is a detailed description of a person's life. It involves more than just basic facts like education, work, relationships, and death; it portrays a person's experience of these life events. Unlike a profile or curriculum vitae (résumé), a biography presents a subject's life story, highlighting various aspects of their life, including intimate details of experience, and may include an analysis of the subject's personality. Biographical works are usually nonfiction, but fiction can also be used to portray a person's life. One in-depth form of biographical coverage is called legacy writing. Works in diverse media, from literature to film, form the genre known as biography. An authorized biography is written with the permission, cooperation, and at times, participation of a subject or a subject's heirs. An autobiography is written by the person themselves, sometimes with the assistance of a collaborator or ghostwriter.

- **Conversation**: naturally occurring spoken interaction. Represents a wide variety of people of different regional origins, ages, occupations, genders, and ethnic and social backgrounds. The predominant form of language use represented is face-to-face conversation, but lso documents many other ways that that people use language in their everyday lives: telephone conversations, card games, food preparation, on-the-job talk, classroom lectures, sermons, story-telling, town hall meetings, tour-guide spiels, and more. Fiction refers to creative works, particularly narrative works, that depict imaginary individuals, events, or places. These portrayals deviate from history, fact, or plausibility. In our data, fiction pertains to written narratives like novels, novellas, and short stories.

- An **interview** is a structured conversation where one person asks questions and another person answers them. It can be a one-on-one conversation between an interviewer and an interviewee. The information shared during the interview can be used or shared with others.

- **News** is information about current events, shared through various media like word of mouth, printing, broadcasting, electronic communication, and witness testimonies. It covers topics such as war, government, politics, education, health, environment, economy, business, fashion, entertainment, sports, and unusual events. Government announcements and technological advancements have accelerated news dissemination and influenced its content.

- A (political) **speech** is a public address given by a political figure or a candidate for public office, usually with the aim of persuading or mobilizing an audience to support their ideas, policies, or campaigns. Political speeches are an essential tool for politicians to communicate their vision, articulate their positions, and connect with voters or constituents.

- A **textbook** is a book containing a comprehensive compilation of content in a branch of study with the intention of explaining it. Textbooks are produced to meet the needs of educators, usually at educational institutions. Schoolbooks are textbooks and other books used in schools. Today, many textbooks are published in both print and digital formats.

- A **vlog**, also known as a video blog or video log, is a form of blog for which the medium is video. The dataset contains transcripts of the speech occurring in the video.

- A travel/**voyage** guide is a wiki providing information for visitors or tourists about a particular place. It typically includes details about attractions, lodging, dining, transportation, and activities. It may also contain maps, historical facts, and cultural insights. Guide wikis cater to various travel preferences, such as adventure, relaxation, budget, or specific interests like LGBTQ+ travel or dietary needs.

- A Wikihow how-to (**whow**) guide is an instructional document that offers step-by-step guidance on accomplishing a specific task or reaching a particular goal. It aims to assist individuals in learning and comprehending the process involved in successfully completing the task. These guides are typically written in a clear and concise manner, simplifying complex processes into manageable steps. They often include detailed explanations, diagrams, illustrations, or examples to enhance understanding. How-to guides cover various topics, such as technical tasks, practical skills, creative endeavors, troubleshooting, and more.

Introduction Topics

The DDC system is a widely used library classification system developed by Melvil Dewey in the late 19th century. The DDC is based on the principle of dividing knowledge (in our case sentences) into ten main classes, each identified by a three-digit number; we only focus on the first two:

1. The ten main classes in the Dewey Decimal Classification system are as follows:

- 000 Computer science, information & general works
- 100 Philosophy & psychology
- 200 Religion
- 300 Social sciences
- 400 Language
- 500 Science
- 600 Technology
- 700 Arts & recreation
- 800 Literature
- 900 History & geography

These higher level classes belong to L1 in the annotation spreadsheet, and we added the NO-TOPIC label (see description below)

2. Each main class is further divided into subclasses using additional digits (10s). For example, in the 500s (natural sciences and mathematics), you'll find 510 for mathematics, 520 for astronomy, 530 for physics, and so on. The system allows for more specific classification of books and materials based on their subject matter.

   See the following page: https://www.oclc.org/content/dam/oclc/dewey/ddc23-summaries.pdf

   This page separates the ten classes above into more finer-grained classes. There is not an explanation for each of them, but usually the name of the label encapsulates the subclass already. Note that the subclasses overwrite the main classes (so you can't pick 400 and 510, then you'd have to change 510 to 500).

   These subclasses belong to L2 in the annotation spreadsheet.

   Note that for each fine-grained class we deem the main number/code (e.g., 100, 200, 300) in L2 as the No-topic/Other category. The "Other" class can only be chosen in the fine-grained label classes (L2). Choosing this means that you believe that the current sentence belongs to a specific class. But the label is not present.

The Dewey Decimal Classification system is used in many libraries around the world to organize their collections and make it easier for users to locate resources. It provides a systematic way of arranging materials and enables efficient browsing and retrieval of information based on subject areas.

Brief explanation of the topic classes (L1)

- 000 Computer science, information & general works is the most general class and is used for works not limited to any one specific discipline, e.g., encyclopedias, newspapers, general periodicals. This class is also used for certain specialized disciplines that deal with knowledge and information, e.g., computer science, library and information science, journalism. Each of the other main classes (100-900) comprises a major discipline or group of related disciplines. Note that in our experiments, we do not consider this a miscellaneous category, we have "No-topic" for this.
- 100 Philosophy & psychology covers philosophy, parapsychology and occultism, and psychology.
- 200 Religion is devoted to religion.
- 300 Social sciences covers the social sciences. Class 300 includes sociology, anthropology, statistics, political science, economics, law, public administration, social problems and services, education, commerce, communications, transportation, and customs.
- 400 Language comprises language, linguistics, and specific languages. Literature, which is arranged by language, is found in 800.
- 500 Science is devoted to the natural sciences and mathematics.
- 600 Technology is technology.
- 700 Arts & recreation covers the arts: art in general, fine and decorative arts, music, and the performing arts. Recreation, including sports and games, is also classed in 700.
- 800 Literature covers literature, and includes rhetoric, prose, poetry, drama, etc. Folk literature is classed with customs in 300.
- 900 History & geography is devoted primarily to history and geography. A history of a specific subject is classed with the subject.
- No topic: For cases where the topic can not be determined, or even guessed. For example for utterances that contain no natural language or do not have enough context.

FAQ

- Should the colors of L1 and L2 in the annotation spreadsheet match?

Yes, apart from that the colours should match, the first number of the class to which the sentence belongs should also match.

For example, a sentence that belongs to Arts (700), is restricted to anything in the 700 class, e.g., a painting (750).

- If a sentence has a clear topic in general, but the L2 category does not match, how do we annotate?

  The fine-grained (L2) topics have the priority, and since they have to match you adjust the main topic accordingly.

- Does my choice of Topic depend on the Genre or vice versa?

  No, by default, annotating for genre and topic should be a separate task and should not influence each other.

- How do we distinguish between something that is in the No-topic (or Others) class and NS (”not sure”)?

  Use the “others” category when you believe the current instance to belong to a class which is not in the listed ones. Mark your choice with “NS” when you have a guess, but you are not confident about it (e.g., because the instance is very short, or you are not familiar with the genre/topic)

  If you are able to find L1, but none of the labels fit for the sentence in L2, you should choose ”Other” (e.g, 000, 100, 200, etc.) in the same colour (class) of L2. The ”Other” class can only be chosen in the fine-grained label classes (L2). Choosing this means that you believe that the current sentence belongs to a specific class. But the label is not present. Otherwise, mark your best guess with “NS”.

- Is it better to label a sentence as ”NO-TOPIC” if there is not a clear label associated with it or are we encouraged to take a guess?

  You are encouraged to take a guess. However, for cases where you have no preference for any of the labels (i.e. a wild guess), label it as NO-TOPIC.

- There is already another ”Other” class in Religion/Language (e.g., 290 Other religion).

  Good catch, imagine this situation. Let’s say the sentence is talking about Buddhism. This falls under 290, because we’re talking about another religion. However, if the sentence is ”vaguely” talking about religion and doesn’t fit within any of the labels, then choose 200 (Other).

- Where do ads/exam questions fit?

  In whichever of the genres you would expect to come across advertisements/exam questions. However, note that the data is scraped from the main information channel of source (i.e., advertisements next to a news text or before a vlog are not included).

- Can we use external resources?

  External resources are allowed, but do not look up the literal sentence.

- How to pick topics (L1/L2) for fiction (genre)?

  Note that the genre and topic tasks should be seen as distinct tasks. So, the genre fiction should not automatically lead to a literature topic label (unless the fiction work is about literature).

- Some utterances seem to be taken from the same text; do we have to give them the same label, or take the contexts into account?

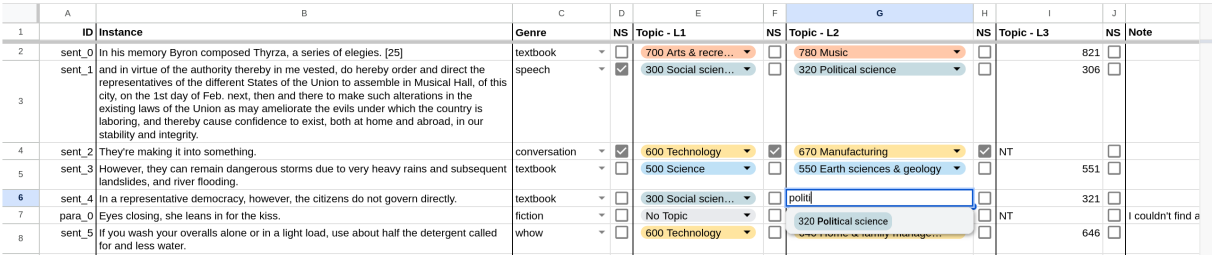
  No, each utterance should be judged independently.

Note for L3:

- For each L2, there is a finer-grained class namely L3. These numbers go in the thousands. Now, try to pick the most likely thousands’ topic:
  - You will have to refer to the PDF (L3-1000.pdf) for the right classes.
  - Please write the class number in the spreadsheet cell. There is no dropdown menu.
- The ”no-topic” option still exists. Use ”NT”;
- You should pick the fine-grained L3 topic that best fits the utterance. This time you don’t have to match the L1-L2 categories, but we ask you to NOT update your previous L1-L2 annotations, and just annotate L3 independently.

G. Annotation Tool

We used Google Spreadheets for annotation. The setup is shown in Figure 16.



	A	B	C	D	E	F	G	H	I	J	
1	ID	Instance	Genre	NS	Topic - L1	NS	Topic - L2	NS	Topic - L3	NS	Note
2	sent_0	In his memory Byron composed Thyrza, a series of elegies. [25]	textbook	☐	700 Arts & recre…	☐	780 Music	☐	821	☐	
3	sent_1	and in virtue of the authority thereby in me vested, do hereby order and direct the representatives of the different States of the Union to assemble in Musical Hall, of this city, on the 1st day of Feb. next, then and there to make such alterations in the existing laws of the Union as may ameliorate the evils under which the country is laboring, and thereby cause confidence to exist, both at home and abroad, in our stability and integrity.	speech	☑	300 Social scien…	☐	320 Political science	☐	306	☐	
4	sent_2	They're making it into something.	conversation	☑	600 Technology	☑	670 Manufacturing	☑	NT	☐	
5	sent_3	However, they can remain dangerous storms due to very heavy rains and subsequent landslides, and river flooding.	textbook	☐	500 Science	☐	550 Earth sciences & geology	☐	551	☐	
6	sent_4	In a representative democracy, however, the citizens do not govern directly.	textbook	☐	300 Social scien…	☐	politi	☐	321	☐	
7	para_0	Eyes closing, she leans in for the kiss.	fiction	☐	No Topic	☐	320 Political science	☐	NT	☐	I couldn't find a
8	sent_5	If you wash your overalls alone or in a light load, use about half the detergent called for and less water.	whow	☐	600 Technology	☐		☐	646	☐	

Figure 16: Example of annotation in Google Spreadsheets. NS = Not Sure