Enhancing Hate Speech Detection on Social Media: Pragmatic-Level Information Integration Using Prompting Techniques

Muhammad Ahmed
22311047

Masters Computer Science
Erasmus

Supervisor: Anaïs Ollagnier

Host Laboratory: I3S Templiers Bat E

Report Recipients: Anne-Laure Simonelli

Disciplinary Jury: Aline Menin

Project Title: Improving Hate Speech Detection Using Pragmatic-Level Information

N° Project: 524

Project Type: Tutorship

Research Project Period: March 2024 – May 2024

Abstract:
The widespread proliferation of hate speech online underscores the need for effective detection methods to ensure online safety and inclusivity. Conventional models often miss the nuanced contextual and intentional aspects of language. This study proposes an innovative approach integrating pragmatic information into hate speech detection using advanced Large Language Models (LLMs) such as BERT and generative artificial intelligence (GenAI) models like Mistral 7B, Llama 2 7B, and GPT-3. By enriching these models with contextual layers, they gain a deeper understanding of language subtleties. Through fine-tuning and prompt engineering, the models show superior performance compared to standard methods across various metrics and domains. They also exhibit robust generalization capabilities on unseen data. This research not only advances Natural Language Processing (NLP) in hate speech detection but also enhances practical efforts in fostering safer digital environments. It emphasizes the significance of incorporating pragmatic elements in automated content moderation for a more nuanced and effective approach against online hate speech.

Keywords: hate speech, pragmatic level information, prompting

1. Introduction:

The digital age has significantly altered how we communicate, especially with the rapid expansion of social media, which has become deeply embedded in our daily interactions. This surge in digital communication has brought numerous benefits, such as enhanced interaction and learning opportunities. However, it has also ushered in substantial challenges, including a disturbing increase in online hate speech and cyberbullying, particularly among adolescents who spend a significant amount of time on various social networking platforms. The negative impact of such online behaviour extends beyond the digital realm and can manifest in real-life violence and discrimination against individuals and communities.

In the field of NLP, there has been an increased focus on methods to detect and mitigate the effects of online hate. Traditional models, while effective to an extent, often fall short due to their inability to capture the nuanced and multifaceted nature of hate speech, which varies significantly across social media platforms and communication contexts. To address these limitations, recent advancements have led to the development of innovative AI techniques such as transformer models and generative AI. Researchers, including (Mayer et al., 2023), have been working to mitigate these risks using transformer models and innovative AI prompting techniques. This focus on detecting and managing online hate speech aims to create a safer online environment for all users.

Our research delves into these complexities by integrating pragmatic analysis with advanced NLP techniques, striving to provide a more comprehensive and context-aware classification system. Hate speech, characterized by its multifaceted nature and expressions of aggression based on race, religion, gender, and other factors, often employs aggressive or subtly insidious language (Ollagnier et al., 2022). Building upon the existing body of knowledge, this study incorporates nuanced linguistic insights, making a threefold **contribution**: (1) addressing the task of context-awareness in online hate detection; (2) exploitation of pragmatic-level information to enrich text representation and linguistic analysis; and (3) in-depth analysis of the impact of injecting topical-awareness.

2. Literature Review

Hate speech detection has evolved significantly, primarily focusing on explicit content. However, detecting nuanced hate speech, often embedded in sarcasm, cultural references, and indirect speech, remains a challenge. (Jahan 2021) emphasizes the intricacies of language that conventional models often overlook. (Liu et al., 2021) introduces the emerging paradigm of prompt-based learning in NLP, offering a promising avenue for nuanced language understanding. Complementing these insights, our research explores the integration of pragmatic information to enhance hate speech detection. We delve into studies like (Toraman et al., 2022), who emphasize large-scale dataset construction for model training, and (Brown et al. 2020), showcasing the potential of transformer models in complex language tasks. This comprehensive review lays the groundwork for our approach, aiming to develop a more contextually aware and sophisticated system for hate speech detection.

2. Related Work

This section delineates the evolutionary trajectory and present status of methodologies and datasets, spotlighting the advancements made and hurdles confronted in the precise identification and analysis of online hate speech.

2.1 Methods for Hate Speech Detection

Early efforts in hate speech detection primarily relied on lexicon-based approaches, which involved identifying predefined hate keywords. However, this method had inherent limitations as it couldn't adapt to the evolving and contextual nature of hate speech language. To overcome this, researchers delved into supervised learning techniques. These methods extracted various features, including linguistic patterns, syntactic structures, n-grams, and part-of-speech tags, from training datasets to enhance detection accuracy (Nobata et al., 2016; Davidson et al., 2017). Semantic understanding saw a significant leap with the integration of word embeddings like GloVe, offering a more nuanced grasp of word meanings and associations (Pennington et al., 2014).

The landscape shifted dramatically with the introduction of deep neural networks, notably Convolutional Neural Networks (CNNs) and Long Short-Term Memory networks (LSTMs). These models could directly learn complex feature representations from data, eliminating the need for manual feature engineering (Kim, 2014). More recently, Transformer architectures, particularly exemplified by models like BERT, have taken center stage in hate speech detection. Their ability to process text with deep contextual understanding has propelled them beyond previous methodologies (Devlin et al., 2019).

The strides made by the NLP community have been remarkable, with the development of semantic frameworks aimed at tackling the intricate and multifaceted nature of hate speech. These frameworks address various aspects, including specific targets (such as targeted groups), nuances (abusive, toxic, dangerous, offensive or aggressive language), and rhetorical devices (slurs, obscenity, offences or sarcasm), among others. The integration of pragmatic level information entails crafting algorithms capable of grasping and interpreting the complex facets of language. This process extends beyond mere literal interpretation, delving into context, connotations, tone, and the underlying intentions of words (Sperber & Wilson, 1986). Such elements are vital for algorithms to accurately unravel the subtleties in human communication.

2.2 Resources for Hate Speech Detection

Multiple surveys have offered comprehensive overviews of existing datasets (Vidgen and Derczynski, 2020; Poletto et al., 2021). The majority of these datasets originate from Twitter and primarily utilize a binary system to classify the presence or absence of hate speech, as initially proposed by (AbdelHamid et al. 2022). However, certain datasets have opted for more intricate annotation schemes, accounting for various manifestations of hate speech, such as hate speech alongside offensive language (Martins et al., 2018; Mathur et al., 2018), or differentiating between sexism and aggressive language (Bhattacharya et al., 2020). The The NLP community has also emphasized the importance of biases and threats in Meitei, Bangla and Hindi (Kumar et al., 2022). In addition to these developments, numerous open shared tasks at NLP-related conferences have further advanced the field.

While the majority of research has concentrated on analyzing specific phenomena and their computational treatment, little emphasis has been placed on examining the pragmatic and syntactic structures inherent in cyberbullying scenarios. An initial attempt to address this gap was made in a previous study (Kumar et al., 2018), which discussed labels delineating discursive roles and impacts. This endeavor was further developed in (Kumar et al. 2022), where the framework was refined and expanded to better encompass the diverse communication strategies employed in such contexts. Expanding upon earlier investigations (Kumar et al., 2018, 2022), my research employs the English Twitter dataset by (Toraman et al. 2022) to explore how pragmatic-level insights can enhance detection accuracy, introducing novel computational perspectives on online hate.

4. Dataset for hate speech detection

As our benchmark dataset, I utilize the standard hate speech Twitter dataset (Toraman et al., 2022), comprising 68,590 English tweets from the years 2020 and 2021. This dataset exclusively consists of original tweets, excluding retweets, replies, or quotes. The tweet contents undergo prefiltering using a curated keyword list determined by the dataset curators, encompassing hashtags and keywords from five distinct topics (referred to as hate domains): religion, gender, racism, politics, and sports. Each tweet is tagged with a single topic. A selection of keywords from the comprehensive list, along with their respective domains, is outlined in Table 1.

The tweets are preclassified as hate speech, offensive, or normal. A text cleaning procedure is applied to standardize the textual content, eliminating extraneous elements like URLs, mentions, and hashtags, while also normalizing case and whitespace. Subsequently, the dataset undergoes meticulous filtering, removing rows with null values and isolating entries. Irrelevant columns are simultaneously removed to streamline the dataset. To achieve consistent sample sizes and address class imbalance, we opted to select 1,615 rows for each label. This choice was informed by the observation that among the labels, "Hate" had the smallest percentage contribution (2%) in our dataset of 68,590 rows as highlighted in Table 2, and 1,615 rows represented the highest count available within this category. This approach ensures uniformity in sample sizes across all labels and effectively mitigates class imbalance.

Domain	**Keywords**
Religion	Christianity, Islam, Judaism, Hinduism, Atheist, belief, church, mosque, Jewish, Muslim, Bible
Gender	LGBTQ, bisexual, female, male, homophobia, gay, lesbian, homosexual, bisexual, transgender
Race	foreigner, refugee, immigrant, Syrian, African, Turk, American, Iranian, Russian, Arab, Greek
Politics	democratic party, republican party, government, white house, president, Trump, Biden, minister
Sports	football, baseball, volleyball, referee, barcelona, real madrid, chelsea, new york knicks, coach

Table 1: Samples from our keyword list.

Domain - label	**Normal - 0**	**Offensive - 1**	**Hate – 2**	**Total**
Religion - 0	10,712	2,369	328	**13,409**
Gender - 1	9,536	3,043	255	**12,834**
Race - 2	12,565	1,631	405	**14,601**
Politics - 3	9,991	2,972	343	**13,306**
Sports - 4	11,340	2,814	286	**14,440**
Total	**54,144 (79%)**	**12,829 (19%)**	**1,617 (2%)**	**68,590**

Table 2: Distribution of tweets in my total dataset.

5. Proposed Framework

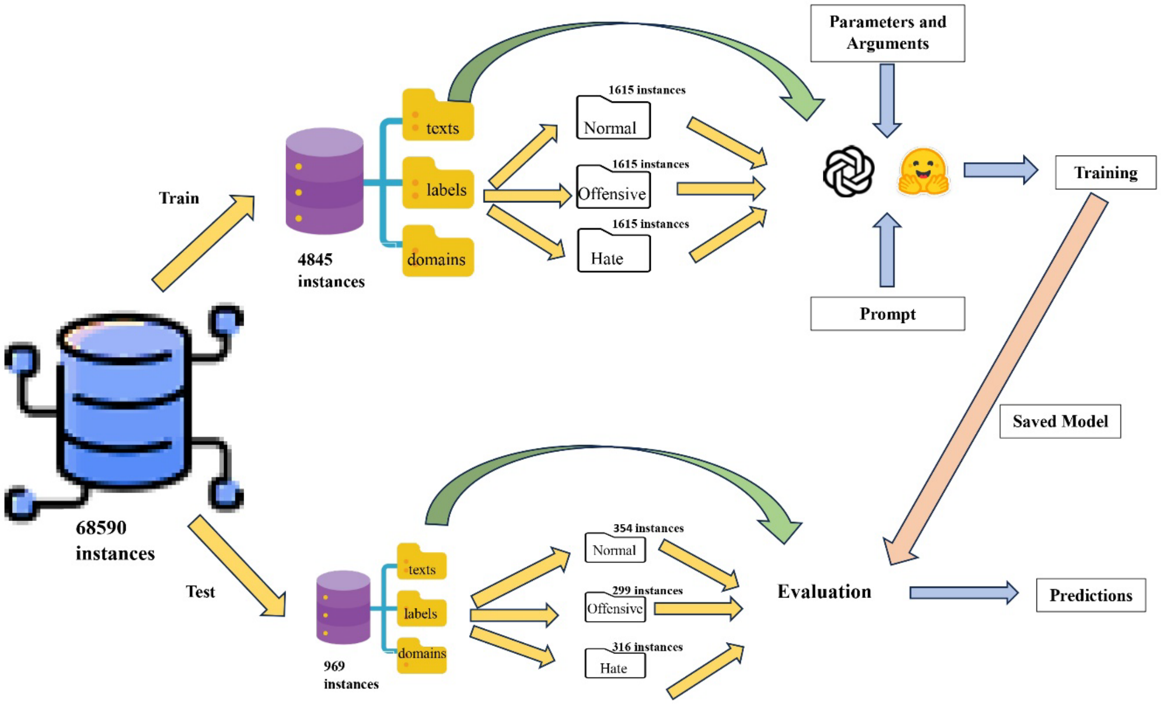


Fig 1: Workflow for multi-class classification

Initially, our training data, consisting of text and labels, is fed separately into four different transformer-based models: BERT, GPT3, Mistral 7B, and Llama 2 7B. Each model processes the text according to its specific prompt, arguments, and parameters, producing a classification. The best-trained model from each is saved and subsequently used to evaluate and predict on the test data, as shown in Figure 1.

Our project incorporated the zero-shot which includes just instructions with no description and examples and manual prompt template engineering technique, a tuning-free prompting approach within the broader scope of hard prompting (Yang et al., 2022). This technique involves meticulously designing prompts to guide the pre-trained model's response towards specific classification objectives. In this method, prompts are deliberately designed to include specific instructions, definitions, and examples that align with the task's objectives. It can also be called Few-shot prompting (Brown et al, 2020) which incorporates pragmatic level information. For instance, in classifying tweets, the prompts in my project provided clear guidelines on what constitutes hate speech, offensive language, and normal speech Table 3. This structured approach directs the model to analyze and categorize text based on these predefined criteria, harnessing its natural language understanding and inference abilities.

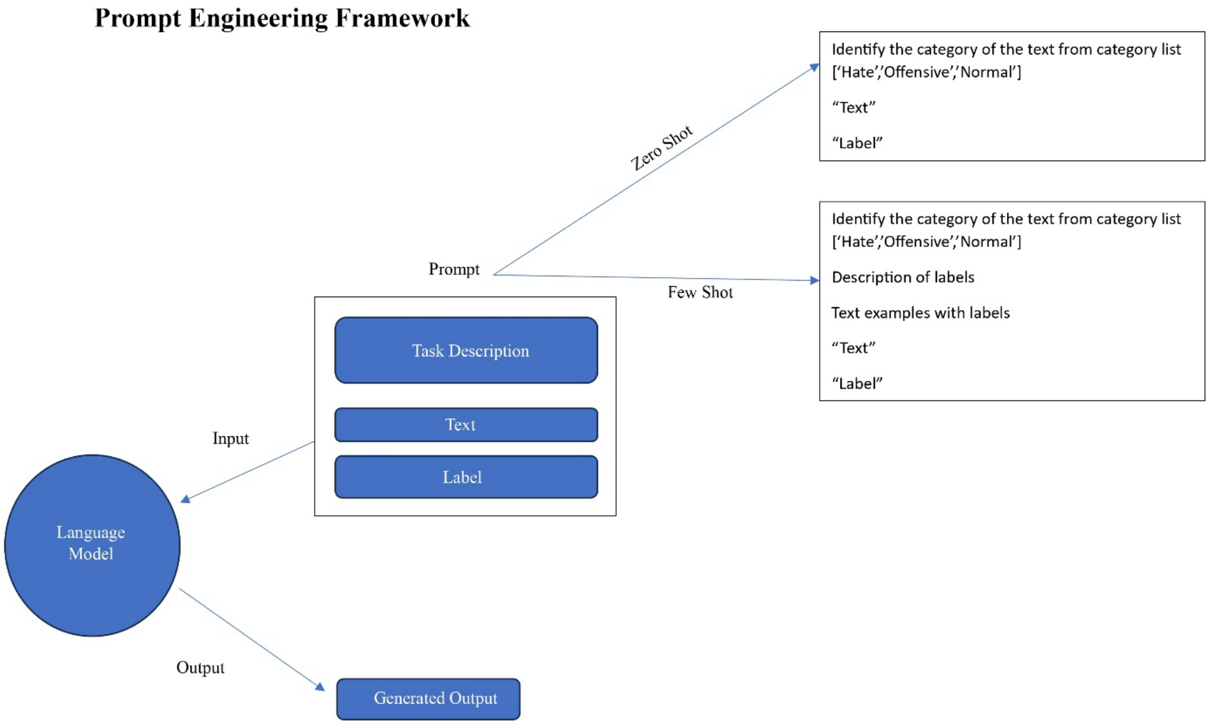

Prompt Engineering Framework
Identify the category of the text from category list ['Hate','Offensive','Normal']
"Text"
"Label"
Zero Shot
Identify the category of the text from category list ['Hate','Offensive','Normal']
Description of labels
Text examples with labels
"Text"
"Label"
Prompt
Few Shot
Task Description
Text
Label
Input
Language Model
Output
Generated Output

Fig2: Prompting Framework

In our approach, the introduction of label description and examples enhances distinction in cyberbullying practices by utilizing advanced manual prompt template engineering, as exemplified in Table 3 (Torman 2020) and Table 4 respectively. This step toward a more nuanced and context-sensitive classification circumvents the need for additional training or model alterations, thereby streamlining the process and ensuring precise and relevant outcomes.

Description of Labels for Few-shot Prompting
Definitions of the labels in the categories_list are as follows:
- Hate Speech: Targets, incites violence against, threatens, or calls for physical damage to an individual or group because of their identifying traits or characteristics.
- Offensive: Humiliates, taunts, discriminates, or insults an individual or group in any form, including text.
- Normal: Tweets that do not fall into the above categories.

Table 3: Labels are described to help the model better understand them in few-shot settings.

Examples for Few-shot Prompting
Consider the following examples to guide your analysis:
1. 'I can't live in a world where gay marriage is legal.' Okay, so die. - Label: Hate
2. 'Good night to the Turks, death to the Americans.' - Label: Hate
3. 'Self-proclaimed atheist doesn't make you a cool kid, bitch.' - Label: Offensive
4. 'After that, the referee, who does not count and apply the 6-second rule, will be dishonest.' - Label: Offensive
5. 'Biden, you're a liar and a cheat and an old idiot.' - Label: Offensive

Table 4: Examples are provided which integrate pragmatic-level information to achieve more nuanced and context-aware classification.

By carefully crafting the structure and content of the prompts, we were able to leverage LLMs extensive language understanding without any modifications to its underlying algorithms. This approach is particularly effective in text classification tasks like distinguishing between hate speech, offensive language, and normal speech, as it capitalizes on the model's inherent capabilities while avoiding the complexities and resource demands of model tuning.

6. Model Integration

Our study integrates advanced LLMs such as Mistral 7B, Llama 2 7B, GPT-3, and BERT to enhance the detection of hate speech on social media platforms. We utilize complex prompts in both zero-shot and few-shot settings, with an emphasis on incorporating pragmatic context. Our methodology focuses on identifying instances of cyberbullying. By adhering to standardized experimental protocols and utilizing high-performance computing resources, we ensure reliable classification results.

6.1 GPT-3

Description: The GPT-3, utilizing the acclaimed Davinci variant of OpenAI's GPT-3 model, is renowned for its exceptional language comprehension and generation skills (Brown et al., 2020). In our framework, this classifier serves a pivotal role by leveraging its refined linguistic analysis capabilities to differentiate between hate speech, offensive language, and normal discourse. This precision is essential for accurately categorizing online content (Bender et al., 2021; Radford et al., 2019).

Configuration: We opted for the text-davinci-003 engine due to its exceptional capabilities. Our approach leverages manual prompt template engineering as a tuning-free technique, operating within the context of hard prompting in both zero-shot and few-shot settings. This specific prompt structure guides the classification into three distinct categories: Hate, Offensive, and Normal.

Role in Research: GPT-3 serves as a flexible and dynamic model, capable of understanding complex language structures and contexts.

6.2 Mistral 7B

Description: The Mistral 7B model, introduced by (Jiang et al. 2023), with its advanced architecture, exhibits a profound grasp of language, making it exceptionally effective for linguistic analysis tasks within our framework. Its open-source nature enables precise differentiation between hate speech, offensive language, and normal discourse, crucial for accurate online content categorization. Boasting an impressive 7.3 billion parameters, Mistral 7B stands as a State-of-the-Art (SOTA) language model, marking a significant leap in natural language understanding and generation.

Configuration: I selected the Unsloth 4-bit Mistral 7B for its increased speed (2.2x faster) and reduced memory usage. The model's integration with Lora adapters requires updating only 1 to 10% of its parameters. It also utilizes Grouped-Query Attention (GQA) for rapid inference and Sliding Window Attention (SWA) for efficient handling of longer sequences, balancing performance and computational cost. Its flexible design enables fine-tuning for various tasks. The Mistral 7B's robust language comprehension capabilities make it ideal for our framework, allowing for manual prompt template engineering to achieve stable, tuning-free prompting, and facilitating classification into Hate Speech, Offensive Language, and Normal Communication categories.

Role in Research: Mistral 7B functions as a versatile and dynamic classifier, capable of accurately interpreting complex linguistic structures and contexts.

6.2 Llama 2 7B

Description: The Llama 2 7B model, developed by Meta (the parent company of Facebook), is designed to handle various language processing tasks with high accuracy. Its architecture, consisting of 7 billion parameters, enables comprehensive text understanding and generation, facilitating nuanced categorization and analysis of online content. The model is open-sourced, allowing for flexible modifications and integrations, which are essential for diverse applications (Touvron et al., 2023).

Configuration: I selected the Unsloth 4-bit Llama 2 7B for its speed (2x faster than the standard model) and reduced memory usage, along with the same Lora adapters used for Mistral 7B. It features a transformer-based architecture that ensures robust language comprehension. Llama 2 is an auto-regressive language model that employs an optimized transformer architecture. The fine-tuned versions leverage supervised fine-tuning (SFT) and reinforcement learning with human feedback (RLHF) to align with human preferences for helpfulness and safety. This adaptability makes it suitable for a variety of tasks, balancing precision and computational efficiency.

Role in Research: Llama 2 7B is a versatile open-sourced language model, useful for a range of research applications. Its comprehensive language understanding capabilities make it an excellent tool for complex text classification, generation, and analysis tasks, enabling effective content categorization and linguistic research.

6.4 BERT

Description: BERT, a transformer-based model, is renowned for effectively processing the context of words within sentences (Devlin et al., 2019). We utilize its base variant, 'bert-base-uncased', optimized for sequence classification tasks.

Configuration: Configured the model for classifying texts into three categories - Hate Speech, Offensive, and Normal, by fine tuning the classification head.

Role in Research: BERT’s proficiency in contextual understanding positions it as an ideal tool for differentiating complex and nuanced hate speech expressions. It serves as a benchmark for assessing the efficacy of transformer-based models in hate speech detection (Wolf et al., 2020).

7. Experimental Settings

We utilize Generative Artificial Intelligence (GenAI) models to classify tweets into 'Hate Speech', 'Offensive', or 'Normal'. Our approach involves two strategies: zero-shot prompting, where the model classifies tweets based solely on its pre-trained knowledge, and few-shot prompting, which incorporates labeled examples for a deeper contextual analysis. For GPT-3 models, responses are limited to 60 tokens for concise output. For Mistral 7B and Llama 2 7B, maintaining a batch size of 2, learning rate of 2e-4, AdamW (8-bit) optimizer, and weight decay of 0.01.

BERT classifier utilizes the BERT tokenizer to convert input text into appropriate tokens, respecting the model's input length restrictions. The model's output logits are then converted into probabilities using a softmax function, and the class with the highest probability is selected as the prediction. Parameters settings include batch size of 32, learning rate of 5e-5, weight decay 0.01 and optimizer AdamW.

We manage computational demands with GPUs like the GeForce MX130 for BERT and the NVIDIA A100 for other models.

8. Experimental Results

Table 6 and Table 7 report the performance of our models, revealing GPT-3 as the top performer, thanks to its ability to interpret meanings and contextual nuances in text. Mistral 7B shows a performance close to that of Llama 2 7B, while BERT struggles to predict accurately for a particular class as shown in heatmap in Fig 3.

GPT-3 significantly outperforms the others, achieving the highest accuracy with balanced metrics across the board. Mistral 7B shows impressive performance among open-source models, with the best results in a few-shot setting, although falling slightly short of Llama 2 7B in a zero-shot setting. BERT trails behind, particularly in classifying the Offensive class accurately (results in Table 5), indicating a lack of contextual understanding. This comparison underscores GPT-3's superior ability in detecting nuanced hate speech accurately.

Confusion matrices were also employed to evaluate the performance of the classification models, providing insight into not only the overall accuracy but also how well each model performs for individual classes. Fig 4 illustrates the models' performances.

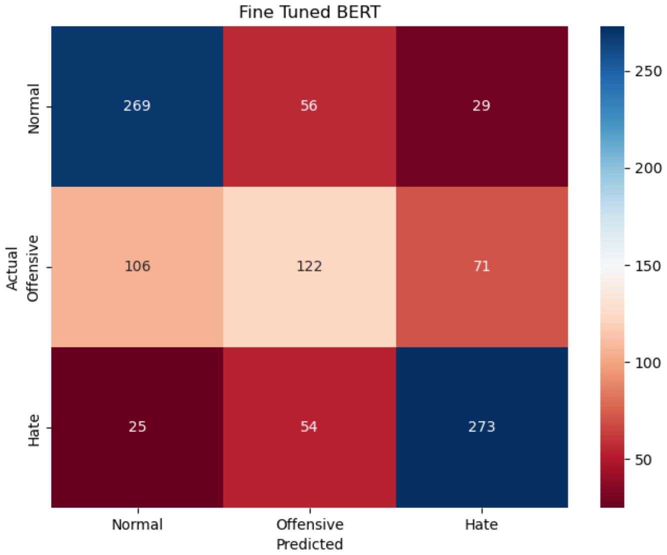


Figure 3: Confusion Matrix of FineTuned BERT

	Precision	Recall	F1-score
Hate	0.7319	0.8639	0.7924
Normal	0.6725	0.7598	0.7135
Offensive	0.6224	0.4080	0.4929
accuracy	0.6852	0.6852	0.6852
macro avg	0.6756	0.6772	0.6663
weighted avg	0.6764	0.6852	0.6711

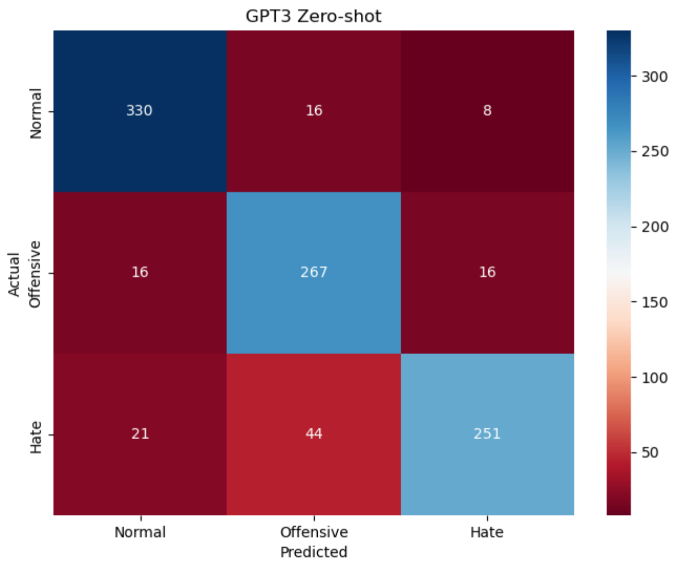
Table 5: Performance of the multi-class BERT model (Hate, Offensive, and Normal) on our dataset.

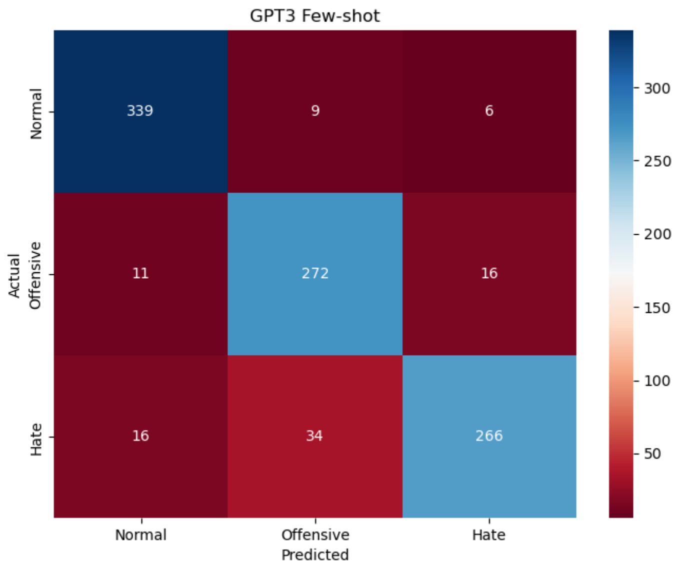
	GPT3	Accuracy	**0.8751**	**Mistral 7B**	Accuracy	0.8204	**Llama 2 7B**	Accuracy	0.8359
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Normal	0.8991	**0.9322**	**0.9154**	**0.8580**	0.8192	**0.8382**	**0.9333**	**0.9096**	**0.9213**
Offensive	0.8165	0.8930	0.8530	0.8439	0.7592	0.7993	0.8208	0.7659	0.7924
Hate	**0.9127**	0.7943	0.8494	0.7680	**0.8800**	0.8201	0.7508	0.8196	0.7837

Table 6: Multi-class performance of GenAI models (Hate, Offensive, and Normal) on our dataset using zero-shot prompting.

	GPT3	Accuracy	**0.9051**	**Mistral 7B**	Accuracy	0.8394	**Llama 2 7B**	Accuracy	0.8080
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Normal	**0.9262**	**0.9576**	**0.9417**	0.8532	**0.8701**	**0.8615**	**0.9399**	**0.8390**	**0.8866**
Offensive	0.8635	0.9097	0.8860	0.7914	0.8651	0.8264	0.7404	0.7726	0.7561
Hate	0.9236	0.8418	0.8808	**0.8759**	0.7816	0.8261	0.7478	0.8070	0.7763

Table 7: Multi-class performance of GenAI models (Hate, Offensive, and Normal) on our dataset using few-shot prompting.





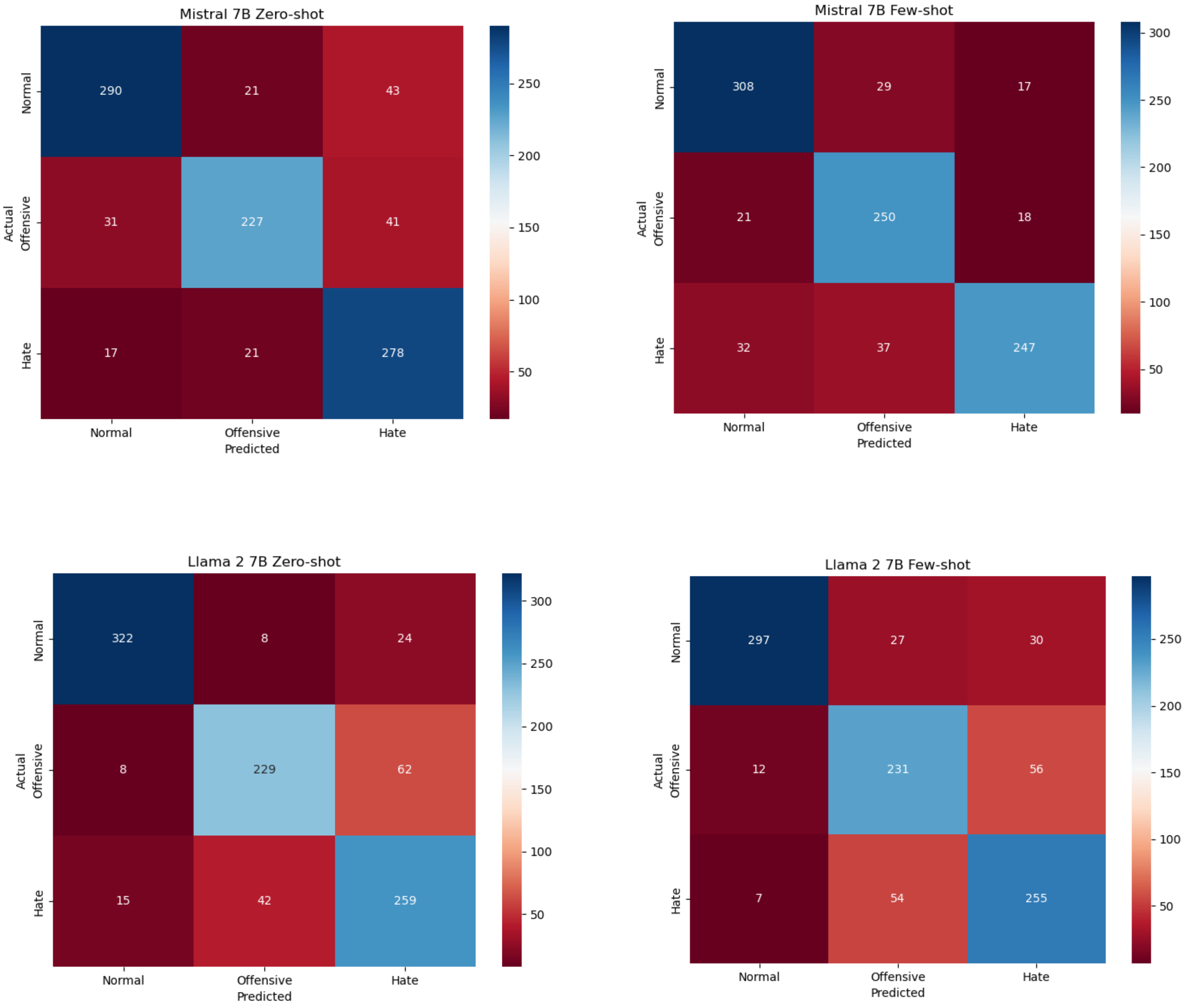


Fig 4: Classification performance of three models: GPT-3, Mistral 7B and Llama 2 7B. Each matrix shows true labels against predicted labels. Diagonal cells represent correct classifications, while off-diagonal cells indicate misclassifications.

BERT performs well for the Normal and Hate classes but struggles significantly with the Offensive class. In a zero-shot setting, Llama 2 7B classifies more instances correctly for the Normal class than Mistral 7B. However, in a few-shot setting, Mistral 7B outperforms Llama 2 7B and BERT, particularly in classifying the Offensive class. GPT-3 achieves the most balanced classification, with significantly fewer misclassifications and a superior ability to categorize tweets correctly. This highlights the efficacy of the model and refined prompt design, incorporating pragmatic information to enhance accuracy and reliability in a few-shot setting.

9. Error Analysis

We provide an error analysis on model predictions in Table 8, comparing the performance of four models—BERT, Llama 2 7B, GPT-3, and Mistral 7B—on five tweets. In the first tweet, labeled 'Normal,' three models (BERT, Llama 2 7B, and GPT-3) correctly identify it, while Mistral 7B misclassifies it as 'Hate,' indicating a potential sensitivity to geopolitical content. In the second tweet, also labeled 'Normal,' BERT, Llama 2 7B, and GPT-3 incorrectly classify it as 'Offensive' or 'Hate,' highlighting their difficulty in distinguishing between personal frustration and negative sentiments, whereas Mistral 7B correctly classifies it. The third tweet, labeled 'Normal,' is accurately classified by all models, demonstrating their ability to discern explicit language in a non-offensive context. In the fourth tweet, labeled 'Hate,' all models correctly identify it, showing a shared ability to recognize hateful content. In the fifth tweet, labeled 'Offensive,' BERT and Llama 2 7B misclassify it as 'Normal,' while GPT-3 and Mistral 7B correctly classify it, demonstrating variability in how each model interprets content involving public figures. This analysis highlights the models' strengths and weaknesses in handling context, identifying negative content, and distinguishing between various forms of sentiment.

#	Label	Tweet	BERT	Llama 2 7B	GPT3	Mistral 7B
1	Normal	US flies two B52 bombers to middle east ahead of possible Iranian attack	**Normal**	**Normal**	**Normal**	Hate
2	Normal	Was on an absolute heater, then I lose 500 dollars on Korean baseball and Costa Rican soccer. I hate my life	Offensive	Hate	Hate	**Normal**
3	Normal	I’m gay but I’ve got the biggest fucking crush on katyperry I have for years!	**Normal**	**Normal**	**Normal**	**Normal**
4	Hate	A gay killing another gay nice	**Hate**	**Hate**	**Hate**	**Hate**
5	Offensive	barack obama would never let the world get this bad trump ain’t shit	Normal	Normal	**Offensive**	**Offensive**

Table 8: Error analysis using model predictions. Correct predictions are given in bold.

10. Conclusion

This study represents advancement in nuanced hate speech detection on social media platforms. By integrating pragmatic information, NLP techniques, and prompt engineering, we achieve superior detection accuracy on different models. Our results show that GenAI models outperform other models, making a substantial contribution to creating safer online spaces.

11. Limitations

While our approach signifies progress, it is not without significant limitations. The model's efficacy relies heavily on the quality and diversity of the dataset, and interpreting the complexities of online communication poses unique challenges. A key limitation is the lack of testing on larger datasets. Additionally, the model's complexity necessitates substantial computational power, and as a result, hyperparameter tuning was not conducted. To address these challenges, future work should focus on enhancing the dataset, optimizing processing efficiency, and utilizing more powerful open-source GenAI models such as Mistral 8B, Gemma 13B, and CodeLlama 34B, which require robust hardware capabilities. Furthermore, there is a need to delve into prompt design, ensuring that future iterations of the model benefit from more refined and well-crafted prompts to improve overall performance.

12. References

1. Mayer, C. W. F., Ludwig, S., & Brandt, S. (2023). Prompt text classifications with transformer models! An exemplary introduction to prompt-based learning with large language models. *Journal of Research on Technology in Education*, *55*(1), 125–141. https://doi.org/10.1080/15391523.2022.2142872
2. Anaïs Ollagnier, Elena Cabrio, Serena Villata, and Catherine Blaya. 2022. CyberAgressionAdo-v1: a Dataset of Annotated Online Aggressions in French Collected through a Role-playing Game. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 867–875, Marseille, France. European Language Resources Association.
3. Jahan, Md Saroar, and Mourad Oussalah. "A systematic review of Hate Speech automatic detection using Natural Language Processing." arXiv preprint arXiv:2106.00742 (2021). https://arxiv.org/abs/2106.00742
4. Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2021). Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. https://doi.org/10.48550/arXiv.2107.13586
5. Cagri Toraman, Furkan Şahinuç, and Eyup Yilmaz. 2022. Large-Scale Hate Speech Detection with Cross-Domain Transfer. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2215–2225, Marseille, France. European Language Resources Association.
6. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language Models are Few-Shot Learners. *arXiv preprint arXiv:2005.14165*. https://doi.org/10.48550/arXiv.2005.14165
7. Nobata, Chikashi, et al. "Abusive Language Detection in Online User Content." Proceedings of the 25th International Conference on World Wide Web, April 2016, pp. 145–153. https://doi.org/10.1145/2872427.2883062.
8. Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated Hate Speech Detection and the Problem of Offensive Language. *Proceedings of the International AAAI Conference on Web and Social Media*, *11*(1), 512-515. https://doi.org/10.1609/icwsm.v11i1.14955
9. Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.
10. Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.
11. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
12. Sanders, R. E. (1988). Dan Sperber and Deirdre Wilson, Relevance: Communication and cognition, Oxford: Basil Blackwell, 1986. Pp. 265. *Language in Society*, *17*(4), 604–609. http://dx.doi.org/10.1017/S004740450001318X
13. Vidgen, B., & Derczynski, L. (2020). Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLoS ONE*, *15*(12), Article e0243300. https://doi.org/10.1371/journal.pone.0243300

14. Poletto, F., Basile, V., Sanguinetti, M. *et al.* Resources and benchmark corpora for hate speech detection: a systematic review. *Lang Resources & Evaluation* **55**, 477–523 (2021). https://doi.org/10.1007/s10579-020-09502-8
15. Abdelhamid, Ibrahim & Yahaya, Hazrati & Ahmad, Zahidah & Zhafri, Muhammad & Nazmi, Muhammad & Ahmad, Nor & Malim, Tanjung & Bahasa, Akademi & Shah, Uitm. (2022). Foreign Language Learning Through Social Media: A Review Study. International Journal of Academic Research in Business and Social Sciences. 12. 1424-1436. http://dx.doi.org/10.6007/IJARBSS/v12-i6/13910
16. Ricardo Martins, Marco Gomes, José João Almeida, Paulo Novais, and Pedro Rangel Henriques. 2018. Hate speech classification in social media using emotional analysis. In 7th Brazilian Conference on Intelligent Systems, BRACIS 2018, São Paulo, Brazil, October 22-25, 2018, pages 61–66. IEEE Computer Society.
17. Puneet Mathur, Rajiv Ratn Shah, Ramit Sawhney, and Debanjan Mahata. 2018. Detecting offensive tweets in hindi-english code-switched language. In Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media, SocialNLP@ACL 2018, Melbourne, Australia, July 20, 2018, pages 18–26. Association for Computational Linguistics.
18. Shiladitya Bhattacharya, Siddharth Singh, Ritesh Kumar, Akanksha Bansal, Akash Bhagat, Yogesh Dawer, Bornini Lahiri, and Atul Kr. Ojha. 2020. Developing a multilingual annotated corpus of misogyny and aggression. In Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, TRAC@LREC 2020, Marseille, France, May 2020, pages 158–168. European Language Resources Association (ELRA).
19. Ritesh Kumar, Shyam Ratan, Siddharth Singh, Enakshi Nandi, Laishram Niranjana Devi, Akash Bhagat, Yogesh Dawer, Bornini Lahiri, Akanksha Bansal, and Atul Kr. Ojha. 2022. The comma dataset V0.2: annotating aggression and bias in multilingual social media discourse. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022, pages 4149–4161. European Language Resources Association.
20. Ritesh Kumar, Aishwarya N. Reganti, Akshit Bhatia, and Tushar Maheshwari. 2018. Aggressionannotated corpus of hindi-english code-mixed data. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018. European Language Resources Association (ELRA).
21. Yang, Z., et al. (2022). Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing.
22. Bender, E. M., et al. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? Proceedings of FAccT.
23. Radford, A., et al. (2019). Language Models are Unsupervised Multitask Learners. OpenAI Blog.
24. Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M. A., Stock, P., Le Scao, T., Lavril, T., Wang, T., Lacroix, T., & El Sayed, W. (2023). Mistral 7B v0.1: A 7-billion-parameter language model for superior performance and efficiency. arXiv:2310.06825. https://doi.org/10.48550/arXiv.2310.06825
25. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Canton Ferrer, C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., ... Scialom, T. (2023). Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288. https://arxiv.org/abs/2307.09288
26. Wolf, T., et al. (2020). Transformers: State-of-the-Art Natural Language Processing. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations.