LANGUAGE UNDERSTANDING IN THE HUMAN MACHINE ERA 2024

Proceedings of 1st LUHME Workshop

Edited by

Rui Sousa-Silva

Henrique Lopes Cardoso

Maarit Koponen

Antonio Pareja-Lora

Márta Seresi

LUHME @ ECAI 2024

Santiago de Compostela, Spain

October 20, 2024

Published by
CLUP - Centro de Linguística da Universidade do Porto (https://doi.org/10.54499/UIDB/00022/2020)
FLUP - Faculdade de Letras da Universidade do Porto

ISBN 978-989-9193-12-3
DOI https://doi.org/10.21747/978-989-9193-12-3/lan

October, 2025

Introduction: Understanding Language in the Human-Machine Era

Large language models (LLMs) have revolutionized the way interactional artificial intelligence (AI) systems are developed, as they are made available to ordinary users. Significant advances have been observed in applications such as conversational AI and machine translation, and their widespread use in the so called human-machine era, where technology is integrated with our senses (Sayers et al., 2021), is undeniable; those models have produced remarkable achievements in several benchmarks (Gao et al., 2021; Hendrycks et al., 2021; Wang et al., 2019; Zhou et al., 2020), and the scientific community has discussed emergent properties (Wei et al., 2022) that result from scaling laws (Kaplan et al., 2020). Nevertheless, state-of-the-art systems are still prone to brittleness in language understanding, which raises doubts about the extent to which such systems can truly understand human language(s) (Mitchell & Krakauer, 2023).

The concept of language understanding has always been controversial (Lyons, 1990; Michael et al., 2023). As contemporary linguistic theories have shown, meaning-making relies not only on form and (immediate) meaning, but also on context. Thus, understanding natural language entails more than observing the form and the meaning withdrawn from that form; instead, harnessing meaning (Bender & Koller, 2020) requires access to grounding. Understanding language is, hence, a very complex task, even for humans (Lyons, 1990). As discourse, pragmatics, and (social) context are particularly relevant for understanding language, how to equip language models with such linguistics-grounded capabilities is yet to be fully understood.

Nevertheless, language models are seemingly capable of generalizing concepts, which could arguably be seen as some kind of meaning understanding (Piantadosi & Hill, 2022), even if modest. Understanding language in the human-machine era is, therefore, a doubly challenging task. Besides understanding the intrinsic capabilities of LLMs, it is increasingly important to investigate the requirements and impact of using such systems in real-world applications. As has been empirically demonstrated, LLMs can be used effectively in various applications, even without sophisticated language understanding skills, but the lack of theories that support these findings raises concerns about which kinds of applications, particularly those dealing directly with human interaction, pose greater risks and ethical concerns. Notable examples include the impact of language technology on teaching and language work. For instance, research is underway into using language models in educational settings, including question-generation (Leite & Lopes Cardoso,

2023); likewise, machine translation (MT) is increasingly ubiquitous, as it is used by both language professionals and general speakers at (apparently) no cost. Yet, as MT systems can take in a limited amount of context, they tend to make mistakes similar to those of human translators, who need to rely on their own knowledge to do their job more accurately.

As the way AI systems are intertwined with human expertise in language understanding is quickly changing, some have raised the question of the role played by language professionals in tasks such as translation. These professionals systematically add value to building next-generation language models that use linguistic and common-sense knowledge to provide more robust systems.

The "Language Understanding in the Human-Machine Era" (LUHME) workshop retrieves, resumes and refocuses the longstanding debate about the role of understanding in natural language use and related applications. In particular, the workshop provides insight into what language understanding is and whether it is required for computational natural language tasks, such as machine translation and natural language generation. Additionally, it furthers the discussion about the role played by language professionals (e.g., linguists, professional translators, and language teachers) in computational natural language understanding.

LUHME brings together researchers interested in the intersection between language understanding and the effective use of language technologies in human-machine interaction to discuss, among others, language understanding in/by LLMs; language grounding; psycholinguistic approaches to language understanding; discourse, pragmatics and language understanding; socio-cultural aspects in understanding language(s); effects of language misunderstanding by computational models; manifestations of language understanding; linguistic theory and language understanding by machines; linguistic, world, and common sense knowledge in language understanding; machine translation and/or interpreting and language understanding; human vs. machine language understanding; role of language professionals in the LLMs era; understanding language and explainable AI.

Rui Sousa-Silva
Henrique Lopes Cardoso
Maarit Koponen
Antonio Pareja-Lora
Márta Seresi

References

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185–5198. https://doi.org/10.18653/v1/2020.acl-main.463

Gao, L., Tow, J., Biderman, S., & et al. (2021). *A framework for few-shot language model evaluation* (Version v0.0.1). Zenodo. https://doi.org/10.5281/zenodo.5371628

Hendrycks, D., Burns, C., Basart, S., & et al. (2021). Measuring massive multitask language understanding. *International Conference on Learning Representations*. https://openreview.net/forum?id=d7KBjmI3GmQ

Kaplan, J., McCandlish, S., Henighan, T., & et al. (2020). Scaling laws for neural language models. *CoRR*, *abs/2001.08361*. https://arxiv.org/abs/2001.08361

Leite, B., & Lopes Cardoso, H. (2023). Towards enriched controllability for educational question generation. In N. Wang, G. Rebolledo-Mendez, N. Matsuda, O. C. Santos, & V. Dimitrova (Eds.), *Artificial intelligence in education* (pp. 786–791). Springer Nature Switzerland.

Lyons, J. (1990). *Language and linguistics: An introduction*. Cambridge University Press.

Michael, J., Holtzman, A., Parrish, A., Mueller, A., Wang, A., Chen, A., Madaan, D., Nangia, N., Pang, R. Y., Phang, J., & Bowman, S. R. (2023). What do NLP researchers believe? results of the NLP community metasurvey. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 16334–16368. https://doi.org/10.18653/v1/2023.acl-long.903

Mitchell, M., & Krakauer, D. C. (2023). The debate over understanding in ai's large language models. *Proceedings of the National Academy of Sciences*, *120*(13), e2215907120. https://doi.org/10.1073/pnas.2215907120

Piantadosi, S. T., & Hill, F. (2022). Meaning without reference in large language models.

Sayers, D., Sousa-Silva, R., Höhn, S., & et al. (2021). *The dawn of the human–machine era: A forecast of new and emerging language technologies* (tech. rep.). EU COST Action CA19102 'Language In The Human–Machine Era'. https://doi.org/https://doi.org/10.17011/jyx/reports/20210518/1

Wang, A., Pruksachatkun, Y., Nangia, N., & et al. (2019). Superglue: A stickier benchmark for general-purpose language understanding systems. In *Procs. 33rd international conference on neural information processing systems*. Curran Associates Inc.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). Emergent abilities of large language models [Survey Certification]. *Transactions on Machine Learning Research*. https://openreview.net/forum?id=yzkSU5zdwD

Zhou, X., Zhang, Y., Cui, L., & Huang, D. (2020). Evaluating commonsense in pre-trained language models. *Procs. 34th AAAI, New York, USA, February 7-12, 2020*, 9733–9740. https://doi.org/10.1609/aaai.v34i05.6523

Organising Committee

Rui Sousa-Silva
University of Porto, Portugal

Henrique Lopes Cardoso
University of Porto, Portugal

Maarit Koponen
University of Eastern Finland, Finland

Antonio Pareja-Lora
Universidad de Alcalá, Spain

Márta Seresi
Eötvös Loránd University, Hungary

Assistants

Andrés-Felipe Herrera-Ramírez
Universidad de Alcalá, Spain

Ana Sofia Meneses Silva
University of Porto, Portugal

Karen de Souza
University of Eastern Finland, Finland

Programme Committee

Aida Kostikova, *Bielefeld University, Germany*
Alex Lascarides, *University of Edinburgh, United Kingdom*
Alípio Jorge, *University of Porto, Portugal*
António Branco, *University of Lisbon, Portugal*
Belinda Maia, *University of Porto, Portugal*
Caroline Lehr, *ZHAW School of Applied Linguistics, Switzerland*
Diana Santos, *Universitetet i Oslo, Norway*
Efstathios Stamatatos, *University of the Aegean, Greece*
Ekaterina Lapshinova-Koltunski, *University of Hildesheim, Germany*
Eliot Bytyçi, *Universiteti i Prishtinës "Hasan Prishtina", Kosovo*
Hanna Risku, *University of Vienna, Austria*
Lynne Bowker, *University of Ottawa, Canada*
Nataša Pavlović, *University of Zagreb, Croatia*
Paolo Rosso, *Universitat Politècnica de València, Spain*
Ruslan Mitkov, *Lancaster University, United Kingdom*
Sule Yildirim Yayilgan, *Norwegian University of Science and Technology, Norway*

Keynote Speakers

Alexander Koller, *Saarland University*
Anders Søgaard, *University of Copenhagen*

Contents

Keynote Speakers

Keynote Speaker

Why Most Are Wrong About LLM Understanding

Anders Søgaard

University of Copenhagen

Abstract: I identify a fallacy common to many LLM no-go theorems of the form: LLMs cannot do X, because they were designed to – or trained to – do Y. I present observations that seem to challenge stochastic parrot or database views of LLMs, as well as arguments for why, contrary to popular belief, structural similarity may be sufficient for grounding.

Bio: Anders Søgaard is in a dual position as Professor of Computer Science and Professor of Philosophy at the University of Copenhagen. He is a recipient of an ERC Starting Grant, a Google Focused Research, a Carlsberg Semper Ardens Advanced, and has won eight best paper awards. He has written more than 300 articles and five academic books.

Keynote Speaker

Untrustworthy and still revolutionary: Some thoughts on how LLMs are changing NLP

Alexander Koller

Saarland University

Abstract: There is no doubt that large language models (LLMs) are revolutionizing the field of natural language processing (NLP) in many ways. There are many doubts on whether this a good thing, whether we will ever be able to overcome their inability to reliably distinguish truth from falsehood, whether there is any place left for pre-LLM models, and how to do good science any more.

I do not have definitive answers on any of these questions, and am personally torn on many of them. In this talk, I will first discuss some recent research on the limitations of LLMs for semantic parsing and on overcoming them through the use of neurosymbolic models. I will then discuss recent work on the extent to which LLMs can capture world knowledge and apply it to planning and reasoning tasks. I will conclude with some general thoughts on science and engineering in NLP in the era of LLMs.

Bio: Alexander Koller is a Professor of Computational Linguistics at Saarland University in Saarbrücken, Germany. His research interests include planning and reasoning with LLMs, syntactic and semantic processing, natural language generation, and dialogue systems. He is particularly interested in neurosymbolic models that bring together principled linguistic modeling and correctness guarantees with the coverage and robustness of neural approaches. Alexander received his PhD from Saarland University and was previously a postdoc at Columbia University and the University of Edinburgh, faculty at the University of Potsdam, and Visiting Senior Research Scientist at the Allen Institute for AI.

Workshop Presentations

Converso: Improving LLM Chatbot Interfaces and Task Execution via Conversational Forms

Gianfranco Demarco, Nicola Fanelli[a,*], Gennaro Vessio[a] and Giovanna Castellano[a]

[a]Department of Computer Science, University of Bari Aldo Moro, Italy

Abstract. Recent advancements in large language models (LLMs) have enabled more autonomous conversational AI agents. However, challenges remain in developing effective chatbots, particularly in addressing LLMs' lack of "statefulness". This paper presents Converso, a novel chatbot framework that introduces a new conversation flow based on stateful conversational forms designed for natural data acquisition through dialogue. Converso leverages LLMs, LangChain, and a containerized architecture to provide an end-to-end chatbot system with Telegram as the user interface. The key innovation in Converso is its implementation of conversational forms, which guide users through form completion via a structured dialogue flow. Converso's chatbots can be linked with multiple forms that are automatically triggered based on the user's intent. Our forms are fully integrated into the LangChain ecosystem, allowing the LLM to use tools for form completion and dynamic validation. Evaluations show that this approach significantly improves task completion rates compared to LLMs alone. Converso demonstrates how specifically designed conversational flows can enhance the capabilities of LLM-based chatbots for practical data collection applications. Our implementation is available at: https://github.com/gianfrancodemarco/converso-chatbot.

1 Introduction

Chatbots have emerged as one of the most widely adopted applications of artificial intelligence (AI) in consumer products and services. These conversational agents directly interact with end users through natural language interfaces, serving various domains such as entertainment, education, information retrieval, e-commerce, and more [1]. Since the early conceptualization of chatbots in the 1960s, numerous approaches have been explored to enhance their capabilities, transitioning from basic pattern-matching techniques to leveraging advanced machine learning algorithms and language models.

Recent advancements in large language models (LLMs) have led to a significant shift, advancing chatbots to new levels of independence and conversational ability [2]. LLMs are complex neural network architectures with billions of parameters, trained on extensive text corpora. The extensive data on which these models are trained, along with additional techniques like reinforcement learning from human feedback (RLHF) [14], enables them to generate natural language responses that closely resemble human communication. However, LLMs still face limitations, including a lack of access to up-to-date knowledge, an inability to perform complex reasoning, and difficulties interacting with external environments [13]. Researchers have developed techniques such as retrieval-augmented generation (RAG) [8] and model-calling capabilities [15] to address these challenges, allowing LLMs to leverage external data sources and tools during inference. Furthermore, open-source frameworks like LangChain [3] have emerged to streamline the development of LLM-based applications, including chatbots. LangChain provides a comprehensive suite of libraries, tools, and templates that facilitate the integration of LLMs, RAG techniques, and external tools, enabling the creation of context-aware and reasoning-capable chatbot systems.

This paper explores the design and implementation of modern chatbot systems leveraging LLMs, the LangChain ecosystem, and related techniques. It introduces Converso, a novel chatbot framework that incorporates a conversation flow based on *conversational forms*, enhancing traditional web forms for data acquisition through natural language interactions. The paper discusses Converso's system architecture, conversational flow, and use cases, highlighting the benefits of integrating LLMs and the LangChain ecosystem. Additionally, an evaluation protocol is presented to assess the effectiveness of conversational forms in improving task completion rates.

The rest of this paper is structured as follows: Section 2 discusses related work, Section 3 presents the Converso framework, Section 4 describes our experimental use cases and presents the results obtained, and Section 5 concludes the paper.

2 Related Work

Chatbots have been an active area of research for a long time, with early systems like ELIZA [18] and PARRY [6] employing pattern-matching techniques with pre-defined rules and responses. Subsequent work incorporated machine learning algorithms for intent classification and entity extraction [1], leading to more advanced chatbots capable of understanding user intents and relevant context.

With recent developments in LLMs like GPT-3 [2] and PaLM [5], there has been renewed interest in leveraging these powerful models for building conversational AI systems. LLMs have demonstrated impressive emergent abilities in few-shot prompting settings [2] and can engage in substantive multi-turn dialogues by conditioning on previous conversation history [17].

However, LLMs still face limitations such as hallucinating incorrect facts [12], being confined to their training data distributions, and lacking mechanisms to interact with external tools or information sources. Several techniques have been proposed to overcome these limitations. For instance, retrieval-augmented generation (RAG) [8] integrates external data retrieval into the language model's generation process to enhance factual accuracy and access to up-to-date information. Tool calling [15, 19] allows LLMs to call and receive

* Corresponding author. Email: nicola.fanelli@uniba.it.

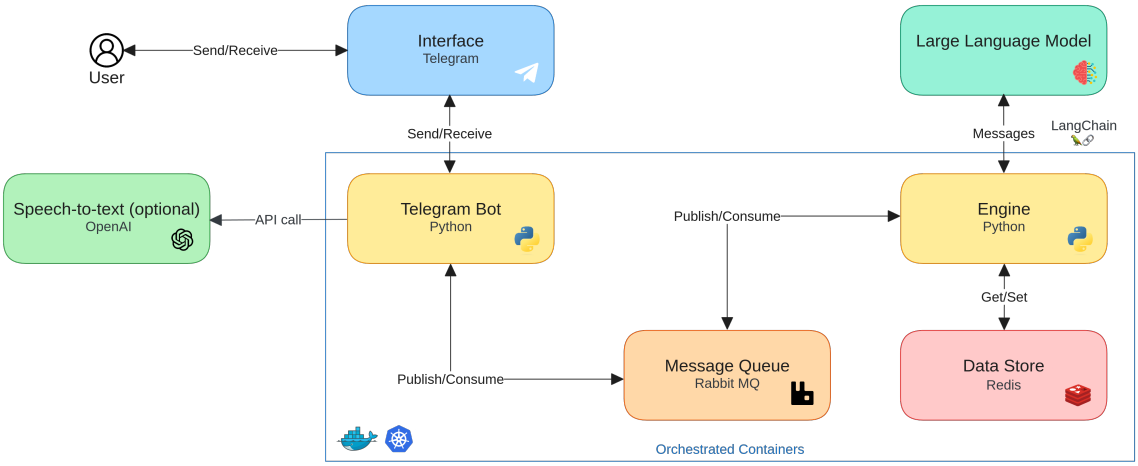


Figure 1: Overview of the system architecture of a chatbot implemented using the Converso framework. The user interacts through Telegram as the interface. User messages are written to the message queue by the Telegram bot, which also consumes the LLM messages to send chatbot responses back to the interface. We also allow for vocal user inputs via OpenAI speech-to-text APIs. The engine is responsible for providing prompts to the LLM via LangChain and maintaining the conversation history in a Redis data store. We use a fully containerized architecture, implementing our components inside Docker containers orchestrated by Kubernetes.

results from external tools/APIs, enabling capabilities beyond pure text generation, like mathematical reasoning [16] and real-world interactions [10].

Such advancements, combined with the ability to influence the behavior of LLMs through prompt engineering [4], have opened a wide range of applications for chatbot systems, such as data acquisition. For example, Hakimov et al. [9] proposed a modular system based on LLMs for form filling, creating a method to evaluate dialogues using user simulation with an LLM. Additionally, some simple (usually closed-source) implementations of forms for data acquisition with LLMs have emerged online.

In contrast to these approaches, we propose a framework for creating chatbots that is fully integrated with the LangChain ecosystem. Our innovative approach does not tie the LLM to a specific form, allowing chatbot developers to specify an arbitrary number of forms that can be invoked based on user intents. This flexibility enables users to interact with the chatbot for various goals beyond executing actions requiring form completion. Furthermore, with Converso, we propose a fully containerized architecture streamlining the chatbot creation process.

3 Our Framework: Converso

In this section, we present Converso, a framework developed as an extension of LangChain that enables building conversational forms for goal-oriented interactions. Converso introduces *stateful* conversational forms that guide users through structured data collection processes, reducing reliance on lengthy conversation histories and mitigating hallucinations or deviations from the intended goals. Converso facilitates the creation of fully containerized chatbot applications, leveraging Kubernetes for container orchestration, an event manager for asynchronous communication, and a Telegram bot interface as the front end. This enables a production-ready, scalable implementation of conversational chatbots that can be seamlessly integrated into existing systems.

3.1 System Architecture

Converso implements a fully functional chatbot system composed of several components. An overview of the system architecture of the Converso project is presented in Fig. 1.

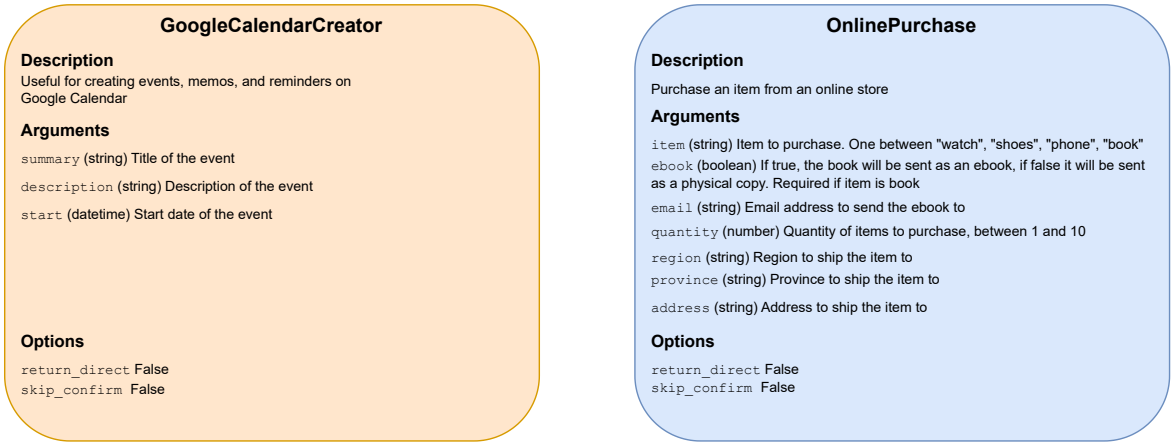
All components of the Converso chatbot are created as Docker containers orchestrated by Kubernetes. Users can interact with the LLM via a Telegram bot interface. RabbitMQ is used as a message broker to enable decoupling and asynchronous request processing. Finally, Redis stores conversation history and other data, such as user credentials.

3.2 Conversation Flow

The conversation flow generates a textual response starting from the user's input. This flow is constructed as a graph using LangGraph, a library for multi-actor interactions within LangChain. The primary components of our conversation flow are the Base Agent and the Error Agent. The Base Agent is responsible for conducting the conversation with the user, while the Error Agent assists by correcting any errors that may occur during the interaction. Agents are specific instances of the LLM, each equipped with unique system prompts that guide their behavior.

The conversation flow consists of the following steps:

1. The user's input and conversation history are injected into a prompt template to create the final input for the model.
2. The chosen LLM, instantiated as the Base Agent, is called with the rendered prompt as input.
3. The model's output is evaluated, with three possible outcomes:
   - The model produces an error, typically a formatting issue for structured outputs. In this case, a new prompt is constructed that includes the conversation history and the error. The Error Agent is then responsible for correcting the error.
   - The model produces the final answer, which is then sent back to the user, concluding the flow.



(a) *GoogleCalendarCreator* for the Personal Assistant use case

(b) *OnlinePurchase* for the Shopping Assistant use case

Figure 2: *FormTools* examples designed for our use cases and implemented using Converso. A *FormTool* includes a description, a set of arguments for the user to complete when prompted by the LLM during the conversation, and options regarding the return mode. The return mode can be either direct, where the result is given directly to the user without further interaction with the LLM, or indirect, where additional processing by the LLM is required. Additionally, the *FormTool* specifies whether a confirmation step is needed.

- The model requests a tool execution. If the execution results in an error, the error is handled as in the first case. Otherwise, the result can be sent directly to the user or modified by the LLM before being sent. The creator of the specific conversation flow can choose the best option.

3.3 Conversational Forms

A key contribution of Converso is the introduction of a multi-agent conversation flow based on conversational forms, where the latter are represented as *FormTools* within the LangChain framework. We implement *FormTools* to encapsulate user intents that require gathering structured data through a multi-turn conversation. This novel approach addresses the limitations of existing chatbots that rely solely on the current message and conversation history, which can lead to hallucinations or goal deviations, especially as the history becomes long [11].

The conversation flow in Converso is extended to recognize user intents that map to specific *FormTools*. When such an intent is detected, the corresponding *FormTool* is activated, and all other *FormTools* are temporarily inhibited. At this point, the Base Agent is replaced by the Form Agent, which uses specifically engineered prompts to drive the conversation to fill the form, prompting the user to provide the required information fields through natural language interactions. The activation mechanism for *FormTools* is inspired by the concept of *semantic frame evocation* [7], where key expressions trigger the activation of the appropriate frame. In Converso, the LLM's recognition of user intents serves as the triggering mechanism, dynamically activating the corresponding *FormTool* and its associated conversational form.

FormTools maintain an internal state that tracks the progress of the data collection process. This state can be inactive (initial state), active (collecting data from the user), or filled (all required data has been provided). The collected data is stored in an internal form object within the *FormTool*.

To enhance user experience, *FormTools* support dynamic validation of user inputs. For example, in a purchase scenario, the available shipping regions can be dynamically updated based on the user's selections, ensuring only valid options are presented. *FormTools* allows developers to define custom logic for form compilation, enabling them to determine the order in which fields should be filled, the actions to be taken based on the values provided, and to implement complex inter-field validation logic. Once all required data has been collected, the *FormTool* can execute its associated action, such as making an API call or performing a specific task. Before execution, if the developer of the specific tool requires the confirmation step, the user is presented with a summary of the collected information for confirmation, ensuring transparency and control over the process. Examples of *FormTools* are presented in Fig. 2.

The conversational forms approach, coupled with the stateful nature of *FormTools*, reduces the dependence on lengthy conversation histories, mitigating the risk of hallucinations or deviations from the intended goals. By abstracting the data collection process into structured forms, Converso simplifies the development of goal-oriented conversational applications while leveraging the power of LLMs.

4 Experiments

4.1 Use Cases

To showcase the functionalities of our framework and provide an implementation guide for developers, we implemented two use cases using Converso.

The first use case involves creating a chatbot as a Personal Assistant. The chatbot is implemented using the containerized system architecture illustrated in Fig. 1. What distinguishes different use cases in the Converso framework is the definition of the tools to use for the specific application, which in the case of the Personal Assistant are:

- The *PythonCodeInterpreter*, *GoogleSearch*, and *GmailRetriever* tools, which are *BaseTools* and take a single argument from the LLM to perform an operation with it (for example, the *PythonCodeInterpreter* takes in a valid Python script expressed as a string and executes it, returning the result to the LLM).
- The *GoogleCalendarRetriever*, *GoogleCalendarCreator*, and *GmailSender* tools which are implemented as *FormTools* to illustrate our conversational forms (Fig. 2a).

The second use case developed using Converso involves the creation of a Shopping Assistant. This example demonstrates how the framework can be used to enhance the shopping functionality of an e-commerce platform. The use case includes dynamic data validation: for instance, only certain regions are available, and once a region is selected, only the provinces within that region are shown. A single *FormTool*, named *OnlinePurchase* (Fig. 2b), is implemented for this use case. Figure 3 illustrates chat examples for both use cases.

4.2 Evaluation Protocol

In this section, we present our evaluation protocol for assessing Converso's performance. Specifically, our evaluation focus is not on the underlying LLM, which can be selected by the developer of a specific chatbot based on its use case. Instead, we aim to determine whether the conversation flow incorporating conversational forms introduced with Converso performs better in real-world scenarios than the basic conversation flow.

Our evaluation framework consists of three main components:

- The *Task Generator*, which uses predefined templates and instructions to create real-world scenarios.
- The *User Simulator*, an LLM that carries out the generated tasks by interacting with the Converso chatbot under evaluation.
- The *Converso System*, which implements either the basic conversation flow or the conversation flow with conversational forms.

Table 1: Prompt guidelines provided to the User Simulator for the evaluation protocol.

Task Type	**Prompt**
All the information contained in the first message (AFM)	State your intent to the system, and then follow its instructions to complete the task. Provide all the necessary data to the system in your first message.
No information contained in the first message (NFM)	State your intent to the system without providing any data, and then follow its instructions to complete the task. For example, "I want to create an event" or "I want to buy something."
Main information contained in the first message (MFM)	State what you want to do, providing only the main information, and then follow its instructions to complete the task. For example, "I want to create an event called Meeting" or "I want to buy a watch."
Confused user (CU)	State your intent to the system without providing any data, and then follow its instructions to complete the task. Act like a very naive user who doesn't know what to do: misspell words, give incorrect information, and then correct it.

Table 2: System prompts given to the different types of agents employed in Converso's chatbots. Information between curly braces is dynamically populated.

Agent	**Prompt**
Base Agent	You are a personal assistant trying to help the user. You always answer in English. The current datetime is {datetime}. Don't use any of your knowledge or information about the state of the world. If you need something, ask the user for it or use a tool to find or compute it.
Error Agent	[Base Agent prompt] + There was an error with your last action. Please fix it and try again. Error: {error}.
Form Agent (information needed)	Help the user fill data for form {form_tool.name}. Ask to provide the needed information. Now you must update the form with any information the user provided or ask the user to provide a value for the field {information_to_collect}. You MUST use the form {form_tool.name} tool to update the stored data every time the user provides one or more values.
Form Agent (confirmation needed)	Help the user fill data for {form_tool.name}. You have all the information you need. Show the user all of the information using bullet points and ask for confirmation: {information_collected}. If the user agrees, call the {form_tool.name} tool one more time with confirm=True. If the user doesn't want to change something, call it with confirm=False.

An evaluation task consists of the following:

- A user guideline, selected from the four listed in Table 1, which provides the User Simulator with instructions on how to behave as a user.
- A target tool, chosen from *GoogleCalendarCreator*, *GoogleCalendarRetriever*, *GmailSender*, *GmailRetriever*, or *OnlinePurchase* as defined in Section 4.1. These tools are adapted to implement the *FormTool* interface for evaluating the conversation flow with conversational forms.
- A target input, selected from 20 randomly generated payloads.

This setup brings to a total of 400 generated evaluation tasks. A task's execution is considered successful if the correct tool is invoked and filled with the correct inputs. In this case, the execution is stopped, and the run is marked as successful. The task is considered unsuccessful if the maximum number of iterations is reached or the models reach a stale phase (typically because the correct tool was called with the wrong inputs).

4.3 Implementation Details

The implementation of the conversation flows is based on LangChain. The Base Agent, Error Agent, and Form Agent are implemented as LangChain agents and connected using LangGraph. We report the system prompts used to tune each agent in Table 2. *FormTools* are implemented as subclasses of *BaseTool*. For all evaluation experiments, we use gpt-3.5-turbo-0125 via OpenAI APIs, both as the User Simulator and the LLM chatbot. The target tools for evaluation are implemented as *StructuredTools* to assess the basic conversation flow and as *FormTools*, managed by the Form Agent, to evaluate our conversation flow with conversational forms. The maximum number of iterations is set to 25.

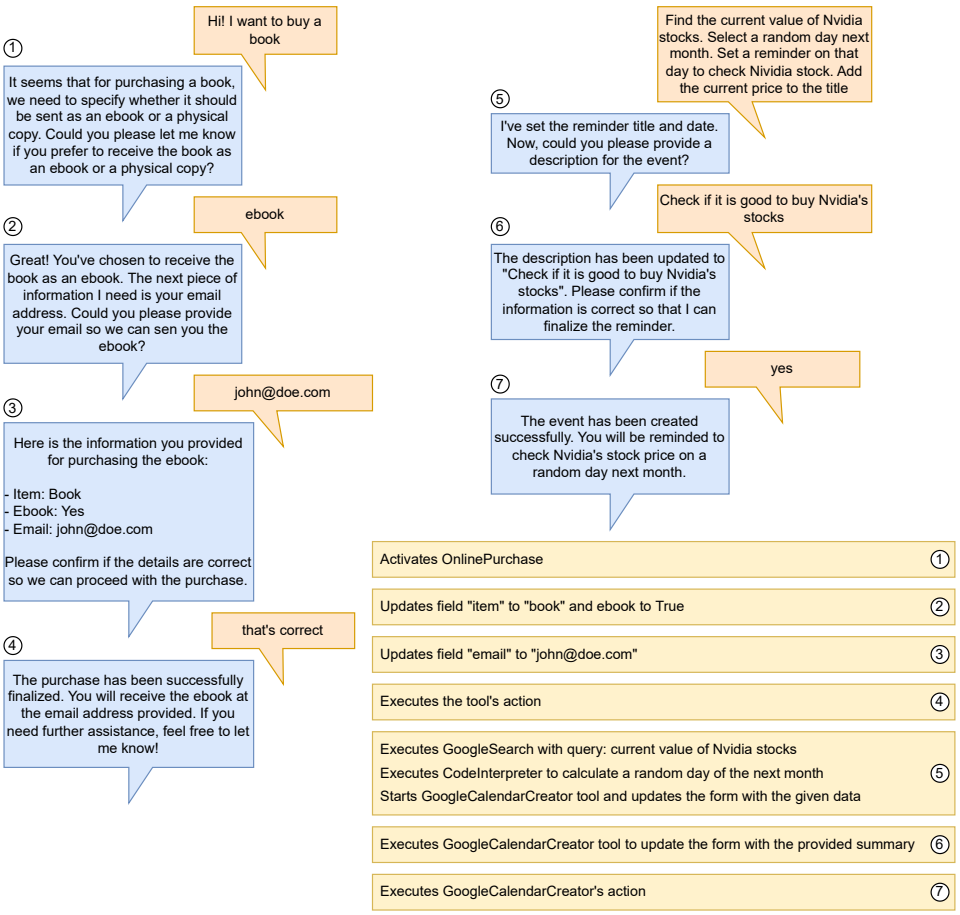


Figure 3: Two examples of chat illustrating our use cases. The chat on the left demonstrates the Shopping Assistant use case, while the chat on the right demonstrates the Personal Assistant use case. Orange messages with the tail pointing to the right represent user messages, while blue messages with the tail pointing to the left represent the chatbot. Actions executed by the chatbot using the connected tools are displayed inside yellow rectangles, which are numbered and connected to the corresponding points in the conversation. Notably, the agent can work with multiple tools simultaneously, allowing it to perform complex combinations of actions.

4.4 Results

Table 3 presents the results obtained using our evaluation protocol. The percentage of tasks executed correctly is 75.7% when using the basic conversation flow. It rises to 92.5% when employing the conversational forms, with a consistent 16.8% increase, showing the overall effectiveness of using *FormTools* with the Form Agent.

Table 3: Evaluation results, with scores expressed as percentages of success for the evaluation tasks defined in Section 4.2.

	Conversation Flow	
Tool	Basic	Ours
GmailRetriever	75.00	100.00
GmailSender	77.50	81.25
GoogleCalendarCreator	81.25	90.00
GoogleCalendarRetriever	90.00	95.00
OnlinePurchase	55.00	96.25
Task Type	Basic	Ours
AFM	91.00	99.00
NFM	73.00	93.00
MFM	62.00	89.00
CU	77.00	89.00
Total	Basic	Ours
	75.7	92.5

Analyzing the results from the perspective of tool usage, we observe that implementing conversational flows improves the use of every tool. Notably, the improvements are nearly double for the *OnlinePurchase* tool, which requires the most significant number of parameters and is, therefore, the most complex. This suggests a correlation between tool complexity and the benefits of using conversational forms. This outcome was expected, as more complex tools require keeping more detailed information in the conversation history, which can lead to goal deviation or hallucinations.

Considering the task type, we observe that the most challenging situation for the basic conversation flow occurs when the User Simulator's first prompt contains only the primary information for using the tool. We investigated this situation qualitatively (Fig. 4). In this case, the Base Agent invokes the correct tool but fills in the remaining information with hallucinated data without confirmation. This poses a risk for real-life applications, where actions could be executed with

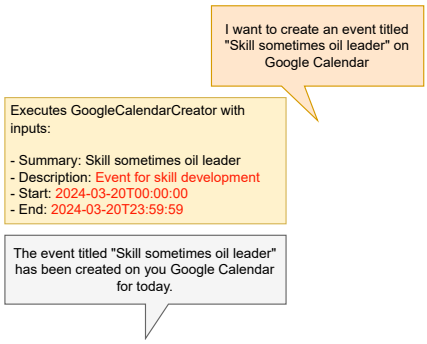


Figure 4: Chat example for an MFM task using the basic conversation flow (chatbot's response is in the gray box). The correct tool is executed but with incorrect inputs (highlighted in red), demonstrating the hallucination problem that occurs when conversational forms are not used.

incorrect inputs without the user's awareness. Conversely, conversational forms address this issue by querying the user for the missing information and requesting confirmation before executing any action.

5 Conclusion

Our work demonstrates the significant advancements in chatbot systems by integrating large language models and modern frameworks like LangChain. The proposed Converso framework, which incorporates conversational forms, showcases the potential to enhance user interactions by transforming traditional data acquisition methods into dynamic, interactive conversations. The evaluation results indicate a marked improvement in task success rates when using conversational forms, particularly with complex tools requiring detailed input.

By addressing the limitations of basic conversation flows—such as the hallucination of data and lack of confirmation—Converso improves accuracy and ensures a safer and more reliable user experience. This is especially crucial in real-world applications where incorrect data could lead to unintended and potentially harmful actions.

Furthermore, the Converso framework's robustness, demonstrated by its ability to handle diverse and complex use cases in the evaluation experiments, underscores its versatility. The consistent performance improvements, nearly doubling success rates in some cases, highlight the effectiveness of integrating stateful interactions and form-based data gathering.

Our work contributes to conversational AI by providing a practical and scalable solution for developing sophisticated chatbot applications. The insights gained from this research pave the way for future innovations in chatbot design, aiming to bridge the gap between human-like interactions and automated systems.

Acknowledgements

The research of Nicola Fanelli is funded by a PhD fellowship within the framework of the Italian "D.M. n. 118/23" - under the National Recovery and Resilience Plan, Mission 4, Component 1, Investment 4.1 - PhD Project "Analisi e valorizzazione del patrimonio artistico digitalizzato mediante tecniche di Intelligenza Artificiale" (CUP H91I23000690007).

References

[1] E. Adamopoulou and L. Moussiades. An Overview of Chatbot Technology. In I. Maglogiannis, L. Iliadis, and E. Pimenidis, editors, *Artificial Intelligence Applications and Innovations - 16th IFIP WG 12.5 International Conference, AIAI 2020, Neos Marmaras, Greece, June 5-7, 2020, Proceedings, Part II*, volume 584 of *IFIP Advances in Information and Communication Technology*, pages 373–383. Springer, 2020.

[2] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language Models are Few-Shot Learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[3] H. Chase. LangChain, Oct. 2022.

[4] B. Chen, Z. Zhang, N. Langrené, and S. Zhu. Unleashing the potential of prompt engineering in Large Language Models: a comprehensive review. *CoRR*, abs/2310.14735, 2023.

[5] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y. Tay, N. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. Garcia, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A. M. Dai, T. S. Pillai, M. Pellat, A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Diaz, O. Firat, M. Catasta, J. Wei, K. Meier-Hellstern, D. Eck, J. Dean, S. Petrov, and N. Fiedel. PaLM: Scaling Language Modeling with Pathways. *J. Mach. Learn. Res.*, 24:240:1–240:113, 2023.

[6] K. M. Colby, F. D. Hilf, S. Weber, and H. C. Kraemer. Turing-like Indistinguishability Tests for the Calidation of a Computer Simulation of Paranoid Processes. *Artif. Intell.*, 3(1-3):199–221, 1972.

[7] C. J. Fillmore and C. Baker. 313 A Frames Approach to Semantic Analysis. In *The Oxford Handbook of Linguistic Analysis*. Oxford University Press, 12 2009. ISBN 9780199544004.

[8] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, Q. Guo, M. Wang, and H. Wang. Retrieval-Augmented Generation for Large Language Models: A Survey. *CoRR*, abs/2312.10997, 2023.

[9] S. Hakimov, Y. Weiser, and D. Schlangen. Evaluating Modular Dialogue System for Form Filling Using Large Language Models. In *Proceedings of the 1st Workshop on Simulating Conversational Intelligence in Chat (SCI-CHAT 2024)*, pages 36–52, 2024.

[10] B. Hu, C. Zhao, P. Zhang, Z. Zhou, Y. Yang, Z. Xu, and B. Liu. Enabling Intelligent Interactions between an Agent and an LLM: A Reinforcement Learning Approach. *CoRR*, abs/2306.03604, 2023.

[11] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang. Lost in the Middle: How Language Models Use Long Contexts. *Trans. Assoc. Comput. Linguistics*, 12:157–173, 2024.

[12] J. Maynez, S. Narayan, B. Bohnet, and R. T. McDonald. On Faithfulness and Factuality in Abstractive Summarization. In D. Jurafsky, J. Chai, N. Schluter, and J. R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1906–1919. Association for Computational Linguistics, 2020.

[13] S. Minaee, T. Mikolov, N. Nikzad, M. Chenaghlu, R. Socher, X. Amatriain, and J. Gao. Large Language Models: A Survey. *CoRR*, abs/2402.06196, 2024.

[14] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike, and R. Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

[15] T. Schick, J. Dwivedi-Yu, R. Dessì, R. Raileanu, M. Lomeli, E. Hambro, L. Zettlemoyer, N. Cancedda, and T. Scialom. Toolformer: Language Models Can Teach Themselves to Use Tools. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neu-*

ral Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023, 2023.

[16] K. Wang, H. Ren, A. Zhou, Z. Lu, S. Luo, W. Shi, R. Zhang, L. Song, M. Zhan, and H. Li. MathCoder: Seamless Code Integration in LLMs for Enhanced Mathematical Reasoning. *CoRR*, abs/2310.03731, 2023.

[17] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, and W. Fedus. Emergent Abilities of Large Language Models. *Trans. Mach. Learn. Res.*, 2022, 2022.

[18] J. Weizenbaum. ELIZA - a computer program for the study of natural language communication between man and machine. *Commun. ACM*, 9(1):36–45, 1966.

[19] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. R. Narasimhan, and Y. Cao. ReAct: Synergizing Reasoning and Acting in Language Models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

A *Grice-ful* Examination of Offensive Language: Using NLP Methods to Assess the Co-operative Principle

Katerina Korre [a,*], **Federico Ruggeri** [b] **and Alberto Barrón-Cedeño** [a]

[a]DIT, University of Bologna
[b]DISI, University of Bologna

Abstract. Natural Language Processing (NLP) can provide tools for analyzing specific intricate language phenomena, such as offensiveness in language. In this study, we employ methods from pragmatics, more specifically Gricean theory, as well as NLP techniques, to analyze instances of online offensive language. We present a comparative analysis between offensive and non-offensive instances with regard to the degree to which the 4 Gricean Maxims (Quality, Quantity, Manner, and Relevance) are flouted or violated. To facilitate our analysis, we employ NLP tools to filter the instances and proceed to a more thorough qualitative analysis. Our findings reveal that offensive and non-offensive speech do not differ significantly when we evaluate with metrics that correspond to the Gricean Maxims, apart from some aspects of the Maxim of Quality and the Maxim of Manner. Through this paper, we advocate for a turn towards mixed approaches to linguistic topics by also paving the way for a modernization of discourse analysis and natural language understanding that encompasses computational methods.

Warning: *This paper contains offensive language that might be triggering for some individuals.*

1 Introduction

Natural Language Processing (NLP) is characterized by creating applications adept at addressing real-world challenges. Among those applications, we find machine translation, text summarization, coreference resolution, and part-of-speech-tagging, to name a few [19]. These rapid technological developments, along with the recent emergence of large language models (LLMs), have brought to the fore the question of how linguistics could benefit from such advancements and contribute to the current wave. This issue is discussed in a recent *Nature* editorial, where it is illustrated that there is a distinction between NLP and Computational Linguistics, with the latter focusing more on the two aforementioned questions. More specifically, "Computational Linguistics traditionally uses computational models to address questions in linguistics and borders the field of Natural Language Processing, which in turn builds models of language for practical applications" [1]. Dupre [6] poses an opposite opinion, claiming that deep learning techniques cannot illuminate linguistic theory, as the former focuses on language performance, while the latter on language competence, which are arbitrarily different.

In this paper, we draw from the distinction between Computational Linguistics and NLP, and we use NLP methods as tools for discourse analysis. Despite the argument that, at least current deep learning techniques are pertinent to theoretical insights in linguistics [6], we believe that deep learning tools can facilitate linguistic analysis. We exemplify this in our paper, by analyzing the structure of offensive language. Offensive language detection is a popular topic in NLP, as its intricate nature, lying within the borders of linguistics, psychology, sociology, and law studies, makes it hard for current models to identify positive instances adequately [35]. Current approaches in NLP view offensive language as a detection task without delving further into the intricate dynamics of an offensive conversation. We believe that a thorough analysis of offensive language requires a pragmatic approach. By examining contextual factors, such as speech acts, perlocutionary effects, politeness strategies, and cultural norms, we can gain a deeper understanding of how and why language becomes offensive. Ignoring these pragmatic aspects would result in an incomplete and potentially flawed analysis of offensive language.

In order to perform discourse analysis on online offensive language from a pragmatic perspective, we employ part of the Gricean theory [12], which outlines four conversational principles —Quality, Quantity, Relevance, and Manner— which ensure that speakers provide truthful, informative, relevant, and clear contributions to conversations. We argue that there is a pattern in the flouting/violation of the maxims when it comes to online offensive language. The most obvious assumption is that offensive language flouts the Maxim of Manner as this type of discourse is inherently not in accordance with this Maxim. In particular, offensive language uses an inappropriate lexicon that is unsuitable for any occasion, leading to uncooperative conversations. On a similar note, Pasa et al. [27] have shown that the sarcasm of hate speech in Instagram comments flouts all four maxims. The authors hypothesize that the main factors driving these violations are the lack of concise and clear information in comments, the cultural value in Western countries that emphasizes the right to free speech, the tendency to seek excessive attention from others, and the ego that boosts self-importance while devaluing others.

Our contribution is two-fold. Inspired from previous endeavors [32, 10], we first translate the Maxims into actual metrics using NLP methods, thus bridging the gap between theoretical and computational linguistics. Secondly, we offer a computationally facilitated discourse analysis on offensive language, showing that such analyses can be semi-automated as the data can be filtered faster, allowing for a more precise examination of specific instances. To our knowledge, this is the first study that equips Gricean theory through computational methods to analyze offensive language.

Our findings indicate that violations or floutings of the Maxims do not differ when comparing offensive and non offensive online dis-

* Corresponding Author. Email: aikaterini.korre2@unibo.it

Table 1: The 4 Gricean Maxims and their corresponding submaxims.

Maxim	Sub-maxims
Maxim of Quality	• Do not say what you believe to be false. • Do not say that for which you lack adequate evidence.
Maxim of Quantity	• Make your contribution as informative as is required for the current purposes of the exchange. • Do not make your contribution more informative than is required.
Maxim of Relevance	• Be relevant.
Maxim of Manner	• Avoid obscurity of expression. • Avoid ambiguity. • Be brief. • Be orderly.

course statistically. The only exception is the Maxim of Manner due to the intense use of profanity and possibly the Maxim of Quality, as in the 'offensive' class, untruthful comments are more frequent. To assess the effectiveness of Maxim-based metrics in discourse analysis, we also conduct a qualitative analysis.

The paper is organized as follows. In the next section, we introduce some essential concepts of theoretical pragmatics, as they are the basis of our metrics and the discourse analysis. In Section 3, we discuss NLP approaches that involve linguistic pragmatic aspects both in terms of human language and artificially-generated language. In Section 4, we describe our methodology, translating the Maxims into metrics and using them as discourse analytical tools. In Section 5, we present our results, discussing them in Section 6. Finally, we summarize our final remarks and potential future work in Section 7, and we close this paper with a presentation of the limitations in Section 8.

2 Theoretical Background

One central point in pragmatics is the work of HP Grice, who formulated several pragmatic theories applicable today to conversation and discourse analysis. These include the *cooperative principle*, according to which the contribution of the conversation "must be such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged" [12]. The cooperative principle outlines the fundamental principle *guiding* communication and it is broken down into multiple sub-principles, or 'Maxims'. More specifically, it suggests that in conversation, participants generally adhere to four Maxims: the Maxim of Quantity (provide just enough information), Quality (speak truthfully), Relevance (be relevant), and Manner (be clear and concise). More information about the Maxims and their explanations, often referred to as submaxims, can be found in Table 1. These Maxims serve as implicit guidelines for effective and efficient communication.

Another essential concept in pragmatics, and upon which we touch in this study, is implicature. Implicature occurs when speakers convey meaning beyond the literal interpretation of their words, relying on context and shared understanding. Implicatures can be further divided into those that are explicitly conveyed (explicature) and those inferred by the listener. Table 2 shows an example of conversation implicature, which illustrates a type of pragmatic inference that arises when words can be arranged on a semantic scale, such as the value judgments *excellent* $>$ *good* $>$ *OK*. Speaker A infers from B's response because if the accommodation had been better than just OK, B could have described it as good; if it had been very good, B could have said excellent. Since B did not use good or excellent, A concludes that the accommodation was merely satisfactory. At the time of the conversation, A might also have observed signs confirming this inference, such as B's unenthusiastic tone or body language indicating discomfort. These contextual clues further help in interpreting implicatures. However, they are unavailable in online language, making the task of interpreting implicatures even harder, especially in cases of sarcasm and irony.

Table 2: Example of conversational implicature taken from Griffiths and Cummins [13].

Speaker	Dialogue
A	What was the accommodation like on the work camp?
B	It was OK.
A	Not all that good, hey?

Implicature is also related to the violation or flouting of the Gricean Maxims. Violation of the 4 Maxims occurs when speakers deviate from the expected norms of communication, potentially causing confusion or misunderstanding, most often unintentionally. In contrast, flouting involves intentionally disregarding the Maxims, giving rise to conversational implicature for rhetorical or humorous effect. For example, when someone asks, "How's the weather?" during a thunderstorm, they flout the Maxim of Relevance by intentionally ignoring the obvious context.

3 Related Work

This section reviews key contributions in computational pragmatics, highlighting advancements in pragmatic inference, model evaluation, and the integration of pragmatic principles into NLP systems. We conclude by providing an overview of existing pragmatic approaches for offensive language detection and analysis.

3.1 Pragmatics in NLP

Jurafsky [16] defines computational pragmatics as the computational study of the relationship between utterances and their context. It examines how utterances relate to actions, discourse, and environmental factors like time and place. Inference is a key focus in computational pragmatics, addressing four main problems: reference resolution, speech act interpretation and generation, discourse structure and coherence, and abduction. Each problem involves inferring missing information from utterances. However, when it comes to the interaction between NLP and pragmatics, the focal point of research lies in the detection of pragmatic effects, either in natural language or artificially generated language. Most approaches are concerned with the capability of the models to detect and/or to understand different pragmatic phenomena (such as irony, metaphor, and sarcasm) in natural language data, including social media posts and user inputs [2, 18, 21, 34]. Another emerging area of pragmatics in NLP is concerned with the ability of the models themselves (mainly LLMs) to actually produce speech intricate enough to mimic human language, including pragmatic language functions [4, 15, 17, 29].

As a broad field of linguistics, different aspects of pragmatics have been exploited or explored in NLP. Among them are also the Gricean Maxims. For instance, Hu et al. [15] present a fine-grained analysis of the pragmatics in the language of humans and LLMs in an attempt

to answer three questions: whether models select pragmatic interpretations of speaker utterances; whether models make similar errors as humans; and whether models use similar linguistic cues to solve the task. They show that certain pragmatic phenomena, such as humor, irony, and Grice's Maxims, involve violating listeners' expectations in some way and for which the LLMs fail to choose pragmatic interpretations. On a similar note, Jwalapuram [17] evaluate computer-generated dialogues according to Grice's Maxims. They use a survey in which the user is asked to rate the system performance on a Likert scale from 1 to 5 for 4 questions that correspond to the Maxims. In this way, they are able to identify: (1) if the system provides substantive responses; (2) if the system is faithful to the factual knowledge it is provided with; (3) if the system is able to understand the user and, therefore, provide relevant replies, and, finally; (4) if the system provides awkward or ambiguous responses. While they report diverse results depending on the generated dialogues, the authors do not go into speculations as to why that could be the case. Sorower et al. [31] present a method for learning rules from natural language texts by addressing the challenge of missing data. They introduce a mention model that addresses the probability of facts being mentioned in the text based on what other facts have already been mentioned and domain knowledge in the form of Horn clause rules and by formalizing Gricean Maxims encoding them as rules in Markov Logic.

Apart from being used as tools for conversation analysis, Gricean Maxims are also used as metrics. Ge et al. [10] propose the task of knowledge-driven follow-up question generation in conversational surveys. They produce a human-annotated dataset and they propose new metrics based on the Gricean Maxims. Freihat et al. [9] use the Maxims for ranking community question answers as hypothesize that linguistics offers a good opportunity to predict the relevance of answers and rank them accordingly. They use different indicators for each Maxim (except quality). Although their approach did not achieve the performance of machine-learning-based approaches, it gave a linguistically motivated solution that can be improved so that it reaches the performance of machine learning methods. Tewari et al. [32] focus specifically on the Maxim of Quantity and they model it as a new metric to assess the informativeness for short texts.

Implicature has also been in the limelight of linguistically motivated NLP research. Benotti and Traum [4] investigate the pragmatic implications of comparative constructions from a computational standpoint, emphasizing the challenges in determining the superiority of one answer over another. Zheng et al. [36] introduce a dataset for recovering implicature and conversational reasoning, showing that model performance improves when a module on implicature is included during training. Similarly, Ruis et al. [29] show that fine-tuning on conversational data or benchmark-level instructions does not produce models with pragmatic understanding. However, fine-tuning on instructions at the example-level paves the way towards more useful models of human discourse.

Understanding pragmatic functions in real-life situations presents a challenge for NLP. Unlike humans, who effortlessly use context and background knowledge to deduce implicatures, NLP models find this process difficult [36]. For example, in many cases incorporating Gricean theory (i.e. the cooperative principle and the 4 Maxims) involves using survey methods, employing humans to evaluate model capabilities with regard to the understanding of pragmatic discourse [17, 32].

3.2 Pragmatic Approaches on Offensive Language Detection and Analysis

Many studies on hate speech, toxic language, offensive language or any other type of harmful language detection typically focus on individual instances, neglecting its inherently conversational nature [28]. This approach might be enough for solely NLP purposes but it limits the exploration of pragmatic analysis of harmful language on a discourse analysis level. One study that takes into account the context of toxicity in online conversations is the one from Madhyastha et al. [23], where they clearly show the significance of context and the effect on annotations. Other studies, such as in the case of Gevers et al. [11], have tried to analyze the structure of hate speech or different linguistic attributes of it, such as length and lexical diversity. Saveski et al. [30] studied the structure of toxic language spread. They show that, at the individual level, toxicity is spread across many low to moderately toxic users. At a dyad level, they observe that toxic replies are more likely to come from users who do not have any social connection nor share many common friends with the poster. At the group level, they find that toxic conversations tend to have larger, wider, and deeper reply trees, but sparser follow graphs.

One of the few works that has pragmatic aspects embedded in the methodology is the work of Upadhyaya et al. [33], where they introduce a dataset for toxic language that includes annotations for speech acts that could reveal information about the stance and that could help further in the toxic language detection. More traditional approaches of discourse analysis include the work of Hidayati and Arifuddin [14] that aims to reveal the types of hate speech on social media based on the criteria developed by Austin, and the meaning of hate speech spoken by individuals to other individuals on Facebook, using qualitative descriptive methods. The results show that hate speech on social media can be classified based on illocutionary acts developed by Austin, into verdictive, behabitives, and expositive. Finally, the work of Parvaresh [26] provides a corpus-assisted analysis of hate language as found on Instagram, focusing on Afghan immigrants. The study reveals that hate speech may lack markedly hateful language and that hate language may revolve around covert ways of expressing hatred.

In this paper, we investigate the potential of using NLP methods to evaluate pragmatic discourse, using the publicly available ToxiChat dataset [3]. We build on previous research to adapt NLP techniques for assessing Gricean Maxims and the cooperative principle. These tools are employed to conduct an advanced discourse analysis of a pragmatically complex discourse type: offensive language.

4 Methodology

For the purposes of this study, we use metrics and NLP tools for each of the 4 Maxims. In this way, we attempt to filter different instances that will be used for a qualitative analysis in a more traditional discourse analysis manner.

4.1 Translating the 4 Maxims into Metrics

The purpose of the cooperative principle and the maxims is to guide effective and efficient communication by encouraging speakers to be informative, truthful, relevant, and clear in their discourse. To quantitatively assess the success of the cooperative principle, we employ metrics and tools commonly used in NLP, aligning each one with a respective maxim. Our approach draws inspiration from prior research that has endeavored to translate these maxims into NLP met-

Table 3: Information about the ToxiChat dataset

Dataset	Source	Participants	Turns	Purpose	Instances
ToxiChat	Reddit	Human + Bot	3	Offensiveness and Stance detection	3,211

Table 4: ToxiChat example.

Turn	Text	Label
1	**Title: [Question] Why do Libertarians get so much flack from the rest of reddit Like seriously I was downvoted when I said "Libertarian is a good one" on a post about third party voting.**	Safe
2	Because the rest of reddit are unironically communists.	Offensive
3	Bullshit most are democrats	Offensive

rics [10, 32], while also introducing new methods tailored to the specific focus of our study.[1]

Maxim of Quality For the Maxim of Quality, we define a text classification approach aimed at detecting deceptive content.

We train a BERT-based text classifier [5] on the Deceptive Opinion Spam Corpus (DOSC) [24, 25].[2] The corpus contains 1,600 customer reviews (both positive and negative) about 20 hotels. Half of the reviews are labeled as deceptive, while the remaining half are labeled as truthful. We group reviews based on the target hotel and build train (reviews of 16 hotels), validation (reviews of 2 hotels), and test (reviews of 2 hotels) splits such that all reviews belonging to a hotel are in the same split. We follow standard practice [5] and fine-tune the BERT-based text classifier for up to five epochs. We consider five different seed runs to ensure a sound evaluation. The classifier achieves an average macro F1-score of 0.926 ± 0.021 on the DOSC corpus test set.

Maxim of Quantity The Maxim of Quantity has been first studied in Tewari et al. [32]. The authors propose informativeness as a metric of the Maxim of Quantity based on syntactic cohesion. They use a dependency parser to transform segments into graphs of syntactic relations, defining syntactic cohesion as the sum of these relations. Syntactic cohesion is computed by comparing two sets of heads and their dependents, with normalized values falling between -1 and 1, indicating optimal, slightly cohesive, or fragmented cohesion. They normalize cohesion, dividing the score by the total number of words in the segment. Informativeness in an instruction sequence is the sum of syntactic cohesion values across all segments, with a normalized score ranging from 0 to 1, indicating under-informative, optimally informative, or over-informative sequences. In this study, we employ the same methodology.

Maxim of Relevance For the Maxim of Relevance, we implement a methodology to assess relevance within conversations using BERT embeddings and cosine similarity. Beginning with data preprocessing, we apply a custom binary relevance calculation function, which uses BERT embeddings to measure the similarity between conversation titles and their two subsequent responses. This process computes relevance scores by capturing the coherence of each response with respect to both the conversation title and the preceding response.

Maxim of Manner Our approach for the Maxim of Manner is inspired by Kiyavitskaya et al. [20] and focuses on assessing the instances in accordance to two main aspects of the Maxim: ambiguity and orderliness. There are many types of ambiguity, such as lexical, syntactic, and pragmatic. However, language models are not sensitive enough to successfully capture such delicate linguistic nuances yet [20, 22]. For that reason, we focus only on lexical ambiguity. We formulate our approach as follows:

Let S be a sentence consisting of words $w_1, w_2, \ldots, w_n$. We define the ambiguity $\text{amb}(w_i)$ of word w_i as the number of senses (synsets) that w_i has in WordNet [7]. The ambiguity of S is computed as

$$\text{amb}_{\text{total}}(S) = \sum_{i=1}^{n} \text{amb}(w_i) \tag{1}$$

Let D be a dataset of sentences, where S_j is a sentence in D. The maximum total ambiguity value is defined as:

$$\max(\text{amb}_{\text{total}}(S)) = \max_{S_j \in D} \text{amb}_{\text{total}}(S_j) \tag{2}$$

The Normalized Ambiguity of a sentence S is then defined as:

$$\text{amb}_{\text{norm\_total}}(S) = \frac{\text{amb}_{\text{total}}(S)}{\max(\text{amb}_{\text{total}}(S))} \tag{3}$$

We also apply a readability metric as a proxy for text obscurity. We use the Flesch readability metric [8], which evaluates the ease of reading a text based on sentence length and word syllable count, providing a score from 0 to 100 (higher scores indicate easier readability and lower scores suggest more complex texts).

Regarding profanity, we use the *better profanity* library,[3] which enables us to identify instances of profanity within the data, thereby automatically violating the Maxim of Manner. The library includes a word list and returns True if any word in the provided string matches a word in the list. By systematically analyzing instances for ambiguity, obscurity, and orderliness, the methodology ensures adherence to principles of clarity and coherence in online discourse.

Table 5: Truthfulness of ToxiChat instances with respect to the offensiveness of each instance. The bars are annotated with the percentages per class (safe/offensive).

Predictions	True	Untrue
Offensive	558 (60%)	371 (40%)
Safe	1609 (70%)	673 (30%)

4.2 Data

To conduct a discourse analysis based on the cooperative principle and the four maxims, we require that the data consist not only of isolated comments but also of dialogues with conversational turns. To our knowledge, there are few datasets containing instances of dialogues with offensive language, and those that do typically offer no more than two turns. Therefore, the data used for this study are sourced from the ToxiChat dataset [3], primarily constructed for stance analysis in online offensive contexts. Details about the data are presented in Table 3, with an illustrative example available in Table 4. In our study, we use only the train set of the dataset, and since we are interested in a pragmatic analysis of natural language, we are only concerned with the turns in the thread that are produced by humans and not the turn produced by the bot.

5 Results

Quality Table 5 shows the results of the BERT based deception classifier, assessing the truthfulness of the instances. Predominantly

[1] Code available at: https://github.com/katkorre/A-Griceful-Examination-of-Offensive-Language.git

[2] We use the `bert-base-uncased` model card from HuggingFace.

[3] https://pypi.org/project/better-profanity/

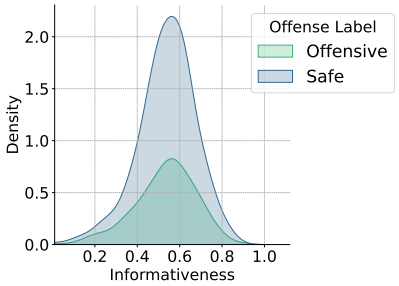


Figure 1: Histogram of Informativeness of ToxiChat instances with respect to the offensiveness of each instance.

in both classes, there are more instances labeled as True rather than Untrue. However, we notice that in the 'safe' class, around 70% of the instances are marked as True, while this percentage drops in the 'offensive' class, with 60% of the cases classified as True. Proportionately, untruthful statements are more likely to appear in offensive language. Therefore, it is more likely that the Maxim of Quality is flouted or violated in offensive contexts.

Quantity The evaluation results of the Maxim of Quantity, based on informativeness, are presented in Figure 1. We observe that, in most cases, both offensive and non-offensive instances consistently achieve a reasonable level of informativeness throughout the thread excerpts, with the majority of the values falling close to 0.5 which indicates an optimally informative instance.

We analyze this further with Table 6 as we proceed to form three thresholds that correspond to three classes of informativeness to compare the offensive against the safe class. A threshold of 0.25 was set to delineate instances deemed 'Under-Informative', indicating low levels of informativeness. A threshold of 0.75 was designated for data instances categorized as 'over-informative', denoting instances that contain redundant information. Finally, the values in-between denote the optimum level of informativeness. Most of the instances are optimally informative. Comparing the 'safe' and 'offensive' classes, there are more under- or over-informative instances in the 'safe' class. The difference in the number of under- or over-informative instances between the 'safe' and 'offensive' classes might be influenced by the uneven distribution of instances across these classes. Since the dataset contains significantly more 'safe' instances than 'offensive' ones, this imbalance can skew the analysis. Instead of normalizing or stratifying in this study, we maintain the raw data characteristics to interpret with a focus on real-world relevance.

Relevance In terms of relevance, our results are shown in Figure 2. Most instances, offensive or not, are deemed relevant by our model. Similar to previous Maxims, relevance is rarely flouted or violated, and when it is, it most frequently occurs in the 'safe' class. An exception is seen in responses to the title, where violations also occur in the 'offensive' class. This suggests that there is likely no correlation between the offensiveness of an instance and its violation of the maxim of relevance. However, it is important to consider the domain of the data, which is sourced from Reddit. Given that Reddit revolves around specific questions and answers, the room for irrelevant responses is limited.

Manner We evaluate the Maxim of Manner in terms of ambiguity, readability and profanity. Figure 3 displays a boxplot of our ambiguity detection results. The two boxes are similar in size, with the 'safe' class showing a slightly larger range of values. The general pictures accounts to the fact that, in terms of ambiguity, 'offensive' and 'safe' dialogue instances do not differ to a significant degree. The picture is similar when calculating readability, with both classes presenting high scores in the Flesch readability metric, with the lower quartile being close to 50 in both cases. The 'safe' class, however, also presents a higher number of outliers that tend to have lower readability. Among those scores there are also negative ones which indicates a very short sentence or an extremely complex one. This could also be due to internet language and formatting. We initially hypothesized that the Maxim of Manner is typically flouted in the context of toxic or offensive language, and our results confirm this through the high frequency of profanity. Offensive language often relies on strong, explicit terms to convey hostility or aggression, which naturally includes a higher frequency of profanities. This is obvious in Figure 5 which shows that instances labeled as offensive contain more profanities compared to those that are labeled as safe.

Table 6: Informativeness thresholds of ToxiChat instances with respect to the offensiveness of each instance.

Category	Optimally Informative	Over-Informative	Under-Informative
Offensive	886 (95.37%)	21 (2.26%)	22 (2.37%)
Safe	2186 (95.79%)	50 (2.19%)	46 (2.02%)

6 Exemplary Discourse Analysis

Quantifying the maxims and examining the results in Section 5 have allowed us to form a more concrete idea and hypothesis, while it also allows us to filter results that would be of discourse analysis interest. To perform a discourse analysis, we first proceed to select instances according to the results of the previous section. For that reason, we look only at the offensive class and we randomly choose one example that violates each maxim, and one example that does not and proceed to compare the instances. The selected examples can be found in Table 7.

Comparing the two examples for the Maxim of Quality, the one that does not violate the maxim, does not contain any information that could potentially be untrue. The use of hedging with 'seem' and the simile introduced with 'like' mitigate the certainty of the author of the comment, despite the fact that it is an offensive comment. The example that violates the maxim, however, is full of potentially false assumptions, such as "he used all the sexual energy into fighting". This information is misleading and does not contribute to an effective cooperative (online) conversation.

Looking at the examples for the Maxim of Quantity, both responses generally adhere to the maxim. Response 1, which looks as an additional comment from the author of the title thread, provides enough information to support its point without overwhelming details, justifying the reasoning to their initial question. Response 2 offers a concise and direct answer. However, it is possible it could be considered slightly under-informative as it does not precisely reply to the initial question. About the second example, that according to the used algorithm violates the maxim of quantity, Response 1 provides an abundance of specific criticisms, making it slightly verbose and less clear due to its structure. That could lead us to the conclusion that it violates the Maxim of Quantity. The second response expresses an opposite opinion from the one presented in the title. It does not answer directly the question. However, with that response 'I like them', we are to lead to the implicature that the author of the response does not support banning 'GenderCritical', contrary to the suggestion in the title.

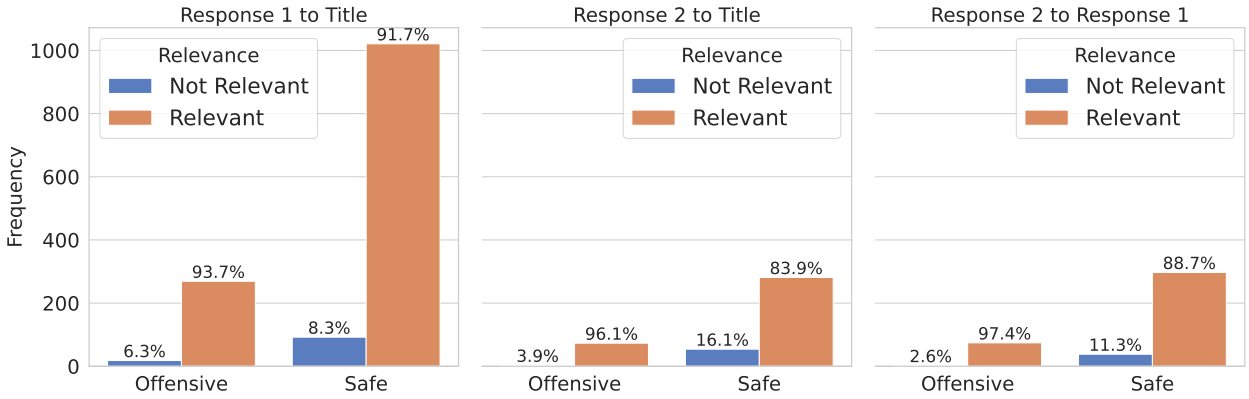


Figure 2: Distribution of relevance scores concerning offensiveness. The first plot shows the distribution of the first replies to the title, the second plot shows the distribution of the second replies to the title, and the third plot shows the distribution of relevance for the second reply to the first reply.

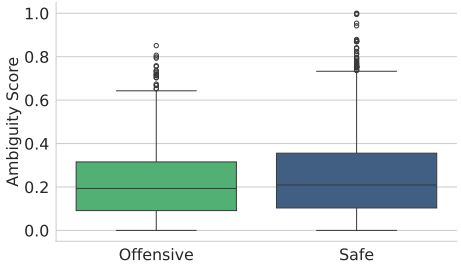


Figure 3: Ambiguity scores of ToxiChat instances with respect to the offensiveness of each instance.

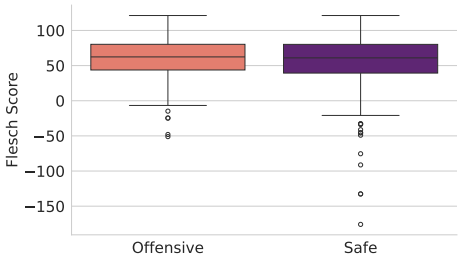


Figure 4: Flesh Readability. Higher scores indicate easier readability, with lower scores suggesting more complex texts.

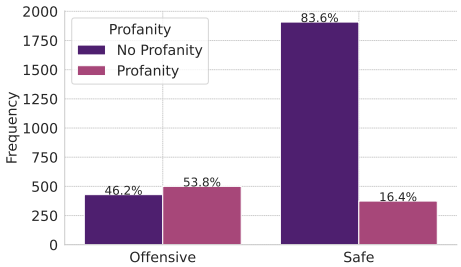


Figure 5: Distribution of profanity in ToxiChat instances with respect to the offensiveness of each instance. The bars are annotated with the percentages per class (safe/offensive).

Table 7: Randomly selected offensive instances that violate the Gricean Maxims or not. For Quantity and Relevance we report 2 or 3 conversation turns, as the metrics take into account the whole conversation. For Quality and Manner, we report the individual instances.

Maxim	Text	Violates
Quality	a lot of you seem like genuinely shitty people	No
	Mike tyson didnt have sex for 5 years Boxing over Bitches He used all the sexual energy into Fighting also not wasting excess Testestrone and using it as fuel I guess he is champ and sometimes sacrifice like that what makes sucessful people different over normal people What are your thoughts	Yes
Quantity	**Title**: Do you get along with the NONcrazy Liberals Leftists; **Response 1**: And before you say there are no noncrazy Democrats read the comment section of this NY post opinion piece URL They despise the fascist farleft takeover of everything as much as we do that comment section alone was a real eyeopener; **Response 2**: I get along with everyone until they start being a dick	No
	Title: Im curious can we get GenderCritical banned; **Response 1**: I mean it hates autistic men male babies and encourages misandry and hatred against men as a whole It even surpassed incels in terms of idiocy and hatred because at least incelsfrom rMGTOW like their daughters to some extent GenderCriticial even hates sons and male children and anything to do with men; **Response 2**: I like them They show the true nature of feminism to everyone I wish they had 1mm members so they cant be brushed off as a minority	Yes
Relevance	**Title**: British women abroad are a FUCKING DISGRACE Morgoths Review morgoths review; **Response**: Tourists are the worst kind of twatthot trash	No
	Title: Have an amazing daynight No updoot required You have no choice You must be happy right now or else I will come over to you and give you a hug No homo lt3 you can accomplish your days; **Response**: yes homo	Yes
Manner	Could someone make hitlers easy bake oven it might still exist but I cant find it	No
	Stoning the whores is the one thing Muslims got right	Yes

Examining the examples for the Maxim of Relevance, the response in the first one does not directly address the specific claim about British women abroad being a disgrace. Instead, it broadens the criticism to include all tourists, which dilutes the specificity of the original statement. This shift can be seen as a deviation from the maxim of relevance because it does not maintain the specific focus introduced by the title. While the response is contextually related (both the title and the response criticize behavior abroad), the failure to address the specific group mentioned (British women) makes it less relevant. Thus, it can be argued that the response violates the maxim of relevance by not directly engaging with the specific claim made in the title. This could therefore be considered a potential error of the model. About the second example, the initial message and the response both adhere to the Maxim of Quantity. The initial message provides enough context and information to be encouraging and humorous, and the response is brief but relevant and clear. The playful nature of the exchange is maintained through both contributions, and neither is too sparse nor too detailed for the context. The response directly addresses the 'No homo' part of the initial message, playfully contradicting it. It provides a relevant and humorous counterpoint to the initial message without adding unnecessary information. Therefore, this could be a false positive error for the algorithm.

Finally, about the Maxim of Manner, the first example is offensive probably towards Jews. It could be considered a slightly ambiguous statement, as it is unclear whether the speaker is making a dark joke, referring to a specific object or concept, or whether they misunderstand the implications of the words they are using. About the last example, the statement is highly offensive and lacks clarity. It uses derogatory language and promotes violence without any regard for decency or ethical considerations.

Despite occasional algorithmic errors in detecting the maxim floutings or violations, the pragmatic discourse analysis facilitated by our approach effectively highlights the nuances in which offensive language interacts with conversational norms. By applying computational algorithms, we can systematically filter and analyze large datasets, revealing patterns in offensive language.

7 Conclusions

In this paper, we adopt a pragmatics-based approach to hate speech analysis, reinforced by NLP methods. We draw from the linguistic theory of the 4 Gricean Maxims and the Co-operative principle, and we employ NLP methods as tools to assess whether the maxims are flouted or violated in offensive online contexts. Our approach provides an essential step before any type of discourse analysis, that allows a better understanding of the data and consequent filtering of instances for qualitative discourse analysis. Our experimental results showed some patterns in the flouting/violations of the maxims in offensive language settings, such as the flouting/violation of the maxim of manner due to ambiguity and profanity. With this paper, we advocate for more mixed approaches that will encompass both computational and traditional linguistics, and which will contribute to better data analysis.

In future work, we aim to dive deeper into the potential of the metrics, particularly when coupled with advanced LLMs. This exploration will contribute to further automate the assessment process of the cooperative principle, potentially enhancing its accuracy. Additionally, we intend to investigate other discourse domains posing challenges to NLP, such as sentiment analysis, humor, and sarcasm detection, which represent pragmatically charged categories. Furthermore, we are interested in examining intersections among the maxims to gain a comprehensive understanding.

8 Limitations

Our work is not without limitations. First of all, the metrics and NLP techniques that we used for the assessment of the maxims is not perfected and should be tested in other settings, as well as evaluated in more contexts, ideally from human experts. Even with NLP models with very high performance, there is always the possibility of error. Therefore, manual examination for discourse analysis is essential. Another limitation relates to the fact that, during the qualitative analysis, we examined each example for only one maxim flouting or violation each time, though it is possible that more than one floutings or violations could co-occur.

Acknowledgements

This research is carried out under the project "RACHS: Rilevazione e Analisi Computazionale dell'Hate Speech in rete", in the framework of the PON programme FSE REACT-EU, Ref. DOT1303118.

References

[1] Language models and linguistic theories beyond words. *Nature Machine Intelligence*, 5(7):677–678, 07 2023. ISSN 2522-5839. doi: 10.1038/s42256-023-00703-8. URL https://doi.org/10.1038/s42256-023-00703-8.

[2] A. A, S. G, S. H R, M. Upadhyaya, A. P. Ray, and M. T C. Sarcasm detection in natural language processing. *Materials Today: Proceedings*, 37:3324–3331, 2021. ISSN 2214-7853. doi: https://doi.org/10.1016/j.matpr.2020.09.124. URL https://www.sciencedirect.com/science/article/pii/S2214785320368164. International Conference on Newer Trends and Innovation in Mechanical Engineering: Materials Science.

[3] A. Baheti, M. Sap, A. Ritter, and M. Riedl. Just say no: Analyzing the stance of neural dialogue generation in offensive contexts. In M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4846–4862, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.397. URL https://aclanthology.org/2021.emnlp-main.397.

[4] L. Benotti and D. Traum. A computational account of comparative implicatures for a spoken dialogue agent. In H. Bunt, editor, *Proceedings of the Eight International Conference on Computational Semantics*, pages 4–17, Tilburg, The Netherlands, Jan. 2009. Association for Computational Linguistics. URL https://aclanthology.org/W09-3704.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, and T. Solorio, editors, *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, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

[6] G. Dupre. (what) can deep learning contribute to theoretical linguistics? *Minds and Machines*, 31(4):617–635, December 2021. ISSN 1572-8641. doi: 10.1007/s11023-021-09571-w. URL https://doi.org/10.1007/s11023-021-09571-w.

[7] C. Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, 1998.

[8] R. Flesch. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233, 1948. doi: 10.1037/h0057532. URL https://doi.org/10.1037/h0057532.

[9] A. Freihat, M. Qwaider, and F. Giunchiglia. Using grice maxims in ranking community question answers. In *Proceedings of The Tenth International Conference on Information, Process, and Knowledge Management (eKNOW 2018)*, Rome, Italy, 03 2018.

[10] Y. Ge, Z. Xiao, J. Diesner, H. Ji, K. Karahalios, and H. Sundaram. What should I ask: A knowledge-driven approach for follow-up questions generation in conversational surveys. In C.-R. Huang, Y. Harada, J.-B. Kim, S. Chen, Y.-Y. Hsu, E. Chersoni, P. A, W. H. Zeng, B. Peng, Y. Li,

and J. Li, editors, *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 113–124, Hong Kong, China, Dec. 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.paclic-1.12.

[11] I. Gevers, I. Markov, and W. Daelemans. Linguistic analysis of toxic language on social media. *Computational Linguistics in the Netherlands Journal*, 12:33–48, Dec. 2022. URL https://www.clinjournal.org/clinj/article/view/146.

[12] P. Grice. Logic and conversation. In P. Cole and J. Morgan, editors, *Syntax and Semantics*, volume 3 of *Speech Acts*, pages 41–58. Academic Press, New York, 1975.

[13] P. Griffiths and C. Cummins. *An Introduction to English Semantics and Pragmatics*. Edinburgh Textbooks on the English Language. Edinburgh University Press, United Kingdom, 2 edition, Dec. 2016. ISBN 9781474412810.

[14] A. Hidayati and Arifuddin. Hate speech on social media: A pragmatic approach. *KnE Social Sciences*, 5(4):308–317, Mar. 2021. doi: 10.18502/kss.v5i4.8690. URL https://knepublishing.com/index.php/KnE-Social/article/view/8690.

[15] J. Hu, S. Floyd, O. Jouravlev, E. Fedorenko, and E. Gibson. A fine-grained comparison of pragmatic language understanding in humans and language models. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4194–4213, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.230. URL https://aclanthology.org/2023.acl-long.230.

[16] D. Jurafsky. Pragmatics and computational linguistics. In L. R. Horn and G. Ward, editors, *The Handbook of Pragmatics*, page 578. Blackwell Publishing Ltd, Malden, MA, USA, 2006. ISBN 978-0-631-22547-8.

[17] P. Jwalapuram. Evaluating dialogs based on Grice's maxims. In V. Kovatchev, I. Temnikova, P. Gencheva, Y. Kiprov, and I. Nikolova, editors, *Proceedings of the Student Research Workshop Associated with RANLP 2017*, pages 17–24, Varna, Sept. 2017. INCOMA Ltd. doi: 10.26615/issn.1314-9156.2017_003. URL https://doi.org/10.26615/issn.1314-9156.2017_003.

[18] J. Karoui, F. Benamara, V. Moriceau, V. Patti, C. Bosco, and N. Aussenac-Gilles. Exploring the impact of pragmatic phenomena on irony detection in tweets: A multilingual corpus study. In M. Lapata, P. Blunsom, and A. Koller, editors, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 262–272, Valencia, Spain, Apr. 2017. Association for Computational Linguistics. URL https://aclanthology.org/E17-1025.

[19] D. Khurana, A. Koli, K. Khatter, et al. Natural language processing: state of the art, current trends and challenges. *Multimed Tools Appl*, 82: 3713–3744, 2023. doi: 10.1007/s11042-022-13428-4.

[20] N. Kiyavitskaya, N. Zeni, L. Mich, and D. M. Berry. Requirements for tools for ambiguity identification and measurement in natural language requirements specifications. *Requirements Engineering*, 13(3): 207–239, 2008. doi: 10.1007/s00766-008-0063-7. URL https://doi.org/10.1007/s00766-008-0063-7.

[21] Y. Li, S. Wang, C. Lin, and F. Guerin. Metaphor detection via explicit basic meanings modelling. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 91–100, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-short.9. URL https://aclanthology.org/2023.acl-short.9.

[22] A. Liu, Z. Wu, J. Michael, A. Suhr, P. West, A. Koller, S. Swayamdipta, N. Smith, and Y. Choi. We're afraid language models aren't modeling ambiguity. In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 790–807, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.51. URL https://aclanthology.org/2023.emnlp-main.51.

[23] P. Madhyastha, A. Founta, and L. Specia. A study towards contextual understanding of toxicity in online conversations. *Natural Language Engineering*, 29(6):1538–1560, 2023. doi: 10.1017/S1351324923000414.

[24] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock. Finding deceptive opinion spam by any stretch of the imagination. In D. Lin, Y. Matsumoto, and R. Mihalcea, editors, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 309–319, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL https://aclanthology.org/P11-1032.

[25] M. Ott, C. Cardie, and J. T. Hancock. Negative deceptive opinion spam. In L. Vanderwende, H. Daumé III, and K. Kirchhoff, editors, *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 497–501, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL https://aclanthology.org/N13-1053.

[26] V. Parvaresh. Covertly communicated hate speech: A corpus-assisted pragmatic study. *Journal of Pragmatics*, 205:63–77, 2023. ISSN 0378-2166. doi: https://doi.org/10.1016/j.pragma.2022.12.009. URL https://www.sciencedirect.com/science/article/pii/S037821662200296X.

[27] T. A. Pasa, Nuriadi, and H. Lail. An analysis of sarcasm on hate speech utterances on just jared instagram account. *Journal of English Education Forum (JEEF)*, 1(1):10–19, Jun. 2021. URL https://jeef.unram.ac.id/index.php/jeef/article/view/94.

[28] J. Pavlopoulos, J. Sorensen, L. Dixon, N. Thain, and I. Androutsopoulos. Toxicity detection: Does context really matter? In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4296–4305, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.396. URL https://aclanthology.org/2020.acl-main.396.

[29] L. Ruis, A. Khan, S. Biderman, S. Hooker, T. Rocktäschel, and E. Grefenstette. The goldilocks of pragmatic understanding: Fine-tuning strategy matters for implicature resolution by llms, 2023.

[30] M. Saveski, B. Roy, and D. Roy. The structure of toxic conversations on twitter. In *Proceedings of the Web Conference 2021*, WWW '21, page 1086–1097, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383127. doi: 10.1145/3442381.3449861. URL https://doi.org/10.1145/3442381.3449861.

[31] M. Sorower, J. Doppa, W. Orr, P. Tadepalli, T. Dietterich, and X. Fern. Inverting grice's maxims to learn rules from natural language extractions. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. URL https://proceedings.neurips.cc/paper_files/paper/2011/file/8c7bbbba95c1025975e548cee86dfadc-Paper.pdf.

[32] M. Tewari, S. Bensch, T. Hellström, and K.-F. Richter. Modelling grice's maxim of quantity as informativeness for short text. In *Proceedings of the 10th International Conference in Languages, Literature, and Linguistics (ICLLL 2020)*, pages 1–7, Japan, 2020. URL https://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-176269.

[33] A. Upadhyaya, M. Fisichella, and W. Nejdl. Toxicity, morality, and speech act guided stance detection. In H. Bouamor, J. Pino, and K. Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4464–4478, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.295. URL https://aclanthology.org/2023.findings-emnlp.295.

[34] C. Van Hee, E. Lefever, and V. Hoste. Exploring the fine-grained analysis and automatic detection of irony on twitter. *Language Resources and Evaluation*, 52(3):707–731, 2018. ISSN 1574-0218. doi: 10.1007/s10579-018-9414-2. URL https://doi.org/10.1007/s10579-018-9414-2.

[35] W. Yin and A. Zubiaga. Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Computer Science*, 7, 2021. URL https://api.semanticscholar.org/CorpusID:231942329.

[36] Z. Zheng, S. Qiu, L. Fan, Y. Zhu, and S.-C. Zhu. GRICE: A grammar-based dataset for recovering implicature and conversational rEasoning. In C. Zong, F. Xia, W. Li, and R. Navigli, editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2074–2085, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.182. URL https://aclanthology.org/2021.findings-acl.182.

Mapping Sentiments: A Journey into Low-Resource Luxembourgish Analysis

Nina Hosseini-Kivanani[a,*], **Julien Kühn**[a] **and Christoph Schommer**[a]

[a]Department of Computer Science, University of Luxembourg

Abstract.

Sentiment analysis (SA) plays a vital role in interpreting human opinions across different languages, especially in contexts like social media, product reviews, and other user-generated content. This study focuses on Luxembourgish, a low-resource language critical to Luxembourg's identity, utilizing advanced deep learning models such as BERT, RoBERTa, LuxemBERTand LuxGPT-2. These models were enhanced with transfer learning, active learning strategies, and context-aware embeddings, enabling effective Luxembourgish processing. These models further improved with context-aware embeddings and were able to accurately detect sentiments, categorizing news comments into positive, negative, and neutral sentiments. Our approach highlights the significant role of human-in-the-loop (HITL) methodologies, which refine model accuracy by aligning automated analyses with human judgment. The findings indicate that LuxemBERT, especially when enhanced with the HITL method involving feedback from 500 and 1000 annotated sentences, outperforms other models in both binary (positive vs. negative) and multi-class (positive, neutral, and negative) classification tasks. The HITL approach not only refined model accuracy but also provided substantial improvements in understanding and processing sentiments and sarcasm, often challenging for automated systems. This study establishes the basis for future research to extend these methodologies to other under-resourced languages, promising improvements in Natural Language Processing (NLP) applications across diverse linguistic landscapes.

Keywords: Human-in-the-loop, Low-resource languages, Luxembourgish, Sentiment analysis, Transfer learning

1 Introduction

Sentiment analysis (SA), a key branch of NLP, automates the extraction of opinions, emotions, and attitudes from texts concerning various entities such as products and organizations [28]. Emerging in the early 2000s and also referred to as opinion mining or sentiment mining, this field primarily aims to classify texts as positive, negative, or neutral [6]. Specialized forms detect whether texts are hateful or offensive [12, 42], and address social issues such as racism using data from social media [17]. Given the significant influence of social media on political elections and marketing, SA has become critically important for businesses and governments [2].

In the area of SA, researchers have explored a range of methods, including both supervised and unsupervised techniques, all showing promising results (e.g., [23]). Early studies indicate that unsupervised models, which use sentiment dictionaries, grammatical analysis, and sentence structure patterns with manually created rules, can perform just as well as traditional supervised methods, such as Support Vector Machines (SVMs) and Naïve Bayes classifiers [34]. This exploration sets the stage for addressing the impact of data scarcity in low-resource languages.

The significant challenges of applying advanced SA techniques become evident in the context of low-resource languages like Luxembourgish. Despite the superior performance of DL models over traditional methods, their dependency on extensive labeled datasets remains a major hurdle, given the high costs and time required for data annotation [18]. To mitigate these challenges, methods such as transfer learning and active learning have been introduced to offer viable solutions (e.g., [1]).

Consider Luxembourg, a trilingual nation of over 590,000 residents, home to the largest population of Luxembourgish speakers. Recognized as the national language in 1984, Luxembourgish is integral to the nation's identity and essential for communication within the country. Originally a Central Franconian dialect, Luxembourgish has evolved into an independent language, becoming essential for all forms of communication within the country. In contexts where all participants are fluent, switching to French or German is generally avoided [14]. The term "low-resource language" refers to languages that lack substantial annotated or digital data [32, 4]. In the NLP field, there has been a significant increase in methods targeting these low-resource languages, broadening the applicability of language models to a more diverse set of languages.

Traditional static word embeddings do not capture contextual variations that influence meaning, which is essential for accurately understanding and analyzing language. To address this limitation, context-aware embeddings like BERT (Bidirectional Encoder Representations from Transformers) [8], Robustly Optimized BERT Approach (RoBERTa) [29], and GPT have been developed. These models offer dynamic, contextual representations that adapt based on the surrounding text, thus providing a more precise reflection of subtle sentiment expressions within texts [24]. By understanding the specific context in which words are used, these advanced models significantly enhance the accuracy of SA, making them indispensable in extracting true sentiment from complex language constructions, such as irony or sarcasm, commonly found in social media and other digital communications. Further, human-in-the-loop (HITL) is particularly valuable as it allows systems to adjust to real-world variables and user-specific needs that may not be fully anticipated at the time of a model's initial training. For instance, in SA, HITL can be instrumental in refining the understanding and classification of language used in different contexts, such as irony or cultural-specific expressions that automated

* Corresponding Author. Email: nina.hosseinikivanani@uni.lu.

systems might misinterpret [44]. This approach not only enhances the performance and trustworthiness of AI systems but also enables them to become more aligned with human values and ethics, a critical consideration as AI becomes more pervasive in everyday life [38].

1.1 Contribution

The main contribution of this paper is the application of state-of-the-art frameworks and the integration of human-in-the-loop components [44] for SA of Luxembourgish. We specifically focus on developing an SA model that classifies Luxembourgish news comments into positive, negative, and neutral sentiment classes at the sentence level. The structure of the paper is as follows:

- The 'Related Works' section provides an overview of the research in this field, highlighting previous approaches to SA in multilingual and low-resource contexts.
- 'Materials and Methodology' describes the proposed models in detail.
  - The 'Dataset' section outlines the description and sourcing of the datasets used.
- Comprehensive discussions of the findings are presented in the 'Evaluation' section, including comparisons with existing models and discussion on the effectiveness of different methodologies.
- The paper concludes with final remarks in the 'Conclusion and Future Work' section, summarizing the implications and potential future directions for SA research in low-resource languages.

2 Related works

SA is a crucial aspect of understanding human opinions across various languages, particularly in the context of social media, product reviews, and other user-generated content. DL methods have shown significant promise in improving SA, especially for low-resourced languages like Luxembourgish. Recent advancements in DL architectures, particularly Transformer-based language models, have led to breakthroughs in SA tasks. These models use pre-trained knowledge to enhance performance on downstream tasks, a method that is particularly effective in contexts where annotated data is scarce [20].

In the area of SA for low-resource languages, the challenges and potential solutions are diverse and multifaceted. For example, translating datasets from resource-rich languages to those with fewer resources, such as Urdu, can often change the meaning of sentiment and cause performance degradation due to polarity shift [13]. This shift can make sentiment classification systems work poorly. This challenge is further compounded in domain adaptation scenarios, such as with Danish, where dramatic performance drops occur when switching domains [9].

Several approaches have proven effective in dealing with these issues. Using methods like transfer learning [22], unsupervised learning, semi-supervised learning, and active learning can significantly improve SA for these languages. Sentiment classification approaches are broadly categorized into supervised [36], semi-supervised [16], and unsupervised [19]. Although most studies use supervised methods, a major challenge remains the lack of well-organized datasets.

Traditionally, SA research has primarily focused on well-resource languages such as English, German, and Chinese. However, the focus has shifted towards investigating SA in low-resource languages in recent years, promoting greater linguistic inclusivity in NLP tools [11, 27]. A study by Pang et al. [36] demonstrated that ML techniques for sentiment classification significantly surpass human-generated benchmarks. They applied three ML models—Naïve Bayes, Support Vector Machine, and Maximum Entropy—to a dataset of movie reviews. Using a 3-fold cross-validation method, they compared the effects of feature presence versus feature frequency. They found that feature presence, which indicates the binary occurrence of a feature, was more effective than feature frequency, which measures how often a feature appears. Of the three classifiers, SVM yielded the best performance.

The application of ML techniques to comments in Bangla from the entertainment sector has shown promising results, with accuracy rates exceeding 75% for sentiment classification [39]. This indicates that even low-resource languages can achieve significant performance in SA tasks with the right methodologies. Similarly, adaptive pretraining and careful selection of source language have been shown to improve SA for African languages, leading to improvements of over 10% F1 score points [40].

The challenges of SA in low-resource languages are substantial but can be overcome, as demonstrated by various studies proposing cutting-edge strategies to enhance precision (e.g., [15]). A systematic review of multilingual SA techniques reveals a growing interest in developing models for such languages, with DL methods particularly recommended [31]. In detail, DL models, particularly those that incorporate attention mechanisms, have been successfully applied to SA in Albanian social media comments, achieving an F1 score of 72.09% [21]. Furthermore, transformer-based models have demonstrated their potential to improve SA for low-resource African languages such as Nigerian Pidgin and Yoruba, achieving top rankings in SemEval-2023 Task 12 [20].

The application of Pre-trained language models like Bidirectional Encoder Representations from Transformers (BERT) and multilingual BERT (mBERT) has also been noteworthy (e.g., [43]). These models can enhance SA tasks without extensive fine-tuning, thus reducing training time and resource consumption while maintaining or even improving accuracy [25]. Transfer learning techniques using pre-trained multilingual models often outperform language-specific models in low-resource settings, showing further improvements after fine-tuning even with a small number of samples [35].

Fawzy et al. [10] address the challenge of SA in Arabic, a low-resource language with diverse dialects and complex linguistic features. They propose an approach that combines a BERT model with a Convolutional Neural Network (CNN) to improve SA accuracy. The model, BERT-CNN, fine-tunes only the last four layers of a pre-trained BERT model, reducing computational requirements while leveraging the CNN as a classification head for enhanced feature extraction. Tested on three Arabic Twitter datasets, the BERT-CNN model not only outperforms existing state-of-the-art models but does so with 50% smaller batch sizes, fewer training layers, and approximately 20% fewer epochs on the datasets.

Human-in-the-loop (HITL) approaches have also shown promise. Human-in-the-loop linguistic Expressions with Deep Learning (HEIDL), a prototype HITL machine learning system, enables higher-level interaction between humans and machines, improving productivity and generalizing models to unseen data [37]. HITL NLP frameworks integrate human feedback to improve NLP models, with promising future studies in integrating human feedback in the development loop [41]. HITL can achieve comparable or better performance than unsupervised domain adaptation (UDA) in person re-identification scenarios when unlabeled target data is infeasible [7].

While manually annotated datasets are essential for training and evaluating NLP models, recent studies have highlighted that even

widely used benchmark datasets often contain many incorrect annotations. This reveals additional challenges in SA for low-resource languages, where data scarcity is compounded by quality concerns [26].

Building on these findings, recent efforts have extended these techniques to low-resource languages, where traditional feature extraction methods face challenges due to sparse data availability. Innovations in transfer learning and unsupervised learning methods are beginning to show promise in overcoming these barriers, enabling more effective SA across a broader spectrum of languages [5]. These developments underline the growing necessity and potential for applying advanced ML techniques to enhance linguistic inclusivity in NLP applications.

In summary, while SA for low-resource languages presents unique challenges, primarily due to the scarcity of sufficient annotated data and linguistic resources, research indicates that these can be effectively addressed with innovative techniques such as adaptive pre-training, DL, cross-lingual techniques, and data augmentation strategies. Emerging methods like mBERT and adversarial learning are proving effective in enhancing the precision and generalizability of SA models for these languages.

3 Materials and Methodolgy

3.1 Dataset

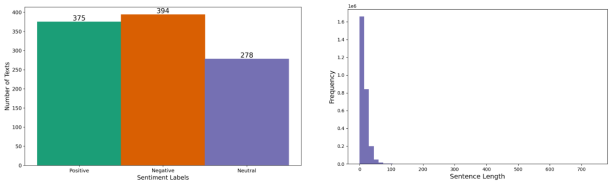


Figure 1: Sentiment Distribution Across the Dataset and Comment Lengths: The leftmost bar graph shows the distribution of sentiments, while the rightmost plot illustrates the distribution of sentence lengths.

The analysis in this study uses a corpus comprising user comments on news articles collected between 2009 and 2018 [1], with the longest sentence consisting of 744 words. The dataset, generously provided by RTL Luxembourg, is invaluable for SA as it encompasses responses to a wide range of topics authored by Luxembourgish speakers from diverse backgrounds and with varying personal histories. Figure 1 displays the number of annotations grouped by sentiment label (1047 comments in total), showing that *negative* comments were more frequent than *positive* or *neurtal* comments.

```
<sentence id="3bs">
  <w id="11" pos="" sen="3">pierre</w>
  <w id="12" pos="" sen="3">je</w>
  <w id="13" pos="" sen="3">suis</w>
  <w id="14" pos="" sen="3">surtout</w>
  <w id="15" pos="" sen="3">persuadé</w>
  <w id="16" pos="" sen="3">que</w>
  <w id="17" pos="" sen="3">L</w>
  <c id="18" pos="" sen="3">.</c>
</sentence>
```

Figure 2: Example of XML sentence structure from the dataset. *"pos"* refers to POS-tagging, which is not provided in this part of the dataset.

The data was provided in a simple XML file (see Figure 2), with each file containing a subset of the sentences. Prior to training the model, the XML file required preprocessing steps on user comments. This process involved transforming the XML data into a Pandas DataFrame [33], which is well-suited for handling such data operations in Python.

The preprocessing steps included:

- Parsing the XML Structure: Identifying and extracting key elements and attributes within the XML structure that contain the relevant information, such as sentence IDs and words.
- Cleaning the Data: Removing any extraneous tags and normalizing text to ensure consistency in sentiment classification.
- Annotating Sentiments: Ensuring the sentiment labels provided by annotators were correctly categorized into negative, neutral, and positive sentiments.

The sentiment labels were initially provided by human annotators and manually categorized. These annotations are critical as they form the basis for training and evaluating the SA models (see Figure 1).

3.2 Models

BERT: BERT is a transformative model in the field of NLP. Developed by Google, BERT has revolutionized how machines understand human language. It is based on the transformer architecture, which relies on attention mechanisms rather than sequence-aligned recurrent processing. This design allows for a more flexible interpretation of sentence structures.

The primary innovation of BERT is its approach to pre-training on a large corpus using only unlabeled data, followed by fine-tuning on smaller specific tasks. Unlike previous models that processed text in a single direction, either from left to right or right to left, BERT processes text bi-directionally. This bidirectional training is fundamental to its success, as it enables the model to capture the context of a word based by considering all surrounding text, both preceding and following. This capability allows BERT to understand the meaning of words within their specific sentence structures, which is a significant advance over traditional methods that often depend on labor-intensive feature engineering.

BERT's versatility is demonstrated in its ability to be fine-tuned with just an additional output layer to produce state-of-the-art results for a range of tasks, including question answering, language inference, and SA. In SA, BERT's ability to analyze the complete context of words makes it exceptionally effective in accurately classifying sentiments. This is particularly useful not just at the sentence level but also for more detailed aspect-level analysis, where the sentiments regarding specific aspects of a product or service are assessed.

The "bert-base-multilingual-cased" model is a variation of BERT designed to handle multiple languages. It retains BERT's powerful bidirectional context analysis while supporting text in various languages. This multilingual capability is particularly valuable for SA in multilingual settings, where it can interpret sentiments across different languages without needing separate models for each language. This feature extends BERT's versatility, allowing for consistent performance and ease of use in global applications.

BERT has consistently outperformed earlier models that relied on embeddings generated from simpler neural networks. Its ability to integrate and understand target-specific information within a text further enhances its performance, enabling more accurate sentiment discernment in complex scenarios. Research indicates that BERT's deep contextual understanding significantly improves performance across various NLP benchmarks, making it an essential tool for researchers and practitioners working with language data [8].

[1] www.rtl.lu.

RoBERTa: RoBERTa, or Robustly Optimized BERT Pretraining Approach, builds upon the foundational concepts of BERT by incorporating several key modifications that significantly improve its effectiveness. Unlike BERT, which is trained for a fixed amount of time on a set dataset size, RoBERTa benefits from training on larger datasets and for longer periods. This extended training allows RoBERTa to develop a more profound understanding of language details and complexities.

A critical enhancement in RoBERTa is the dynamic adjustment of the masking pattern during the pre-training phase. Whereas the masked language model (MLM) task in BERT randomly masks 15% of the tokens once at the beginning of training, which remains the same for every training epoch. In contrast, RoBERTa recalculates and randomizes the masks throughout the training process. This dynamic masking prevents the model from merely memorizing the masked positions, instead fostering more robust and generalizable language representations.

RoBERTa also adopts byte-level Byte-Pair Encoding (BPE) as its tokenization method, enhancing its ability to handle a more compact and efficient vocabulary. This approach is particularly beneficial for processing languages with rich morphology or those that use compound words, as it can decompose words into more frequently occurring subwords or bytes. By simplifying the vocabulary size and complexity, RoBERTa can process text data more quickly and with fewer resources than BERT.

Moreover, RoBERTa removes the next-sentence prediction (NSP) task, which BERT originally used. This decision is based on evidence that NSP does not significantly contribute to model performance on downstream tasks. Instead, RoBERTa focuses on optimizing the MLM objective, which has been shown to improve outcomes directly across a wide range of NLP benchmarks. This focused approach particularly benefits tasks requiring deep contextual understanding, such as SA, question answering, and natural language inference.

RoBERTa's performance demonstrates the importance of iterative improvements and optimizations in model pre-training strategies. It has outperformed BERT model and its other variants across various NLP benchmarks, establishing RoBERTa as one of the most potent models for tackling complex language processing challenges [29].

XLM-RoBERTa extends RoBERTa's capabilities to a multilingual setting. XLM-RoBERTa is designed to handle multiple languages simultaneously while supporting cross-lingual tasks. This model is especially valuable for SA in multilingual environments, where it can interpret and classify sentiments across different languages without the need for separate models for each language. XLM-RoBERTa retains RoBERTa's dynamic masking and BPE tokenization advantages, ensuring robust performance across diverse linguistic contexts.

LuxemBERT: LuxemBERT [30] is a state-of-the-art transformer-based language model specifically designed for the Luxembourgish language. It builds on the architecture of BERT (Bidirectional Encoder Representations from Transformers), which is renowned for its powerful bidirectional context understanding. This capability is particularly advantageous for addressing the challenges associated with low-resource languages like Luxembourgish.

LuxemBERT adopts BERT's robust framework, which features multiple layers of transformer encoders. These encoders use self-attention mechanisms to process input text sequences, allowing the model to discern complex dependencies and contextual relationships within the text. The bidirectional nature of LuxemBERT allows it to consider the context of each word from both preceding and following texts, thereby enhancing the accuracy of language representations.

A key strength of LuxemBERT is its adaptation to the Luxembourgish context through pre-training on language-specific corpora. Additionally, it uses transfer learning approaches, fine-tuned to specific downstream tasks relevant to Luxembourgish, such as SA, named entity recognition (NER), and text classification. This dual approach of pre-training and fine-tuning ensures that LuxemBERT can effectively generalize from limited data, maintaining high performance across diverse NLP applications.

LuxGPT-2[2]: LuxGPT-2 is an advanced transformer-based language model, developed using the GPT-2 (Generative Pre-trained Transformer 2) architecture, and specifically adapted for text generation in the Luxembourgish. Unlike traditional models that process text linearly, Lux-GPT-2 understands and generates text bi-directionally, where it predicts the next word in a sequence considering the entire context provided by all preceding words. This bidirectional approach is particularly effective in handling the subtleties of language that are critical for realistic and coherent text generation.

LuxGPT-2 was pre-trained on a substantial and varied corpus consisting of 711 MB of Luxembourgish text, which includes diverse sources such as RTL.lu news articles, parliamentary speeches, Wikipedia entries, and various web crawls. This extensive training set provides a rich linguistic foundation, allowing LuxGPT-2 to learn and reproduce the unique syntactic and semantic patterns of the Luxembourgish language.

The model's training process involved transfer learning techniques, where the model was initially conditioned on an English-based model to establish a broad understanding of linguistic structures. It then received further training (fine-tuning) to adapt these structures to Luxembourgish specifics. This phase included gradual layer freezing, a technique where lower layers of the model are incrementally locked as they stabilize, allowing the training focus to shift toward the upper layers that are responsible for capturing more complex and abstract language features.

Following its pre-training, LuxGPT-2 demonstrates remarkable versatility by being adaptable for fine-tuning on smaller, task-specific datasets. This flexibility enables LuxGPT-2 to excel in various NLP tasks, including SA, text summarization, and question-answering. Its ability to generate contextually rich, grammatically correct, and semantically detailed Luxembourgish text positions it as an essential resource for applications demanding high-quality Luxembourgish text output.

HITL: The concept of human-in-the-loop (HITL) has become increasingly relevant in various domains, particularly in fields like ML and artificial intelligence (AI). HITL methodologies are designed to integrate human judgment into the loop of automated systems, facilitating a dynamic interaction where humans provide real-time corrections and feedback.

HITL is instrumental in enhancing the reliability of AI systems. By incorporating human judgment, these systems can perform complex decision-making tasks with greater precision. Human intervention helps to refine AI responses by correcting errors that ML models may not identify on their own due to limitations in training data or inherent biases in their algorithms.

Involving human judgment in AI systems helps mitigate the risk associated with biases that are often present in the training data. Humans can identify and correct biased AI decisions in real-time, enhancing the fairness and impartiality of automated decisions. This real-time corrective feedback not only improves the model's current accuracy in the short term but also influences its learning trajectory, promoting better generalization and reliability in subsequent applications.

[2] https://huggingface.co/

Table 1: Performance metrics (accuracy, precision, recall) across BERT, RoBETa, LuxemBERT, and LuxGPT-2 models across two classification scenarios: Binary classification with positive and negative labels, and Multi-class classification with positive, neutral, and negative labels).

Models	Accuracy	Precision	Recall
BERT - Binary classification	57	60	62
RoBERTa - Binary classification	59	66	65
LuxemBERT - Binary classification	55	62	66
LuxGPT-2 - Binary classification	49	59	64
BERT - Multi-class classification	60	69	73
RoBERTa - Multi-class classification	64	71	72
LuxemBERT - Multi classification	67	73	72
LuxGPT-2 - Multi classification	35	49	52
HITL-multi (LuxemBERT-after training 500 sentences)	70	77	73
HITL-multi (LuxemBERT-after training 1000 sentences)	**75**	**78**	**77**

3.3 Training

To train our SA model, we used various Python libraries that support data manipulation, visualization, and ML. Key libraries included:

- NumPy: For numerical operations.
- Pandas: For managing data frames, allowing for seamless data manipulation and preparation.
- Matplotlib and Seaborn: For creating informative visualizations to analyze data trends and model performance.
- regex: For handling regular expressions.
- xml.etree.ElementTree: For parsing XML files.

Our model was trained using the TensorFlow framework. The dataset was divided into training (70%), validation (10%), and test (20%) sets through *stratified sampling* function from Scikit-learn [3]. This approach ensured that the distribution of data across each set matched that found in the original dataset, which is particularly important in studies with imbalanced classes. To further address the class imbalance, we incorporated class weights into the models. By assigning higher weights to the minority class, we ensured that the model penalized misclassifications of the minority class more heavily, thereby improving the overall balance and performance of the model across all classes.

3.4 Evaluation Metrics and their Implications

To evaluate the effectiveness of the classification techniques, specific metrics such as accuracy, precision, and recall were used:

- **Terminology for All Metrics:**
  - **TP (True Positives)** is the number of positive instances that the model correctly identifies.
  - **TN (True Negatives)** is the number of negative instances that the model correctly identifies.
  - **FP (False Positives)** are instances that the model incorrectly predicted as positive.
  - **FN (False Negatives)** are positive instances that the model fails to identify.
- **Accuracy** is a commonly used metric for evaluating classification models. It signifies the ratio of correctly classified instances to the total number of instances within the dataset. Specifically in text classification, accuracy measures the proportion of texts that are accurately categorized. A higher accuracy value indicates a more precise model. The formula for accuracy is expressed as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Precision** assesses the proportion of true positive predictions (correctly classified instances) among all positive predictions made by the model. It is a crucial metric that measures the model's ability to minimize false positives. It is determined by dividing the number of true positives by the total of true positives and false positives. A higher precision value suggests that the model makes fewer false positive errors, which is particularly valuable in scenarios where the cost of a false positive is high.

$$\text{Precision} = \frac{TP}{TP + FP}$$

- **Recall**, also referred to as sensitivity or the true positive rate, measures the proportion of true positives that are correctly identified out of the total sum of true positives and false negatives. Essentially, recall quantifies how effectively the model identifies all positive samples within the dataset. A higher recall value indicates a smaller number of false negatives. This metric evaluates the model's capability to detect all relevant instances without missing positive ones. In essence, recall measures how well the model captures all positive samples in the dataset.

$$\text{Recall} = \frac{TP}{TP + FN}$$

These metrics not only provide a comprehensive overview of the model's accuracy but also help understand its performance in terms of specificity and sensitivity, guiding further refinements.

4 Results and Discussion

Table 1 presents the performance metrics—accuracy, precision, and recall—of selected language models in binary (negative vs. positive) and multi-class (positive, neutral, and negative) classification scenarios. The evaluated models include BERT, RoBERTa, LuxemBERT, and LuxGPT-2, along with outcomes from a HITL setup involving iterative feedback loops on 500 and 1000 sentences, specifically designed to refine and enhance LuxemBERT's learning algorithms. This

[3] https://scikit-learn.org/stable/

setup enabled targeted improvements in the model's understanding of nuanced sentiments, particularly in complex linguistic contexts.

In binary classification tasks, RoBERTa outperforms BERT, achieving higher accuracy (59% vs. 57%), precision (66% vs. 60%), and recall (65% vs. 62%). This reflects RoBERTa's ability to efficiently differentiate between positive and negative classes. LuxemBERT, while showing comparable recall, experiences a slight drop in precision and accuracy, indicating sensitivity due to its architectural differences. LuxGPT-2, with significantly lower scores in accuracy (49%) and precision (59%), suggests potential configuration deficiencies for binary tasks.

The multi-class classification results reveal that LuxemBERT excels, particularly in a multilingual setting, achieving the highest metrics across accuracy (67%), precision (73%), and recall (72%). This indicates its ability to handle more complex label distinctions. RoBERTa maintains robust performance, although it slightly lags behind LuxemBERT, highlighting the subtle differences in its processing capabilities. In contrast, LuxGPT-2 shows considerable underperformance in this scenario, indicating that enhancements may be necessary to improve its adaptability to multi-class contexts.

The Iterative training process of LuxemBERT with human annotations under the HITL methodology demonstrates significant improvements, with accuracy increasing to 70% and 75% with 500 and then 1000 sentences, respectively. This gradual improvement underscores the value of integrating human feedback into the training process, enhancing both the accuracy and reliability of the model. Following the incorporation of human feedback, significant performance improvements were observed across the models. Detailed metrics such as accuracy, precision, and recall for each model iteration are summarized in Table 1.

As detailed in Table 1, the LuxemBERT model showed superior performance in handling complex expressions after training with human-annotated data. One noteworthy example of HITL's impact is its correction of the misinterpretation of the Luxembourgish expression "*Dat war awer eppes!*". Thanks to the expert annotations, the model could adjust its training algorithms to understand that, despite the presence of "*awer*" (but), the expression conveyed a positive sentiment. This correction was a direct result of the iterative training and feedback process unique to our HITL methodology. Compared to a baseline LuxemBERT model trained on the same initial dataset but without human feedback, our HITL-enhanced LuxemBERT model showed 8% improvement in detecting complex sentiments such as irony, which are often misunderstood by traditional SA tools. Human annotators played a crucial role, particularly in correcting sentiments related to cultural idioms like "*Ech si gréng hannert den Oueren.*" Initially labeled as neutral by the model, annotators clarified that this typically expresses a negative sentiment about one's lack of experience. Incorporating these corrections reduced the model's error rate in similar contexts. The HITL approach led to substantial improvements in LuxGPT2-s ability to discern sarcasm, a sentiment often misinterpreted by automated systems without localized training inputs.

One of the primary challenges was the scalability of human annotations, as recruiting enough native speakers trained in linguists was time-consuming and costly. Despite efforts to mitigate bias, the dominance of certain dialects within the annotated data occasionally skewed the model's performance on regional variations of Luxembourgish. Our findings emphasize the role of HITL methodologies in enhancing the performance of SA models for low-resource languages, with potential applications in content moderation, customer service bots, and social media analytics for improved accuracy and cultural relevance [3].

Integrating HITL-enhanced models with multilingual platforms like Google Translate or customer relationship management (CRM) systems could further enhance adaptability and accuracy across different languages and dialects. This approach not only improves technological inclusivity for low-resource language speakers but also underscores the need for ethical considerations in handling sensitive sentiment data.

The broader impact of our research extends to enhancing the digital inclusion of minority language speakers by developing technology that accurately understands and processes their language. However, as we collect and use sensitive sentiment data, it is imperative to adhere to stringent data privacy laws and ethical guidelines to protect individual privacy and prevent biases that could inadvertently arise from data misinterpretation. Despite BERT's good cross-lingual performance on high-resource languages, it struggles with low-resource languages, indicating a need for more efficient pretraining techniques or more data [43].

5 Limitations and Challenges

A limitation of our study is the dependency on a sufficient number of trained human annotators, which poses scalability challenges. This dependency could limit the application of our methods in larger-scale environments or in cases where such resources are scarce. Furthermore, the iterative training process, while effective, requires substantial computational resources, which may not be feasible in all application scenarios. Future research should, therefore, aim to optimize these processes to balance accuracy with operational efficiency better.

6 Conclusion and Future Work

This study introduces a method that can be useful for low-resource languages by adopting an existing model to a new context. Training a new model for a specific language can be expensive and time-consuming, taking days or even weeks, depending on the available computing power.

In our research, we focused on Luxembourgish, a language considered resource-scarce, comparing various BERT-based models alongside a HITL strategy. To address the scarcity of data, we implemented HITL strategy, which demonstrated improvements in SA of news comments. While not all models showed statistical significance, the HITL approach on LuxemBERT model consistently outperforms the BERT-based alternatives.

Future work should consider integrating semi-supervised learning techniques to better use unlabeled data. This approach could potentially expand the model's training dataset without requiring extensive human annotation. Additionally, applying the HITL methodology to other under-resourced languages could provide insights into the generalizability of this method across diverse linguistic landscapes.

Acknowlegments

We would like to thank Christoph Purschke (**Faculty of Humanities, Education and Social Sciences**, University of Luxembourg) for sharing the data and Aria Nourbakhsh (**Faculty of Science, Technology, and Medicine**, University of Luxembourg) for his invaluable assistance with brainstorming for the paper.

References

[1] S. A. A. Asli, B. Sabeti, Z. Majdabadi, P. Golazizian, R. Fahmi, and O. Momenzadeh. Optimizing annotation effort using active learning strategies: A sentiment analysis case study in persian. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2855–2861, 2020.

[2] M. Birjali, M. Kasri, and A. Beni-Hssane. A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, 226:107134, 2021.

[3] A. Chaudhary, J. Xie, Z. Sheikh, G. Neubig, and J. G. Carbonell. A little annotation does a lot of good: A study in bootstrapping low-resource named entity recognizers. 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 5164–5174, 2019.

[4] C. Cieri, M. Maxwell, S. Strassel, and J. Tracey. Selection criteria for low resource language programs. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4543–4549, 2016.

[5] F. Daneshfar. Enhancing low-resource sentiment analysis: A transfer learning approach. *Passer Journal of Basic and Applied Sciences*, 6(2): 265–274, 2024.

[6] N. C. Dang, M. N. Moreno-García, and F. De la Prieta. Sentiment analysis based on deep learning: A comparative study. *Electronics*, 9(3): 483, 2020.

[7] R. Delussu, L. Putzu, G. Fumera, and F. Roli. Online domain adaptation for person re-identification with a human in the loop. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 3829–3836. IEEE, 2021.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[9] J. Elming, B. Plank, and D. Hovy. Robust cross-domain sentiment analysis for low-resource languages. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 2–7, 2014.

[10] M. Fawzy, M. W. Fakhr, and M. A. Rizka. Sentiment analysis for arabic low resource data using bert-cnn. In *2022 20th International Conference on Language Engineering (ESOLEC)*, volume 20, pages 24–26. IEEE, 2022.

[11] R. R. R. Gangula and R. Mamidi. Resource creation towards automated sentiment analysis in telugu (a low resource language) and integrating multiple domain sources to enhance sentiment prediction. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*, 2018.

[12] A. Gaydhani, V. Doma, S. Kendre, and L. Bhagwat. Detecting hate speech and offensive language on twitter using machine learning: An n-gram and tfidf based approach. *arXiv preprint arXiv:1809.08651*, 2018.

[13] A. Ghafoor, A. S. Imran, S. M. Daudpota, Z. Kastrati, . Abdullah, R. Batra, and M. A. Wani. The impact of translating resource-rich datasets to low-resource languages through multi-lingual text processing. *IEEE Access*, 9:124478–124490, 2021. doi: 10.1109/ACCESS.2021.3110285.

[14] P. Gilles and C. Moulin. Luxembourgish. *Germanic standardizations: Past to present*, pages 303–329, 2003.

[15] V. Girija, T. Sudha, and R. Cheriyan. Analysis of sentiments in low resource languages: Challenges and solutions. In *2023 IEEE International Conference on Recent Advances in Systems Science and Engineering (RASSE)*, pages 1–6. IEEE, 2023.

[16] A. B. Goldberg and X. Zhu. Seeing stars when there aren't many stars: Graph-based semi-supervised learning for sentiment categorization. In *Proceedings of TextGraphs: The first workshop on graph based methods for natural language processing*, pages 45–52, 2006.

[17] E. Greevy and A. F. Smeaton. Classifying racist texts using a support vector machine. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 468–469, 2004.

[18] M. A. Hedderich, L. Lange, H. Adel, J. Strötgen, and D. Klakow. A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568, 2021.

[19] X. Hu, J. Tang, H. Gao, and H. Liu. Unsupervised sentiment analysis with emotional signals. In *Proceedings of the 22nd international conference on World Wide Web*, pages 607–618, 2013.

[20] N. Hughes, K. Baker, A. Singh, A. Singh, T. Dauda, and S. Bhattacharya. Bhattacharya_lab at semeval-2023 task 12: A transformer-based language model for sentiment classification for low resource african languages: Nigerian pidgin and yoruba. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1502–1507, 2023.

[21] Z. Kastrati, L. Ahmedi, A. Kurti, F. Kadriu, D. Murtezaj, and F. Gashi. A deep learning sentiment analyser for social media comments in low-resource languages. *Electronics*, 10(10):1133, 2021.

[22] Z. Kastrati, L. Ahmedi, A. Kurti, F. Kadriu, D. Murtezaj, and F. Gashi. A deep learning sentiment analyser for social media comments in low-resource languages. *Electronics*, 10(10):1133, 2021.

[23] G. Kaur and A. Sharma. A deep learning-based model using hybrid feature extraction approach for consumer sentiment analysis. *Journal of big data*, 10(1):5, 2023.

[24] S. Khan and T. Yairi. A review on the application of deep learning in system health management. *Mechanical Systems and Signal Processing*, 107:241–265, 2018.

[25] Y. Kit and M. M. Mokji. Sentiment analysis using pre-trained language model with no fine-tuning and less resource. *IEEE Access*, 10:107056–107065, 2022.

[26] M. Laurer, W. Van Atteveldt, A. Casas, and K. Welbers. Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and bert-nli. *Political Analysis*, 32(1):84–100, 2024.

[27] T. A. Le, D. Moeljadi, Y. Miura, and T. Ohkuma. Sentiment analysis for low resource languages: A study on informal indonesian tweets. In *Proceedings of the 12th Workshop on Asian Language Resources (ALR12)*, pages 123–131, 2016.

[28] B. Liu. *Sentiment analysis and opinion mining*. Springer Nature, 2022.

[29] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[30] C. Lothritz, B. Lebichot, K. Allix, L. Veiber, T. F. D. A. Bissyande, J. Klein, A. Boytsov, A. Goujon, and C. Lefebvre. Luxembert: Simple and practical data augmentation in language model pre-training for luxembourgish. In *Proceedings of the Language Resources and Evaluation Conference, 2022*, pages 5080–5089, 2022.

[31] K. R. Mabokela, T. Celik, and M. Raborife. Multilingual sentiment analysis for under-resourced languages: a systematic review of the landscape. *IEEE Access*, 11:15996–16020, 2022.

[32] A. Magueresse, V. Carles, and E. Heetderks. Low-resource languages: A review of past work and future challenges. *arXiv preprint arXiv:2006.07264*, 2020.

[33] W. McKinney et al. pandas: a foundational python library for data analysis and statistics. *Python for high performance and scientific computing*, 14(9):1–9, 2011.

[34] S. Momtazi et al. Fine-grained german sentiment analysis on social media. In *LREC*, volume 12, pages 1215–1220, 2012.

[35] A. Nugumanova, Y. Baiburin, and Y. Alimzhanov. Sentiment analysis of reviews in kazakh with transfer learning techniques. In *2022 International Conference on Smart Information Systems and Technologies (SIST)*, pages 1–6. IEEE, 2022.

[36] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. *arXiv preprint cs/0205070*, 2002.

[37] P. Sen, Y. Li, E. Kandogan, Y. Yang, and W. Lasecki. Heidl: Learning linguistic expressions with deep learning and human-in-the-loop. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 135–140, 2019.

[38] B. Shneiderman. Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction*, 36(6):495–504, 2020.

[39] N. Sultana, R. Sultana, R. I. Rasel, and M. M. Hoque. Aspect-based sentiment analysis of bangla comments on entertainment domain. In *2022 25th International Conference on Computer and Information Technology (ICCIT)*, pages 953–958. IEEE, 2022.

[40] M. Wang, H. Adel, L. Lange, J. Strotgen, and H. Schütze. Nlnde at semeval-2023 task 12: Adaptive pretraining and source language selection for low-resource multilingual sentiment analysis. *ArXiv*, abs/2305.00090, 2023. doi: 10.48550/arXiv.2305.00090.

[41] Z. J. Wang, D. Choi, S. Xu, and D. Yang. Putting humans in the natural language processing loop: A survey. In *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 47–52, 2021.

[42] H. Watanabe, M. Bouazizi, and T. Ohtsuki. Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. *IEEE access*, 6:13825–13835, 2018.

[43] S. Wu and M. Dredze. Are all languages created equal in multilingual bert? In *5th Workshop on Representation Learning for NLP, RepL4NLP*

2020 at the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, pages 120–130. Association for Computational Linguistics (ACL), 2020.

[44] X. Wu, L. Xiao, Y. Sun, J. Zhang, T. Ma, and L. He. A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, 135:364–381, 2022.

Navigating Opinion Space: A Study of Explicit and Implicit Opinion Generation in Language Models

Chaya Liebeskind[a,*,1] **and Barbara Lewandowska-Tomaszczyk**[b,1]

[a]Department of Computer Science, Jerusalem College of Technology, Israel
[b]Department of Language and Communication, University of Applied Sciences in Konin, Poland

Abstract. The paper focuses on testing the use of conversational Large Language Models (LLMs), in particular chatGPT and Google models, instructed to assume the role of linguistics experts to produce opinionated texts, which are defined as subjective statements about animates, things, events or properties, in contrast to knowledge/evidence-based objective factual statements. The taxonomy differentiates between Explicit (Direct or Indirect), and Implicit opinionated texts, further distinguishing between positive and negative, ambiguous, or balanced opinions. Examples of opinionated texts and instances of explicit opinion-marking discourse markers (words and phrases) we identified, as well as instances of opinion-marking mental verbs, evaluative and emotion phraseology, and expressive lexis, were provided in a series of prompts. The model demonstrated accurate identification of Direct and Indirect Explicit opinionated utterances, successfully classifying them according to language-specific properties, while less effective performance was observed for prompts requesting illustrations for Implicitly opinionated texts.To tackle this obstacle, the Chain-of-Thoughts methodology was used. Requested to convert the erroneously recognized opinion instances into factual knowledge sentences, LLMs effectively transformed texts containing explicit markers of opinion. However, the ability to transform Explicit Indirect, and Implicit opinionated texts into factual statements is lacking. This finding is interesting as, while the LLM is supposed to give a linguistic statement with factual information, it might be unaware of implicit opinionated content. Our experiment with the LLMs presents novel prospects for the field of linguistics.

1 Introduction

The present paper aims to discuss testing results with reference to the use of conversational Large Language Model (LLM), in particular chatGPT and Google models, instructed to assume the role of linguistics expert in our testing exercises.

2 ChatGPT applications

2.1 Language-focused LLM applications

Language has been the first system and communication medium which has been subject to Artificial Intelligence applications.

* Corresponding Author. Email: liebchaya@gmail.com
[1] Equal contribution.

2.1.1 Translation

is the first linguistic skill that was the forerunner of other machine-instructed applications. Researchers perform various studies to apply LLMs to machine translation tasks and evaluate their performance. One of the most recent ones uses automatic retrieval or human feedback as supervision signals to enhance the LLM's translation through in-context learning [23].

2.1.2 Language education

Since the first attempts of its systematic studies, language education has been an object of investigation and applicational attempts by various types of e-learning Computer-Assisted Language Learning systems. At present, LLMs are particularly widely used in education generally, and in language education in particular, proved to be especially relevant for teachers to devise work plans, curricula, language exercises and testing.

A significant step in the development of **automatised linguistic application systems** has been performed since first attempts to collect large language corpora [4]. Compiled by Henry Kučera and W. Nelson Francis at Brown University, in Rhode Island, USA, the so-called Brown Corpus, contained 500 samples of, predominantly written, American English (ca one million words). Since then, corpus linguistics, aided in the following years by computational linguistics, has begun its career in linguistics and its applications, particularly in lexicology, morpho-syntax, and discourse studies and, with the development of spoken corpora – in phonetics and phonology. The findings have been applied to lexicography, and more recently, to the identification of figurative meanings and implicit senses in semantics.

LLMs have paved the way towards communicative natural conversation applications, not to mention the areas of multilanguage and multimodal applications.

ChatGPT's naturally occurring **conversational skills**, fluent, human like, and coherent, are particularly attractive to the millions of users. And yet, although e.g., ChatGPT's conversational behaviour is considered structurally correct in the majority of tests, it equally often happens to be pragmatically unconventional, due to some excessive length, not following what are considered 'conversational routines', lacking context-sensitivity and conventional pragmatic competences.

2.1.3 Academic editing

Academic editing is by far most frequently used application of chatGPT [1]. It is used both lecturers in humanities and in STEM, as well as in student writing. There it may well serve language education objectives. There are also attempts to make ChatGPT write original poems, though its products typically lack refinement and finesse.

During last few months lexicographic testing, comments, descriptions, research projects, have also appeared in larger numbers, particularly relevant to **dictionary making and ontoterminological system building** [15, 12].

There are also attempts to automatically identify in the corpora rather **vague and implicit categories of meaning** (e.g., [5]). One of such categories, opinions, the topic of the present paper, is not a particularly frequent object to be satisfactorily identified by means of LLMs.

3 Language of Opinions

Language of opinionated texts is characterized by some properties which make it potentially distinguishable from fact-based statements. On the other hand, the class of implicit opinionated texts is particularly problematic to identify outside of context, because of the absence of a set of criterial, ever present universal markers. Therefore, we assume that some of the tasks to identify this and some other categories of opinion text, will also cause identification and illustration problems for LLM models.

3.1 Definitions of Opinion

In the paper by [7] definitions of opinion were scrutinized and the conclusion was reached that opinion is a subjective statement, containing judgement about THINGS (Human/Animal), which can be expressed in language or multimodally, about OBJECTS (people or things), EVENTS or PROPERTIES (Lewandowska-Tomaszczyk et al. 2023:461). The property of truth concerning expressing of an opinions is suspended – it is not known whether what is proposed is true or not [2].

To reach a contextually based definition, we proposed a cognitive-social understanding of opinion, perceiving it not as a single word, or sentence, but rather as an Opinion (Speech) Event (Lewandowska-Tomaszczyk et al. 2023:471) considered a semiotic act, which is embedded in a social-cultural context, and expresses an opinion holder's judgement on a person, animal, property or event. One additional caveat must be added to the definition of opinions, with reference to the emotional and evaluative language used in opinionated texts, dubbed as 'private state' expressions. Wiebe et al. [21] view private states in terms of their functional components as "states of experiencers holding attitudes optionally toward targets. For example, for the private state expressed in the sentence John hates Mary, the experiencer is John, the attitude is hate, and the target is Mary" (Wiebe etc. p.4). If the private states (including emotions, beliefs, etc.) are expressed with reference to direct experiences (e.g., I love Mary) they either cannot be treated as prototypical opinions or can fall out entirely of the definitional characteristics of opinions. Opinions can be private states expressions though expressed only towards so-called nested (or linguistically embedded) constructions, hence e.g., 'I love John' is not an opinion, 'I love skiing', can be considered a marginal opinion, while 'Russia fears war escalation' is an opinion due to the fact that the experience of the target 'other X fears war escalation' is a nested Speech Event.

In that paper we also proposed a typology of context-immersed opinions postulating a basic distinction between Explicit, which can be Direct or Indirect as opposed to Implicit opinionated texts. These categories are further subdivided into positive and negative opinions, ambiguous in this respect, and balanced opinions.

3.2 Taxonomy of Opinions

1. **Explicit – introduced by semantically transparent structural/semantic opinionated markers:** *Syntactic framing* imposes the order of linguistic elements used an opinion and together with *Semantic framing* identifies degrees of certainly and conviction by particular Agents: e.g., My/Our opinion is. . . ./According to me. . . Lexical framing is marked by relevant lexical items, as e.g., Cognitive verbs (e.g., I think, I believe, I feel), Modifiers (adjectives slow, adverbs slowly) that express evaluation or judgement (e.g., good/bad, worthy, valuable; slowly), in the three comparison degrees: positive pretty, comparative prettier (than), and superlative the prettiest (of. . . .), as well as expressions that convey personal feelings or experiences (e.g., I/they. . . love, I/they hate, I/they enjoy. . .).
2. **Explicit indirect opinion markers:** Opinions may be reinforced with persuasive language, such as rhetorical questions, appeals to authority, and emotional appeals [16], often accompanied by offensive and vulgar language. **Indirectly conveyed opinions:** he said/I've heard. Those opinionated texts which are introduced by means of unambiguous opinion markers such as 'I think/I don't think/I do not think', 'in my opinion' or 'according to me' or else by indirect Explicit Opinionated Texts heard/repeated from outside sources or via intermediaries. Contrasted with pragmatically expressed opinions, which are context-identifiable are Implicit Opinions.
3. **Implicit Opinions** Implicit opinions are typically used unaccompanied by any explicit opinion markers. However, they may include reference to targets that are vague.

4 Computational opinion identification and GPT at work

4.1 Previous attempts

In a report by Pew Research Center, Mitchell et al. [13] propose that in real life it is political awareness, digital savviness and trust in the media that all play large roles in the ability to distinguish between factual and opinion news statement. In digital methodology, the situation is not so simple. Rather modest numbers of publications focusing on the topic of opinion as opposed to factual knowledge statements is not direct. Rather they uncover opinions by the identification of opinion holders (e.g., [6]), or else most of the efforts focus on opinion mining that can analyse opinions from many information sources automatically and extract opinions, along with determining primarily their positive or negative (or else neutral) polarities, holders, strength, and possibly targets, typically by heuristic rule based and machine learning based methods.

A particularly problematic issue in opinion research, characteristic in fact of all language study, is the identification of vagueness and implicit language. There have been numerous attempts towards achieving this goal. It is also particularly important for the purposes of our study to investigate methods of identifying uncertainly, implicitness and vagueness in textual data as an important category of opinionated texts.

Original works by Wiebe et al. [22] with collaborators [20, 21] laid foundations on the development of a gold standard dataset for subjectivity classifications, subjectivity, which is a criterial property of opinionated statements (Lewandowska-Tomaszczyk et al. 2023). As the next step, there have been attempts at rather indirect ways to get to the sense of the concept of opinion. Yu and Hatzivassiloglou [24] investigated it via looking answers to opinion questions and, in this way, identify the polarity of opinion sentences. There are attempts at uncovering distinctions between general and specific types of text e.g., Louis and Nenkova [11] who investigated identifying general and specific sentences in news articles by exploring the feasibility of using existing annotations of discourse relations as training data for a general/specific classifier. This tool relies on classes of features that capture lexical and syntactic information, as well as word specificity and polarity. Dinu et al. [3] proposed an entirely different approach: hermeneutic introspection towards the intrinsic vagueness of analyzed texts, particularly for research on historical documentation. The author also presented limitations of annotation approaches in this respect.

In an extensive, detailed study on textual uncertainty Zerva [25] examined options of its automatic identification in to provide a more informative weighting of extracted knowledge, representing the confidence of the author in a statement. The author develops a set of uncertainty cues, grouped according to category Strong/Weak speculation where such words and expressions as WEAK certainty is represented by such forms as indicate, suggest, speculate, while admission to lack of knowledge by the words such as unknown/unclear; strong: hypothesize, propose, potent, while the medium level as there is evidence/it is known to be). Particular word clouds were generated by using the relative frequencies of cues in the corpora. The author used an adaptation of subjective logic theory in order to frame each event mention as an opinion model, in this way capturing potentially varying classification of uncertainty schemes.

In recent papers context-focused considerations have been applied e.g., Lian et al. [10] propose an approach of the F_vague detector to automatically detect vagueness in the text. According to their analysis, a large part of individual vague sentences have at least one clarifying sentence in the documents. The experiments showed good performance of high recall and precision.

With the advent of LLM generative tools, attempts at their use to identify and generate linguistically complex utterances have risen, e.g., in their paper on the identification of implicit toxicity in texts, Wen et al. [18] show that LLMs generate implicit toxic outputs that are exceptionally difficult to detect via simply zero-shot prompting.

4.1.1 Our approach

We implemented the chain-of-thought prompting (CoT) methodology [17]. CoT enhances the reasoning capacity of LLMs by incorporating systematic step-by-step reasoning procedures into the demonstration. CoT prompting enhances the model's comprehension of the question's complexities and the process of reasoning. In addition, the model produces a series of logical stages, providing us with a clear understanding of the model's cognitive process, hence improving its interpretability.

4.2 Prompts

White et al. [19] provided a comprehensive collection of efficient engineering methods, organized in a pattern format, that have been applied to address typical challenges encountered during interactions with LLMs. We used the following patterns in our experiment:

1. The Persona Pattern – we asked the LLMs to act as a linguistic expert, i.e., somebody who uses and knows the language very well, and provide outputs that such a persona would.
2. The Reflection Pattern - we successfully accomplished the objective of the reflection pattern, which involves prompting the model to automatically explain the rationale behind provided replies to the user. This was achieved by integrating the persona pattern with a request to provide a range of diverse examples that exemplify various linguistic phenomena.
3. The Cognitive Verifier Pattern - Research literature has established that LLMs demonstrate improved reasoning abilities when a question is broken into sub-questions, with their respective replies merged to form the overall solution to the original question [26]. Therefore, we attempted to apply this pattern as well. The description of explicit and indirect explicit opinionated texts encompasses various illustrative instances. For example, in explicit opinionated texts lexical framing is characterized by the use of pertinent lexical items, such as cognitive verbs and modifiers. We executed two queries. Initially, we solicited instances of lexical framing including cognitive verbs, and subsequently, we asked cases of lexical framing with modifiers. Nevertheless, the LLMs were unsuccessful in achieving the separation and, as a result, produced a combination of both types for the two queries. By employing a comprehensive and intricate definition of direct opinionated texts, encompassing a wide range of examples, the LLM models demonstrated superior performance. It not only generated more effective examples without repetition, but also categorized them based on the language phenomena they showed.
4. The Context Manager Pattern – we specify context for a conversation with the LLMs. We have enhanced the model by providing extra context, including a description of the category of opinionated text together with integrated relevant examples, instead of simply requesting examples based on a specific category name. The LLMs were then requested to provide examples that adhere to the category's rule.

It is observed that the utilization of the template pattern, which enables the user to specify a template for the output, was unnecessary in this case, as the bulleted list was already obtained in response to the examples request.

In addition, we incorporated an emotional stimulus into our prompt based on prior research [8] indicating that LLMs possess emotional intelligence and that their performance can be enhanced by the use of emotional prompts.

Given that LLMs tend to be chatty and have a tendency to engage in a 'question and answer' format by inventing their own questions [14], we explicitly urged the LLMs to avoid such behavior.

In one session, we used the conversational LLMs, to execute the following conversation:

1. We requested the LLMs to provide instances for the initial category of explicit opinionated text, based on the concept of the category that was explained with illustrations.
2. Following the LLMs' successful presentation of accurate examples demonstrating its comprehensive grasp of the category, the LLMs were then presented with the subsequent description of indirect explicit opinionated text and tasked with providing appropriate examples for this newly introduced category.
3. The LLMs provided accurate illustrations and received the definition of the final classification of implicit opinionated text.

4. The LLMs provided some incorrect examples and offered new examples of explicit direct and indirect opinionated content. In order to discern between factual and opinionated content, we requested the LLMs to transform the generated instances from the previous stage into factual statements.

```
system = ("You are a linguistic expert.") 1
prompt_text = '''
        2 Context: {context}
        3 Examples: {example}
        4 Task: {task}
        5 My job depends on how good
          and diverse these examples are.
          Don't be chatty,
        6 give me only the output format I asked for
          '''
prompt = ChatPromptTemplate.from_messages([
                            ("system", system),
                            ("human", prompt_text)])
```

Figure 1. Prompt Structure

Figure 1 illustrates the structure of our prompt. #1 is the persona pattern "You are a linguistic expert". #2 is the context, the category definition presented in the taxonomy. #3 includes examples. This option was exclusively utilized for the implicit category. Other categorized examples were presented in the context. #4 is the task: "Give me 50 diverse examples that represent different linguistic phenomena of opinionated sentences which follow this rule." (the rule, category definition, is provided in the context). #5 is the emotional prompt "My job depends on how good and diverse these examples are", and #6 is an instruction to avoid chatty behavior: "Don't be chatty, give me only the output format I asked for".

By applying the specified prompting method, we effectively provided examples for each of the preset categories of opinionated text, which proved challenging for corpus linguistic techniques [7].

4.3 Results

In this section we report the results of two popular conversational LLMs: OpenAI's ChatGPT-4[2] and Google's Gemini[3]. Both models were given identical prompts. Next, we detail the examples extracted for each category in the taxonomy of opinions.

4.3.1 Explicit opinions

Both LLMs successfully extracted 50 accurate examples as requested. Nevertheless, Gemini autonomously categorized them into distinct linguistic phenomena. Table 1 displays the various categories along with two examples for each category. In the semantic framing category, all the examples consisted of the first person singular/plural. This type can be considered as a peripheral opinion type. The fundamental semantic framework is centered around the concept of a person (excluding the first person singular/plural). The expected structure would be: You/He/She/It/They... strongly believe(s) (present simple or past simple tense). The ten examples of expressions of personal feelings or experiences are not opinions. At most, they can be considered extremely peripheral opinions, as they do not convey a nested target, as previously explained. While all the examples of the other linguistic phenomena category are correct, the majority of them do not establish a distinct category but rather offer supplementary illustrations for existing categories.

[2] https://chat.openai.com
[3] https://deepmind.google/technologies/gemini/

Table 1. Categories of explicit opinions by Gemini

Type	Ex1	Ex2
General	"In my view, the United States should withdraw from the Paris Agreement."	"I am convinced that the Earth is flat."
Semantic Framing	"I am somewhat skeptical of the claims that the new drug is effective."	"I am indifferent to the outcome of the election."
Lexical Framing	"The new tax policy is a disaster."	"College tuition is too expensive."
Expressions of Personal Feelings or Experiences	"I love my country."	"I am grateful for all that I have."
Other Linguistic Phenomena	"I believe that the media is biased against conservatives, don't you?"	"According to me, the best way to solve the climate crisis is to invest in renewable energy, and I am confident that it will work."

4.3.2 Explicit indirect opinions

As requested, both LLMs extracted fifty accurate examples successfully. However, while ChatGPT utilized rhetorical questions to emphasize opinions in all the extracted examples (such as "Can't you see the blatant injustice in our legal system?", "How can we deny the suffering of the homeless people in our city?", and "He said that our company's stocks are plummeting. Isn't it a sign of weak leadership?"), Gemini extracted a diverse range of examples that employed various forms of persuasive language and indirectly conveyed opinions. Gemini additionally offered a clarification, enclosed in parentheses, for each of the examples.

- "Isn't it obvious that this new policy is a disaster waiting to happen?" (Rhetorical question)
- "If you don't support this bill, you"re clearly heartless and don't care about the poor." (Emotional appeal)
- "My boss told me that if I don't meet my sales target this month, I'm out of a job." (Text heard/repeated from an intermediary)
- "This movie is so bad, it's almost funny." (Sarcasm)
- "If you don't like it here, then leave." (Appeal to ignorance)
- "If you don't vote for me, you're throwing your vote away." (Appeal to fear)
- "She's not exactly a beauty queen, but she's got a great personality." (Backhanded compliment)
- "I'm not saying he's a bad person, but I wouldn't trust him with my wallet." (Damning with faint praise)
- "I'm not ageist, but I don't think older people should be allowed to drive." (Ageism)
- "You're so stupid, you don't even know what you're talking about." (Ad hominem attack)

4.3.3 Implicit opinions

Despite being requested to provide 50 examples, the LLM only provided 25 and 30 examples for ChatGPT and Gemini respectively. All

Table 2. Implicit opinionated examples converted to facts

Type	Opinion	Fact
Implicit → Implicit	"Her research is groundbreaking - it could revolutionize the field."	"Her research has significant potential to make a major impact on the field."
	"That politician just talks the talk, but never delivers on their promises."	"The politician's pronouncements often fail to translate into concrete actions or outcomes."
Explicit Indirect → Explicit Indirect	"What a stunning view from this mountaintop!"	"The panorama from the mountaintop offers breathtaking visuals."
	"You should definitely come to the party tonight, it'll be a blast!"	"Many people are looking forward to the party tonight."
Implicit → Explicit Indirect	"The movie was just meh, not really worth the hype."	"The movie received mixed reviews."
	"I wouldn't mind trying that new restaurant - everyone says it's amazing!"	"Many people are praising the new restaurant."
	"He's certainly got a way with words, that's for sure."	"The movie received mixed reviews."
	"Her artwork is so bold and daring, I love it"	"Her artwork is characterized by its use of vivid colors and unconventional techniques."

the examples provided by ChatGPT consisted of explicit opinionated texts containing clear markers of opinion. The erroneous instances of implicit opinionated texts were effectively transformed into factual statements. For example, "I think that the movie was fantastic." was converted to "The movie received positive reviews", "I suppose the concert ended late" was converted to "The concert ended at midnight", and "My opinion is that the law should be revised" was converted to "The law is under review.".

Gemini excelled in producing implicit opinionated texts. 28 of the 30 examples were accurate. Additionally, the typology identification gave promising results. The generated distinction between the categories of Implicit and Explicit Indirect types can be presented in the following format: out of the 30 exemplary instances, generated as Implicit Opinions to be converted to Factual statements, less numerous instances are Implicit to Implicit opinions (3 examples), and Explict Indirect converted to synonymous Explicit Indirect opinions. The most numerous category are Implicit opinions converted to Explicitly Indirect ones (25 instances), which can be considered a big step towards the full clarification of opinion typology content. Examples are presented in Table 2. Some of the examples may be considered taxonomically ambiguous due to independent reasons: the missing reference to the contextually-anchored Opinion Event context that would disambiguate the taxonomy type. As discussed in the first sections of the present paper (p. 2), opinion is proposed to be defined as an event with the identification of opinion holder, its sources, target, effects, relation to evidence data, etc. With no such reference available, options to identify opinionated samples from factual statements are lower. This is not unique to LLM system's performance. Similarly, human language users experience identical problems with implicit opinion identification. The reason is that in terms of language, implicit opinions most often adopt a linguistic form identical to that used for factual statements, i.e., with the absence of evident linguistic clues that would make it possible to differentiate between the two types. The context-free utterance such as e.g., "It is raining", when said to another person on the phone, cannot be verified by the addressee as to its factual content. The sentence is referentially ambiguous between two conflicting scenarios. It can either convey a factual statement uttered in the outdoors context in a heavy rain, or else it can be produced in a cosy room, when the speaker sees water falling from the roof outside. The statement "It is raining" in the latter context is an implicit opinion, a shorthand for the complete (opinionated) form "I think it is raining". Both in the former and the latter scenarios, the contextual information on the event would clarify the ambiguity. One can thus conjecture that LLM systems would exploit its fuller taxonomic options when provided with the information on the Opinion Event contextual clues. Our next section will focus on attempts to refine a series of prompts towards making the opinion event contextual clues more transparent to incorporate such them into a system of contextual information transfer.

Contextual clues The search for entity context has recently been an important topic of inquiry in computational linguistics. Although existing corpus-based methods have achieved some progress, they still rely on annotated corpora to obtain the context patterns through the position of the entity in a sentence. There are attempts to overcome the annotation dilemma some authors such as Li et al. [9] explore an approach based on Entity Set Expansion (ESE) and its newer version GAPA (GenerAted PAtterns) to expand target entities, which are corpus-independent to automatically generate high-quality context patterns for entities.

To generate sufficient contextual clues by ChatGPT and Gemini to identify and produce implicit opinions we refined our prompts to urge the LLMs to generate some context for Implicit Opinion Events with the sentences provided.

Prompt: Generate a few sentences of context to make the following sentence an opinionated text, the context should include the sentence itself. The sentences are provided in the first column of Table 3:

5 Discussion

It is important to note that LLMs are nondeterministic models, meaning they exhibit different behaviors each time they are executed. Executing the identical prompt can yield varying instances. In addition, we did not optimize the prompts for any of the models. Customizing the prompt for each LLM individually may lead to improved outcomes. However, provision of more detailed, contextually specific condition types for each possible implicit opinion scenario might be considered time and effort consuming, especially when a series of similar forms would need to be sequentially generated, with reference to distinct objects or domains. A cumulative prompt admitting variance would be needed in such cases which would constitute an adequate prompting signal for the system. This level of opinion analysis is typically influenced by the context around the sentence and is considered decisive for applications dealing with various types of text - from social networks, political comments and casual conversations. Its aim is to determine language users' opinions about various phenomena, people, objects, properties and states and possibly confront them with the actual states of affairs in the outside reality. In order to perform this level of analysis, it is essential to identify both the linguistic context and the actual situational scene entities mentioned and their respective properties related to the words produced. From there, the judgment concerning a distinction between factual

Table 3. Implicit opinionated examples in context

Sentence	ChatGPT	Gemini
It is raining.	"It is raining outside," she muttered with a sigh, annoyed by the incessant downpour that seemed to dampen her plans every time she dared to make them. In her opinion, rainy days were nothing but a nuisance, disrupting outdoor activities and leaving everything soggy and gloomy. She couldn't fathom why some people found solace or romance in the pitter-patter of raindrops; to her, it was simply an inconvenience that warranted staying indoors with a hot cup of tea and a good book.	I peeked out the window and groaned. It is raining. Perfect. With this downpour, there's no way I'll be able to wear my new sandals today. Looks like it's boots again.
Her research is groundbreaking - it could revolutionize the field.	"Her research is groundbreaking - it could revolutionize the field," exclaimed Professor Thompson, his eyes gleaming with excitement as he discussed the latest findings in the faculty lounge. In his opinion, her work represented a paradigm shift, challenging long-held assumptions and opening up new avenues of exploration that could redefine the boundaries of their discipline. He couldn't help but feel a sense of awe at the sheer innovation and potential impact of her discoveries, envisioning a future where her ideas would shape the trajectory of their field for years to come.	For years, scientists have been searching for a way to create a more sustainable energy source. Many have tried and failed, but Dr. Amelia Rose's latest research is different. Her research is groundbreaking - it could revolutionize the field by creating a way to harness solar energy even more efficiently. With this new technology, we could see a dramatic reduction in our reliance on fossil fuels.
That politician just talks the talk, but never delivers on their promises.	"That politician just talks the talk, but never delivers on their promises," grumbled Sarah, her frustration evident as she flipped through the channels, each one featuring the same empty rhetoric and hollow pledges. In her opinion, this particular politician was a master of spin, adept at crafting persuasive speeches and making grandiose commitments during campaign season, only to backtrack or conveniently forget once elected. She couldn't understand how so many people continued to support someone who consistently failed to follow through on their word, lamenting the state of politics and longing for leaders who prioritized integrity and accountability over empty platitudes.	The election cycle is in full swing, and promises are being thrown around like confetti. But one politician in particular, [politician's name], seems to be all talk and no action. That politician just talks the talk, but never delivers on their promises. It's a shame, because they have some good ideas, but their lack of follow-through makes them untrustworthy in my eyes.
The movie was just meh, not really worth the hype.	"The movie was just meh, not really worth the hype," sighed Emily, disappointed as she left the theater with her friends. In her opinion, the film fell short of its lofty expectations, failing to deliver the gripping storyline and memorable performances that had been promised in the trailers and early reviews. She couldn't help but feel let down by the hype surrounding it, wishing she had spent her time and money on a more captivating cinematic experience instead.	Everyone was raving about this new movie, calling it a masterpiece. I was excited to see it, but after finally sitting down to watch it, I was left feeling disappointed. The movie was just meh, not really worth the hype. The plot was predictable, the characters lacked depth, and the special effects were underwhelming. I'd rather have spent my time watching something else.
I wouldn't mind trying that new restaurant - everyone says it's amazing.	"I wouldn't mind trying that new restaurant - everyone says it's amazing," remarked Mark, his curiosity piqued by the enthusiastic reviews he had heard from friends and colleagues. In his opinion, the widespread acclaim surrounding the eatery hinted at an exceptional dining experience just waiting to be savored. He couldn't resist the allure of exploring new culinary delights, eager to indulge his taste buds and see if the restaurant truly lived up to its stellar reputation.	Dating nights have been feeling a little repetitive lately. We keep going to our usual Italian place, and while it's good, I'm craving something new. I wouldn't mind trying that new restaurant on Elm Street - everyone says it's amazing. It's a little pricier than we usually go for, but for a special occasion, it could be perfect!
He's certainly got a way with words, that's for sure.	"He's certainly got a way with words, that's for sure, "chuckled Sarah, impressed by her colleague's eloquence during the presentation. In her opinion, his ability to articulate complex ideas with such clarity and flair was nothing short of remarkable. She couldn't help but admire his skillful communication style, which effortlessly captivated their audience and left a lasting impression.	I wasn't sure about attending the presentation at first, but then I heard who the speaker was. He's certainly got a way with words, that's for sure. People say he can take the driest topic and make it fascinating. I figured even if the content wasn't the most exciting, at least his delivery would be entertaining.
Her artwork is so bold and daring, I love it.	"Her artwork is so bold and daring, I love it," exclaimed Mia, her eyes sparkling with admiration as she gazed at the vibrant canvases adorning the gallery walls. In her opinion, the artist's fearless exploration of color and form challenged conventional boundaries, evoking a sense of excitement and inspiration. She couldn't help but feel drawn to the raw energy and emotion infused into each brushstroke, finding herself captivated by the bold creativity on display.	Galleries these days are filled with safe, predictable pieces. I crave art that pushes boundaries and makes a statement. Her artwork is so bold and daring, I love it! The way she uses color and texture is unlike anything I've seen before. It's truly refreshing to see an artist unafraid to take risks.

statements and opinions generally and possible implicit opinions, targeted at each of the identified entity or property, can be determined. It has been shown in the present paper that in the case of LLM generative tools more explicit prompts at the training phase that would indicate contextual variants of opinion scenario types might constitute materials leading to more satisfactory outcomes. In this way, a more effective functioning of the current models in this areas of their activity were obtained, particularly for implicit opinions, especially with Gemini LLMs.

References

[1] W. Castillo-González, C. O. Lepez, and M. C. Bonardi. Chat gpt: a promising tool for academic editing. *Data Metadata*, 1:23, 2022.

[2] J. Corcoran and I. S. Hamid. Investigating knowledge and opinion. *The Road to Universal Logic: Festschrift for 50th Birthday of Jean-Yves Béziau Volume I*, pages 95–126, 2015.

[3] A. Dinu, W. v. Hahn, and C. Vertan. On the annotation of vague expressions: a case study on romanian historical texts. *Proceedings of the LT4DHCSEE in conjunction with RANLP*, pages 24–31, 2017.

[4] W. N. Francis and H. Kucera. A standard corpus of present-day edited american english, for use with digital computers. *Brown University, Providence*, 2, 1964.

[5] M. Gupta, S. S. Bharti, and S. Agarwal. Implicit language identification system based on random forest and support vector machine for speech. In *2017 4th International Conference on Power, Control & Embedded Systems (ICPCES)*, pages 1–6. IEEE, 2017.

[6] L.-W. Ku, C.-Y. Lee, and H.-H. Chen. Identification of opinion holders. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 14, Number 4, December 2009*, 2009.

[7] B. Lewandowska-Tomaszczyk, C. Liebeskind, A. Baczkowska, J. Ruzaite, A. Dylgjeri, L. Kazazi, and E. Lombart. Opinion events: Types and opinion markers in english social media discourse. *Lodz Papers in Pragmatics*, 19(2):447–481, 2023.

[8] C. Li, J. Wang, Y. Zhang, K. Zhu, W. Hou, J. Lian, F. Luo, Q. Yang, and X. Xie. Large language models understand and can be enhanced by emotional stimuli. *arXiv preprint arXiv:2307.11760*, 2023.

[9] Y. Li, S. Huang, X. Zhang, Q. Zhou, Y. Li, R. Liu, Y. Cao, H.-T. Zheng, and Y. Shen. Automatic context pattern generation for entity set expansion. *IEEE Transactions on Knowledge and Data Engineering*, 2023.

[10] X. Lian, D. Huang, X. Li, Z. Zhao, Z. Fan, and M. Li. Really vague? automatically identify the potential false vagueness within the context of documents. *Mathematics*, 11(10):2334, 2023.

[11] A. Louis and A. Nenkova. Automatic identification of general and specific sentences by leveraging discourse annotations. In *Proceedings of 5th international joint conference on natural language processing*, pages 605–613, 2011.

[12] J. Miloš and R. Michael. The end of lexicography? can chatgpt outperform current tools for post-editing lexicography? In *Proceedings of the eLex 2023 Conference: Electronic Lexicography in the 21st Century. Brno: Lexical Computing*, pages 508–523, 2023.

[13] A. Mitchell, J. Gottfried, M. Barthel, and N. Sumida. Distinguishing between factual and opinion statements in the news. *Pew Research Center*, 2018.

[14] H. Pearce, B. Tan, P. Krishnamurthy, F. Khorrami, R. Karri, and B. Dolan-Gavitt. Pop quiz! can a large language model help with reverse engineering? *arXiv preprint arXiv:2202.01142*, 2022.

[15] G. P. Rees and R. Lew. The effectiveness of openai gpt-generated definitions versus definitions from an english learners' dictionary in a lexically orientated reading task. *International Journal of Lexicography*, page ecad030, 2023.

[16] W. R. Roberts. Rhetoric by aristotle. *HyperText Presentation © 1996 Procyon Publishing*, 1996. URL http://libertyonline.hypermall.com/Aristotle/Rhetoric/Rhetoric.html.

[17] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

[18] J. Wen, P. Ke, H. Sun, Z. Zhang, C. Li, J. Bai, and M. Huang. Unveiling the implicit toxicity in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1322–1338, 2023.

[19] J. White, Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. Elnashar, J. Spencer-Smith, and D. C. Schmidt. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*, 2023.

[20] J. Wiebe, T. Wilson, R. Bruce, M. Bell, and M. Martin. Learning subjective language. *Computational linguistics*, 30(3):277–308, 2004.

[21] J. Wiebe, T. Wilson, and C. Cardie. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39:165–210, 2005.

[22] J. Wiebe et al. Learning subjective adjectives from corpora. *Aaai/iaai*, 20(0):0, 2000.

[23] X. Yang, R. Zhan, D. F. Wong, J. Wu, and L. S. Chao. Human-in-the-loop machine translation with large language model. In *Proceedings of Machine Translation Summit XIX, Vol. 2: Users Track*, pages 88–98, 2023.

[24] H. Yu and V. Hatzivassiloglou. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 129–136, 2003.

[25] C. Zerva. *Automatic identification of textual uncertainty*. The University of Manchester (United Kingdom), 2019.

[26] D. Zhou, N. Schärli, L. Hou, J. Wei, N. Scales, X. Wang, D. Schuurmans, C. Cui, O. Bousquet, Q. Le, et al. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*, 2022.