Identifying Hate Speech Online Using LLM-based Sentiment Analysis

A Scoping Review

Duha Altorky, Sana Ghazal

College of Computing and Information Technology, University of Doha for Science and Technology

Abstract

Online platforms have made it possible to communicate with people around the world, but they have also made hate speech easier to spread, which makes strong detection systems necessary. This scoping review examines developments in online hate speech detection, emphasizing the use of sentiment analysis, large language models (LLMs), and natural language processing (NLP). The ability of hate speech detection systems to comprehend and produce text that resembles that of a person has been improved by LLMs, such as GPT-based models. Innovations that increase the precision and scalability of hate speech identification are highlighted in the review, including machine-generated datasets, chain-of-thought reasoning, and fine-tuning methods. It also examines difficulties with multilingual detection, interpretability, and model generalization across various datasets.

1. Introduction

The rapid growth of social media and online platforms has created a surge in the emergence of hate speech as a significant societal issue. The prevalent spread of hateful content online has significant consequences, since it not only promotes prejudice and discrimination, but also incites violence and division in the real world. In order to promote inclusive communication, make digital environments safer, and lessen the negative effects of online hate speech on society, it is imperative that such content is mitigated and moderated.

1.1 Natural Language Processing & Large Language Models:

Amongst the fastest growing and most prevalent fields of Artificial Intelligence nowadays is Natural Language Processing (NLP), which is essentially an intersection of computer science, AI and linguistics. [1] NLP "employs computational techniques for the purpose of learning, understanding, and producing human language content" [2], and evidently, Large Language Models (LLMs) pose as a significant advancement in NLP. LLMs are computational models with the ability to generate and understand human-like text, and perhaps the most widely used and well-known example of LLMs are the Generative Pre-trained Transformers (GPTs), which were first introduced in 2018 by OpenAI [3] and are now the basis for ChatGPT, arguably the biggest breakthrough in practical applications of NLP and LLMs yet.

1.2 Sentiment Analysis

Sentiment Analysis is the computational study of people's opinions, sentiments, emotions, and attitudes towards entities such as products, services, issues, events, topics, and their attributes expressed in written text [4]. It involves using natural language processing (NLP), text analysis, computational linguistics and statistics to identify and classify the sentiment of the text as positive, negative, or neutral [5]

1.3 Online Hate speech

The United Nations defines hate speech as "any kind of communication in speech, writing or behavior, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, color, descent, gender or other identity factor" [6]. With the exponential growth of social media, the existence of hate speech in online spaces is inevitable, and detecting and filtering hateful content in those spaces is urgent now more than ever, especially when the presence of such content influences behavior and conflicts in the real world. Thankfully, the growth of social media is accompanied by the growth of technology that helps with regulating social media and detecting said hateful content in those spaces. Sentiment Analysis has revolutionized the detection of hate speech online and is helping take the toll of monitoring harmful hateful content off humans.

2. Methodology

For this scoping review, we conducted comprehensive systematic research across multiple databases and platforms including Google Scholar, PubMed, ArXiv, IEEE Xplore, Springer and PubMed. We used a set of keywords for our search of papers, the keywords include combinations of "Hate speech detection", "Identifying hate speech", "Hate speech online", "Hate speech sentiment analysis", "Large language models sentiment analysis", "LLM sentiment analysis", "Large language model hate speech" and "Large language model hate speech online". Our initial research process highlighted a limited number of studies in our target domain, this can be attributed to the fast evolving and fresh nature of LLMS and their applications in sentiment analysis. These search conditions resulted in an initial collection of 33 papers collected with Zotero.

We established inclusion and exclusion criteria to further ensure that the papers collected are relevant to our review domain.

Table 1: Inclusion and exclusion criteria

Inclusion Criteria	Exclusion Criteria
Papers that focus on the use of large language models	Papers that discuss the use of sentiment analysis but not using LLMs
Papers that focus on the use of large language models for sentiment analysis	Papers that don't mention identifying hate speech online
Papers that mention detecting and identifying online hate speech	Papers that discuss detecting hate speech in languages other than English or Arabic
Papers that focus on identifying hate speech in English or Arabic only	Papers published more than 5 years ago
Papers that report key performance metrics for LLMs in hate speech detection	Duplicated Papers
Papers published in the last 5 years	Paywalled papers

3. Findings

We remained with 20 papers after filtering the papers through our criteria. We discarded 2 papers that were paywalled and we remained with 18 papers in the final set. We decided to divide the papers among the team members, each member reviewed 9 papers.

3.1 Publication dates

Figure 1 below shows the publication date of each paper that was included in the final set in our collection.

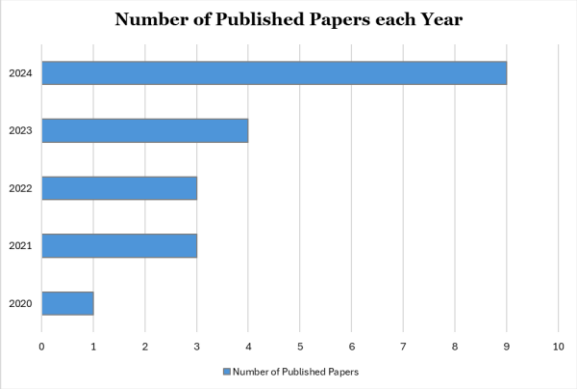

Figure 1: the distribution of publication dates of each paper

We can see that the majority of papers were published this year, and all the papers were published in the last 4 years. This can be attributed to the exponential advancement and rapid adaptation of LLMS in various applications in the last couple of years, specifically given that advanced models like GPT-4 were launched only last year [7]. Additionally, the COVID-19 Pandemic led to a surge in online presence and consequently increased amounts of hateful content [8], which prompted more research in effort to develop more reliable and effective hate speech detection systems.

3.2 Models used

By going through all selected papers, we can see that BERT is the most used model in the observed case studies, followed by GPT (by all its variations). 2 papers used LLaMA followed by a single paper using the Flan-T5 model. The figure below shows the distribution of model usage across the papers.

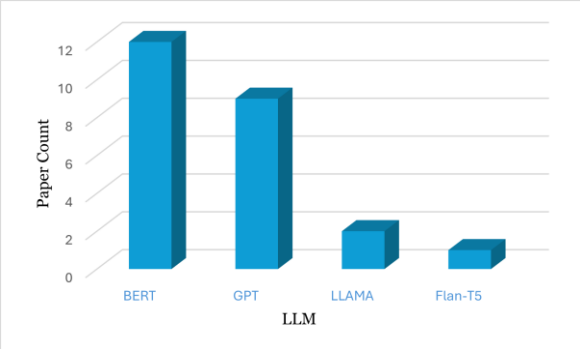

Figure 2: distribution of models used in the studies collected

The popularity of BERT (Bidirectional Encoder Representations from Transformers) in the domain of sentiment analysis can be attributed to its advanced architecture and its ability to capture nuanced semantic information in text [9]. BERT has the advantage of bidirectional training over other LLMs, which allows it to consider the context of a word based on all its surroundings rather than just the preceding or following words, this allows the model to get full understanding of the sentiment in a specific piece of text [10]. Research has shown that BERT outperforms traditional models in sentiment classification tasks, demonstrating superior accuracy and effectiveness in capturing grammatical and semantic features of text [11]

3.3 Datasets

The datasets used in the collected papers vary in sizes, source and format. We made sure in the exclusion phase of the review that the data used is in English (or Arabic), so we are able to understand the nature and context of the data used. Most of the data was collected from Twitter. Twitter has a public API that allows the researchers to collect tweets [12]. Besides, tweets represent a short text format and very often reflect real-time reactions; this makes Twitter an ideal source for studying linguistic nuances, including hate speech. We will go through the details of each data source in the following sections.

3.3.1 Hate targets

The datasets used in the collected papers often specify hate targets, which are crucial for categorizing and analyzing hate speech. These targets typically include various identity groups and societal categories that are frequently subjected to hateful or offensive language. Understanding hate targets helps improve the precision of hate speech detection models by accounting for context-specific nuances. Below are common categories of hate targets identified in the datasets:

1. Race and Ethnicity: Many datasets contain annotations for hate speech targeting racial or ethnic groups, reflecting the prevalence of racial slurs and discriminatory language on platforms like Twitter.
2. Religion: Hate speech directed at specific religious groups, often including Islamophobia, anti-Semitism, or anti-Christian sentiments, is a frequent focus of analysis.
3. Gender and Sexual Orientation: Datasets often include examples of misogyny, homophobia, and transphobia, which are common forms of hate speech online.
4. Political Affiliation: Some datasets capture hate speech directed at political ideologies or figures, which has become increasingly common in polarized online environments.
5. Disability and Appearance: Instances of hate speech targeting individuals based on physical or mental disabilities, or body shaming, are also frequently annotated in datasets.

The granularity and labeling of hate targets in datasets significantly impact the model's ability to generalize across different contexts.

3.3.2 Synthetic Data

Synthetic Data allows researchers to create diverse and scalable datasets, including subtle, implicit, or adversarial hate speech patterns. Examples like ToxiGen provide benchmarks for testing model robustness, while generative models like GPT-3 are often used to create customized datasets for pre-training or evaluation. Despite its benefits in mitigating bias and enhancing dataset diversity, synthetic data may lack realism and can lead to overfitting or ethical concerns about misuse. In the reviewed papers, synthetic data complements real-world datasets, expanding the scope of hate speech detection research

3.4 Overview of existing models Below is a table summarizing publicly available models for hate speech detection found in the collected papers, detailing the base architectures, training strategies, fine-tuning techniques, performance metrics, and data sources. These models represent a diverse range of approaches and capabilities tailored for specific datasets and hate speech detection challenges.

Table 2: hate speech detection frameworks with details on each model

Model	Base Architecture	Training Strategy	Fine-tuning	Performance Metrics	Data Sources
HATEGUARD [13]	BERT	Pre-trained on large datasets, fine-tuned for hate speech detection	Yes	Accuracy, F1-score	Twitter, Reddit
HateTinyLLM [14]	LLaMA	Fine-tuned on hate speech detection tasks	Yes	Accuracy, Precision, Recall	Synthetic data
COVID-HateBERT [15]	BERT	Pre-trained on large corpora, fine-tuned for COVID-19-related hate speech detection	Yes	Accuracy, Precision, Recall	Twitter
ToxiGen [16]	GPT	Pre-trained on vast datasets, fine-tuned for toxic and adversarial speech detection	Yes	Accuracy, Precision, F1-score	Twitter, Reddit
SHIELD [17]	BERT	Fine-tuned on annotated hate speech datasets	Yes	Accuracy, Precision, Recall	Twitter, Reddit, Wikipedia & 4chan

3.5 Comparing performance

In this section, we will conduct analysis and comparison of the performances of models mentioned in the collection of papers used in this scoping review. Below is a table of models mentioned in the papers that were accompanied by performance metrics to report on their accuracy. This provides quantitative benchmarks to evaluate how effective the models are at detecting hate speech under various conditions and datasets and diverse types of architectures

Table 3: analysis of models' performance by datasets, metrics and comparison notes

Model	Dataset	Data Sources	Training Strategy	Accuracy	Precision	Recall	F1-Score	Notes	Source
BERT	HateXplain	Twitter, Reddit	Fine-tuning	0.6909	0.6909	0.6909	0.6909	Baseline model 1	[18]
DM-RES				0.7273	0.7273	0.7273	0.7273	Outperformed BERT 1	[19]
HateBERT	Hate Speech and Offensive Language dataset	Twitter	Fine-tuning	-	0.681	0.674	0.677	Specialized for hate speech 1	[15]
COVID-HateBERT	COVID-19 related hate speech dataset	Social media	Pre-training + Fine-tuning	0.847	-	-	-	Specific to COVID-19 hate speech	[20]
GPT-3	Dynamically generated dataset	-	Few-shot learning	0.78	-	-	-	Used for hate speech generation and detection	[21]
XLM-RoBERTa	HASOC 2019 dataset	Twitter, Facebook	Fine-tuning	0.8090	0.8087	0.8090	0.8088	Multilingual model	[22]
mBERT				0.7843	0.7845	0.7843	0.7844	Multilingual model	
HateTinyLLM	HateXplain, HASOC 2021	Twitter, Reddit	Fine-tuning	-	-	-	0.83	Lightweight model	[14]
ALBERT	ToxiGen dataset	Synthetic data	Fine-tuning	-	0.781	0.781	0.781	Used with ToxiGen dataset	[16]
RoBERTa				-	0.805	0.805	0.805	Best performance on ToxiGen	

3.5.1 Analysis

The various models for hate speech detection have large gaps concerning their architecture, strategy, and dataset on which these have been trained. Thus, the DM-RES model gave very superior performance on the HateXplain dataset and outperformed the BERT baseline by an F1-score of 0.7273 against that of BERT at 0.6909, showing its better capability in the detection of hate speech [18]. RoBERTa achieved the highest, 0.805 F1-score on the ToxiGen dataset, hence proving its robustness against adversarial, as well as implicit hate

speech [16]. Specialized models, like COVID-HateBERT, reached an accuracy as high as 0.847 in detecting pandemic-related hate speech [20], whereas HateBERT resulted in lower scores, showing its shortcomings outside generalized hate speech contexts [15]. Lightweight models such as HateTinyLLM secured the highest F1-score (0.83) across several datasets, merging efficiency with robust performance [14], while GPT-3's few-shot learning capabilities provided competitive accuracy (0.78), though at the cost of significant resource demands [21].

4. Discussion

This scoping review highlights current developments in online hate speech detection using large language models (LLMs). Throughout the recent literature review, LLMs continuously outperformed both conventional machine learning and earlier deep learning techniques. They were able to analyze vast amounts of data effectively and comprehend contextual subtleties, which led to increased accuracy, robustness, and adaptability. A comprehensive discussion of the findings, including insights from the studies evaluated, is provided below.

4.1 Transformative Capabilities of LLMs

LLMs such as GPT-3, BERT, XLNet, and RoBERTa have solidified their roles as advanced tools for detecting hate speech across various datasets and methodologies. According to recent studies, transformer-based models frequently outperform conventional methods and produce remarkable outcomes with little feature engineering. The HATEGUARD framework, which uses chain-of-thought (CoT) reasoning to curb new waves of hate online, is a unique addition. This method places a strong emphasis on reasoning-based decision-making and outperforms traditional tools by dynamically adapting prompts to new hatred targets. [13] [18] [26]

LLMs have an advantage in zero-shot and few-shot learning settings, which makes them especially useful in situations where annotated data is limited. [30] While sentiment analysis tasks can benefit from the use of LLMs, their efficacy in structured or complex sentiment evaluations remains restricted, indicating that domain-specific data is required for further development.

4.2 Addressing Data Scarcity Through Generative Techniques

Data scarcity is a persistent challenge in hate speech detection. Recent advancements have focused on generative approaches to mitigate this limitation. The potential of deep generative modeling was demonstrated by creating a dataset of one million hate and non-hate speech sequences using GPT-2. Training a BERT-based model on this generated data led to significant performance improvements across diverse datasets, showcasing the scalability of such approaches. Similarly, the TOXIGEN dataset emphasizes balanced and implicit toxicity generation, enhancing the performance of existing toxicity classifiers. [16] [29] Augmentation strategies reinforce the

utility of LLMs in addressing dataset imbalances. By integrating GPT-3's generative capabilities with BERT-cosine similarity, these strategies improve model accuracy and robustness, particularly in cross-dataset generalization tasks. [23]

4.3 Advancements in Interpretability and Fine-Tuning Techniques

While LLMs excel in performance, interpretability remains a critical focus area. The SHIELD framework leverages LLM-extracted rationales to enhance the interpretability of hate speech detection models without compromising accuracy [27]. This approach bridges the gap between model predictions and human understanding, ensuring alignment with human-annotated rationales.

Additionally, lightweight fine-tuning methods, such as LoRA and adapter techniques, have been explored to optimize smaller LLMs for resource-constrained environments. fine-tuning specific layers in models like opt-1.3B achieved competitive results with reduced computational overhead, offering practical solutions for smaller organizations. [14]

4.4 Multilingual and Domain-Specific Applications

The effectiveness of LLMs in multilingual contexts and domain-specific tasks has also garnered attention. monolingual models excel on datasets in their respective languages, while multilingual transformers perform well across similar scripts. However, imbalances in multilingual datasets remain a concern, necessitating augmentation techniques like SMOTE or ADASYN. [22]

Domain-specific adaptations, such as COVID-HateBERT, highlight the importance of tailoring LLMs to emerging contexts. Studies demonstrated that pretraining BERT on COVID-19-related tweets improved detection accuracy significantly, offering insights into the adaptability of LLMs for specialized tasks. [15]

4.5 Challenges and Future Directions

Despite their transformative potential, LLMs face several challenges, including data imbalances, high computational costs, and ethical considerations. Imbalanced datasets hinder model generalization, particularly for subtle or emerging forms of hate speech. studies emphasized the need for comprehensive and diverse datasets to overcome these limitations [24] [25]. Additionally, the computational demands of training models like GPT-3 pose accessibility barriers, especially for smaller organizations. [18]

Ethical concerns further complicate the deployment of LLMs. biases in training data can perpetuate discriminatory outcomes, raising questions about accountability and fairness. [25] Incorporating transparency mechanisms, such as explainable AI techniques, is vital for addressing these concerns. Frameworks like SHIELD and chain-of-thought reasoning provide promising avenues for enhancing interpretability and fairness.

4.6 Implications for Stakeholders

The advancements in LLM-based hate speech detection have far-reaching implications for social media platforms, researchers, and policymakers. Automated moderation systems powered by LLMs can enhance the accuracy and efficiency of content moderation [28]. However, balancing effective moderation with free speech and cultural sensitivities remains a challenge. Policymakers and researchers must collaborate to develop ethical frameworks that address bias, fairness, and accountability in LLM-driven systems.

The integration of large language models into hate speech detection marks a significant leap in addressing online toxicity. While models like GPT-3, BERT, and domain-specific adaptations demonstrate remarkable accuracy and adaptability, challenges related to data, computational demands, and ethics persist. Future research should prioritize creating balanced datasets, optimizing resource-efficient models, and fostering interdisciplinary collaborations to ensure the responsible and effective deployment of LLMs. As online hate speech evolves, continuous innovation will be essential to maintaining safe and inclusive online spaces while upholding freedom of expression.

5. Conclusion

This scoping review provides a comprehensive analysis of advancements in detecting online hate speech, with a focus on the role of large language models (LLMs) and emerging methodologies. The findings demonstrate that LLMs, including domain-specific and multilingual adaptations, have significantly outperformed traditional and earlier deep learning approaches in terms of accuracy, robustness, and contextual understanding. Techniques such as chain-of-thought reasoning, fine-tuning, and the creation of synthetic datasets have further expanded the capabilities of LLMs, addressing challenges related to dataset scarcity and evolving hate speech patterns.

Despite these advancements, several challenges persist. Issues such as dataset imbalances, computational costs, and ethical considerations, including biases and transparency, highlight the complexities of deploying LLMs at scale. The need for interpretability and fairness in automated hate speech detection systems underscores the importance of interdisciplinary collaboration and ethical AI frameworks.

The implications for stakeholders are significant, offering opportunities to enhance online content moderation, foster more inclusive digital spaces, and establish ethical AI practices. Future research should prioritize improving the interpretability and efficiency of LLMs, developing balanced and diverse datasets, and addressing ethical challenges to ensure that technological progress aligns with societal values. This review emphasizes the transformative potential of LLMs in combating online hate speech while acknowledging the ongoing need for innovation and responsible application.

References

[1] R. Németh and J. Koltai, “Natural language processing,” *Intersections*, vol. 9, no. 1, pp. 5–22, Apr. 2023, doi: https://doi.org/10.17356/ieejsp.v9i1.871

[2] J. Hirschberg and C. D. Manning, “Advances in natural language processing,” *Science*, vol. 349, no. 6245, pp. 261–266, Jul. 2015, doi: https://doi.org/10.1126/science.aaa8685

[3] A. Radford, “Improving language understanding with unsupervised learning,” *Openai.com*, 2018. https://openai.com/index/language-unsupervised/

[4] A. Ligthart, C. Catal, and B. Tekinerdogan, “Systematic reviews in sentiment analysis: a tertiary study,” *Artificial Intelligence Review*, vol. 54, Mar. 2021, doi: https://doi.org/10.1007/s10462-021-09973-3

[5] DataRobot, “Introduction to Sentiment Analysis: What is Sentiment Analysis?,” *DataRobot*, Oct. 10, 2024. https://www.datarobot.com/blog/introduction-to-sentiment-analysis-what-is-sentiment-analysis

[6] United Nations, “What is hate speech?,” *United Nations*, 2024. https://www.un.org/en/hate-speech/understanding-hate-speech/what-is-hate-speech

[7] OpenAI, “GPT-4,” *Openai.com*, 2023. https://openai.com/index/gpt-4/

[8] V. Cotik *et al.*, “A study of Hate Speech in Social Media during the COVID-19 outbreak,” *OpenReview*, Jul. 2020.

[9] M. S. Sayeed, V. Mohan, and K. S. Muthu, “BERT: A Review of Applications in Sentiment Analysis,” *HighTech and Innovation Journal*, vol. 4, no. 2, pp. 453–462, Jun. 2023, doi: https://doi.org/10.28991/HIJ-2023-04-02-015.

[10] A. Areshey and H. Mathkour, “Transfer Learning for Sentiment Classification Using Bidirectional Encoder Representations from Transformers (BERT) Model,” *Sensors*, vol. 23, no. 11, pp. 5232–5232, May 2023, doi: https://doi.org/10.3390/s23115232

[11] B. V. P. Kumar and Dr. M. Sadanandam, “A Fusion Architecture of BERT and RoBERTa for Enhanced Performance of Sentiment Analysis of Social Media Platforms,” *International Journal of Computing and Digital Systems*, vol. 15, no. 1, pp. 51–66, Jan. 2024, doi: https://doi.org/10.12785/ijcds/150105

[12] X, “Twitter API Documentation,” *X.com*, 2024. https://developer.x.com/en/docs/x-api

[13] Nishant Vishwamitra *et al.*, “Moderating New Waves of Online Hate with Chain-of-Thought Reasoning in Large Language Models,” *2022 IEEE Symposium on Security and Privacy (SP)*, vol. 35, pp. 788–806, May 2024, doi: https://doi.org/10.1109/sp54263.2024.00181

[14] T. Sen, A. Das, and M. Sen, “HateTinyLLM: Hate Speech Detection Using Tiny Large Language Models,” *arXiv (Cornell University)*, Apr. 2024,

doi: https://doi.org/10.48550/arxiv.2405.01577

[15] M. Li *et al.*, "COVID-HateBERT: a Pre-trained Language Model for COVID-19 related Hate Speech Detection," *IEEE Xplore*, Dec. 01, 2021. https://ieeexplore.ieee.org/abstract/document/9680128

[16] T. Hartvigsen, S. Gabriel, H. Palangi, M. Sap, D. Ray, and E. Kamar, "ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection," *ACLWeb*, May 01, 2022. https://aclanthology.org/2022.acl-long.234/

[17] J. S. Malik, G. Pang, and Anton, "Deep Learning for Hate Speech Detection: A Comparative Study," Feb. 2022, doi: https://doi.org/10.48550/arxiv.2202.09517

[18] S. Gupta, Sachin Lakra, and M. Kaur, "Study on BERT Model for Hate Speech Detection," Nov. 2020, doi: https://doi.org/10.1109/iceca49313.2020.9297560

[19] H. Saleh, A. Alhothali, and K. Moria, "Detection of Hate Speech using BERT and Hate Speech Word Embedding with Deep Model," *Applied Artificial Intelligence*, vol. 37, no. 1, Feb. 2023, doi: https://doi.org/10.1080/08839514.2023.2166719

[20] A. F. Y. Chao, C.-S. Wang, B.-Y. Li, and H.-Y. Chen, "From Hate to Harmony: Leveraging Large Language Models for Safer Speech in Times of COVID-19 Crisis," *Heliyon*, vol. 10, no. 16, pp. e35468–e35468, Jul. 2024, doi: https://doi.org/10.1016/j.heliyon.2024.e35468

[21] K.-L. Chiu, A. Collins, and R. Alexander, "Detecting Hate Speech with GPT-3 *," Mar. 2022. Accessed: Jun. 12, 2024. [Online]. Available: https://arxiv.org/pdf/2103.12407

[22] K. Ghosh and Dr. Apurbalal Senapati, "Hate speech detection: a comparison of mono and multilingual transformer model with cross-language evaluation," *ACL Anthology*, pp. 853–865, Oct. 2022, Accessed: Nov. 28, 2024. [Online]. Available: https://aclanthology.org/2022.paclic-1.94

[23] M. S. Jahan, M. Oussalah, D. R. Beddia, J. kabir Mim, and N. Arhab, "A Comprehensive Study on NLP Data Augmentation for Hate Speech Detection: Legacy Methods, BERT, and LLMs," *Semantic Scholar*, 2024, doi: https://doi.org/10.48550/ARXIV.2404.00303

[24] D. Kikkisetti *et al.*, "Using LLMs to discover emerging coded antisemitic hate-speech in extremist social media," *arXiv (Cornell University)*, Jan. 2024, doi: https://doi.org/10.48550/arxiv.2401.10841

[25] T. Kumarage, A. Bhattacharjee, and J. Garland, "Harnessing Artificial Intelligence to Combat Online Hate: Exploring the Challenges and Opportunities of Large Language Models in Hate Speech Detection," *arXiv.org*, 2024. https://arxiv.org/abs/2403.08035v1

[26] S. Mukherjee and S. Das, "Application of Transformer-Based

Language Models to Detect Hate Speech in Social Media," *Journal of Computational and Cognitive Engineering*, vol. 1, no. 1, Mar. 2022, doi: https://doi.org/10.47852/bonviewjcce2022010102

[27] A. Nirmal, A. Bhattacharjee, P. Sheth, and H. Liu, "Towards Interpretable Hate Speech Detection using Large Language Model-extracted Rationales," *arXiv (Cornell University)*, Mar. 2024, doi: https://doi.org/10.48550/arxiv.2403.12403

[28] R. Pan, J. A. García-Díaz, and R. Valencia-García, "Comparing Fine-Tuning, Zero and Few-Shot Strategies with Large Language Models in Hate Speech Detection in English," *Computer Modeling in Engineering & Sciences*, vol. 140, no. 3, pp. 2849–2868, 2024, doi: https://doi.org/10.32604/cmes.2024.049631

[29] T. Wullach, A. Adler, and E. M. Minkov, "Towards Hate Speech Detection at Large via Deep Generative Modeling," *IEEE Internet Computing*, pp. 1–1, 2020, doi: https://doi.org/10.1109/mic.2020.3033161

[30] H. Zhang, M. Botler, and J. P. Kooman, "Deep Learning for Image Analysis in Kidney Care," *Advances in Kidney Disease and Health*, Dec. 2022, doi: https://doi.org/10.1053/j.akdh.2022.11.003

Contribution

	Duha	Sana
Abstract	✓	
1. Introduction	✓	
1.1 Natural Language Processing & Large Language Models		✓
1.2 Sentiemnt Analysis		✓
1.3 Online hate speech		✓
2. Methodology		✓
3. Findings		✓
3.1 Publication dates		✓
3.2 Models used		✓
3.3 Datasets		✓
3.3.1 Hate Targets		✓
3.3.2 Synthetic Data		✓
3.4 Overview of existing models		✓
3.5 Comparing performance		✓
3.5.1 Analysis		✓
4. Discussion	✓	
4.1 Transformative Capabilities of LLMs	✓	
4.2 Addressing Data Scarcity Through Generative Techniques	✓	
4.3 Advancements in Interpretability and Fine-Tuning Techniques	✓	
4.4 Multilingual and Domain-Specific Applications	✓	
4.5 Challenges and Future Directions	✓	
4.6 Implications for Stakeholders	✓	
5. Conclusion	✓	
References	✓	✓
Tables and Figures		✓