Automatic Counter-Narrative Generation

María Estrella Vallecillo-Rodríguez

Computer Science Department, SINAI, CEATIC, Universidad de Jaén, 23071, Spain

Abstract

Social networks have become essential platforms for communication, information sharing, and personal expression. However, their openness and lack of regulation have also enabled the spread of harmful content, including hate speech, misinformation, and offensive stereotypes. These issues are exacerbated by the anonymity and virality inherent to online environments, which allow such content to proliferate rapidly and without accountability. In this context, Natural Language Processing (NLP) offers valuable tools to address the growing volume and impact of harmful messages online. This doctoral thesis explores the use of Large Language Models (LLMs) for the automatic generation of fact-based counter-narratives (CN) responses that adapt to different context situations designed to directly challenge and deconstruct harmful content while promoting empathy, inclusion, and critical reflection. By leveraging the capabilities of advanced generative models, the proposed system aims to support healthier digital discourse, reduce the burden on human content moderators, and contribute to broader efforts in combating misinformation and hate speech across social networks.

Keywords

Hate-speech, Counter-narrative Generation, Argumentation, Natural Language Generation

1. Justification of the proposed research

Social networks have become an integral part of daily life, enabling users to share experiences, ideas, and thoughts, while accessing real-time information and connecting with diverse communities. These platforms support the creation of digital identities with virtually unlimited possibilities.

However, their influence is not entirely beneficial. Often lacking sufficient regulation, social networks can enable irresponsible use that threatens various aspects others' lives. A major contributing factor is the anonymity they offer, coupled with the rapid spread of content. The ease of publishing without consequences and the potential for instant, widespread visibility often encourages harmful behavior.

A particularly concerning example is the spread of hate speech. The United Nations defines it as "any kind of communication in speech, writing or behaviour, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are" [1]. In digital environments, such content can quickly gain traction due to algorithms designed to maximize engagement. This fosters ideological bubbles and disproportionately targets marginalized communities, who become victims of hate speech, misinformation, and extremist narratives.

In response, some platforms have implemented moderation strategies, such as removing offensive content or blocking repeat offenders. However, these measures are not always effective, often raising concerns about freedom of expression. Some experts argue that censorship can backfire, attracting sympathy and reinforcing the censored message. In addition, moderation is typically performed by people regularly exposed to harmful content, which poses serious mental health risks. While some platforms are reducing reliance on human moderators, their absence leaves gaps in user protection.

In this context, the automatic generation of counter-narratives (CNs) has emerged as a promising strategy. The Council of Europe defines a counter-narrative as "a short and direct reaction to hateful messages used to directly de-construct, discredit and demystify violent extremist messages" [2]. This approach safeguards freedom of expression, promotes dialogue, and empowers affected communities, contributing to more inclusive and resilient digital spaces. It may also reduce the burden on human moderators and limit user exposure to harmful content.

Doctoral Symposium on Natural Language Processing, 25 September 2025, Zaragoza, Spain.

mevallec@ujaen.es (M. E. Vallecillo-Rodríguez)

0000-0001-7140-6268 (M. E. Vallecillo-Rodríguez)

This thesis proposes a system for generating fact-based counter-narratives that promote empathy, respect, and tolerance. We will leverage Large Language Models (LLMs), which can handle complex tasks with high accuracy, understanding instructions from minimal examples. In parallel, we will explore lightweight generative models trained on high-quality datasets, which have shown strong performance in specific applications. If successful, this system could extend to other areas of responsible AI, such as combating disinformation, fake news, and polarization on social media.

2. Related work

To combat hate speech on social media, automatic generation of CNs or counterspeech has gained attention. These responses are intended to challenge stereotypes and promote inclusion. CNs can be argumentative, offering factual rebuttals, or non-argumentative, simply rejecting offensive content.

Several studies have evaluated the effectiveness of CN. For example, Munger [3] and Mathew et al. [4] show that exposure to CNs reduces the use of racist slurs. At a theoretical level, Benesch [5] identifies CNs as a promising strategy against online abuse.

Efforts have also focused on the development of datasets. The CONAN corpus [6] combines multilingual data, context-aware messages, dialogues, and targeted hate speech. More recent datasets extend to languages such as Basque, Spanish, and Chinese [7, 8, 9, 10, 11], although most rely heavily on CONAN's structure, limiting linguistic diversity and generalization.

Regarding generation methods, comparative studies by Qian et al. [12] explore sequence-to-sequence models, variational autoencoders, and reinforcement learning. Other works employ LLMs [13, 14], integrate external knowledge [15], or regulate Transformer attention to improve generalization [16].

The Evaluation of the generated CNs is also a key challenge. Even though manual evaluation is still common [17, 10], automated approaches are gaining traction. Notably, [18] propose a multi-dimensional framework based on NGO guidelines, and [19] use pairwise LLM comparisons aligned with human preferences. Other studies assess tone, accessibility, and ethical risks [20], highlighting that LLM outputs, while empathetic, tend to be verbose and less accessible. Emotionally guided prompts improve results, but concerns about safety and effectiveness persist.

Community-driven efforts, such as the CS4OA and Multilingual Counterspeech workshops [21], and shared tasks such as RefutES [22], are essential for addressing shared challenges. A recurring issue is the generation of CNs that are both direct and argumentative. Bonaldi et al. [23] note that safety filters may weaken the argumentative force, and that targeting implicit stereotypes with well-reasoned arguments enhances quality. Similarly, Furman et al. [17] show that using a small, focused set of examples can match the quality of outputs trained on full datasets. To address generic outputs, Baez Santamaria et al. [24] propose modeling dialogue history to produce more personalized CNs.

As this thesis also intersects with argumentation theory, it is important to consider foundational concepts. Hate speech and argumentative discourse function within a communicative act [25], comprising logic (premises and conclusions), dialectic (rules of interaction), and rhetoric (persuasion and ethics) [26, 27]. In computational linguistics, systems now identify argumentative components [28] or apply Event Argument Extraction (EAE) to detect events such as dehumanization or incitement [29]. Although there is growing interest in classifying persuasion types in social media [30, 31], systems that adapt to different communicative contexts remain scarce. Only a few works [17, 32, 23] address the argumentative weaknesses in hate speech, but none offer the contextual depth proposed in this thesis (see Section 5).

3. Hypothesis and objectives

We propose the following hypothesis: a Large Language Model can be used to respond to hate-speech messages by retrieving and generating solid, truthful arguments in Spanish that enrich the discourse. In addition, the system must be adaptable to different contextual situations and be able to establish a tone and style appropriate to the ongoing conversation.

With this hypothesis in mind, the following objectives are established:

- Analyze and characterize counter-narrative strategies in terms of language use and identify where argumentation can be effectively applied.
- Know in depth the different types of text used to offend, misinform, or promote stereotypes.
- Investigate existing resources on argumentation, counter-argumentation, and counter-narratives, and develop new resources specifically for Spanish.
- Conduct experiments using prompting techniques, multi-agent setups, or model adaptation (e.g., fine-tuning), leveraging both existing and newly developed datasets.
- Participate in evaluation campaigns to assess and improve the developed systems.
- Disseminate results through academic publications and propose the organization of shared tasks related to this research.

4. Methodology and proposed experiments

This thesis project is structured over an estimated three-year timeline. Throughout this period, the candidate is expected to participate in seminars, workshops, and conferences to present progress, exchange ideas with fellow researchers, and apply the developed systems in relevant academic contexts. These activities aim to validate the research and contribute to the broader scientific community. The publication of results in specialized journals is also expected. To test the main hypothesis and meet the outlined objectives, the following stage-based plan is proposed:

4.1. First Year: Review of the State of the Art and Existing Resources

The first year will focus on a comprehensive review of the state-of-the-art in automatic CN generation. This will involve identifying current challenges, effective methodologies, and the characteristics of existing datasets in different domains.

Additionally, relevant work in argument mining will be analyzed to assess how argumentation frameworks can be integrated into CN systems, thereby enhancing their persuasive effectiveness.

By the end of this stage, the expected outcomes are:

- A theoretical framework for analyzing and designing argument-based CNs.
- A methodology for evaluating the quality of generated CNs and counter-arguments.
- A survey and characterization of existing datasets suitable for use in subsequent experiments.

4.2. Second Year: Resource Development and System Implementation

In the second year, existing and newly created linguistic resources will support the development of a system for generating argument-based CNs, with a focus on iterative refinement and evaluation.

The intended results of this stage include:

- A set of Spanish-language linguistic resources (corpora, datasets, databases) specifically tailored to CN generation.
- New resources collected from social media or other approved sources.
- A formal design specification of the proposed system architecture.
- A functional system prototype prepared for iterative enhancement in subsequent phases.

4.3. Third Year: Experimentation and Evaluation

The final year will focus on validating the system through comparative analyses with existing approaches, using standard metrics for text and argument quality. Results will guide targeted improvements and be shared through academic publications and events.

The project is expected to conclude with the following outcomes:

- A detailed experimental plan, including methodologies and associated hypotheses.
- A critical evaluation of the results in relation to the initial research questions.
- Final system refinements based on empirical findings.
- Scientific publications in peer-reviewed journals and conference proceedings.
- Participation in academic events to present results and receive expert feedback.

5. Current state of research

Currently, several experiments have been conducted on the automatic generation of CNs in Spanish, with the goal of addressing hate speech and the stereotypes that often accompany it. These efforts have led to the creation of two dedicated corpora, which serve as foundational resources for future research. Additionally, argumentative approaches supported by up-to-date knowledge bases are being explored to construct messages capable of countering misinformation and manipulative discourse. In this context, techniques such as Retrieval-Augmented Generation (RAG) [33], as well as reasoning strategies based on LLMs such as Chain of Thought (CoT) and Tree of Thought (ToT) are under consideration. These methods aim to enhance the effectiveness of CNs through informed and structured reasoning. Finally, there has been active participation in evaluation campaigns and workshops related to the automatic generation of CNs and responsible reasoning in language models.

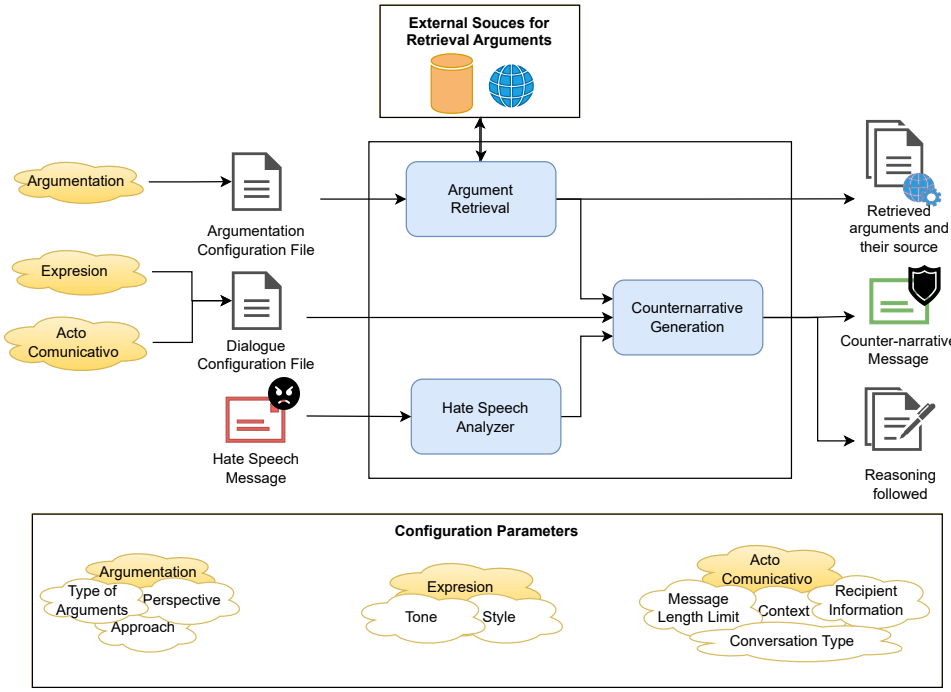

Figure 1: Scheme of the system prototype to be developed in this thesis with its inputs, components and outputs.

A preliminary prototype of the system has been designed (currently under development and subject to change as research progresses and large language models evolve). As shown in Figure 1, the system receives an offensive message along with configuration files that guide the language models in generating personalized CNs. These configurations specify tone, style, relationship with the recipient, communicative context (formal or informal), message length, type of interaction (single exchange or part of a dialogue), and the argumentative approach, which may involve addressing the most influential aspects of the offensive message or targeting its weakest points. The system comprises three main modules. The first analyzes the offensive message, identifying the targeted group, the attacked aspects, and the level of aggressiveness. The second retrieves relevant arguments, which are used by the third component to generate the CN. The output includes the generated CN, the arguments used along with their sources, and a file detailing the reasoning followed by the system at each stage.

5.1. Generated datasets

Among the corpora generated for Spanish, we find:

- **CONAN-SP** [9], based on CONAN-KN [13], contains 238 HS-CN pairs translated with DeepL and generated with GPT-3.5 using three prompting strategies. Each pair was manually evaluated for offensiveness, stance, and informativeness.
- **CONAN-MT-SP** [10], based on CONAN-MT, includes 5003 pairs translated into Spanish and CNs generated with GPT-4 using FSL prompting. Each instance includes human evaluations in six dimensions (offensiveness, stance, truthfulness, required editing, etc.), as well as a comparison between human and model.

5.2. Developed systems

Different prompting strategies based on Few-Shot Learning (FSL) and Zero-Shot Learning (ZSL) [34] have been tested for dataset generation. Additionally, the efficient training strategy QLoRA [35] has been applied to fine-tune a model that has served as a baseline in the RefutES task. Currently, research continues on multi-stage prompting strategies such as CoT and Tree-of-Thought. As future work, methods like LOMO [36] and the incorporation of external information via RAG will be explored.

Finally, as a result of the state-of-the-art analysis and in response to a common limitation in existing CN generation systems, such as their overly generic outputs, we are developing a novel framework designed to generate more nuanced and context-aware responses. This framework is based on a configurable prompt structure that allows for the generation of more nuanced and context-sensitive responses. The user can adjust tone, style, target audience, argumentative approach, minimum response length, type of evidence, emphasis (logical, ethical, emotional, analogical), and other parameters. This system includes mechanisms to prevent conflict escalation and ensure that the response is coherent, logical, verifiable, and non-offensive.

5.3. Participation in Shared Tasks

To evaluate the behavior of different LLMs and their adaptation through prompts, participation has been made in three shared tasks:

- **Multilingual detoxification (PAN Lab 2024)** [37]: The CoT-SC strategy [38] was employed to generate three neutral versions of a toxic message and select the best one based on automatic metrics.
- **Oppositional Author Analysis (PAN Lab 2024)** [39]: LLMs such as LLaMA3 and GPT-3.5 were used, fine-tuned with specific instructions to classify texts as critical vs. conspiratorial and to detect oppositional narrative elements.
- **Retrieval-Augmented Debating (Touché Lab 2025)** [1]: A multi-stage system was proposed where different models select relevant arguments from a database to support a counter-narrative structured in five rounds of debate.

5.4. Organization of scientific events

I have been part of the organizing committee for:

- **RefutES 2024**: A task in IberLEF focused on generating automatic CNs in Spanish in response to offensive messages targeting vulnerable groups. A baseline was established using ZSL and QLoRA on LLaMA2-13B-chat.

[1] https://touche.webis.de/clef25/touche25-web/retrieval-augmented-debating.html

- **The First Workshop on Multilingual Counterspeech Generation (MCG) 2025**: Organized at COLING, this workshop aims to bring together the scientific community to promote the development of systems in low-resource languages, propose new evaluation methods, and evaluate LLMs in Spanish, Basque, Italian, and English through the introduction of a shared task. It is proposed that the systems be capable of generating reasoned and specific CNs, evaluated using traditional metrics and the JudgeLM method for final ranking.

6. Research Elements Proposed for Discussion

Being at the early stage of my research, several key issues remain open for discussion and further exploration. These issues span linguistic, technical, and ethical dimensions, and will guide the theoretical and experimental development of this thesis:

- **Adaptation and performance of LLMs in Spanish.** What adaptation techniques are most effective in enhancing the accuracy and coherence of texts generated in Spanish? What are the main limitations of LLMs when generating argumentative texts in Spanish, and how might these be overcome—through fine-tuning, prompt engineering, or other strategies?
- **Generation of counter-arguments.** What kinds of messages are suitable for counter-argumentation? What typologies of counter-arguments exist, and which are appropriate for addressing harmful or offensive speech responsibly? Can counter-arguments be tailored to different user profiles? How can their quality and persuasive power be reliably evaluated?
- **Access to and integration of external information.** When should external data be used in counter-argument generation, and how can it be reliably retrieved, filtered, and integrated to ensure factual accuracy?
- **Responsibility and ethics in text generation.** How can biases present in LLMs be mitigated, particularly when generating content in Spanish? What safeguards are needed to ensure that counter-arguments do not reproduce stereotypes or misinformation? What frameworks can guide the responsible deployment of these systems?
- **Evaluation and validation of system outputs.** What metrics are most appropriate for assessing the relevance, effectiveness, and ethical soundness of generated counter-narratives? Can human-in-the-loop approaches help refine model performance? Are LLMs reliable as evaluators of counter-narratives, as in the "LLM-as-a-judge" paradigm? What kinds of biases may arise when relying on LLMs for evaluation?
- **User adaptation and personalization.** How can the system adapt counter-narratives to different target audiences, such as victims, aggressors, or bystanders? What level of personalization is appropriate, and how can it be achieved without introducing ethical risks?

7. Acknowledgments

This work would not have been possible without the guidance and constant support of my supervisors, Arturo Montejo-Ráez and María Teresa Martín-Valdivia, to whom I am deeply grateful for sharing their expertise, time and trust throughout this process. I also extend my gratitude to the doctoral program at my beloved University of Jaén, as well as the Centro de Estudios Avanzados en Tecnologías de la Información y Comunicación (CEATIC), for providing the resources that made this research feasible. This work has been developed in the context of the CONSENSO project (PID2021-122263OB-C21), funded by the Spanish Government's Plan Nacional I+D+i.

Declaration on Generative AI

During the preparation of this work, the author used GPT-4o and DeepL for grammar correction, translation and spelling checks. After using these tools, the author carefully reviewed and edited the content as needed and takes full responsibility for the final version of the publication.

References

[1] United Nations, What is hate speech?, in: STRATEGY AND PLAN OF ACTION ON HATE SPEECH, United Nations, 2019. URL: https://www.un.org/en/genocideprevention/documents/UN%20Strategy%20and%20Plan%20of%20Action%20on%20Hate%20Speech%2018%20June%20SYNOPSIS.pdf.

[2] A. de Latour, N. Perger, R. Salaj, C. Tocchi, P. Viejo-Otero, We can! taking action against hate speech through counter and alternative narratives, in: Toolkit for human rights speech, European Union and Council of Europe, 2017, pp. 77–86. URL: https://rm.coe.int/wecan-eng-final-23052017-web/168071ba08.

[3] K. Munger, Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment, Political Behavior 39 (2017) 629–649. URL: https://doi.org/10.1007/s11109-016-9373-5. doi:10.1007/s11109-016-9373-5.

[4] B. Mathew, N. Kumar, Ravina, P. Goyal, A. Mukherjee, Analyzing the hate and counter speech accounts on twitter, 2018. arXiv:1812.02712.

[5] S. Benesch, Countering Dangerous Speech: New Ideas for Genocide Prevention, 2014. URL: https://papers.ssrn.com/abstract=3686876. doi:10.2139/ssrn.3686876.

[6] M. Guerini, Counter-narratives datasets to fight hate speech, 2023. URL: https://github.com/marcoguerini/CONAN, original-date: 2019-05-30T09:48:42Z.

[7] H. Bonaldi, M. E. Vallecillo-Rodríguez, I. Zubiaga, A. Montejo-Raez, A. Soroa, M.-T. Martín-Valdivia, M. Guerini, R. Agerri, The first workshop on multilingual counterspeech generation at COLING 2025: Overview of the shared task, in: H. Bonaldi, M. E. Vallecillo-Rodríguez, I. Zubiaga, A. Montejo-Ráez, A. Soroa, M. T. Martín-Valdivia, M. Guerini, R. Agerri (Eds.), Proceedings of the First Workshop on Multilingual Counterspeech Generation, Association for Computational Linguistics, Abu Dhabi, UAE, 2025, pp. 92–107. URL: https://aclanthology.org/2025.mcg-1.10/.

[8] M. Bennie, D. Zhang, B. Xiao, J. Cao, C. X. Liu, J. Meng, A. Tripp, PANDA - paired anti-hate narratives dataset from Asia: Using an LLM-as-a-judge to create the first Chinese counterspeech dataset, in: H. Bonaldi, M. E. Vallecillo-Rodríguez, I. Zubiaga, A. Montejo-Ráez, A. Soroa, M. T. Martín-Valdivia, M. Guerini, R. Agerri (Eds.), Proceedings of the First Workshop on Multilingual Counterspeech Generation, Association for Computational Linguistics, Abu Dhabi, UAE, 2025, pp. 1–12. URL: https://aclanthology.org/2025.mcg-1.1/.

[9] M. E. Vallecillo-Rodríguez, A. Montejo-Raéz, M. T. Martín-Valdivia, Automatic counter-narrative generation for hate speech in spanish, Procesamiento del Lenguaje Natural 71 (2023) 227–245.

[10] M.-E. Vallecillo-Rodríguez, M.-V. Cantero-Romero, I. Cabrera-De-Castro, A. Montejo-Ráez, M.-T. Martín-Valdivia, CONAN-MT-SP: A Spanish corpus for counternarrative using GPT models, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italy, 2024, pp. 3677–3688. URL: https://aclanthology.org/2024.lrec-main.326.

[11] J. Bengoetxea, Y.-L. Chung, M. Guerini, R. Agerri, Basque and Spanish counter narrative generation: Data creation and evaluation, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 2132–2141. URL: https://aclanthology.org/2024.lrec-main.192/.

[12] J. Qian, A. Bethke, Y. Liu, E. Belding, W. Y. Wang, A benchmark dataset for learning to intervene in online hate speech, 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), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 4755–4764. URL: https://aclanthology.org/D19-1482. doi:10.18653/v1/D19-1482.

[13] Y.-L. Chung, S. S. Tekiroğlu, M. Guerini, Towards knowledge-grounded counter narrative generation for hate speech, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Association for Computational Linguistics, Online, 2021, pp. 899–914. URL:

https://aclanthology.org/2021.findings-acl.79. doi:10.18653/v1/2021.findings-acl.79.
[14] S. Tekiroglu, H. Bonaldi, M. Fanton, M. Guerini, Using pre-trained language models for producing counter narratives against hate speech: a comparative study, in: Findings of the Association for Computational Linguistics: ACL 2022, 2022, pp. 3099–3114.
[15] M. Ashida, M. Komachi, Towards automatic generation of messages countering online hate speech and microaggressions, in: Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH), Association for Computational Linguistics, Seattle, Washington (Hybrid), 2022, pp. 11–23. URL: https://aclanthology.org/2022.woah-1.2. doi:10.18653/v1/2022.woah-1.2.
[16] H. Bonaldi, G. Attanasio, D. Nozza, M. Guerini, Weigh Your Own Words: Improving Hate Speech Counter Narrative Generation via Attention Regularization, in: Y.-L. Chung, H. Bonaldi, G. Abercrombie, M. Guerini (Eds.), Proceedings of the 1st Workshop on CounterSpeech for Online Abuse (CS4OA), Association for Computational Linguistics, Prague, Czechia, 2023, pp. 13–28. URL: https://aclanthology.org/2023.cs4oa-1.2.
[17] D. Furman, P. Torres, J. Rodríguez, D. Letzen, M. Martinez, L. Alemany, High-quality argumentative information in low resources approaches improve counter-narrative generation, in: H. Bouamor, J. Pino, K. Bali (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore, 2023, pp. 2942–2956. URL: https://aclanthology.org/2023.findings-emnlp.194/. doi:10.18653/v1/2023.findings-emnlp.194.
[18] J. Jones, L. Mo, E. Fosler-Lussier, H. Sun, A multi-aspect framework for counter narrative evaluation using large language models, 2024. URL: https://arxiv.org/abs/2402.11676. arXiv:2402.11676.
[19] I. Zubiaga, A. Soroa, R. Agerri, A LLM-based ranking method for the evaluation of automatic counter-narrative generation, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2024, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 9572–9585. URL: https://aclanthology.org/2024.findings-emnlp.559/. doi:10.18653/v1/2024.findings-emnlp.559.
[20] M. K. Ngueajio, F. M. P. del Arco, Y.-L. Chung, D. B. Rawat, A. C. Curry, Think like a person before responding: A multi-faceted evaluation of persona-guided llms for countering hate, 2025. URL: https://arxiv.org/abs/2506.04043. arXiv:2506.04043.
[21] H. Bonaldi, M. E. Vallecillo-Rodríguez, I. Zubiaga, A. Montejo-Ráez, A. Soroa, M. T. Martín-Valdivia, M. Guerini, R. Agerri (Eds.), Proceedings of the First Workshop on Multilingual Counterspeech Generation, Association for Computational Linguistics, Abu Dhabi, UAE, 2025. URL: https://aclanthology.org/2025.mcg-1.0/.
[22] M. E. V.-R. y María Victoria Cantero-Romero y Isabel Cabrera-de-Castro y Luis Alfonso Ureña-López y Arturo Montejo-Ráez y María Teresa Martín-Valdivia, Overview of refutes at iberlef 2024: Automatic generation of counter speech in spanish, Procesamiento del Lenguaje Natural 73 (2024) 449–459. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6630.
[23] H. Bonaldi, G. Damo, N. B. Ocampo, E. Cabrio, S. Villata, M. Guerini, Is safer better? the impact of guardrails on the argumentative strength of LLMs in hate speech countering, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 3446–3463. URL: https://aclanthology.org/2024.emnlp-main.201/. doi:10.18653/v1/2024.emnlp-main.201.
[24] S. Baez Santamaria, H. Gomez Adorno, I. Markov, Contextualized graph representations for generating counter-narratives against hate speech, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2024, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 7664–7674. URL: https://aclanthology.org/2024.findings-emnlp.450/. doi:10.18653/v1/2024.findings-emnlp.450.
[25] S. Gutiérrez, El discurso argumentativo. una propuesta de análisis (????) 45–66. URL: https://biblat.unam.mx/es/revista/escritos-revista-del-centro-de-ciencias-del-lenguaje/articulo/el-discurso-argumentativo-una-propuesta-de-analisis, number: 27.
[26] R. Morado, Funciones básicas del discurso argumentativo (????). URL: https://revistas.uam.es/ria/article/view/8195. doi:10.15366/ria2013.6.007, number: 6.

[27] H. W. Simons, Persuasion in Society, 2 ed., Routledge, ???? doi:10.4324/9780203933039.
[28] M. Alshomary, H. Wachsmuth, Conclusion-based counter-argument generation, in: A. Vlachos, I. Augenstein (Eds.), Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 957–967. URL: https://aclanthology.org/2023.eacl-main.67/. doi:10.18653/v1/2023.eacl-main.67.
[29] X. Wang, H. Peng, Y. Guan, K. Zeng, J. Chen, L. Hou, X. Han, Y. Lin, Z. Liu, R. Xie, J. Zhou, J. Li, MAVEN-ARG: Completing the puzzle of all-in-one event understanding dataset with event argument annotation, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 4072–4091. URL: https://aclanthology.org/2024.acl-long.224/. doi:10.18653/v1/2024.acl-long.224.
[30] J. Piskorski, N. Stefanovitch, G. Da San Martino, P. Nakov, SemEval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multi-lingual setup, in: A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, E. Sartori (Eds.), Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 2343–2361. URL: https://aclanthology.org/2023.semeval-1.317/. doi:10.18653/v1/2023.semeval-1.317.
[31] D. Dimitrov, F. Alam, M. Hasanain, A. Hasnat, F. Silvestri, P. Nakov, G. Da San Martino, SemEval-2024 task 4: Multilingual detection of persuasion techniques in memes, in: A. K. Ojha, A. S. Doğruöz, H. Tayyar Madabushi, G. Da San Martino, S. Rosenthal, A. Rosá (Eds.), Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 2009–2026. URL: https://aclanthology.org/2024.semeval-1.275/. doi:10.18653/v1/2024.semeval-1.275.
[32] D. A. Furman, P. Torres, J. A. Rodríguez, L. Alonso Alemany, D. Letzen, V. Martínez, Which argumentative aspects of hate speech in social media can be reliably identified?, in: J. Bonn, N. Xue (Eds.), Proceedings of the Fourth International Workshop on Designing Meaning Representations, Association for Computational Linguistics, Nancy, France, 2023, pp. 136–153. URL: https://aclanthology.org/2023.dmr-1.13/.
[33] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. tau Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021. URL: https://arxiv.org/abs/2005.11401. arXiv:2005.11401.
[34] W. Wang, V. W. Zheng, H. Yu, C. Miao, A survey of zero-shot learning: Settings, methods, and applications, ACM Trans. Intell. Syst. Technol. 10 (2019). URL: https://doi.org/10.1145/3293318. doi:10.1145/3293318.
[35] T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer, QLoRA: Efficient Finetuning of Quantized LLMs, 2023. URL: http://arxiv.org/abs/2305.14314. doi:10.48550/arXiv.2305.14314, arXiv:2305.14314 [cs].
[36] K. Lv, Y. Yang, T. Liu, Q. Gao, Q. Guo, X. Qiu, Full parameter fine-tuning for large language models with limited resources, 2024. arXiv:2306.09782.
[37] D. Dementieva, D. Moskovskiy, N. Babakov, A. A. Ayele, N. Rizwan, F. Schneider, X. Wang, S. M. Yimam, D. Ustalov, E. Stakovskii, A. Smirnova, A. Elnagar, A. Mukherjee, A. Panchenko, Overview of the multilingual text detoxification task at pan 2024, in: CEUR Workshop Proceedings, CEUR-WS.org, 2024. URL: https://pan.webis.de/clef24/pan24-web/text-detoxification.html.
[38] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, D. Zhou, Self-consistency improves chain of thought reasoning in language models, 2023. URL: https://arxiv.org/abs/2203.11171. arXiv:2203.11171.
[39] D. Korenčić, B. Chulvi, X. B. Casals, M. Taulé, P. Rosso, F. Rangel, Overview of the oppositional thinking analysis pan task at clef 2024, in: G. Faggioli, N. Ferro, P. Galuvakova, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, 2024. URL: https://pan.webis.de/clef24/pan24-web/oppositional-thinking-analysis.html.