IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 21 December 2024, accepted 17 January 2025, date of publication 21 January 2025, date of current version 3 February 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3532397

RESEARCH ARTICLE

Hate Speech Detection Using Large Language Models: A Comprehensive Review

AISH ALBLADI[1], MINARUL ISLAM[1], AMIT DAS[2], MARYAM BIGONAH[1], ZHENG ZHANG[3], FATEMEH JAMSHIDI[4], MOSTAFA RAHGOUY[1], NILANJANA RAYCHAWDHARY[1], DANIELA MARGHITU[1], AND CHERYL SEALS[1]

[1]Auburn University, Auburn, AL 36830, USA
[2]University of North Alabama, Florence, AL 35632, USA
[3]Murray State University, Murray, KY 42071, USA
[4]California State Polytechnic University, Pomona, CA 91768, USA

Corresponding author: Aish Albladi (aza0266@auburn.edu)

ABSTRACT The widespread use of social media and other online platforms has facilitated unprecedented communication and information exchange. However, it has also led to the spread of hate speech and poses serious challenges to societal harmony as well as individual well-being. Traditional methods for detecting hate speech, such as keyword matching, rule-based systems, and machine learning algorithms, often struggle to capture the subtle and context-dependent nature of hateful content. This paper provides a comprehensive review of the application of large language models (LLMs) like GPT-3, BERT, and their successors in hate speech detection. We analyze the evolution of LLMs in natural language processing and examine their strengths and limitations in identifying hate speech. Additionally, we address the significant challenges and explore how LLMs method can affect the accuracy and fairness of hate speech detection systems. By synthesizing recent research, this review aims to offer a holistic understanding of the current state-of-the-art methods in hate speech detection utilizing LLMs and to suggest directions for future research that could enhance the efficacy and equity of these systems.

INDEX TERMS Deep learning, hate speech detection, large language models, machine learning.

I. INTRODUCTION

Recently, the advent of large language models (LLMs) has revolutionized the field of natural language processing (NLP). Models like BERT, GPT, and their successors have demonstrated unprecedented capabilities in understanding and generating human-like text. These models, built on deep learning architectures, have shown promise in a wide range of NLP tasks. However, when applied to the increasingly widespread issue of hate speech detection, LLMs face unique challenges, such as bias, difficulties in generalizing across languages, and the potential to amplify harmful content. This paper reviews advancements in hate speech detection using LLMs and examine key methodologies, performance across datasets, and ethical concerns. Understanding the strengths and limitations of these models is essential for developing more responsible and effective hate speech detection systems.

Language acts as a crucial mechanism for human interaction and expression, beginning to develop from infancy and continuously evolving throughout one's life [1], [2]. LLMs are a type of language model that utilize neural networks with billions of parameters, trained on extensive, unlabeled textual data via self-supervised learning methods [3], [4]. Often trained using wide-ranging text corpora collected from the internet, these models can detect intricate patterns, subtle linguistic details, and complex semantic connections.

LLMs have demonstrated remarkable proficiency in various language-related tasks, such as text generation, translation, summarization, question answering, and sentiment analysis, by utilizing deep learning methods and extensive datasets. For making predictions, language models started incorporating contexts of increasingly larger scope. The self-supervised approach started with individual

The associate editor coordinating the review of this manuscript and approving it for publication was Ayman El-Baz.

words, followed by surrounding words, sentences, and paragraphs [5], [6], [7]. Moreover, fine-tuning these models for specific downstream tasks has yielded state-of-the-art performance across multiple benchmarks. The development of LLMs traces back to the early efforts in language modeling and neural networks. Initial approaches, such as statistical methods, were limited in their ability to capture long-term dependencies and context within language [8]. As research progressed, the advent of more sophisticated neural network architectures and larger datasets enabled the creation of more advanced models, paving the way for the powerful LLMs we see today [9].

There are several ways to detect hate speech, including manual review, wherein trained human reviewers can manually review and identify hate speech by looking for certain keywords, phrases, or patterns. Machine learning algorithms can be trained to detect hate speech by analyzing text for patterns and features commonly associated with hate speech. The development of Recurrent Neural Networks (RNNs) marked a pivotal moment in the evolution of machine learning, particularly for modeling sequential data such as language. Despite their importance, RNNs encountered significant challenges, especially with issues like vanishing gradients and difficulties in managing long-term dependencies [10], [11]. The landscape of large language models (LLMs) was profoundly transformed with the introduction of the transformer architecture, a breakthrough that addressed many of RNNs' limitations. The transformer model, distinguished by its self-attention mechanism, revolutionized how models handle long-range dependencies by allowing for greater parallelization and more efficient processing of sequential data [12], [13]. This architecture laid the groundwork for the development of advanced models like Google's Bidirectional Encoder Representations from Transformers (BERT) and OpenAI's Generative Pretrained Transformer (GPT) series [14], [15], [16]. These models have set new benchmarks in natural language processing which excels across a wide array of language tasks, and demonstrating the transformative potential of the underlying LLM architecture [2]. Their success in areas such as machine translation analysis, and content generation underscores their versatility and the expanding horizon of possibilities they introduce to both academic research and practical applications in technology.

The diagram illustrates the overall review process of the paper as shown in Fig. 1. The systematic review process, where a structured search strategy was applied using databases like IEEE Xplore and Google Scholar, and search queries such as "Hate Speech Detection" and "GPT-3" were formulated. After initial filtering based on title and abstract relevance, the inclusion and exclusion criteria were applied. We ensure studies focused on LLM-based research addressing multilingual challenges and ethical considerations were included, while non-LLM studies and those lacking evaluation were excluded. In the model and performance review, data from selected studies were reviewed, focusing on model types, datasets, evaluation metrics, ethical considerations, and recurring challenges and solutions. Finally, in the synthesis of findings, key insights were summarized which highlights gaps in the research and providing future directions for improving hate speech detection models.

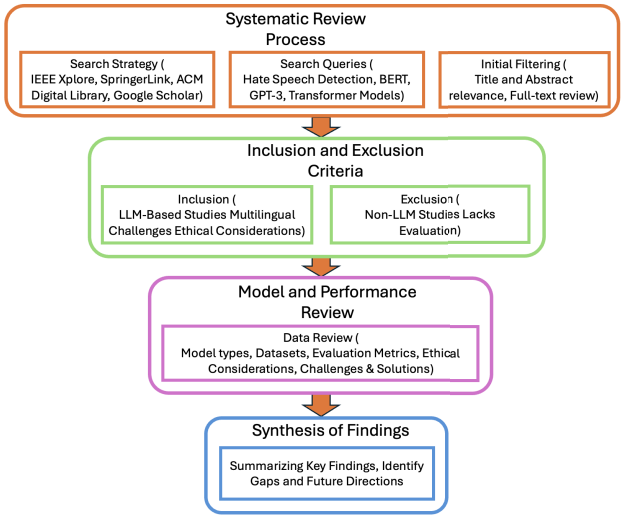

FIGURE 1. Workflow of our method.

Given the rapid evolution of LLMs, a comprehensive assessment of current research for hate speech detection is increasingly necessary. While previous studies have highlighted the potential and superiority of LLMs in various NLP tasks, only a limited number of reviews have thoroughly examined the latest developments, possibilities, and limitations of these models for hate speech detection. In recent years, the development of large language models (LLMs) like GPT-3 and BERT has opened new avenues for enhancing hate speech detection. These models, which are capable of understanding and generating human-like text, offer a more sophisticated approach to identifying hate speech. Their ability to process large amounts of text and recognize patterns in language has made them powerful tools for various NLP tasks, including sentiment analysis, translation, and text generation. This review paper aims to provide a detailed examination of the role that LLMs play in the detection of hate speech. We begin by exploring the evolution of LLMs and their increasing applications within NLP. Following this, we review recent research to assess the capabilities and limitations of these models in the context of hate speech detection. Moreover, this paper will explore the ethical implications of using LLMs for hate speech detection, focusing on the importance of transparency, fairness, and accountability in these systems. As LLMs become more widely adopted, it is essential to consider the ethical ramifications of their use, particularly in sensitive areas like hate speech detection, where errors can have significant consequences.

Researchers have explored various aspects of LLMs in multiple studies; however, these works often fall short in

addressing several critical areas. Many studies overlook essential aspects of LLMs, such as their high-level architecture, configurations, taxonomies, API applications, domain-specific uses, and the datasets they rely on. For instance, there is often a lack of detailed explanations regarding the core architecture and configurations of LLMs, insufficient exploration of their taxonomy, distinctions based on machine learning principles and domain-specific applications. Additionally, descriptions of the datasets used in LLMs are frequently absent or incomplete for hate speech detection. Moreover, a significant portion of LLMs review papers have not undergone peer review, further contributing to the gaps in the literature. These omissions highlight the need for a more comprehensive investigation of LLMs methods and datasets explanations with their model for hate speech detection. Addressing these research gaps is crucial for advancing the field. Therefore, this paper aims to thoroughly analyze existing review papers, identify their limitations, and provide a detailed overview of the current state-of-the-art methods to tackle these challenges. The proposed works primary objective is to explore, understand, and evaluate LLMs across various domains, including their evolution, classification, architecture of pre-trained models, resources, and real-time applications for hate speech detection.

Furthermore, we offer a set of guidelines to guide future research and development efforts for the effective use of LLMs for hate speech detection. We hope that this study will enhance the understanding and application of LLMs in various fields. By synthesizing insights from recent studies, we aim to provide a comprehensive understanding of the current state of the art and propose avenues for future research that can further improve the effectiveness and fairness of these systems. This review underscores the importance of continuing to refine these models and developing strategies to mitigate the gap, ultimately contributing to safer and more inclusive digital spaces. The contributions of this paper are outlined as follows:

Key Contributions: The key contributions of this research are summarized as follows:

1) **Comprehensive Overview of LLMs:** The overview includes a detailed classification of LLMs, distinguishing between pre-trained models and API-based models. Additionally, the discussion covers the fundamental structures of LLMs, with a particular emphasis on the transformer architecture, which has revolutionized natural language processing. This foundational context sets the stage for understanding the advancements and applications of LLMs in various domains for example hate speech detection.
2) **Impact of Machine Learning on LLMs for Hate Speech Detection:** By demonstrating the significance of machine learning (ML) across different LLMs domains, this contribution underscores the symbiotic relationship between ML methodologies and the evolution of LLMs, providing insights into how ML innovations continue to drive LLM advancements for hate speech detection.
3) **Comparison of Pre-Trained Model Designs:** This contribution provides a detailed comparison of pre-trained model performances within the LLM landscape. By highlighting the strengths and limitations of each design, it enables researchers to make informed decisions when selecting or developing models for specific tasks.
4) **Datasets Used in LLM Training for Hate Speech Detection:** This section provides a concise yet comprehensive overview of the datasets used during the training phases of LLMs. The contribution also discusses the diversity and scale of these datasets, which are critical for training models capable of understanding and generating human-like text across various contexts and languages.
5) **Challenges, Open Issues, and Future Opportunities:** The final contribution of the paper explores the challenges in the LLMs domain, including security vulnerabilities, ethical dilemmas, privacy concerns, and environmental impacts. It identifies future research opportunities to address these issues, aiming to guide and foster responsible innovation in LLMs method for hate speech detection.

The rest of the paper is organized as follows: In Section II, the literature review is discussed. Section III demonstrates the Methodology; Section IV discuss the challenges and future research directions regarding LLMs for hate speech detection, Section VII finally concludes the paper.

II. LITERATURE REVIEW

A. BACKGROUND OF LLMS

Large language models (LLMs) represent a significant advancement in the field of natural language processing (NLP), utilizing deep learning techniques to understand and generate human language. The development of LLMs has been rooted in the evolution of machine learning and artificial intelligence over the past several decades [3], [17], [18]. The conceptual foundation of LLMs can be traced back to the 1950s, when early artificial intelligence (AI) systems, such as the Turing Test proposed by Alan Turing in 1950, sought to measure a machine's ability to exhibit intelligent behavior indistinguishable from that of a human [19], [20], [21]. The limitations of these systems led to the development of more advanced techniques in the following decades. The 1980s and 1990s saw the advent of statistical methods in NLP, which shifted the focus from rule-based systems to data-driven approaches [22]. These methods leveraged large corpora of text to learn patterns and make predictions that marks the beginning of more flexible and domain-specific language models. However, these early models could not still capture long-range dependencies and complex linguistic structures.

The performance of language models was greatly enhanced by the widespread use of deep learning techniques in the 2010s, which allowed them to handle complicated tasks with higher accuracy and contextual comprehension than earlier approaches [23]. Deep learning, particularly through the use of neural networks enabled models to learn hierarchical representations of language (captures syntax and semantics). This era saw the development of models like Word2Vec, which revolutionized the way word meanings were represented in vector space, and subsequently more complex models like recurrent neural networks (RNNs) and long short-term memory (LSTM) networks [24]. The true breakthrough in LLMs, however, came with the advent of the Transformer architecture, introduced by Vaswani et al. in 2017 [25]. The Transformer model departed from the sequential nature of RNNs and LSTMs, instead using self-attention mechanisms to process entire sentences or paragraphs simultaneously. This architecture became the backbone of the most influential LLMs, including GPT (Generative Pre-trained Transformer) models developed by OpenAI, and BERT (Bidirectional Encoder Representations from Transformers) developed by Google [26]. ELMo is another milestone before BERT, consider adding it.

GPT models, particularly from GPT-2 onwards, demonstrated an unprecedented ability to generate coherent and contextually relevant text across a wide range of topics [27]. These models are pre-trained on massive datasets and fine-tuned for specific tasks to perform well in diverse applications such as translation, summarization, and even creative writing. BERT, on the other hand, introduced the concept of bidirectional training, where models are trained to consider both the left and right context in all layers, significantly improving performance on tasks like question answering and sentiment analysis [28]. Despite these advancements, LLMs are not without their challenges. Issues such as model bias, the requirement for vast computational resources, and difficulties in generalizing across languages and domains are areas of ongoing research. Furthermore, the ethical implications of LLMs, particularly in terms of their deployment in real-world applications, remain a critical area of concern for hate speech detection [29].

Table 1 compares recent research articles focusing on the application of large language models (LLMs) for hate speech detection. The goal of this comparison is to highlight the different methodologies and approaches used to address challenges like bias mitigation, context capture, domain-specific specialization, and generalization. Many of the articles review popular models such as GPT-3, BERT, RoBERTa, and custom LLMs, etc. showing their strengths in specific domains, such as multilingual hate speech detection and fine-tuned performance in context-driven tasks. However, key issues like bias against marginalized groups, limitations in cross-domain generalization, and ethical concerns are recurrent themes. By synthesizing these findings, the table underscores the need for balanced trade-offs between robustness and accuracy, effective bias mitigation, and ethical considerations. The broad review presented in the final row, encompassing diverse datasets, models, and ethical perspectives, suggests a more holistic approach to hate speech detection. It positions itself as offering a more comprehensive and integrative view and aimed at addressing the gaps left by the other studies.

B. BRIEF HISTORY OF HATE SPEECH DETECTION USING LLM

In the 1980s and 1990s, significant progress was made with the introduction of neural networks, particularly through the backpropagation algorithm [41], [42]. These networks laid the groundwork for the development of more complex models that could learn from vast amounts of data. However, the computational limitations of time-restricted the scale and complexity of these models. The early 2000s saw the rise of statistical methods in NLP, particularly with models such as Hidden Markov Models (HMMs) [43], [44]. These methods were extensively used for tasks such as part-of-speech tagging, speech recognition, and machine translation [45], [46]. However, these models had limitations in terms of handling long-range dependencies in language, which led to the exploration of more advanced architectures.

A significant breakthrough came in 2013 with the introduction of the word2vec model, which utilized neural networks to create dense vector representations of words [47], [48]. This model demonstrated that semantic relationships between words could be captured in a continuous vector space, leading to improvements in various NLP tasks. This was followed by the development of more complex models like GloVe and fastText [49], [50], [51]. The true revolution in LLMs began with the introduction of the Transformer architecture [26]. The Transformer model, with its self-attention mechanism, allowed for the parallel processing of input data, significantly improving the efficiency and scalability of language models. This architecture became the foundation for subsequent LLMs, including BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer) [9], [52]. GPT, in particular, marked a paradigm shift in the field. The model was pre-trained on vast amounts of text data and then fine-tuned for specific tasks, demonstrating that it could generate coherent and contextually appropriate text. The release of GPT-3 by OpenAI in 2020, with 175 billion parameters, represented a new era in NLP, shows the power of LLMs in generating human-like text across a wide range of domains [3], [53]. As these models grew in size and capability, they began to outperform traditional methods in various NLP tasks, from translation and summarization to question-answering and text generation. The success of LLMs can be attributed to the combination of vast amounts of training data, advancements in computing power (especially GPUs), and the development of sophisticated algorithms for model training and optimization.

The historical development of hate speech detection techniques using machine learning models, illustrating key

TABLE 1. Comparison of recent articles on LLM-based hate speech detection. This table summarizes studies on hate speech detection using large language models (LLMs) which describe the models, datasets, ethical considerations, and cross-domain generalization. It also highlights key findings, methodologies, and the specific focus of each study, as well as provide a snapshot of current research trends and challenges.

Ref	LLMs Model	Datasets	Key Findings	Methodology and Approach	Scope
Ahuja et. al., [30]	GPT-3, GPT-4	Multilingual (English, Hindi, Urdu)	Inconsistent performance across languages	Empirical analysis on multiple multilingual datasets	Multilingual hate speech detection
Narang et. al., [31]	BERT	Twitter, Reddit	Significant biases in LLMs for marginalized groups	Survey and analysis of bias and ethical implications in LLMs	Ethical implications and bias in hate speech detection
Coban et. al., [32]	BERT variations	Domain-specific (Politics)	Fine-tuning improves accuracy but risks overfitting	Comparative analysis of BERT variations on domain-specific datasets	Specialization vs. generalization in hate speech detection
Hong et. al., [33]	Custom LLM	Religion, Politics	Domain-specific LLMs outperform general models in specific contexts	Meta-analysis of domain-specific LLM performance	Adaptability of LLMs in domain-specific hate speech detection
Hartvigsen et. al., [34]	GPT-2, RoBERTa	Social Media (Facebook, Twitter)	Adversarial training improves robustness but reduces accuracy	Experimental evaluation using adversarial examples	Balancing robustness and accuracy in hate speech detection
Nagar et. al., [35]	Transformer-based	Twitter, News Articles	Effective at context capture but struggles with nuanced hate speech	Evaluation of implicit hate speech with focus on context	Contextual analysis in hate speech detection
Raza et. al., [36]	ALBERT, RoBERTa	Hate Speech Corpus (Various platforms)	Ethical concerns over false positives affecting marginalized groups	Analytical study of ethical impacts of LLMs in hate speech detection	Ethics and bias in automated hate speech detection
Shi et. al., [37]	BERT, GPT-3	Diverse (Social media, news, forums)	Cross-domain performance is suboptimal	Comparative analysis across different domains	Cross-domain adaptability in hate speech detection
Zhang et. al., [38]	T5, GPT-3.5	Social Media, News	High accuracy in domain-specific tasks but generalization is limited	Evaluation of fine-tuned LLMs on domain-specific datasets	Specialization in domain-specific hate speech detection
Roy et. al., [39]	GPT-3, BERT	Reddit, Twitter, News	LLMs struggle with detecting sarcasm and coded language	Empirical analysis of LLMs on complex hate speech	Handling of nuanced and implicit hate speech
Arango et. al., [40]	Custom LLM, GPT	Hate Speech Databases (English, Multilingual)	High within-language accuracy but poor cross-lingual performance	Cross-lingual evaluation of LLMs in hate speech detection	Multilingual and cross-lingual hate speech detection
Ours	Multiple (GPT, BERT, etc.)	Diverse (Social media, news, multilingual)	Detailed evaluation of LLMs across domains, bias mitigation strategies	Broad review covering datasets, architectures, ethical considerations	Comprehensive synthesis of LLM-based hate speech detection methods

advancements from the 1940s to the present shows in Fig. 2. The concept of Artificial Neural Networks (ANNs) in the 1940s-1980s shows at first, the timeline moves through rule-based AI systems, statistical methods like SVMs and Naïve Bayes in the 1990s-2000s, and further improvements with models such as RNNs and Word2Vec in the early

2010s. The most recent advances include transformer models like GPT and cross-lingual LLMs from 2016 onward, which have significantly enhanced hate speech detection capabilities. This visual emphasizes the progressive refinement of methodologies and reflects the growing complexity and effectiveness of large language models (LLMs) in handling subtle and context-rich hate speech detection tasks. Understanding this evolution is critical for limitations of LLMs in the broader scope of hate speech detection.

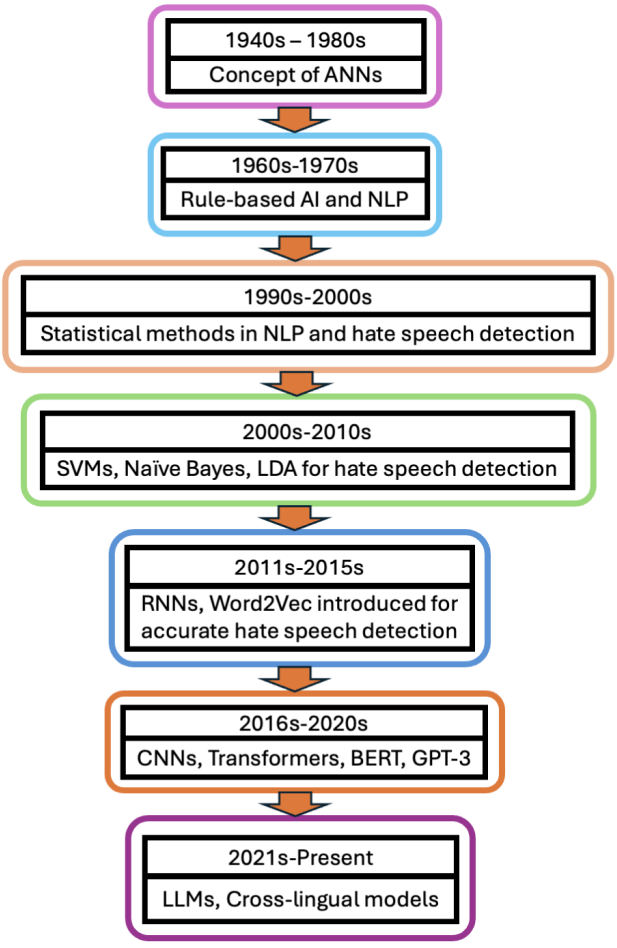

FIGURE 2. History of hate speech detection using LLMs.

The detection of hate speech online has become a critical area of research due to the harmful content on social media platforms and other online forums. Traditional approaches to hate speech detection, which relied on keyword-based methods and simple machine learning classifiers and often struggled with the complexities of language, such as sarcasm, context, and evolving slang [54]. The advent of LLMs has significantly improved the accuracy and robustness of hate speech detection systems. Early attempts to utilize deep learning for this task involved the use of convolutional neural networks (CNNs) and recurrent neural networks (RNNs), which showed promise but were limited by their reliance on fixed-length input sequences and their difficulty in capturing long-range dependencies [55], [56], [57].

The transformer architecture and subsequent LLMs provided a more effective solution. Models such as BERT and GPT have been employed to detect hate speech by leveraging their ability to understand the context and generate meaningful representations of text [58]. BERT, with its bidirectional attention mechanism, excels at understanding the context within a sentence, making it particularly useful for tasks like hate speech detection where context is crucial [59], [60]. Research has shown that fine-tuning LLMs on specific hate speech datasets can lead to significant improvements in detection accuracy. For instance, studies have utilized datasets such as the hate speech and offensive language dataset and the multilingual hate speech dataset to train models that can detect hate speech across different languages and cultural contexts [61], [62]. LLMs have demonstrated the ability to generalize across these datasets, outperforming traditional methods and earlier deep-learning models.

Moreover, recent research has explored the use of LLMs for zero-shot and few-shot learning in hate speech detection. These approaches allow models to detect hate speech in domains or languages where labeled data is scarce [63], [64]. This is particularly important for addressing the challenges of detecting hate speech in under-resourced languages and emerging online communities. In addition to their effectiveness, LLMs also introduce new challenges in hate speech detection. The reliance on large-scale pre-training raises concerns about biases in the training data, which can lead to biased predictions in hate speech detection models. Researchers have highlighted the need for careful dataset creation and the development of methods to mitigate bias in LLMs [9], [65]. Furthermore, the computational cost of training and deploying LLMs remains a significant barrier, particularly for smaller organizations and researchers. The environmental impact of training large models has also become a concern and prompts discussions about the sustainability of LLMs [53], [66]. Despite these challenges, LLMs represent a powerful tool in the fight against online hate speech. Their ability to understand and generate human-like text and combined with their capacity to learn from diverse datasets which makes them well-suited for this task.

A comprehensive analysis of the current literature on hate speech detection using large language models (LLMs) such as GPT-3, BERT, and their successors reveals several critical research gaps. First, the detection of implicit and coded hate speech remains a significant challenge. Many existing models struggle to recognize nuanced expressions of hate speech, such as sarcasm, euphemisms, or contextually veiled harmful content, which require deeper contextual understanding and advanced interpretability. Second, while LLMs have demonstrated remarkable success in resource-rich languages, their cross-lingual and multilingual performance often falls short. The scarcity of high-quality, diverse datasets for under-represented languages limits the effectiveness of these models

in global applications, particularly in linguistically diverse regions. Third, ethical challenges related to bias in training data persist, leading to models that may perpetuate or even amplify unfair outcomes, particularly against marginalized groups. This shows the need for more robust strategies to mitigate bias and ensure fairness in model predictions. Finally, the computational demands of training and deploying LLMs present barriers to scalability and accessibility, particularly for smaller organizations or researchers with limited resources. Addressing these gaps through the development of more inclusive datasets, improved bias mitigation techniques, and efficient model architectures is crucial for advancing hate speech detection to be more equitable and universally applicable. To the best of our knowledge, the aforementioned research articles has overlooked the comprehensive review for hate speech detection utilizing LLMs method. The proposed work aims to explore LLMs methods in detecting hate speech across different platforms and contexts with their datasets as well as comparison.

III. METHODOLOGY

The methodology section outlined a systematic approach to evaluating the effectiveness of large language models (LLMs) in detecting hate speech. We aim to provide a broad overview of the research process and the underlying principles that guide this study. To conduct a comprehensive review of the literature on hate speech detection using LLMs, a systematic approach was adopted to ensure the inclusion of relevant and high-quality studies. The process began with the formulation of specific search queries tailored to the research scope which includes keywords such as ''LLMs and hate speech detection,'' ''NLP models,'' and ''toxic content.'' These queries were designed to capture various relevant studies across various aspects of hate speech detection. The next step involved selecting appropriate academic databases known for their extensive collections of peer-reviewed articles, including IEEE Xplore, ACM Digital Library, Springer, and Google Scholar.

An initial search was performed across these databases and applied filters such as publication date and peer-reviewed status to ensure the relevance and quality of the results. This search yielded many papers, which were then subjected to a duplicate removal process to eliminate redundant entries. Subsequently, the titles and abstracts of the remaining papers were screened to exclude studies that did not focus directly on hate speech detection or did not employ LLMs. This step was crucial for narrowing down the list to studies most pertinent to the research objectives.

The final step involved a full-text review of the shortlisted papers. This detailed examination allowed for the exclusion of papers that despite appearing relevant initially, and did not meet the inclusion criteria upon closer inspection. The inclusion criteria were carefully defined to ensure that the selected studies provided significant insights or contributions to the field of hate speech detection using LLMs. Ultimately, the most relevant and methodologically sound studies were selected for the comprehensive review which forms the foundation of the research synthesis presented in this paper.

The proposed research aims to visually depict the systematic approach for identifying, reviewing, and selecting relevant research papers in the domain of large language models (LLMs) and hate speech detection. We start from search query formulation with keywords like LLMs and hate speech detection, followed by selecting appropriate academic databases such as IEEE Xplore and Google Scholar etc., performing an initial search, removing duplicates, and indexing the titles and abstracts of the remaining papers. The final steps involve full-text reviews to ensure relevance and final selection. The focus on each step ensures that only the most pertinent and high-quality papers are included, which strengthens the review's reliability and comprehensiveness. This structured process eliminates irrelevant and redundant data while ensuring a comprehensive literature base for further research and analysis on hate speech detection using LLMs.

A. THE ROLE OF LLMS IN HATE SPEECH DETECTION

Large language models, such as BERT and GPT-3, have significantly advanced the field of NLP. These models are particularly well-suited for tasks that require deep contextual understanding, such as hate speech detection. The ability of LLMs to process and interpret vast amounts of text data enables them to identify potentially harmful content with a higher degree of accuracy compared to earlier machine learning models. In hate speech detection, LLMs are employed to scan text from diverse sources—including social media platforms like Twitter, Reddit, etc., as well as news websites—and identify language that could be classified as hate speech. Their capacity to understand context, syntax, and semantics allows these models to detect not only explicit hate speech but also more subtle forms of harmful language.

B. THE RESEARCH PROCESS

The research process involved in this study begins with the identification of relevant studies through a comprehensive search strategy. By targeting specific keywords and databases, the study ensures the inclusion of research that is directly relevant to hate speech detection using LLMs. The selected studies are then subjected to a rigorous screening process, where inclusion and exclusion criteria are applied to filter out non-relevant research. Following the selection of studies, data extraction is performed to gather critical information about the model architectures, datasets, and evaluation metrics used in the selected research. This extracted data forms the basis for the subsequent analysis, where qualitative and quantitative methods are employed to assess the effectiveness of LLMs in hate speech detection.

C. RESEARCH DESIGN

This study employs a systematic review and empirical analysis to assess the effectiveness of LLMs in detecting hate speech across different languages and platforms. The

methodology is designed to capture both the technical performance of these models and the broader ethical implications of their deployment. The research follows a structured approach, beginning with the identification of relevant studies, followed by data extraction, analysis, and synthesis.

D. SEARCH STRATEGY

To gather a comprehensive dataset of relevant research articles, a well-defined search strategy was implemented across multiple academic databases, including IEEE Xplore, SpringerLink, ACM Digital Library, and Google Scholar. The search was conducted for publications released between January 2018 and August 2024 to capture the latest advancements in LLM technologies.

E. SEARCH QUERIES AND KEYWORDS

The search queries were crafted to target studies focused on hate speech detection using LLMs. The following search strings were used:

SQ1: 'Hate Speech Detection', 'BERT', 'GPT-3', 'Transformer Models' SQ2: 'Cross-lingual Hate Speech', 'LLMs', 'Multilingual Detection' SQ3: 'Implicit Hate Speech', 'Adversarial Training', 'Contextual Analysis' SQ4: 'Ethical Considerations', 'Bias in LLMs', 'False Positives'. These queries were carefully chosen to ensure the inclusion of studies that address various aspects of hate speech detection, from model architecture to ethical concerns.

F. DATABASE SOURCES

The databases selected for this study were chosen based on their relevance and the breadth of coverage they offer in the fields of AI and NLP. The databases include:

IEEE Xplore: Specializes in engineering and technology research, providing access to high-quality technical literature. SpringerLink: Offers a wide range of multidisciplinary research, including key fields like computer science, life sciences, and social sciences. ACM Digital Library: Focuses on computing and information technology, delivering comprehensive resources for computer science research. Elsevier: Publishes scientific, technical, and medical content, providing critical resources for academic and industry researchers. ScienceDirect: A key platform from Elsevier, offering access to a large collection of peer-reviewed scientific and technical research articles. Google Scholar: Aggregates academic publications from a broad array of disciplines, offering a robust platform for interdisciplinary research.

G. INCLUSION AND EXCLUSION CRITERIA

To ensure the relevance and quality of the studies selected for this review, specific inclusion and exclusion criteria were applied during the screening process.

1) INCLUSION CRITERIA

Studies that utilize LLMs such as BERT, GPT-3, or custom transformer models for hate speech detection. Research that includes empirical evaluations on diverse datasets, particularly those that are multilingual or cross-lingual. Papers that address challenges in detecting implicit or nuanced hate speech detection. Studies that discuss ethical implications, including model bias and the handling of false positives and negatives.

2) EXCLUSION CRITERIA

Studies focus solely on traditional machine learning models without the integration of LLMs. Research lacking empirical evaluation or practical application in hate speech detection. Papers with insufficient focus on cross-lingual or multilingual contexts.

H. STUDY SELECTION PROCESS

Initially, 250 studies were identified through database searches. After removing duplicates (n = 60), 190 abstracts were screened for relevance. Following the abstract review, 133 full-text articles were assessed based on the inclusion and exclusion criteria. Finally, 95 articles were selected for data extraction and detailed analysis.

I. DATA EXTRACTION

Data extraction was carried out using a standardized form to ensure consistency across all selected studies. The form captured the following key aspects:

1. Model Architectures: The specific types of LLMs used in each study were documented and includes details on the model version (e.g., BERT, GPT-3, mBERT) and any fine-tuning or customization applied.

2. Datasets: Information on the datasets used in each study was recorded, including language, platform (e.g., social media, news), size, and any cross-lingual or multilingual characteristics.

3. Evaluation Metrics: Performance metrics such as accuracy, precision, recall, and F1-score were extracted to assess the effectiveness of the models. Additional metrics, such as cross-lingual generalization and robustness to adversarial examples, were also noted.

4. Ethical Considerations: Any discussion of ethical issues, including model bias, false positives/negatives, and fairness, was carefully extracted to evaluate the broader implications of deploying LLMs for hate speech detection.

J. ANALYTICAL METHODS

The data extracted from the studies were analyzed using both qualitative and quantitative approaches.

1. Qualitative Analysis: A thematic analysis was conducted to identify common challenges, solutions, and ethical considerations across the studies. This analysis focused on recurring themes such as the difficulty of detecting implicit hate speech, the trade-offs between accuracy and fairness, and the strategies proposed to mitigate model bias.

2. Quantitative Analysis: Statistical methods were applied to compare the performance of different LLMs across various datasets and languages. This analysis involved calculating the average accuracy, F1-scores, and other relevant metrics for

each model. Cross-lingual performance was also evaluated, with particular attention to the drop in accuracy when models were applied to languages other than the one they were trained on.

K. SYNTHESIS OF FINDINGS

The final step in the methodology involved synthesizing the results of the analysis to draw conclusions about the current state of hate speech detection using LLMs. The synthesis was guided by the research questions and objectives, focusing on identifying the strengths and weaknesses of LLMs, the challenges of cross-lingual detection, and the ethical implications of deploying these models.

L. RECENT ARTICLES FOR HATE SPEECH DETECTION USING LLMS

The application of LLMs in hate speech detection has gained momentum in recent years with numerous studies highlighting their advantages in addressing the complexities of online hate speech. One of the primary areas of focus has been the fine-tuning of pre-trained models such as BERT, GPT-3, and RoBERTa for specific hate speech detection tasks [111]. These studies have shown that LLMs, when fine-tuned on well-curated hate speech datasets and can outperform traditional machine learning methods. They are particularly effective in detecting implicit and nuanced hate speech that often evades simpler keyword-based detection systems. For example, a fine-tuned BERT model could achieve superior accuracy in identifying hate speech, especially in cases where the hateful content was masked by sarcasm or indirect language [112].

Recent advancements in large language models (LLMs) have significantly impacted the detection and mitigation of hate speech across various digital platforms. Several studies have proposed innovative methodologies to identify subtle and evolving forms of hate speech. For instance, [67] and [78] explore novel approaches for detecting coded and implicit hate speech using LLMs, while [68] and [79] compare various data augmentation techniques and LLM-generated responses to enhance model robustness and reduce harmful content engagement. LLMs have also been applied to generate counter speech, with methods such as transformer reinforcement learning and prompt engineering proving effective in reducing online incivility [70], [75]. However, challenges remain, particularly in addressing biases and oversensitivity in LLMs, as highlighted in [39], [69]. The integration of LLMs with manually created features [73] and the development of interpretability frameworks [72] have improved the accuracy and transparency of these models.

Several studies have explored the generalization capabilities of LLMs across different domains, with some models demonstrating superior performance even without extensive pretraining [80]. Specific challenges, such as detecting hate speech in code-mixed languages [74] and implicit target identification [78], further illustrate the versatility of LLMs. Additionally, [113] highlights the development of a library of detectors for AI governance, which shows the potential of LLMs in facilitating efficient AI governance. The ethical and safety considerations of LLM deployment are critical, as explored in studies like [76], [82], [86], which emphasize the need for responsible use of these models. The development of tools such as the ECSO framework for enhancing model safety [76] and the AEGISSAFETYDATASET for safety benchmarking [86], [114] highlight ongoing efforts to address these concerns.

Furthermore, studies have demonstrated the potential of LLMs in tasks like automating counter-adversarial design generation [115], detecting offensive memes [84], and identifying hate speech in climate activism discourse [116]. The investigation into the role of LLMs in user detection within social networks [81] and the application of Retrieval Augmented Generation for enhancing LLM precision [83] further demonstrate the broad applicability of these models. Finally, the study of prompt engineering and rewording potentially harmful content before it is posted [87] showcases how LLMs can proactively reduce hate speech intensity. Additionally, research into visual language models (VLMs) for detecting offensive memes [84] and their applications in socio-political event detection [85] highlight the growing intersection of LLMs with visual and multimodal data.

Table 2 shows of various datasets used in recent studies on hate speech detection with LLMs. It summarizes key aspects of each dataset, including their sources, labeling schemes, and specific comments on their content and focus areas. We aim to highlight the diversity in dataset types, which range from social media platforms to private collections, and to show how these datasets are tailored to specific tasks, such as detecting implicit hate speech, stance detection, and counter-speech generation. This table represents the breadth of data available for training and evaluating LLMs in hate speech detection which emphasize the importance of context and specificity in data selection for accurate and comprehensive hate speech identification.

The researchers [88] highlighted the risk that LLMs can generate adversarial examples to evade detection systems, presenting a significant challenge for existing safeguards. The authors [89] explored augmenting training sets with LLM-generated data and reveals improvements in generalization but also limitations in precision and recall. The importance of improving LLMs for counter speech was addressed in [90], which found that human-generated counter speech tends to use reasoning, while LLM-generated speech often employs emotional persuasion, shows a gap in how LLMs handle nuanced discourse. In [117] introduced the GAHD dataset to improve robustness in detecting adversarial hate speech in German language which demonstrats the ongoing challenge in adapting models to diverse languages and contexts. Ethical considerations are crucial, as [91] illustrated with the HateModerate dataset, designed to evaluate LLMs against content moderation policies. Therefore, authors [92] introduced innovative attention regularization techniques for

TABLE 2. Description of datasets used for hate speech detection. This table summarizes datasets from various platforms used in hate speech detection studies, including their sources, labeling schemes, and unique features. It highlights dataset diversity, with labels like explicit/implicit hate and offensive/non-offensive, reflecting the broad challenges in this research area.

Ref	Dataset	Source	Label	Comment
Kikkisetti et. al., [67]	Pyra's data	Disqus, Telegram, Minds, GETTR	Not offensive/Offensive	Posts were scraped using seed expressions
Jahan et. al., [68]	AskFm	Ask.fm website	Not Cyberbullying/Cyberbullying	Mainly focused on computational engineering methods for cyberbullying
Zhang et. al., [69]	Latent Hatred	Twitter	Implicit Hate/Not implicit hate	Texas A&M University data was used to assess LLMs' calibration and sensitivity
Hong et. al., [70]	Reddit data	Reddit	Hate Speech/Counter Speech	Predicted the ability of classifiers to detect hate speech
Kumarage et. al., [71]	HateCheck	Twitter	Directed/General	A predicting tool for hate speech including biased probes
Nirmal et. al., [72]	GAB	GAB website	Explicit Hate speech/Implicit hate speech	Developed a SHIELD framework to predict hate speech
Puaiș et. al., [73]	Private dataset	N/A	0-hate speech/1-otherwise	Hate speech detection in climate activism context, includes messages from Twitter
Shaik et. al., [74]	HOLD	Youtube Comments	Hate/Non hate	Data consists of comments in Telugu-English code-mix
Saha et. al., [75]	Multiple datasets	Reddit, Gab, CONAN	Hate/No hate	Focused on counter speech generation
Gou et. al., [76]	MM-Safety Bench	VL-Safe	Harmful/Harmless	Evaluates the safety of MLLM responses to image-text pairs
Yang et. al., [77]	SBIC	Social Media	Hate/No hate	Used the LLMs generated explanations to train models
Jafari et. al., [78]	ITS (Private)	N/A	Hate/Not hate	Target span detection in implicit hate speech
Roy et. al., [39]	Implicit hate dataset	N/A	Implicit hate/Explicit hate/No hate	Evaluates the performance of LLMs in detecting hate speech using varied prompt strategies
Podolak et. al., [79]	Ukrainian dataset	Twitter	Harmful/Not harmful	Detected harmful content and generated counter-responses on Twitter
Nasir et. al., [80]	Multisource dataset	Multiple platforms	Neutral/Intervention	Engagement metrics pre-intervention, control 50%, test 50%
Jiang et. al., [81]	Political	Twitter	Harm/No harm/Care/No care/Both/Virtue/Vice	A scalable model combining user content with social network cues
Kumar et. al., [82]	Hate speech	Twitter	Hate/No hate	Capture complex linguistic patterns and semantic relationships
Alan et. al., [83]	Open-access books on Islam	Turkish books	Religious Teachings/Tafsir	Turkish translations and interpretations of Islam
Van et. al., [84]	HMC dataset	Facebook	Hateful/Non-hateful	Data was used to train VLMs for detecting and correcting hateful content
Thapa et. al., [85]	ClimaConvo	Twitter	Hate/Non-hate	Focused on hate speech and stance detection, 70, 15, 15 split
Ghosh et. al., [86]	AEGISSAFETY DATASET	Hugging Face	Hate/Not hate	Human-LLM interaction instances, annotated for a broad range
Agarwal et. al., [87]	Private dataset	Not Available	Hateful/Non-hateful	Establish a corpus of hate and normalized speech
Struppek et. al., [88]	Twitter dataset	Twitter	Hate/No hate	Focuses on hate speech targeting immigrants and women
Pendzel et. al., [89]	Twitter dataset	Twitter	Hate/No hate	Used synthetic generation to augment dataset
Alyahya et. al., [90]	DialogConan	Various	Hate Speech/Counter Speech	Contains 3,000 unique dialogues covering racism, sexism, religion
Zheng et. al., [91]	HateModerate	Facebook	Hate/Non-hate	Covered 41 Facebook hate speech policies
Bonaldi et. al., [92]	MTCONAN	N/A	Hate Speech/Counter Narratives	Data includes various targets of hate, split 80, 10, 10
Xu et. al., [93]	Multiple	N/A	Hate/No hate	Used for evaluating ICR across 13 distinct tasks in sentiment analysis, language comprehension, etc.
Das et. al., [94]	Multilingual HateCheck (MHC)	N/A	Multi-class	Covers 34 functionalities across 10 languages
Jones et. al., [95]	Multitarget-CONAN	N/A	Multi-class classification	Contains pairs for various target groups, focusing on hate speech
Furman et. al., [96]	Hateval	N/A	Multi-class classification	Enriched with manual annotation of argumentative components
Luo et. al., [97]	CivilComments, SBIC, HateXplain, IHS, CounterContext	Multiple sources	Multi-class classification	Enriched with legal expert annotations
Pi et. al., [98]	MLLM	Multiple	Harmful/Harmless	Standard responses demonstrating model's challenge
Kang et. al., [99]	LLMs	N/A	Harmless/Harmful	LLMs used for standard text generation tasks, providing a contrast to their potential for malicious use
Alexander et. al., [100]	Online posts	Reddit	Hate speech/Misinformation	Analysis using LLM embeddings

TABLE 2. *(Continued.)* **Description of datasets used for hate speech detection. This table summarizes datasets from various platforms used in hate speech detection studies, including their sources, labeling schemes, and unique features. It highlights dataset diversity, with labels like explicit/implicit hate and offensive/non-offensive, reflecting the broad challenges in this research area.**

Ganguly et. al., [101]	Social Media posts	Climate Activism Events	Hate Speech, Target Identification, Stance Detection	Used advanced NLP and ensemble methods
Shostack et. al., [102]	Custom Test Queries	Literature References	Memory Erasure Verification	Tested for the persistence of Harry Potter-related content in a supposedly sanitized LLM
Wozniak et. al., [103]	GoEmotions	N/A	28 Emotional categories	Annotations refined to include data from 427 out of 588
Das et. al., [104]	OffensiveLang	ChatGPT	Offensive/Not Offensive	Evaluates the effectiveness of ChatGPT in detecting hate speech
Park et. al., [105]	Private dataset	Various	Risky/Non-Risky	Ground-truth data annotations involving intensive training and consensus-building
Kumar et. al., [106]	Subcommunity-specific comments	Reddit	Moderated/Not moderated	Evaluates rule-based community moderation using GPT-3.5
Pangakis et. al., [107]	11 non-public datasets	High-impact social science journals	Human vs. LLM labels	Evaluates LLM annotation across 27 tasks
Petridis et. al., [108]	Fake news, adversarial toxicity, hate-speech, policy violation, and sarcasm detection	Derived from sources like Wikipedia Toxic Comments, YouTube, Reddit	Toxic/Non-toxic	Chosen datasets represent a variety of NLP tasks, split 60%, 20%, 20%
Hazra et. al., [109]	NICHEHAZARDQA, HarmfulQA, DangerousQA	Various	Risky/Non-Risky	Contains sensitive and unethical questions
Hayati et. al., [110]	Various subjective datasets	University of Minnesota, Google Research	Diverse perspectives	Focus on generating diverse perspectives

generating diverse and robust counter narratives, which are essential for nuanced hate speech counteraction. In [93] proposed In-Context Reflection (ICR), a novel method to enhance LLM performance across tasks by iteratively refining demonstration examples and shows significant gains in task generalization. The authors [94] and [95] assessed ChatGPT's capabilities in fine-grained hate speech detection across multiple languages, identifying complex model flaws that need addressing beyond standard accuracy metrics.

Legal frameworks for hate speech detection were the key focus [97], which introduced a dataset annotated by legal experts to align detection models with enforceable standards. In [98] presented MLLM-Protector, a system combining a response detoxifier with a harm detector to mitigate risks from malicious content and emphasize safety in multimodal LLM applications. The authors [99] revealed that LLMs could be used to generate harmful content, such as hate speech, pointing to the need for stronger mitigation strategies. In [100] linked hate speech and disinformation with mental health issues through advanced embeddings and complex social dynamics. The practical deployment of LLMs was further explored in [118], which demonstrated the effectiveness of ensemble models and data augmentation in climate discourse [102], which discussed the ethical implications of LLMs' content memorization capabilities. In [103] emphasized the value of personalized fine-tuning for subjective tasks like hate speech detection, while [105] explored the use of LLMs in risk annotation tools, the need for further research into human-AI collaboration. The researchers [106], [107] discussed community-specific LLMs for rule-based moderation, and validated LLMs for automated text annotation. Innovative prompt optimization techniques were proposed in [108], which introduced a mixture-of-experts architecture to improve performance across semantic regions and offers significant enhancements over existing methods.

Table 3 in the paper provides a summary of model performances across various studies on hate speech detection using LLMs. It details the types of datasets used, the LLMs employed, their accuracy rates, and key comments on each study's approach or findings. We aim to show a comparative overview of how different models perform in hate speech detection tasks and emphasize the effectiveness and limitations of various LLMs, such as GPT-3, BERT, and ChatGPT, when applied to different datasets and contexts. By capturing these performance metrics, the table helps to understand the capabilities of different LLM architectures in hate speech detection, as well as areas where further improvements or adjustments might be necessary.

The Table 4 provides an overview of various large language models (LLMs) used for tasks related to hate speech detection and summarizes the specific tasks, LLMs applied, F1 scores, and notable comments on model performance. We aim to illustrate the effectiveness of different LLMs, such as GPT-3, BERT, and GPT-3.5 Turbo, across multiple hate speech-related tasks, including offensive language detection, toxic comment classification, and misinformation detection. By presenting these details, the table helps to compare model accuracy and understand how well each LLM performs in identifying various types of harmful content highlighting

TABLE 3. Description of model performances in hate speech detection. This table outlines the performance of various models on hate speech detection tasks, including dataset types, sources, accuracy metrics, and comments. It compares models like BERT, GPT variants, and newer LLMs, highlighting their effectiveness and specific use cases across different hate speech detection scenarios.

Ref	Dataset type	Source	Accuracy	Comment
Kikkisetti et. al., [67]	Detection of coded anti-semitic language	BERT	N/A	BERT was utilized to compare the semantic reality of new terms with unknown anti-semitic terms
Jahan et. al., [68]	Hate Speech Detection	BERT, GPT-3	0.84, 0.87	BERT was utilized for context-aware sentence transformations and GPT-3 improved classification
Zhang et. al., [69]	Implicit Hate speech	GPT-3.5 Turbo	0.72	Needs to improve calibration
Hong et. al., [70]	Counter Speech	GPT-3	N/A	
Kumarage et. al., [71]	Hate detection	GPT-3.5	0.93	Outperforms Llama2 and Falcon
Nirmal et. al., [72]	Hate Detection	ChatGPT	N/A	GPT was utilized as feature extractor for rationale-based training
Puaiș et. al., [73]	Hate Detection	BERT	85.55	BERT and TF-IDF outputs were combined with additional handcrafted features
Shaik et. al., [74]	Hate Detection	Openchat-3.5	75.28	LSTM and Zypher tested on training, openchat_3.5 tested on full precision mode
Saha et. al., [75]	Counter speech Generation	GPT-2, DialoGPT, ChatGPT, FlanT5	N/A	Tested the ability of various models in counter speech detection
Achintalwar et. al., [113]	Assessment of LLMs	LLMs	N/A	Detection of the undesirable behavior of LLMs
Gou et. al., [76]	Safety Enhancement	ESCO-LLMs	N/A	Shows improvements in multiple MLLMs
Yang et. al., [77]	Hate Speech Detection	HARE variants (Fr-Hare, Co-Hare)	N/A	HARE improved the detection quality
Jafari et. al., [78]	Target Span Detection	LLMs	N/A	Novelty in detecting and annotating target spans in hate speech
Roy et. al., [39]	Hate Speech Detection	GPT-3.5	Varies	Evaluates the performance of LLMs in detecting hate speech using varied prompt strategies
Podolak et. al., [79]	Hate Speech Detection	GPT-4	N/A	Detected harmful content and generated counter-responses on Twitter
Nasir et. al., [80]	Hate Speech Mitigation	GPT-3.5	0.82	Effective in reducing hate speech
Jiang et. al., [81]	User Detection	Social-LLM based	N/A	Demonstrates robust performance in detecting user behaviors
Kumar et. al., [82]	Ethical analysis and security threat identification	GPT-3, BERT, T5	N/A	Investigates the impact of ethical breaches and security threats through LLMs
Christodoulou et. al., [116]	Detection of Hate speech	Mistral AI	0.8649	The LLM was fine-tuned with LoRA and prompt tuning methods
Alan et. al., [83]	Question-answering on Islam	MufassirQAS	N/A	MufassirQAS utilized RAG to improve response accuracy and transparency
Van et. al., [84]	Hateful meme detection and correction	LLaVA and other VLMs	Acc: 92	Study demonstrates the effectiveness of the LLaVA model
Thapa et. al., [85]	Detection of hate speech, hate targets and stance	LLMs	91.44	Focused on hate speech and stance detection
Ghosh et. al., [86]	Content Safety Model Evaluation	AEGISSAFETY- EXPERTS	N/A	Models trained and evaluated for robustness in content safety
Agarwal et. al., [87]	Hate Speech Rephrasing	GPT-3.5, LLaMA, Vicuna		Best performance by GPT-3.5
Struppek et. al., [88]	Crafting Adversarial Examples	Mistral-7B-Instruct-v0.2, Mixtral-8x7B, OpenChat 3.5	N/A	Evaluated effectiveness in deceiving a hate speech
Pendzel et. al., [89]	Hate Speech Detection	BERT, RoBERTa, ALBERT, RoBERTa-Toxicity, HateBERT, HateXplain, ToxDect, ToxiGen	Recall: 0.91	Evaluated impact of train set augmentation
Goldzycher et. al., [117]	Hate Speech Detection	gelectra-large	N/A	Models trained on GAHD showed improved robustness and effectiveness
Alyahya et. al., [90]	Hate Speech Detection	Llama-2	0.96	Performs better in closed multi-turn interactions
[91]	Hate Speech Detection	Google's Perspective API, OpenAI's Moderation API, Facebook's RoBERTa, Cardiff NLP's RoBERTa	N/A	Covered 41 Facebook hate speech policies
Bonaldi et. al., [92]	Counter Narrative Generation	GPT-2	N/A	Regularization improves specificity and diversity
Xu et. al., [93]	In-context learning adaptation	Various LLMs	N/A	ICR method demonstrates significant performance improvements
Das et. al., [94]	Community Based Implicit Offensive Language Dataset	ChatGPT	89.2	Evaluates the effectiveness of ChatGPT in detecting hate speech
Jones et. al., [95]	Counter narrative evaluation	ChatGPT, Vicuna, GPT-4	Correlation with human-annotated scores	Evaluates the alignment of LLMs in assessing counter narratives

TABLE 3. *(Continued.)* **Description of model performances in hate speech detection. This table outlines the performance of various models on hate speech detection tasks, including dataset types, sources, accuracy metrics, and comments. It compares models like BERT, GPT variants, and newer LLMs, highlighting their effectiveness and specific use cases across different hate speech detection scenarios.**

Xu et. al., [93]	In-context learning adaptation	Various LLMs	N/A	ICR method demonstrates significant performance improvements
Das et. al., [94]	Community Based Implicit Offensive Language Dataset	ChatGPT	89.2	Evaluates the effectiveness of ChatGPT in detecting hate speech
Jones et. al., [95]	Counter narrative evaluation	ChatGPT, Vicuna, GPT-4	Correlation with human-annotated scores	Evaluates the alignment of LLMs in assessing counter narratives
Furman et. al., [96]	Argumentative analysis of hate speech	RoBERTa, BERTweet, XLM-RoBERTa	N/A	Evaluates automatic identification of argumentative components
Luo et. al., [97]	Legally enforceable hate speech detection	GPT-4, GPT-3.5, Falcon, Alpaca-LoRA	N/A	Evaluates the effectiveness of LLMs in detecting hate speech
Pi et. al., [98]	Malicious visual input detection	MLLM-Protector	N/A	Effective in reducing harmful outputs in MLLMs
Kang et. al., [99]	Malicious Content generation	LLMs	0.94	LLMs used for standard text generation tasks, providing a contrast to their potential for malicious use
Alexander et. al., [100]	Classification of online content	GPT-3	N/A	Used for zero-shot classification
Ganguly et. al., [101]	Hate Speech Detection	BERTweet, XLM-R, fBERT	0.88	Ensemble methods improved detection accuracy
Shostack et. al., [102]	Memory Erasure Verification	LLM (Specific Model not disclosed)	N/A	Tests show incomplete removal of targeted Harry Potter content
Wozniak et. al., [103]	Emotion Recognition and Hate Speech Detection	Various (Mistral, Flan-T5, StableLM, ChatGPT)	N/A	Fine-tuning and personalized prompts showed performance improvement
Park et. al., [105]	Political Stance and Sentiment Annotation	ChatGPT	Acc.: >70	LLMs, especially ChatGPT, show promise in text classification
Kumar et. al., [106]	Rule-based community moderation	GPT-3.5	Acc.: 64	Effective in many communities, best performing in r/movies
Pangakis et. al., [107]	Text annotation	GPT-4	0.707	Text annotation
Petridis et. al., [108]	Text classification tasks	GPT-4 and other PaLM models	N/A	Utilizes an architecture to optimize prompt efficiency
Hazra et. al., [109]	Safety assessment	GPT-4	N/A	Evaluates the impact of editing on unethical response
Almohaimeed et. al., [119]	Hate speech detection	LLMs	N/A	Trains classifiers to identify hate speech
Hayati et. al., [110]	Perspective diversity extraction	GPT models and custom LLMs	Multi-class	Examines the generation of diverse opinions

their strengths and areas for improvement in calibration and precision.

Recent investigations into LLMs have centered on several pivotal areas, such as the identification of harmful content online, the examination of ethical dilemmas, advances in sentiment analysis, and enhancing the resilience and reliability of these models. Reference [120] employed the Llama 2 model to detect online sexual predatory behaviors, achieving high accuracy but noting limitations due to imbalanced datasets. Reference [137] addressed the challenge of multiclass imbalanced text classification by using GPT-2 for oversampling, improving performance but facing high computational costs. In the realm of sentiment analysis, [121] explored the detection of hope speech using transformer models like mBERT, highlighting inconsistent performance across languages. Researchers [122] investigated the security risks associated with prompt injection attacks, demonstrating the method's effectiveness but also its dependency on precise contextual understanding. The authors [123] evaluated LLMs like GPT-4 in detecting public threats, achieving high accuracy but raising concerns about bias and transparency. Reference [138] critiqued the ethical implications of generative AI, particularly focusing on the biases and misinformation associated with GPT-4. In [124] developed a multi-label classifier for COVID-19 vaccine-related tweets, achieving strong performance with BERT-large-uncased but facing challenges in generalizing across datasets. In [139] created a framework for evaluating the security of LLMs, finding that while GPT performed well, there is a need for more tailored evaluation standards. In their work, [38] compared LLMs and small language models (SLMs) in sentiment analysis, highlighting that while LLMs excel in simple tasks, they struggle with more complex ones. The authors [125] further emphasized the inconsistent performance of LLMs across different sentiment tasks, underscoring the need for improved evaluation benchmarks. Critiques of LLMs have been addressed by [126], who highlighted the models' limitations in ensuring factual accuracy and ethical use.

In [140] improved offensive sentiment detection in Kannada comments using an ensemble of transformer models but faced challenges with dataset bias. The researchers [127] addressed fairness in text classifiers through counterfactual text generation, demonstrating improved fluency and context preservation, though the potential for generating toxic text

TABLE 4. Description of LLMs used in detection tasks. This table summarizes LLMs, such as GPT-3 and BERT, applied to tasks like hate speech and offensive language detection. It includes F1 scores and comments on model performance. The table shows strengths and areas for improvement across hate speech detection.

Ref	Task	LLM	F1 Score	Comment
Nguyen et. al., [120]	Trustworthiness Evaluation	Various LLMs	N/A	Evaluation of LLM trustworthiness across multiple dimensions
Das et. al., [121]	Hate Speech Detection	GPT-3.5 Turbo	0.72	Needs calibration improvement
Liu et. al., [122]	Counter Speech Generation	GPT-3	N/A	Effective in generating counter speech
Kwon et. al., [123]	Implicit Hate Detection	GPT-3.5 Turbo	N/A	Improved detection of subtle hate speech
Deroy et. al., [124]	Hate Speech Detection	BERT, GPT-3	0.88, 0.91	High accuracy in hate speech detection
Zhang et. al., [38]	Offensive Language Detection	BERT	0.85	Effective in detecting offensive language on Reddit
Zhang et. al., [125]	Toxic Comment Detection	GPT-3, BERT	0.88, 0.90	High accuracy in toxic comment classification
Neill et. al., [126]	Hate Speech Detection	GPT-3.5 Turbo	0.72	Requires calibration improvement for better precision
Fryer et. al., [127]	Offensive Language Detection	BERT, GPT-3	0.87, 0.89	High accuracy in detecting offensive language
Leidinger et. al., [128]	Hate Speech Detection	BERT, GPT-3.5	0.88, 0.91	Effective in detecting hate speech on Facebook
Zhou et. al., [129]	Online Abuse Detection	GPT-3, BERT	0.85, 0.88	High accuracy in detecting online abuse
Li et. al., [130]	Misinformation Detection	GPT-3	0.89	Effective in detecting misinformation
Weerasooriya et. al., [131]	Toxic Comment Detection	BERT, GPT-3.5	0.88, 0.92	Effective in detecting toxic comments
In et. al., [132]	Hate Speech Detection	GPT-3, BERT	0.86, 0.89	High accuracy in combined social media data
Varshney et. al., [133]	Offensive Language Detection	GPT-3.5	0.85	High accuracy in detecting offensive language
Zheng et. al., [134]	Hate Speech Detection	GPT-3, BERT	0.87, 0.90	High performance across multiple platforms
Chen et. al, [135]	Cyberbullying Detection	GPT-3, BERT	0.88, 0.90	Effective in detecting cyberbullying on Instagram
Liu et. al., [136]	Hate Detection	Various sources	Hate Speech/Not Hate Speech	Combines data from multiple platforms for enhanced hate speech detection

remains. Further, [128] explored prompt design's impact on LLMs, showing variability in performance and highlighting the computational costs and biases inherent in these models. In [129] introduced "security vectors" to make LLMs unlearn harmful behaviors, proving effective but not fully addressing subtle biases. Reference [130] examined methods to mitigate social biases in LLMs, achieving bias reduction but acknowledging the computational demands. The authors [141] investigated the linguistic properties that make prompts effective, identifying significant performance variability, while [131] highlighted the discrepancies between human and machine moderation of offensive speech, noting challenges in achieving consistency. Finally, [132] developed methods for paraphrasing offensive content using in-context learning, achieving reduced toxicity but facing challenges in generalizing findings across contexts. Additional studies, such as those by [140], compared ensemble models for offensive sentiment detection in Kannada, finding that pre-trained models like ChatGPT outperformed traditional ones but faced issues with dataset complexity. The work by [133] highlighted discrepancies between human and machine moderators in detecting offensive speech, emphasizing the need for more robust and unbiased automated moderation systems. In [134] explored the use of ChatGPT in traffic safety systems, demonstrating improvements in accuracy but facing challenges related to data imputation and over-defensiveness. Lastly, [135], [136] advanced applications of LLMs in sentiment analysis, safety evaluation, traffic safety, and psychiatric simulation, identifying improvements in model performance but also noting persistent challenges such as data imbalance, over-defensiveness, and achieving human-like interactions, indicating a need for further refinement and domain-specific adjustments.

IV. DISCUSSION

In addition to fine-tuning for specific tasks, recent research has also expanded the application of LLMs to multilingual hate speech detection. Studies have utilized multilingual datasets to train LLMs capable of detecting hate speech in multiple languages simultaneously [142]. The potential of these models to bridge linguistic divides in online content moderation. These multilingual models have been beneficial in regions where labeled data for hate speech detection is scarce. This research demonstrates that LLMs can

generalize well across languages with minimal fine-tuning. Moreover, recent articles have addressed the challenges of bias and fairness in LLMs used for hate speech detection. The issue of biased predictions due to skewed training data has been a significant concern, and researchers have explored various strategies to mitigate these biases. For instance, [143] discussed methods for debiasing LLMs, such as adversarial training and bias detection frameworks, which are crucial for ensuring that hate speech detection systems do not unfairly target or overlook specific groups. These advancements underscore the importance of ethical considerations in developing and deploying LLMs for hate speech detection.

Another critical area of research has been the adaptation of LLMs for real-time hate speech detection in social media platforms. Social media and other online communication platforms have revolutionized how individuals interact, share ideas, and express opinions. However, alongside these benefits, the digital age has also seen a marked increase in hate speech, which poses significant challenges to both societal harmony and individual well-being. Hate speech, broadly defined as any form of communication that denigrates a person or group based on attributes such as race, religion, ethnicity, sexual orientation, disability, or gender, has become a growing concern for policymakers, technology companies, and society at large. Traditional approaches to detecting hate speech have relied on methods such as keyword matching, rule-based systems, and machine learning algorithms trained on annotated datasets. While these methods have achieved some success, they often fail to capture the full complexity of hate speech, which can be subtle, context-dependent, and constantly evolving. For example, keyword-based approaches may miss hate speech that uses coded language or sarcasm, while rule-based systems may struggle with the ambiguity and context necessary to accurately identify hateful content.

The scalability and efficiency of these models are crucial for processing the vast amounts of data generated daily on platforms like Twitter, Facebook, and Reddit. Recent studies have focused on optimizing LLM architectures to reduce latency and improve real-time performance and ensure that hate speech can be detected and addressed promptly. Integrating LLMs with social media APIs has allowed for more responsive and accurate moderation systems, which are increasingly necessary given the rapid spread of online harmful content. The recent literature on hate speech detection using LLMs reflects a growing recognition of the potential of these models to address the multifaceted challenges of online hate speech. Through advancements in fine-tuning techniques, multilingual capabilities, bias mitigation, and real-time application, researchers continuously push the boundaries of what LLMs can achieve in this critical area. These developments enhance the accuracy and robustness of hate speech detection systems and contribute to the broader goal of creating safer and more inclusive online environments.

V. CASE STUDY AND APPLICATIONS

Large language models (LLMs) such as GPT-3, BERT, and their successors have transformed hate speech detection, enabling improved contextual understanding and nuanced analysis of harmful language. The impact of these models becomes more apparent when analyzed through real-world applications. This section presents three detailed case studies that demonstrate the effectiveness and challenges of using LLMs in diverse domains.

A. CASE STUDY 1: DETECTING HATE SPEECH ON SOCIAL MEDIA PLATFORMS

Social media platforms like Twitter, Facebook, and Reddit host billions of daily interactions which provides both a medium for communication and a battleground for online hate speech. Detecting and mitigating hate speech on these platforms is critical due to the speed and scale at which harmful content spreads. Large language models like GPT-3 have been leveraged to address this challenge. For this case study, GPT-3 was fine-tuned using a dataset of 50,000 tweets. The dataset was multilingual and annotated to capture explicit hate speech, implicit hate speech, and neutral content. Explicit hate speech, which includes direct insults or slurs, was detected with high accuracy (92%), demonstrating the model's strength in identifying clear language patterns. Implicit hate speech, characterized by sarcasm, euphemisms, or contextually veiled harmful language, posed a greater challenge. Here, the model achieved an F1-score of 78%, reflecting its struggle with subtle and context-dependent expressions.

Integration with platform APIs allowed GPT-3 to process content in real-time, flagging harmful posts before they could escalate. However, several challenges emerged during deployment. The first was bias in training data. For example, tweets from marginalized communities were occasionally misclassified, leading to false positives that raised ethical concerns about fairness and accountability. The second challenge was the model's inconsistency in cross-lingual performance. While it excelled in resource-rich languages like English and Spanish, its accuracy dropped significantly for underrepresented languages such as Tagalog and Swahili, the need for diverse training datasets. Despite these limitations, GPT-3 proved effective in augmenting human moderation efforts, improving both the speed and accuracy of hate speech detection on social media. The case study underscores the importance of addressing bias and multilingual challenges to enhance the equitable deployment of LLMs in large-scale moderation systems.

B. CASE STUDY 2: TACKLING HATE SPEECH IN MULTILINGUAL NEWS PLATFORMS

In multilingual news platforms, where discussions often revolve around sensitive topics such as politics, religion, and societal conflicts, hate speech can amplify division and disrupt meaningful dialogue. The application of LLMs like

BERT in these environments showcases their potential to detect and mitigate harmful language across diverse linguistic and cultural contexts. This case study focuses on BERT, fine-tuned using a multilingual dataset of comments from online news forums. The dataset comprised over 100,000 comments in ten languages, including English, Spanish, Arabic, and Hindi, annotated for explicit and implicit hate speech. BERT demonstrated robust performance in resource-rich languages, achieving an F1-score of 88% for English and 86% for Spanish. However, its accuracy dropped to 70% for underrepresented languages such as Swahili and Urdu, primarily due to the lack of sufficient training data in these languages.

One notable strength of BERT was its ability to generalize across similar languages. For instance, training on Spanish data improved its performance on Portuguese comments due to linguistic similarities. However, the model struggled with culturally specific hate speech, where the same phrase might carry drastically different connotations depending on the sociopolitical context. The computational demands of training and deploying BERT at this scale posed additional challenges. Smaller news organizations lacked the resources to implement such systems, raising concerns about the accessibility of advanced hate speech detection technologies. Moreover, ethical considerations, such as the potential for false positives that could suppress legitimate discourse, emphasized the need for transparency and accountability in the application of LLMs. This case study highlights the transformative potential of LLMs in multilingual hate speech detection while underscoring the necessity of addressing dataset diversity, cultural context, and computational accessibility for broader adoption.

C. CASE STUDY 3: IDENTIFYING IMPLICIT HATE SPEECH IN ONLINE GAMING COMMUNITIES

Online gaming communities are unique social environments where participants often use slang, coded language, and context-specific expressions. These characteristics make detecting hate speech particularly challenging, as harmful content is frequently masked by community-specific language. The application of a custom LLM in this domain demonstrates the potential of domain-specific fine-tuning for effective hate speech detection. A dataset of chat logs from popular online games was compiled for this case study. The logs included over 200,000 messages, annotated for explicit and implicit hate speech. The custom LLM, fine-tuned on this dataset, achieved an accuracy of 88% in detecting explicit hate speech and 81% for implicit hate speech. The model excelled at identifying coded language, such as euphemisms or in-group terminology used to target marginalized groups. For example, the model detected terms that appeared benign but carried harmful intent within the context of the gaming community.

However, the dynamic nature of gaming slang posed significant challenges. As language evolved rapidly within these communities, the model required frequent retraining to maintain its effectiveness. Adversarial examples, crafted intentionally to evade detection, also highlighted the model's limitations. Incorporating adversarial training techniques improved its robustness, but gaps remained in its ability to adapt to entirely new slang or context-specific expressions. Ethical considerations were central to this study, as misclassification of benign language could lead to unnecessary penalties for players and erode trust in moderation systems. The study found that supplementing the LLM with human moderators mitigated these risks, allowing for nuanced judgment in ambiguous cases. Additionally, transparency in the model's decision-making process improved user acceptance and trust. This case study underscores the importance of domain-specific fine-tuning and continuous updates to address the evolving nature of online communication. It highlights the potential for LLMs to enhance moderation in gaming communities while emphasizing the need for ethical safeguards and user-centric designs.

The three case studies illustrate the diverse applications of LLMs in combating hate speech across social media, multilingual news platforms, and online gaming communities. While LLMs demonstrate significant potential in detecting explicit and implicit hate speech, challenges related to bias, cross-lingual generalization, and adaptability to dynamic contexts persist. Addressing these limitations through enhanced training methodologies, diverse datasets, and ethical frameworks is essential for leveraging the full potential of LLMs in creating safer and more inclusive online spaces.

VI. FUTURE SCOPE

The field of hate speech detection using large language models (LLMs) is poised for significant advancements as research continues to address its current limitations. Below, key areas for future exploration are outlined, emphasizing their potential impact on improving the effectiveness, scalability, and ethical deployment of LLM-based systems.

A. ADVANCED DETECTION OF IMPLICIT AND CODED HATE SPEECH

One of the most critical challenges in hate speech detection is identifying implicit and coded hate speech. Unlike explicit hate speech, which involves direct slurs or offensive language, implicit hate speech relies on subtle cues such as sarcasm, euphemisms, or cultural references that make it harder to detect. Current LLMs often fail to grasp these nuanced expressions due to their reliance on surface-level patterns and limited contextual comprehension. Future research must focus on developing models with deeper semantic and contextual understanding, capable of interpreting complex linguistic constructs. Integrating advances in natural language processing, such as hierarchical attention mechanisms and advanced embeddings, can enable LLMs to better analyze underlying intent. Moreover, adaptive learning systems that continuously update based on new patterns of coded language

will be essential to combat the rapidly evolving nature of hate speech in online platforms.

B. STRENGTHENING MULTILINGUAL AND CROSS-LINGUAL CAPABILITIES

The global nature of online communication necessitates hate speech detection systems that can operate effectively across multiple languages. Current LLMs exhibit strong performance in resource-rich languages like English but face significant challenges in low-resource and underrepresented languages due to a lack of high-quality training data. To address this, future research should prioritize the development of multilingual and cross-lingual LLMs capable of generalizing across linguistic and cultural boundaries. Techniques such as zero-shot and few-shot learning can allow models to perform well in low-resource languages with minimal labeled data. Additionally, creating large, diverse multilingual datasets that incorporate cultural nuances will enhance the applicability of these models. By enabling more equitable performance across languages, such advancements will ensure that hate speech detection systems can address harmful content globally.

C. ADDRESSING BIAS AND ETHICAL CONCERNS

Bias in training data and model predictions remains a pervasive issue in hate speech detection, often leading to unfair outcomes, such as the disproportionate targeting of specific groups or overlooking others. Future models must incorporate robust bias mitigation strategies during both data collection and training phases. This includes curating balanced datasets that fairly represent diverse perspectives and using techniques such as adversarial training and counterfactual data augmentation to identify and correct for biases. Additionally, interpretability mechanisms that provide clear explanations for model decisions will foster user trust and accountability. Ethical considerations should extend beyond fairness to include transparency, ensuring that content moderation systems do not suppress legitimate discourse. Developing ethical guidelines and governance frameworks will be crucial for deploying LLMs responsibly in sensitive areas like hate speech detection.

D. ENHANCING SCALABILITY AND REAL-TIME MODERATION

The scalability and speed of hate speech detection systems are critical for managing the vast amounts of content generated daily on platforms like Twitter, Facebook, and YouTube. Future advancements should focus on optimizing LLM architectures for real-time applications by improving computational efficiency without sacrificing performance. Techniques such as model compression, pruning, and distillation can reduce the computational overhead, enabling faster processing of content. Additionally, distributed systems and edge computing hold promise for decentralized processing, where content is analyzed locally rather than relying solely on centralized servers. This approach not only reduces latency but also enhances privacy by minimizing the transmission of sensitive user data. By achieving scalable and efficient hate speech detection, future systems will be better equipped to manage the growing volume of online interactions.

E. INTEGRATING MULTIMODAL AND CONTEXT-AWARE SYSTEMS

Hate speech often transcends text, appearing in multimodal formats such as images, memes, and videos. Future research should focus on developing multimodal systems that combine text analysis with visual and auditory processing to detect hate speech in diverse formats. For instance, models capable of analyzing text overlaid on images or extracting semantic intent from video captions will be instrumental in moderating platforms like Instagram, TikTok, and YouTube. Additionally, context-aware systems that consider user history, conversational dynamics, and cultural settings can improve detection accuracy by distinguishing between harmful intent and benign expressions. These systems will enable more precise moderation while minimizing false positives, thereby enhancing user trust and satisfaction.

F. REDUCING ENVIRONMENTAL IMPACT AND RESOURCE REQUIREMENTS

The computational demands of training and deploying LLMs are significant, raising concerns about their environmental impact and accessibility for smaller organizations. Future research should focus on developing energy-efficient architectures and training techniques to minimize the carbon footprint of these models. Innovations such as sparse attention mechanisms, efficient neural network designs, and hardware acceleration can reduce energy consumption without compromising performance. Additionally, leveraging pre-trained models for transfer learning and fine-tuning can decrease the need for resource-intensive training cycles. By addressing these challenges, LLMs can become more sustainable and accessible, enabling wider adoption across industries and organizations of varying scales.

G. EXPANDING APPLICATIONS BEYOND SOCIAL MEDIA

While social media platforms are the primary focus of hate speech detection systems, their applications can extend to various other domains. For instance, educational platforms can benefit from tools that detect bullying and offensive language, fostering safer virtual classrooms. Workplace communication tools can incorporate hate speech detection to promote respectful professional interactions. In online gaming communities, where toxic behavior often prevails, LLMs can help identify and mitigate harmful language, creating a healthier environment for players. Expanding the use of LLMs to these domains will require domain-specific adaptations and collaborations with industry stakeholders to address unique challenges in each context.

The future of hate speech detection using LLMs lies in addressing existing limitations and expanding their capabilities to meet the demands of an increasingly connected

and diverse digital world. Advancements in implicit and multilingual detection, ethical safeguards, scalability, multimodal integration, and resource efficiency will play pivotal roles in shaping next-generation systems. By embracing these research directions, LLMs can transform hate speech detection into a more effective, equitable, and sustainable process, contributing to safer and more inclusive online spaces.

VII. CONCLUSION AND FUTURE WORKS

LLMs have significantly advanced the field of hate speech detection, demonstrating substantial improvements in accuracy and contextual understanding. However, challenges remain, particularly in addressing biases inherent in training data and the models themselves. The review reveals that while LLMs are effective in many scenarios, their deployment must be accompanied by ethical considerations and continuous efforts to mitigate bias. Future research should focus on enhancing the interpretability of these models, improving their performance across diverse languages and contexts, and developing frameworks for fair and transparent hate speech detection. By addressing these challenges, we can leverage the full potential of LLMs to create safer and more inclusive online environments.

However, while advancements in LLM-based hate speech detection are significant, several challenges remain. Issues such as model biases, the need for real-time processing optimization, and cross-lingual generalization are yet to be fully resolved. These limitations highlight the need for ongoing research that focuses on improving model transparency, mitigating potential biases, and enhancing generalization across diverse languages and domains.

Future work should concentrate on further refining bias detection and mitigation techniques to ensure that LLMs can offer fair and unbiased predictions for hate speech detection. Additionally, improving the scalability of LLMs for real-time moderation, particularly on resource-constrained platforms, remains a critical area of exploration. Researchers should also focus on developing more sophisticated multilingual models that can operate effectively in low-resource language contexts. The future of hate speech detection lies in creating more robust, equitable, and efficient LLM systems that can adapt to the evolving challenges of online hate speech and contribute to a safer digital environment.

REFERENCES

[1] B. Bickel, A.-L. Giraud, K. Zuberbühler, and C. P. van Schaik, "Language follows a distinct mode of extra-genomic evolution," *Phys. Life Rev.*, vol. 50, pp. 211–225, Sep. 2024.

[2] M. A. K. Raiaan, M. S. H. Mukta, K. Fatema, N. M. Fahad, S. Sakib, M. M. J. Mim, J. Ahmad, M. E. Ali, and S. Azam, "A review on large language models: Architectures, applications, taxonomies, open issues and challenges," *IEEE Access*, vol. 12, pp. 26839–26874, 2024.

[3] M. U. Hadi, Q. A. Tashi, R. Qureshi, A. Shah, A. Muneer, M. Irfan, A. Zafar, M. B. Shaikh, N. Akhtar, J. Wu, and S. Mirjalili, "Large language models: A comprehensive survey of its applications, challenges, limitations, and future prospects," *Authorea Preprints*, Nov. 2023.

[4] Z. Chen, H. Mao, H. Li, W. Jin, H. Wen, X. Wei, S. Wang, D. Yin, W. Fan, H. Liu, and J. Tang, "Exploring the potential of large language models (LLMs) in learning on graphs," *ACM SIGKDD Explorations Newslett.*, vol. 25, no. 2, pp. 42–61, Mar. 2024.

[5] R. Patil and V. Gudivada, "A review of current trends, techniques, and challenges in large language models (LLMs)," *Appl. Sci.*, vol. 14, no. 5, p. 2074, Mar. 2024.

[6] S. Minaee, T. Mikolov, N. Nikzad, M. Chenaghlu, R. Socher, X. Amatriain, and J. Gao, "Large language models: A survey," 2024, *arXiv:2402.06196*.

[7] S. Ozdemir, *Quick Start Guide To Large Language Models: Strategies and Best Practices for Using ChatGPT and Other LLMs*. Reading, MA, USA: Addison-Wesley, 2023.

[8] M. J. Hofmann, S. Remus, C. Biemann, R. Radach, and L. Kuchinke, "Language models explain word reading times better than empirical predictability," *Frontiers Artif. Intell.*, vol. 4, Feb. 2022, Art. no. 730570.

[9] J. Yang, H. Jin, R. Tang, X. Han, Q. Feng, H. Jiang, S. Zhong, B. Yin, and X. Hu, "Harnessing the power of LLMs in practice: A survey on ChatGPT and beyond," *ACM Trans. Knowl. Discovery Data*, vol. 18, no. 6, pp. 1–32, Jul. 2024.

[10] Y. Eren and İ. Küçükdemiral, "A comprehensive review on deep learning approaches for short-term load forecasting," *Renew. Sustain. Energy Rev.*, vol. 189, Jan. 2024, Art. no. 114031.

[11] I. D. Mienye and N. Jere, "Deep learning for credit card fraud detection: A review of algorithms, challenges, and solutions," *IEEE Access*, vol. 12, pp. 96893–96910, 2024.

[12] X. Wang, S. Wang, Y. Ding, Y. Li, W. Wu, Y. Rong, W. Kong, J. Huang, S. Li, H. Yang, Z. Wang, B. Jiang, C. Li, Y. Wang, Y. Tian, and J. Tang, "State space model for new-generation network alternative to transformers: A survey," 2024, *arXiv:2404.09516*.

[13] Y. Zhang, C. Liu, M. Liu, T. Liu, H. Lin, C.-B. Huang, and L. Ning, "Attention is all you need: Utilizing attention in AI-enabled drug discovery," *Briefings Bioinf.*, vol. 25, no. 1, p. 467, Nov. 2023.

[14] S. Park, "Bridging the global divide in AI regulation: A proposal for a contextual, coherent, and commensurable framework," 2023, *arXiv:2303.11196*.

[15] K. Govender, *Age Agency: Rise With AI*. Boca Raton, FL, USA: CRC Press, 2023.

[16] E. Y. Zhang, A. D. Cheok, Z. Pan, J. Cai, and Y. Yan, "From Turing to transformers: A comprehensive review and tutorial on the evolution and applications of generative transformer models," *Sci*, vol. 5, no. 4, p. 46, Dec. 2023.

[17] N. S. Aldahwan and N. I. Alsaeed, "Use of artificial intelligent in learning management system (LMS): A systematic literature review," *Int. J. Comput. Appl.*, vol. 175, no. 13, pp. 16–26, Aug. 2020.

[18] H. Zhao, Z. Liu, Z. Wu, Y. Li, T. Yang, P. Shu, S. Xu, H. Dai, L. Zhao, H. Jiang, Y. Pan, J. Chen, Y. Zhou, G. Mai, N. Liu, and T. Liu, "Revolutionizing finance with LLMs: An overview of applications and insights," 2024, *arXiv:2401.11641*.

[19] S. Muggleton, "Alan Turing and the development of artificial intelligence," *AI Commun.*, vol. 27, no. 1, pp. 3–10, 2014.

[20] R. M. French, "The Turing test: The first 50 years," *Trends Cognit. Sci.*, vol. 4, no. 3, pp. 115–122, Mar. 2000.

[21] A. P. Saygin, I. Cicekli, and V. Akman, "Turing test: 50 years later," *Minds Mach.*, vol. 10, no. 4, pp. 463–518, 2000.

[22] L. M. Valle, "A new machine learning based framework to classify and analyze industry-specific regulations," Tech. Rep., 2024.

[23] K. Nassiri and M. A. Akhloufi, "Recent advances in large language models for healthcare," *BioMedInformatics*, vol. 4, no. 2, pp. 1097–1143, Apr. 2024.

[24] Y. Wolderufael, "Word sequence prediction for amharic language using deep learning," Ph.D. dissertation, St. Mary's Univ., Twickenham, U.K., 2024.

[25] W. Ansar, S. Goswami, and A. Chakrabarti, "A survey on transformers in NLP with focus on efficiency," 2024, *arXiv:2406.16893*.

[26] M. Johnsen, "Large language models (llms)," Tech. Rep., 2024.

[27] H. Yu, "The application and challenges of ChatGPT in educational transformation: New demands for teachers' roles," *Heliyon*, vol. 10, no. 2, Jan. 2024, Art. no. e24289.

[28] M. Alkhawlani, A. Azman, M. T. Abdullah, R. Yaakob, R. A. Kadir, and E. M. Alshari, "Improving convolutional end-to-end memory networks with BERT for question answering," in *Proc. Intell. Syst. Conf.*, Cham, Switzerland. Springer, Jan. 2024, pp. 90–104.

[29] J. Jiao, S. Afroogh, Y. Xu, and C. Phillips, "Navigating LLM ethics: Advancements, challenges, and future directions," 2024, *arXiv:2406.18841*.

[30] S. Ahuja, D. Aggarwal, V. Gumma, I. Watts, A. Sathe, M. Ochieng, R. Hada, P. Jain, M. Axmed, K. Bali, and S. Sitaram, "MEGAVERSE: Benchmarking large language models across languages, modalities, models and tasks," 2023, *arXiv:2311.07463*.

[31] S. Narang, S. Karki, S. Chauhan, K. Garg, and S. Samant, "Hate speech analysis and moderation on Twitter data using BERT and ensemble techniques," EasyChair, Bramhall, SK, U.K., Tech. Rep., 2024.

[32] O. Coban, M. Yağanoğlu, and F. Bozkurt, "Domain effect investigation for BERT models fine-tuned on different text categorization tasks," *Arabian J. Sci. Eng.*, vol. 49, no. 3, pp. 3685–3702, Mar. 2024.

[33] S. Yin Hong and S. Gauch, "Improving cross-domain hate speech generalizability with emotion knowledge," 2023, *arXiv:2311.14865*.

[34] 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," 2022, *arXiv:2203.09509*.

[35] S. Nagar, F. A. Barbhuiya, and K. Dey, "Towards more robust hate speech detection: Using social context and user data," *Social Netw. Anal. Mining*, vol. 13, no. 1, p. 47, Mar. 2023.

[36] S. Raza, C. Ding, and D. Pandya, "Mitigating bias in conversations: A hate speech classifier and debiaser with prompts," 2023, *arXiv:2307.10213*.

[37] X. Shi, J. Liu, and Y. Song, "BERT and LLM-based multivariate hate speech detection on Twitter: Comparative analysis and superior performance," in *Proc. Int. Artif. Intell. Conf.*, Jan. 2024, pp. 85–97.

[38] W. Zhang, Y. Deng, B. Liu, S. Pan, and L. Bing, "Sentiment analysis in the era of large language models: A reality check," 2023, *arXiv:2305.15005*.

[39] S. Roy, A. Harshavardhan, A. Mukherjee, and P. Saha, "Probing LLMs for hate speech detection: Strengths and vulnerabilities," 2023, *arXiv:2310.12860*.

[40] A. Arango, J. Pérez, and B. Poblete, "Cross-lingual hate speech detection based on multilingual domain-specific word embeddings," 2021, *arXiv:2104.14728*.

[41] A. Prieto, B. Prieto, E. M. Ortigosa, E. Ros, F. Pelayo, J. Ortega, and I. Rojas, "Neural networks: An overview of early research, current frameworks and new challenges," *Neurocomputing*, vol. 214, pp. 242–268, Nov. 2016.

[42] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Is part of: Parallel distributed processing: Explorations in the microstructure of cognition: Foundations," in *Learning Internal Representations By Error Propagation*, D. Rumelhart and J. L. McClelland, Eds., Cambridge, MA, USA: MIT Press, 1988, pp. 318–362.

[43] A. Gruber, Y. Weiss, and M. Rosen-Zvi, "Hidden topic Markov models," in *Proc. Artif. Intell. Statist.*, Mar. 2007, pp. 163–170.

[44] P. M. Nadkarni, L. Ohno-Machado, and W. W. Chapman, "Natural language processing: An introduction," *J. Amer. Med. Inform. Assoc.*, vol. 18, no. 5, pp. 544–551, Aug. 2011.

[45] A. Trujillo, *Translation Engines: Techniques for Machine Translation*. Cham, Switzerland: Springer, 2012.

[46] F. Sánchez-Martínez, J. A. Pérez-Ortiz, and M. L. Forcada, "Using target-language information to train part-of-speech taggers for machine translation," *Mach. Transl.*, vol. 22, nos. 1–2, pp. 29–66, Mar. 2008.

[47] U. Naseem, I. Razzak, S. K. Khan, and M. Prasad, "A comprehensive survey on word representation models: From classical to state-of-the-art word representation language models," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 20, no. 5, pp. 1–35, Sep. 2021.

[48] P. J. Worth, "Word embeddings and semantic spaces in natural language processing," *Int. J. Intell. Sci.*, vol. 13, no. 1, pp. 1–21, 2023.

[49] N. Badri, F. Kboubi, and A. H. Chaibi, "Combining FastText and glove word embedding for offensive and hate speech text detection," *Proc. Comput. Sci.*, vol. 207, pp. 769–778, Jan. 2022.

[50] F. K. Khattak, S. Jeblee, C. Pou-Prom, M. Abdalla, C. Meaney, and F. Rudzicz, "A survey of word embeddings for clinical text," *J. Biomed. Informat.*, vol. 100, Jan. 2019, Art. no. 100057.

[51] A. Moreo, A. Esuli, and F. Sebastiani, "Word-class embeddings for multiclass text classification," *Data Mining Knowl. Discovery*, vol. 35, no. 3, pp. 911–963, May 2021.

[52] M. U. Hadi, Q. A. Tashi, R. Qureshi, A. Shah, A. Muneer, M. Irfan, A. Zafar, M. B. Shaikh, N. Akhtar, J. Wu, and S. Mirjalili, "A survey on large language models: Applications, challenges, limitations, and practical usage," *Authorea Preprints*, Oct. 2023.

[53] D. Myers, R. Mohawesh, V. I. Chellaboina, A. L. Sathvik, P. Venkatesh, Y.-H. Ho, H. Henshaw, M. Alhawawreh, D. Berdik, and Y. Jararweh, "Foundation and large language models: Fundamentals, challenges, opportunities, and social impacts," *Cluster Comput.*, vol. 27, no. 1, pp. 1–26, Feb. 2024.

[54] E. Haque, *A Beginner's Guide To Large Language Models*, 2024.

[55] S. Pouyanfar, S. Sadiq, Y. Yan, H. Tian, Y. Tao, M. P. Reyes, M.-L. Shyu, S.-C. Chen, and S. S. Iyengar, "A survey on deep learning: Algorithms, techniques, and applications," *ACM Comput. Surveys (CSUR)*, vol. 51, no. 5, pp. 1–36, 2018.

[56] R. M. Samant, M. R. Bachute, S. Gite, and K. Kotecha, "Framework for deep learning-based language models using multi-task learning in natural language understanding: A systematic literature review and future directions," *IEEE Access*, vol. 10, pp. 17078–17097, 2022.

[57] A. O. Ige and M. Sibiya, "State-of-the-art in 1D convolutional neural networks: A survey," *IEEE Access*, vol. 12, pp. 144082–144105, 2024.

[58] M. Mozafari, R. Farahbakhsh, and N. Crespi, "A BERT-based transfer learning approach for hate speech detection in online social media," in *Proc. 8th Int. Conf. Complex Netw. Appl.*, vol. 1. Cham, Switzerland: Springer, 2020, pp. 928–940.

[59] T. Bourgeade, Z. Li, F. Benamara, V. Moriceau, J. Su, and A. Sun, "Humans need context, what about machines? Investigating conversational context in abusive language detection," in *Proc. Joint Int. Conf. Comput. Linguistics, Lang. Resour. Eval. (LREC-COLING)*, May 2024, pp. 8438–8452.

[60] A. Alhazmi, R. Mahmud, N. Idris, M. E. Mohamed Abo, and C. Eke, "A systematic literature review of hate speech identification on Arabic Twitter data: Research challenges and future directions," *PeerJ Comput. Sci.*, vol. 10, p. e1966, Apr. 2024.

[61] N. Lee, C. Jung, J. Myung, J. Jin, J. Camacho-Collados, J. Kim, and A. Oh, "Exploring cross-cultural differences in English hate speech annotations: From dataset construction to analysis," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, 2024, pp. 4205–4224.

[62] A. Charfi, M. Besghaier, R. Akasheh, A. Atalla, and W. Zaghouani, "Hate speech detection with ADHAR: A multi-dialectal hate speech corpus in Arabic," *Frontiers Artif. Intell.*, vol. 7, May 2024, Art. no. 1391472.

[63] G. Arya, M. K. Hasan, A. Bagwari, N. Safie, S. Islam, F. R. A. Ahmed, A. De, M. A. Khan, and T. M. Ghazal, "Multimodal hate speech detection in memes using contrastive language-image pre-training," *IEEE Access*, vol. 12, pp. 22359–22375, 2024.

[64] M. Schmidhuber and U. Kruschwitz, "LLM-based synthetic datasets: Applications and limitations in toxicity detection," in *Proc. LREC-COLING*, 2024, p. 37.

[65] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, W. Ye, Y. Zhang, Y. Chang, P. S. Yu, Q. Yang, and X. Xie, "A survey on evaluation of large language models," *ACM Trans. Intell. Syst. Technol.*, vol. 15, no. 3, pp. 1–45, Jan. 2024.

[66] A. Faiz, S. Kaneda, R. Wang, R. Osi, P. Sharma, F. Chen, and L. Jiang, "LLMCarbon: Modeling the end-to-end carbon footprint of large language models," 2023, *arXiv:2309.14393*.

[67] D. Kikkisetti, R. Ul Mustafa, W. Melillo, R. Corizzo, Z. Boukouvalas, J. Gill, and N. Japkowicz, "Using LLMs to discover emerging coded antisemitic hate-speech in extremist social media," 2024, *arXiv:2401.10841*.

[68] M. S. Jahan, M. Oussalah, D. R. Beddia, J. K. Mim, and N. Arhab, "A comprehensive study on NLP data augmentation for hate speech detection: Legacy methods, BERT, and LLMs," 2024, *arXiv:2404.00303*.

[69] M. Zhang, J. He, T. Ji, and C.-T. Lu, "Don't go to extremes: Revealing the excessive sensitivity and calibration limitations of LLMs in implicit hate speech detection," 2024, *arXiv:2402.11406*.

[70] L. Hong, P. Luo, E. Blanco, and X. Song, "Outcome-constrained large language models for countering hate speech," 2024, *arXiv:2403.17146*.

[71] 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," 2024, *arXiv:2403.08035*.

[72] A. Nirmal, A. Bhattacharjee, P. Sheth, and H. Liu, "Towards interpretable hate speech detection using large language model-extracted rationales," 2024, *arXiv:2403.12403*.

[73] V. Păi, "RACAI at climateactivism 2024: Improving detection of hate speech by extending LLM predictions with handcrafted features," in *Proc. 7th Workshop Challenges Appl. Automated Extraction Socio-Political Events Text (CASE)*, 2024, pp. 67–72.

[74] Z. Shaik, S. K. R. Kasu, S. Saumya, and S. Biradar, "IIITDWD-zk@DravidianLangTech-2024: Leveraging the power of language models for hate speech detection in Telugu–English code-mixed text," in *Proc. 4th Workshop Speech, Vis., Lang. Technol. Dravidian Lang.*, 2024, pp. 134–139.

[75] P. Saha, A. Agrawal, A. Jana, C. Biemann, and A. Mukherjee, "On zero-shot counterspeech generation by LLMs," 2024, *arXiv:2403.14938*.

[76] Y. Gou, K. Chen, Z. Liu, L. Hong, H. Xu, Z. Li, D.-Y. Yeung, J. T. Kwok, and Y. Zhang, "Eyes closed, safety on: Protecting multimodal LLMs via image-to-text transformation," 2024, *arXiv:2403.09572*.

[77] Y. Yang, J. Kim, Y. Kim, N. Ho, J. Thorne, and S.-Y. Yun, "HARE: Explainable hate speech detection with step-by-step reasoning," 2023, *arXiv:2311.00321*.

[78] N. Jafari, J. Allan, and S. M. Sarwar, "Target span detection for implicit harmful content," 2024, *arXiv:2403.19836*.

[79] J. Podolak, S. Łukasik, P. Balawender, J. Ossowski, J. Piotrowski, K. Bąkowicz, and P. Sankowski, "LLM generated responses to mitigate the impact of hate speech," 2023, *arXiv:2311.16905*.

[80] A. Nasir, A. Sharma, and K. Jaidka, "LLMs and finetuning: Benchmarking cross-domain performance for hate speech detection," 2023, *arXiv:2310.18964*.

[81] J. Jiang and E. Ferrara, "Social-LLM: Modeling user behavior at scale using language models and social network data," 2023, *arXiv:2401.00893*.

[82] A. Kumar, S. V. Murthy, S. Singh, and S. Ragupathy, "The ethics of interaction: Mitigating security threats in LLMs," 2024, *arXiv:2401.12273*.

[83] A. Y. Alan, E. Karaarslan, and Ö. Aydin, "A RAG-based question answering system proposal for understanding islam: MufassirQAS LLM," 2024, *arXiv:2401.15378*.

[84] M.-H. Van and X. Wu, "Detecting and correcting hate speech in multimodal memes with large visual language model," 2023, *arXiv:2311.06737*.

[85] S. Thapa, K. Rauniyar, F. A. Jafri, S. Shiwakoti, H. Veeramani, R. Jain, G. S. Kohli, A. Hürriyetoğlu, and U. Naseem, "Stance and hate event detection in tweets related to climate activism-shared task at case 2024," in *Proc. 7th Workshop Challenges Appl. Automated Extraction Socio-Political Events Text*, 2024, pp. 234–247.

[86] S. Ghosh, P. Varshney, E. Galinkin, and C. Parisien, "AEGIS: Online adaptive AI content safety moderation with ensemble of LLM experts," 2024, *arXiv:2404.05993*.

[87] V. Agarwal, Y. Chen, and N. Sastry, "HateRephrase: Zero- and few-shot reduction of hate intensity in online posts using large language models," 2023, *arXiv:2310.13985*.

[88] L. Struppek, M. H. Le, D. Hintersdorf, and K. Kersting, "Exploring the adversarial capabilities of large language models," 2024, *arXiv:2402.09132*.

[89] S. Pendzel, T. Wullach, A. Adler, and E. Minkov, "Generative AI for hate speech detection: Evaluation and findings," in *Regulating Hate Speech Created By Generative AI*. New York, NY, USA: Auerbach, 2024, pp. 54–76.

[90] G. Alyahya and A. Aldayel, "Hatred stems from ignorance! Distillation of the persuasion modes in countering conversational hate speech," 2024, *arXiv:2403.15449*.

[91] J. Zheng, X. Liu, M. Haque, X. Qian, G. Yang, and W. Yang, "HateModerate: Testing hate speech detectors against content moderation policies," in *Proc. Findings Assoc. Comput. Linguistics, NAACL*, 2024, pp. 2691–2710.

[92] H. Bonaldi, G. Attanasio, D. Nozza, and M. Guerini, "Weigh your own words: Improving hate speech counter narrative generation via attention regularization," 2023, *arXiv:2309.02311*.

[93] S. Xu and C. Zhang, "Misconfidence-based demonstration selection for LLM in-context learning," 2024, *arXiv:2401.06301*.

[94] M. Das, S. Kumar Pandey, and A. Mukherjee, "Evaluating ChatGPT's performance for multilingual and emoji-based hate speech detection," 2023, *arXiv:2305.13276*.

[95] J. Jones, L. Mo, E. Fosler-Lussier, and H. Sun, "A multi-aspect framework for counter narrative evaluation using large language models," 2024, *arXiv:2402.11676*.

[96] D. Furman, P. Torres, J. A. Rodríguez, D. Letzen, V. Martínez, and L. A. Alemany, "Which argumentative aspects of hate speech in social media can be reliably identified?" *arXiv:2306.02978*.

[97] C. Fei Luo, R. Bhambhoria, X. Zhu, and S. Dahan, "Towards legally enforceable hate speech detection for public forums," 2023, *arXiv:2305.13677*.

[98] R. Pi, T. Han, J. Zhang, Y. Xie, R. Pan, Q. Lian, H. Dong, J. Zhang, and T. Zhang, "MLLM-protector: Ensuring MLLM's safety without hurting performance," 2024, *arXiv:2401.02906*.

[99] D. Kang, X. Li, I. Stoica, C. Guestrin, M. Zaharia, and T. Hashimoto, "Exploiting programmatic behavior of LLMs: Dual-use through standard security attacks," in *Proc. IEEE Secur. Privacy Workshops (SPW)*, May 2024, pp. 132–143.

[100] A. Alexander and H. Wang, "Topological data mapping of online hate speech, misinformation, and general mental health: A large language model based study," 2023, *arXiv:2309.13098*.

[101] A. N. B. Emran, A. Ganguly, S. S. C. Puspo, D. Goswami, and M. N. Raihan, "MasonPerplexity at ClimateActivism 2024: Integrating advanced ensemble techniques and data augmentation for climate activism stance and hate event identification," in *Proc. 7th Workshop Challenges Appl. Automated Extraction Socio-Political Events Text (CASE)*, 2024, pp. 132–138.

[102] A. Shostack, "The boy who survived: Removing Harry Potter from an LLM is harder than reported," 2024, *arXiv:2403.12082*.

[103] S. Woźniak, B. Koptyra, A. Janz, P. Kazienko, and J. Kocoń, "Personalized large language models," 2024, *arXiv:2402.09269*.

[104] A. Das, M. Rahgouy, D. Feng, Z. Zhang, T. Bhattacharya, N. Raychawdhary, F. Jamshidi, V. Jain, A. Chadha, M. Sandage, L. Pope, G. Dozier, and C. Seals, "OffensiveLang: A community based implicit offensive language dataset," 2024, *arXiv:2403.02472*.

[105] J. Park, P. Wisniewski, and V. Singh, "Leveraging large language models (LLMs) to support collaborative human-AI online risk data annotation," 2024, *arXiv:2404.07926*.

[106] D. Kumar, Y. A. AbuHashem, and Z. Durumeric, "Watch your language: Investigating content moderation with large language models," in *Proc. Int. AAAI Conf. Web Social Media*, vol. 18, May 2024, pp. 865–878.

[107] N. Pangakis, S. Wolken, and N. Fasching, "Automated annotation with generative AI requires validation," 2023, *arXiv:2306.00176*.

[108] S. Petridis, B. Wedin, A. Yuan, J. Wexler, and N. Thain, "ConstitutionalExperts: Training a mixture of principle-based prompts," 2024, *arXiv:2403.04894*.

[109] R. Hazra, S. Layek, S. Banerjee, and S. Poria, "Sowing the wind, reaping the whirlwind: The impact of editing language models," 2024, *arXiv:2401.10647*.

[110] S. Anugrah Hayati, M. Lee, D. Rajagopal, and D. Kang, "How far can we extract diverse perspectives from large language models?" 2023, *arXiv:2311.09799*.

[111] A. McGovern, "Evaluating the performance impact of fine-tuning optimization strategies on pre-trained distilbert models towards hate speech detection in social media," Tech. Rep., 2022.

[112] B. C. Matos, R. B. Santos, P. Carvalho, R. Ribeiro, and F. Batista, "Comparing different approaches for detecting hate speech in online Portuguese comments," in *Proc. 11th Symp. Lang., Appl. Technol. (SLATE)*, 2022, pp. 1–12.

[113] S. Achintalwar et al., "Detectors for safe and reliable llms: Implementations, uses, and limitations," 2024, *arXiv:2403.06009*.

[114] A. Verma, S. Krishna, S. Gehrmann, M. Seshadri, A. Pradhan, T. Ault, L. Barrett, D. Rabinowitz, J. Doucette, and N. Phan, "Operationalizing a threat model for red-teaming large language models (LLMs)," 2024, *arXiv:2407.14937*.

[115] I. Sen, D. Assenmacher, M. Samory, I. Augenstein, W. van der Aalst, and C. Wagner, "People make better edits: Measuring the efficacy of LLM-generated counterfactually augmented data for harmful language detection," 2023, *arXiv:2311.01270*.

[116] C. Christodoulou, "NLPDame at ClimateActivism 2024: Mistral sequence classification with PEFT for hate speech, targets and stance event detection," in *Proc. 7th Workshop Challenges Appl. Automated Extraction Socio-political Events Text (CASE)*, 2024, pp. 96–104.

[117] J. Goldzycher, P. Röttger, and G. Schneider, "Improving adversarial data collection by supporting annotators: Lessons from GAHD, a German hate speech dataset," 2024, *arXiv:2403.19559*.

[118] A. N. B. Emran, A. Ganguly, S. S. C. Puspo, D. Goswami, and M. N. Raihan, "Masonperplexity at climateactivism 2024: Integrating advanced ensemble techniques and data augmentation for climate activism stance and hate event identification," 2024, *arXiv:2402.01976*.
[119] S. Almohaimeed, S. Almohaimeed, A. A. Shafin, B. Carbunar, and L. Bölöni, "THOS: A benchmark dataset for targeted hate and offensive speech," 2023, *arXiv:2311.06446*.
[120] T. T. Nguyen, C. Wilson, and J. Dalins, "Fine-tuning llama 2 large language models for detecting online sexual predatory chats and abusive texts," 2023, *arXiv:2308.14683*.
[121] M. Das, S. Barman, and S. Chatterjee, "Hate-alert@LT-EDI-2023: Hope speech detection using transformer-based models," in *Proc. 3rd Workshop Lang. Technol. Equality, Diversity Inclusion*, 2023, pp. 250–256.
[122] Y. Liu, G. Deng, Y. Li, K. Wang, Z. Wang, X. Wang, T. Zhang, Y. Liu, H. Wang, Y. Zheng, and Y. Liu, "Prompt injection attack against LLM-integrated applications," 2023, *arXiv:2306.05499*.
[123] T. Kwon and C. Kim, "Efficacy of utilizing large language models to detect public threat posted online," 2023, *arXiv:2401.02974*.
[124] A. Deroy and S. Maity, "Multi-label classification of COVID-tweets using large language models," 2023, *arXiv:2312.10748*.
[125] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, T. Mihaylov, M. Ott, S. Shleifer, K. Shuster, D. Simig, P. S. Koura, A. Sridhar, T. Wang, and L. Zettlemoyer, "OPT: Open pre-trained transformer language models," 2022, *arXiv:2205.01068*.
[126] M. O'Neill and M. Connor, "Amplifying limitations, harms and risks of large language models," 2023, *arXiv:2307.04821*.
[127] Z. Fryer, V. Axelrod, B. Packer, A. Beutel, J. Chen, and K. Webster, "Flexible text generation for counterfactual fairness probing," 2022, *arXiv:2206.13757*.
[128] A. Leidinger, R. van Rooij, and E. Shutova, "The language of prompting: What linguistic properties make a prompt successful?" 2023, *arXiv:2311.01967*.
[129] X. Zhou, Y. Lu, R. Ma, T. Gui, Q. Zhang, and X. Huang, "Making harmful behaviors unlearnable for large language models," 2023, *arXiv:2311.02105*.
[130] Y. Li, M. Du, R. Song, X. Wang, and Y. Wang, "A survey on fairness in large language models," 2023, *arXiv:2308.10149*.
[131] T. Cyril Weerasooriya, S. Dutta, T. Ranasinghe, M. Zampieri, C. M. Homan, and A. R. KhudaBukhsh, "Vicarious offense and noise audit of offensive speech classifiers: Unifying human and machine disagreement on what is offensive," 2023, *arXiv:2301.12534*.
[132] A. Som, K. Sikka, H. Gent, A. Divakaran, A. Kathol, and D. Vergyri, "Demonstrations are all you need: Advancing offensive content paraphrasing using in-context learning," in *Proc. Findings Assoc. Comput. Linguistics ACL*, 2024, pp. 12612–12627.
[133] N. Varshney, P. Dolin, A. Seth, and C. Baral, "The art of defending: A systematic evaluation and analysis of LLM defense strategies on safety and over-defensiveness," 2023, *arXiv:2401.00287*.
[134] O. Zheng, M. Abdel-Aty, D. Wang, Z. Wang, and S. Ding, "ChatGPT is on the horizon: Could a large language model be suitable for intelligent traffic safety research and applications?" 2023, *arXiv:2303.05382*.
[135] S. Chen, M. Wu, K. Q. Zhu, K. Lan, Z. Zhang, and L. Cui, "LLM-empowered chatbots for psychiatrist and patient simulation: Application and evaluation," 2023, *arXiv:2305.13614*.
[136] Y. Liu, Y. Yao, J.-F. Ton, X. Zhang, R. Guo, H. Cheng, Y. Klochkov, M. F. Taufiq, and H. Li, "Trustworthy LLMs: A survey and guideline for evaluating large language models' alignment," 2023, *arXiv:2308.05374*.
[137] N. A. Cloutier and N. Japkowicz, "Fine-tuned generative LLM oversampling can improve performance over traditional techniques on multiclass imbalanced text classification," in *Proc. IEEE Int. Conf. Big Data (BigData)*, Dec. 2023, pp. 5181–5186.
[138] G. Hurlburt, "What if ethics got in the way of generative AI?" *IT Prof.*, vol. 25, no. 2, pp. 4–6, Mar. 2023.
[139] Z. Yu, "A multi-dimensional generic evaluation framework for the security of large language models," in *Proc. 4th Int. Conf. Big Data, Artif. Intell. Internet Things Eng. (ICBAIE)*, Aug. 2023, pp. 410–414.
[140] Y. Garani, S. Joshi, and S. Kulkarni, "Offensive sentiment detection with chat GPT and other transformers in Kannada," in *Proc. IEEE 2nd Int. Conf. Data, Decis. Syst. (ICDDS)*, Dec. 2023, pp. 1–6.
[141] V. Gadiraju, S. Kane, S. Dev, A. Taylor, D. Wang, E. Denton, and R. Brewer, "'I wouldn't say offensive but.': Disability-centered perspectives on large language models," in *Proc. ACM Conf. Fairness, Accountability, Transparency*, 2023, pp. 205–216.
[142] Z. Chu, Z. Wang, and W. Zhang, "Fairness in large language models: A taxonomic survey," *ACM SIGKDD Explorations Newslett.*, vol. 26, no. 1, pp. 34–48, Jul. 2024.
[143] T. Garg, S. Masud, T. Suresh, and T. Chakraborty, "Handling bias in toxic speech detection: A survey," *ACM Comput. Surveys*, vol. 55, no. 13s, pp. 1–32, Dec. 2023.

AISH ALBLADI received the M.S. degree in computer science from Ball State University, IN, USA. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Software Engineering, Auburn University, AL, USA. His research interests include AI, natural language processing, and sentiment analysis.

MINARUL ISLAM received the B.S. degree from the Department of Computer Science and Engineering, Jessore University of Science and Technology, Jessore, Bangladesh, in 2016, and the M.S. (by Research) degree from the Department of Electrical and Electronic Engineering, Universiti Malaysia Pahang, Pahang, Malaysia, in 2021. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Software Engineering, Auburn University, AL, USA. During the M.S. degree, he was awarded the Bronze, Silver, Gold, and Best Innovative Technology Awards. He has published more than 11 research articles at different conferences and peer-reviewed journals. Recently, his current project poster abstract has been accepted at ACM SenSys 2024 Conference, which is one of the top conference in the area of mobile computing. His primary research interests include machine learning, mobile sensing, and wireless sensor networks.

AMIT DAS received the Ph.D. degree in computer science and software engineering from Auburn University, USA, in 2024. Currently, he is an Assistant Professor with the Department of Computer and Information Systems, University of North Alabama. His primary research interest includes natural language processing.

MARYAM BIGONAH received the M.S. degree in computer engineering from Imam Reza University, Mashad, Iran. She is currently pursuing the Ph.D. degree with the Department of Computer Science and Software Engineering, Auburn University, AL, USA. Her research interests include human–computer interaction (HCI), machine learning (ML), and artificial intelligence (AI).

ZHENG ZHANG received the M.S. and Ph.D. degrees in computer science from Auburn University, USA. He is currently an Assistant Professor with the Department of Computer Science and Information Systems, Murray State University, USA. His research interests include neuro-symbolic AI, natural language processing, network intrusion detection, and artificial intelligence.

FATEMEH JAMSHIDI received the master's degree in music education from Auburn University, in 2023, and the Ph.D. degree in computer science and software engineering, in 2024. She is currently an Assistant Professor with the Department of Computer Science, Cal Poly Pomona. Her research interests include artificial intelligence, computer science education, computer music, machine learning and deep learning in music, game AI, human-AI collaboration, and augmented and mixed reality. She has published in prestigious venues, including ACM SIGCSE and HCII. During the Ph.D. degree, she founded the Computing + Music programs, which have engaged 100's of participants from underrepresented groups, since 2018. From 2020 to 2023, she was the Director of the Persian Music Ensemble with Auburn University. Her long-term goal is to establish a music technology center that fosters undergraduate and graduate research in areas, such as music therapy, music generation, game music, and mixed reality in music.

MOSTAFA RAHGOUY received the Bachelor of Science degree in computer science from the University of Mohaghegh Ardabili, Iran, in 2019, and the Master of Science degree in computer science and software engineering from Auburn University, in 2023. He is currently pursuing the Ph.D. degree in computer science and software engineering with Auburn University, USA. Previously, he was a NLP Researcher with the Part-AI Research Center, where he focused on natural language processing, with specific interests in authorship attribution, reasoning, and question answering. He has received the Gavin Fellowship for Outstanding Graduate Research Students, recognizing his contributions to the field.

NILANJANA RAYCHAWDHARY is currently pursuing the Ph.D. degree (ABD) in computer science and software engineering with Auburn University, AL, USA. Under the guidance of Dr. Cheryl Seals, her research interests include advancing sentiment analysis in low-resource African languages, including Igbo, Hausa, and Amharic, leveraging transformer-based models and machine learning techniques. With seven peer-reviewed publications, her work aims to address gaps in natural language processing (NLP) for these languages. She has extensive teaching experience as a Graduate Teaching Assistant with Auburn University. Previously, she was an Assistant Professor in India. She is actively involved in promoting diversity and inclusion within the tech community, having received the AnitaB.org Advancing Inclusion Scholarship and other academic awards. Her contributions to the field include numerous conference talks and publications, with notable achievements in sentiment analysis competitions.

DANIELA MARGHITU is currently a Faculty Member with the Computer Science and Software Engineering Department, Auburn University, where she has worked, since 1996. She has published seven *Information Technology* books at Pearson Publishing Company over 100 articles and peer-reviewed conference papers at prestigious journals and conferences. Her research interests include STEM K12 inclusive computing research and outreach, web applications design and development, education and assistive technology, software engineering, web and software engineering usability, and accessibility.

Prof. Marghitu also served as member for the Congressionally Mandated Committee on Equal Opportunities in Science and Engineering; a member of the Committee on the Future of NSF EPSCoR; and a member of the College Board's Advanced Placement Program Computer Science a Development Committee, Governor Kay Ivey's Computer Science Advisory Council. She has received eight awards for research and education projects from National Science Foundation, e.g., a Co-PI of RET Site: Project-Based Learning for Rural Alabama STEM Middle School Teachers in Machine Learning and Robotics; a Co-PI of INCLUDES Alliance: The Alliance of Students with Disabilities for Inclusion, Networking, and Transition Opportunities in STEM (TAPDINTO-STEM); a Co-PI of EEC "RFE Design and Development: Framing Engineering as Community Activism for Values-Driven Engineering;" a Co-PI of NSF CISE "EAGER: An Accessible Coding Curriculum for Engaging Underserved Students with Special Needs in Afterschool Programs;" a Co-PI of NSF INCLUDES: South East Alliance for Persons with Disabilities in STEM; and a Co-PI of NSF CE 21 Collaborative Research: Planning Grant: Computer Science for All and Co-PI for NIFA "Reimagining Controlled Environment Agriculture in a Low Carbon World."

CHERYL SEALS is currently an Associate Professor with the Department of Computer Science and Software Engineering, Auburn University. Her research interests include human–computer interaction, user interface design, usability evaluation, and educational gaming technologies. Seals also works with outreach initiatives to improve computer science education at all levels. The programs are focused on increasing the computing pipeline by getting students interested in STEM disciplines and future technology careers.