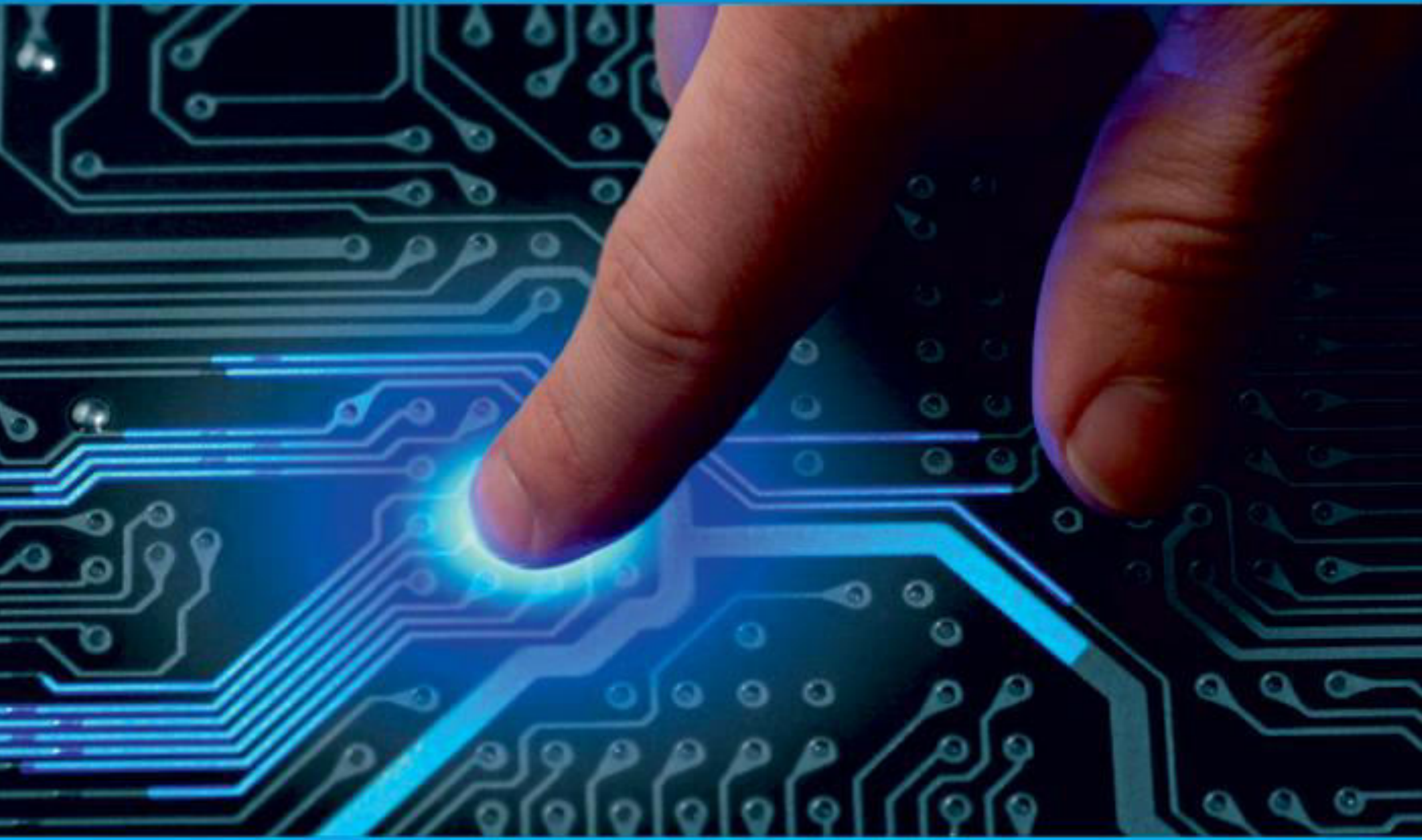
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 12, December 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.625

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Multilingual Counter Speech Generation

Avyukth Potnuru[1], Ayush Samuel Ajith[2], Naheel N Akhtar[3], Hasan Raza B A[4], Likhith S R[5]

Student, Department of Computer Science and Engineering, Presidency University, Bengaluru, India[1-4]

Assistant Professor, Department of Computer Science and Engineering, Presidency University, Bengaluru, India[5]

ABSTRACT*:* In this project we aim to focus on one such solution - Multilingual counter-speech generation using Generative AI, which allows for state-of-the-art natural language processing techniques to successfully detect hate speech, analyze its sentiment and toxicity and generate culturally relevant, informative, and appropriate counter narratives that can curb the mindset of people. The system is designed to be able to process user input, detect the language of the hate speech to ensure global coverage, and classify it into target groups such as religion, caste, color, or gender. Techniques such as TF-IDF vectorization and cosine similarity allow us to identify relevant examples from curated datasets to inform the analysis. The sentiment and toxicity scores allow for the system to process the severity and intention behind the inputted hate speech. To ease the user's interaction with the model, we have included a user interface created using the Gradio library. Through surveying and tone analysis, the project emphasizes cultural sensitivity, linguistic accuracy, and the ethical implications of AI-driven counter-speech.

KEYWORDS*:* Natural Language Processing; Counter-narrative; Toxicity Analysis; Prompt Engineering; Surveying

I. INTRODUCTION

Hate speech refers to any form of communication that gives rise to hatred, violence, or discrimination against individuals or groups based on characteristics such as race, ethnicity, nationality, religion, gender, sexual orientation, or disability etc. The evolution of the internet and social media has drastically transformed the usage of hate speech, enabling it to spread more rapidly and widely than before. Addressing online hate speech is a complex challenge that requires ongoing research, dialogue, and collaborative efforts among stakeholders to effectively tackle its root causes and mitigate its detrimental effects on society.

One of the significant factors contributing to the rise of online hate speech is the safety net provided by social media and other online forums. This often gives confidence to individuals to express hateful sentiments they might otherwise refrain from sharing in public. In response to the growing concern over online hate speech, various organizations, civil society groups, and social media platforms have initiated campaigns to educate users, implement reporting mechanisms, and enforce policy changes.

II. LITERATURE REVIEW

In [1], the study evaluates GPT-2, DialoGPT, FlanT5, and ChatGPT for counterspeech generation in zero-shot settings using datasets like CONAN and Gab. ChatGPT excels in quality metrics, but toxicity rises with model size. Manual prompts often enhance type-specific counterspeech. The study highlights LLMs' potential and the need for better prompting and ethical safeguards. [2] evaluates GPT-3 and GPT-4 for generating counternarratives (CNs) to counter hate speech (HS) in Spanish, using an adapted version of the CONAN Multitarget corpus. Results show that GPT models often outperform human-generated CNs, demonstrating their effectiveness for HS mitigation and creating a valuable Spanish-language CN resource. Zhu and Bhat proposed in [3], a pipeline combining generative modeling, grammaticality filtering, and relevance-based selection to improve diversity and contextual relevance in counterspeech generation. Their approach outperforms traditional models on benchmark datasets, highlighting the importance of modular strategies for effective counterspeech. The comparative study in [4] examines pre-trained language models (e.g., GPT-2, BART) for CN generation, finding that autoregressive models with stochastic decoding produce the most relevant and diverse outputs. It also highlights the importance of target similarity and proposes automatic post-editing to refine CN quality. In [5], the researchers explored automatic counter narrative generation to combat hate speech in Spanish using large language models. Their system combined Mistral-Instruct, Zephyr, and Command-R models with JudgeLM for evaluation. Their findings showed that fine-tuned models outperformed zero-shot approaches, though they noted challenges in ensuring the truthfulness of generated responses despite strong performance on other metrics. In [6], the authors evaluated three LLM approaches for counterspeech generation: fine-tuned GPT-2, zero-shot GPT-3, and ChatGPT. Through human evaluation of 1,740 tweet-response pairs, they found that while all models could generate relevant counterspeech, ChatGPT and GPT-3 performed most consistently, with ChatGPT being most preferred by users (40.9%). The study revealed that response quality, rather than perceived effectiveness, drove user preferences. In [7], the authors developed COUNTERGEDI, a system that generates controlled counterspeech by guiding DialoGPT using generative discriminators (GEDI). The approach enables control over politeness, toxicity, and emotional content, showing significant improvements in attribute scores (15% for politeness, 6% for detoxification) while maintaining output relevance across three datasets.

III. PROPOSED METHODOLOGY

A multilingual dataset containing hate speech and corresponding target groups, such as women, migrants, and people of color (POC), was curated, cleaned, and loaded to predict and generate counter-narratives. The dataset includes hate speech (HS) in English, Spanish, Italian, and Basque, along with contextual background information. Text preprocessing involved cleaning the data by removing punctuation, special characters, and extra whitespaces, followed by lowercasing to ensure case-insensitive vectorization. A language detection tool, using the polyglot library, identified the language of hate speech inputs, mapping them to the four supported languages, with unsupported languages defaulting to English.

Hate speech labeling and classification involved transforming input text and dataset examples into vectorized forms using TF-IDF vectorization, allowing for comparison through cosine similarity. Cosine similarity scores were calculated to identify the most relevant examples, with the top 5 rows selected based on this similarity. A threshold of 0.7 was applied to filter strong matches; if no strong matches were found, the most frequent target category from the top 5 rows was selected. The identified hate speech was then assigned a target category, such as JEWS, POC, or LGBT+.

Toxicity scoring involved sentiment analysis using the Hugging Face sentiment-analysis pipeline to evaluate the sentiment of hate speech inputs and derive toxicity values. A custom metric mapped negative sentiment scores directly to toxicity, while positive and neutral sentiment scores were inverted to reflect lower toxicity values, helping gauge the intent behind the hate speech. Scores close to 1 indicated high toxicity and strongly negative sentiment, values near 0.5 suggested neutrality, and scores approaching 0 represented positive sentiment.

Counter-narrative generation utilized carefully designed prompts with OpenAI's GPT-3.5-turbo to create responses in English, Spanish, Italian, and Basque. Two modes—one-shot and few-shot—were tested to guide counter-narrative generation. The one-shot mode used a single example as context, while the few-shot mode employed multiple examples across the four languages to enhance output quality. As few-shot prompting provided far more reliable and consistent counter-narratives, the project implemented this mode. Multilingual capability ensured counter-narratives were generated in the detected input language, with flexibility to extend support for additional languages based on future use cases.

Integration and user interaction were facilitated through a Gradio-based interface, enabling seamless user engagement with the system. Users could input hate speech text, and the interface displayed the generated counter-narrative along with the calculated toxicity score. Public accessibility was ensured by deploying the Gradio interface with URL sharing, allowing remote access for testing and demonstrations. Evaluation and validation focused on assessing model performance, ensuring linguistic accuracy and cultural sensitivity in generated counter-narratives. Toxicity scores were validated against human judgments to evaluate the sentiment analysis pipeline's effectiveness. User feedback was also gathered via surveying to measure the quality of counter-narratives and iteratively enhance the system.

IV. RESULTS

The survey results indicate distinct preferences in the ranking of responses for different target groups subjected to hate speech. The noticeable trend from the data suggested that from each individual hate speech's counter-narrative selection, the "Combative" option was favored over "Informative" and "Sarcastic", with its highest percentage being

45.1% for target group: WOMEN. This suggests a preference for strong, assertive counter-speech in majority of the target groups.

However, for an overall evaluation, the consensus was that "Informative" is the preferred tone for counter-narratives, boasting 57% of the votes. "Combative" was held 28% of the votes, and "Sarcastic" was least favored with a measly 15% of the votes.

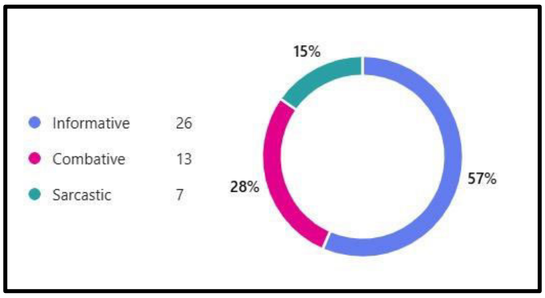


Fig 4.1: Overall preference of tone

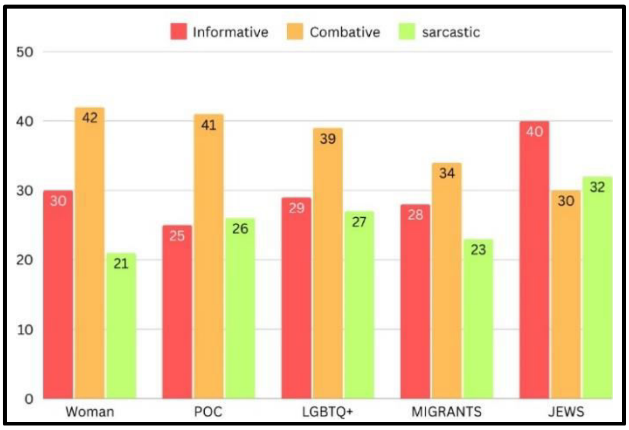


Fig 4.2: Distribution of preferred tone across target groups

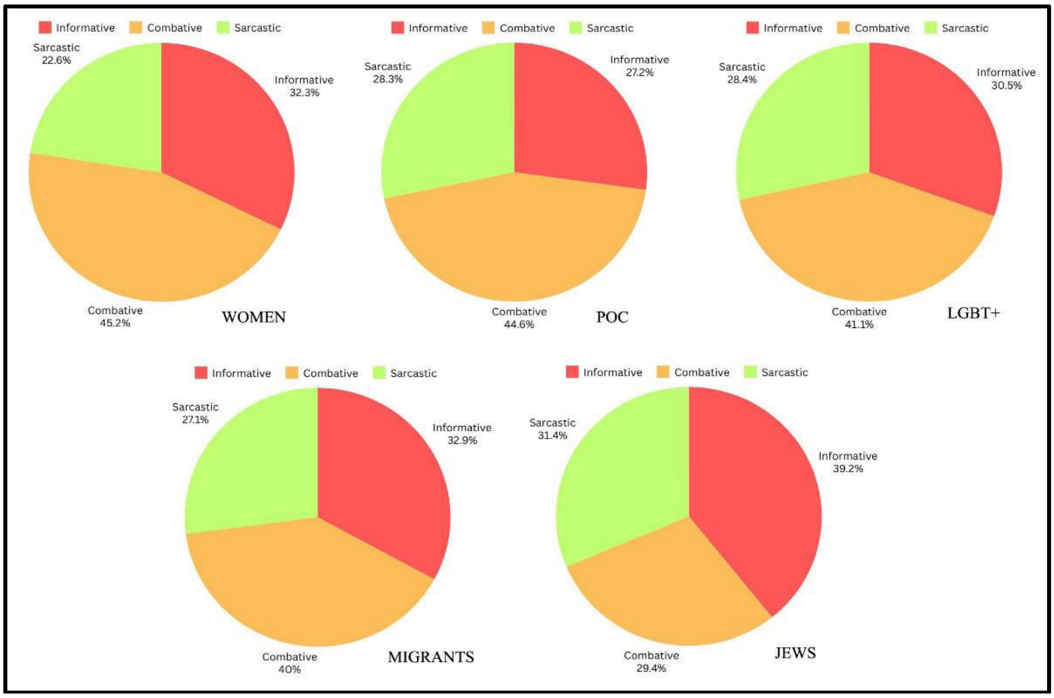


Fig 4.3: Percentage distribution of preferred tones

V. CONCLUSION AND FUTURE WORK

This project successfully addresses the need for multilingual counter-speech generation by leveraging advanced AI models, sentiment analysis, and dynamic language mapping techniques. The integration of tools like Hugging Face's pipelines, TF-IDF vectorization, and OpenAI's GPT-3.5-turbo ensures a robust and adaptive framework capable of analyzing and responding to hate speech across various languages and cultural contexts. The system's ability to classify hate speech targets and assess toxicity levels ensures that the counter-narratives are not only relevant but also tailored to the severity and specific audience, enhancing the overall impact and effectiveness of counter-speech in mitigating online hate.

Currently the project is created to focus on 4 languages - English, Spanish, Italian and Basque, but can be extended to include many other languages. Furthermore, other LLM models such as Claude 3.5 Sonnet, PolyLM, mt5, PaLM2, etc., can be implemented to analysis the variations in counter-narratives generated. The project scope can also be highly focused on the intention behind the hate speech by using IntentCONANv2 dataset.

REFERENCES

[1] Saha, Punyajoy, Aalok Agrawal, Abhik Jana, Chris Biemann, and Animesh Mukherjee. "On Zero-Shot Counterspeech Generation by LLMs." arXiv preprint arXiv:2403.14938 (2024).
[2] Rodríguez, María Estrella Vallecillo, Maria Victoria Cantero Romero, Isabel Cabrera De Castro, Arturo Montejo Ráez, and María Teresa Martín Valdivia. "CONAN-MT-SP: A Spanish Corpus for Counternarrative Using GPT Models." In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pp. 3677-3688. 2024.
[3] Zhu, Wanzheng, and Suma Bhat. "Generate, prune, select: A pipeline for counterspeech generation against online hate speech." arXiv preprint arXiv:2106.01625 (2021).
[4] Tekiroglu, Serra Sinem, Helena Bonaldi, Margherita Fanton, and Marco Guerini. "Using pre-trained language models for producing counter narratives against hate speech: a comparative study." arXiv preprint arXiv:2204.01440 (2022).
[5] Zubiaga, Irune, Aitor Soroa, and Rodrigo Agerri. "Ixa at refutes 2024: Leveraging language models for counter narrative generation." In IberLEF (Working Notes). CEUR Workshop Proceedings. 2024.
[6] Zheng, Yi, Björn Ross, and Walid Magdy. "What makes good counterspeech? a comparison of generation approaches and evaluation metrics." In Proceedings of the 1st Workshop on CounterSpeech for Online Abuse (CS4OA), pp. 62-71. 2023.
[7] Saha, Punyajoy, Kanishk Singh, Adarsh Kumar, Binny Mathew, and Animesh Mukherjee. "CounterGeDi: A controllable approach to generate polite, detoxified and emotional counterspeech." arXiv preprint arXiv:2205.04304 (2022).

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details