International Journal of Computer Sciences and Engineering
Vol.**13**, Issue.**2**, pp.**49-56**, February **2025**
ISSN: 2347-2693 (Online)
Available online at: www.ijcseonline.org

ISSN: 2347-2693 (E)

Research Article

Leveraging Artificial Intelligence and Machine Learning in Online Threat Detection

S. Shamroukh[1*], **T. Johnson**[2]

[1]Dept. of Information Technology and Decision Science, The University of North Texas, Denton, U.S.A
[2]Dept. of Analytics, Harrisburg University of Science and Technology, Harrisburg, U.S.A

[*]*Corresponding Author:* ✉ *Tel.: 940-565-2812*

Received: 19/Dec/2024; **Accepted**: 20/Jan/2025; **Published**: 28/Feb/2025. **DOI**: https://doi.org/10.26438/ijcse/v13i2.4956

Abstract: This literature review examines the roles of artificial intelligence (AI), machine learning (ML), and large language models (LLMs) in identifying and interpreting online threats. As AI and ML technologies advance, their use in analyzing vast online data for potential threats has grown significantly. The review systematically evaluates current methodologies for detecting and assessing threats, particularly in social media and online forums, which are both information hubs and sources of harmful content. Key findings highlight the effectiveness of BERT-based models in hate speech detection across languages and platforms, emphasizing their interpretability and transparency advantages over traditional neural networks. Models like GPT-4 further expand threat identification capabilities, detecting cyber threats and abusive language, with implications for public safety and mental health monitoring. Challenges remain, particularly in handling noisy, diverse, and imbalanced social media datasets. Domain-specific word embeddings and ensemble techniques, such as combining BERT with TextCNN and BiLSTM, show promise in improving detection accuracy in complex data environments. The review advocates for continued focus on hybrid and ensemble models to address data complexities and calls for future research to enhance model transparency and address ethical concerns like data privacy. Given the rise in digital communication, real-time threat detection is crucial for public safety, national security, and violence prevention. This review consolidates findings on the efficacy of AI and ML in detecting online threats, identifies recurring challenges, and outlines research gaps to guide future advancements. By synthesizing recent studies, it provides a structured analysis of the current capabilities and limitations of AI and ML in online threat monitoring, contributing to a foundational understanding of how these technologies can evolve to enhance societal safety.

Keywords: Artificial Intelligence (AI), Online Threat Detection, Machine Learning (ML), Large Language Models (LLMs)

1. Introduction

This paper is a literature review that offers a comprehensive study of the current literature on using AI, ML, and large language models (LLMs) in identifying and interpreting online threats. As advancements in AI and ML technologies have rapidly evolved, their applications in analyzing vast amounts of online data have also expanded. This review will systematically examine the research literature on AI and ML methodologies applied to detecting and assessing potential threats posted on social media platforms and online forums, which have become both hubs of information sharing and potential venues for harmful content.

With the increasing reliance on digital communication, the ability to identify threats in real time is critical to maintaining public safety, ensuring national security, and preventing acts of violence. Recent developments in LLMs have created new avenues for understanding and interpreting language nuancedly, including context and cultural slang. However, despite their potential, many challenges remain in effectively harnessing these technologies for threat detection. This literature review will critically evaluate studies focused on the accuracy, limitations, and ethical considerations associated with using AI, ML, and LLMs in this field.

The primary purpose of this review is three-fold: First, to consolidate and evaluate the effectiveness of these technologies in detecting threats in the public domain; Second, to determine the problems and unsolved questions so far in the literature; Third, to reveal the gaps where future research may lead to better and more reliable solutions. By bringing together key findings from various studies, this review will provide a structured analysis of the state of AI and ML-based threat detection research, offering insights into these technologies' current capabilities and limitations.

Through a critical synthesis of the literature, this paper will contribute to a deeper understanding of how AI and ML can enhance online threat monitoring and help establish a foundation for future advancements.

By examining these areas, this literature review will offer a cohesive view of how AI, ML, and LLMs are currently used for threat detection, what limitations exist, and how these technologies may evolve to serve society better in the future.

A. Background and Motivation

With the increasing accessibility of social media and digital communication, the risk of threats emerging online has grown significantly, presenting new challenges for public safety and security. Threats to public safety in online spaces can vary widely—from direct, explicit messages signaling imminent violence to more subtle language patterns or phrases that may indicate underlying risk factors, such as mental distress or harmful intent. Identifying these threats in real time is a complex because of the enormous amount of data being created every day on social media platforms, messaging apps, forums, and other online locations. This volume renders manual monitoring impractical and underscores the need for automated systems that can detect potential threats swiftly and accurately as they emerge.

AI and ML technologies have proven effective in fields like healthcare and finance, where they analyze massive datasets to identify patterns and make predictions that can save lives or mitigate financial losses. Applying similar methodologies to public safety, AI and ML can help detect patterns in online content that may signify threats, making it possible to intervene before situations escalate. Large Language Models (LLMs), trained on vast and diverse language datasets, bring particular strengths to threat detection efforts. Their ability to understand the nuances of language, regional slang, cultural references, and even subtle changes in tone enables them to identify potentially harmful content with higher precision. This understanding is invaluable for parsing complex online interactions and distinguishing between benign and genuinely threatening messages, making LLMs a powerful tool in advancing real-time public safety solutions.

B. Problem Statement

In a world where social media and online platforms are a significant part of our lives, they can also become a way to spread harmful information like public threats. The ability to detect these threats in real-time is crucial for ensuring public safety and national security and preventing violence. However, identifying threatening content online poses significant challenges due to the vast volume of data, language variability, and ambiguous or coded messages. Traditional monitoring methods are no longer sufficient to manage the scale and complexity of online communications.

Advancements in AI, ML, and LLMs offer promising solutions by automating the detection and analysis of potential threats. Despite these advancements, the efficacy of these technologies in accurately identifying threats remains uncertain. Challenges include the risk of false positives or negatives, ethical concerns regarding privacy and censorship, and limitations in understanding linguistic nuances or cultural context within social media content.

These issues are addressed by this study which carries out a thorough review of the literature regarding the application of AI, ML and LLMs for online threat detection. As for the purpose of this paper, it is intended to assess the effectiveness of these technologies for real-time threat detection, to define the main problems and to determine where more work is required to enhance the effectiveness and to make sure that the application is ethical. By reviewing the strengths, weaknesses, and possibilities of the AI-based threat detection tool in a systematic way, this research is intended to lay a foundation for the development of better, accurate, and responsible approaches to improving public safety in the digital environment.

C. Purpose and Scope

The purpose of this literature review is to comprehensively analyze and combine the research available on the use of AI, ML, and LLMs for identifying threats to the public in online environments, especially on social media. This review is intended to establish how effective these technologies are at finding potential threats, what the main methodologies are, and what challenges and ethical issues are involved. Based on the current developments and constraints, this paper aims to contribute new insights that may help to shape further studies and the improvement of the accuracy and accountability of AI-based threat detection solutions.

The scope of this review encompasses:

1. Technological Approaches: An in-depth analysis of AI, ML, and LLM methodologies, Including natural language processing (NLP) sentiment analysis, and neural network architectures, specifically for threat detection in online content.
2. Application Across Platforms: A comparison of the efficacy of threat detection methods across various social media and online platforms, noting how platform-specific features (e.g., short-form vs. long-form content, anonymity) impact the effectiveness of these technologies.
3. Challenges and Limitations: Identification of the primary challenges, including the risk of false positives/negatives, handling of linguistic ambiguity, and scalability issues, alongside limitations in detecting complex or coded threats.
4. Ethical and Privacy Considerations: An exploration of ethical concerns, particularly related to privacy, potential biases in algorithmic predictions, and implications for civil liberties and public trust.
5. Future Directions: A discussion of gaps in the existing research and suggestions for future studies, with the aim of developing robust, ethical, and adaptable AI-based threat detection systems.

Through this structured review, the study will offer a comprehensive overview of the current landscape, serving as a resource for researchers, policymakers, and developers aiming to enhance public safety while balancing ethical considerations in digital threat monitoring.

D. Organization of the Review

This comprehensive literature review will be organized by themes to explore specific aspects of threat detection, including:

• Algorithmic Approaches: An examination of different ML and AI algorithms used for identifying threats, such as NLP techniques, neural networks, and classification models.

• LLM Capabilities and Challenges: A focus on the role of LLMs in analyzing social media content, their ability to recognize linguistic nuances, and the challenges they face in understanding context and cultural references.

• Ethical and Privacy Concerns: An analysis of ethical considerations in monitoring public threats, addressing potential privacy issues, the risk of bias in algorithms, and the implications for civil liberties.

• Effectiveness Across Platforms: A comparison of how these technologies perform across different social media and online platforms, considering factors like data availability, platform-specific language use, and content formats.

2. Theoretical Framework and Conceptual Background

The theoretical framework for this literature review is grounded in sociotechnical systems theory, NLP, and ML theories related to pattern recognition and classification. This framework provides a basis for understanding how AI, ML, and LLMs can be utilized to detect online threats, recognizing both the technical aspects of detection algorithms and the complex social context of online communication.

A. Sociotechnical Systems Theory

• Definition: Sociotechnical systems theory posits that technological systems are embedded within and influenced by broader social systems. In the context of online threat detection, this theory underscores that AI and ML technologies are not simply technical tools but are deeply influenced by human behavior, ethical considerations, and social dynamics.

• Application: This theory guides the understanding of online threats as socially constructed phenomena. For instance, expressions of intent or harmful speech often depend on cultural, political, and situational factors that require contextual sensitivity in AI models. By viewing AI and ML threat detection as part of a sociotechnical system, this framework provides insights into the need for ethical safeguards and transparency.

B. Natural Language Processing (NLP) and Linguistic Theory

• Definition: NLP theory focuses on the development of algorithms capable of understanding and processing human language. Linguistic theories, particularly those around semantics and pragmatics, also inform NLP's ability to interpret the meaning, intent, and nuances in text.

• Application: NLP is essential for AI-driven threat detection as it allows systems to analyze language patterns, sentiment, and specific keywords or phrases associated with threats. Linguistic theory aids in addressing challenges such as sarcasm, coded language, and evolving slang, which are crucial for detecting subtle or veiled threats.

C. Machine Learning Theory: Pattern Recognition and Classification

• Definition: Machine learning theory, particularly pattern recognition, underpins how algorithms are trained to detect and classify potential threats. Pattern recognition involves identifying and learning from known threat patterns, while classification is essential for categorizing content as threatening or non-threatening.

• Application: In threat detection, ML models rely on large datasets of labeled threat and non-threat content to recognize patterns associated with harmful intent. Classification techniques (e.g., logistic regression, support vector machines, neural networks) allow the model to categorize online content accurately. This theory is fundamental for creating automated systems that distinguish between harmless posts and those indicating genuine threats.

D. Large Language Models (LLMs) and Contextual Understanding

• Definition: LLMs are advanced NLP models trained on massive datasets, capable of generating and interpreting human language with high levels of contextual awareness.

• Application: LLMs offer enhanced capabilities in detecting complex or subtle threats by using deep contextual understanding. They help address limitations in traditional NLP approaches by handling idiomatic expressions, implicit meanings, and nuanced context that may signify a threat. The ability of LLMs to process larger contextual scopes makes them particularly effective in identifying ambiguous language or inferred threats in social media and online content.

E. Ethics of Surveillance and Privacy Concerns

• Definition: Ethical theories related to surveillance, privacy, and data protection are essential for evaluating the implications of using AI and ML for public threat detection. Privacy theories address the balance between public safety and individual rights.

• Application: Applying these ethical theories to AI-driven threat detection highlights the potential risks of overreach, discrimination, and bias in monitoring social media. Ethical principles guide the responsible deployment of these technologies, emphasizing transparency, accountability, and user consent to avoid infringing on civil liberties while ensuring public safety.

These theoretical foundations allow for a holistic view of online threat detection, emphasizing that technological solutions must be contextualized within their social and ethical implications. The framework will support this literature review in evaluating both the technical efficacy and broader societal impacts of AI, ML, and LLMs in detecting public threats online. This multidimensional approach provides a comprehensive lens for assessing the current capabilities and limitations in this rapidly evolving field.

3. Methodology

This literature review seeks to combine current studies on the application of AI, ML, and LLMs for identifying public threats online, especially on social media platforms. The methodology involves well-defined criteria for source selection, a systematic search strategy, and a clear approach to analyzing the data to ensure that the process of combining relevant literature is comprehensive and organized.

A. Criteria for Inclusion/Exclusion

To maintain rigor and relevance in this review, sources were selected based on the following criteria:

- **Inclusion Criteria**:
  - **Publication Type** peer reviewed journal articles, conference proceedings and reports from reputable institutions or government bodies.
  - **Scope and Relevance**: Studies focused on AI, ML, and LLMs specifically applied to online threat detection, content moderation, or analyzing social media data.
  - **Date of Publication**: Published within the past 10 years to capture recent advancements, with emphasis on sources from the last five years to include the latest developments in LLMs.
  - **Language**: English-language publications to ensure consistency in data interpretation.
  - **Methodological Quality**: Studies that demonstrate a clear methodology, robust data analysis, and conclusions supported by empirical data or well-supported theoretical arguments.
- **Exclusion Criteria**:
  - **Irrelevant Focus**: Studies focusing on AI and ML applications outside of threat detection or online data analysis, such as medical or financial AI applications, will be excluded.
  - **Non-Peer-Reviewed Sources**: Opinion pieces, non-academic blog posts, and promotional articles without rigorous peer review.
  - **Outdated Research**: Studies over 10 years old that may not reflect the current state of AI, ML, and LLM advancements, unless they provide foundational theory or context.

B. Search Strategy

A comprehensive and structured search strategy will be employed to locate relevant literature across multiple academic and industry-focused databases.

- **Databases and Resources**:
  - **Academic Databases**: PubMed, IEEE Xplore, Scopus, ACM Digital Library, and Google Scholar for peer-reviewed articles and conference proceedings.
  - **Industry Reports and White Papers**: Reputable institutions such as the Pew Research Center, OpenAI, and governmental reports relevant to AI and public safety.
  - **Government and Institutional Reports**: National security and cyber safety publications from agencies such as the Department of Homeland Security, FBI, and the European Union Agency for Cybersecurity (ENISA).
- **Keywords and Search Terms**:
  - Keywords: "AI in threat detection," "machine learning for public safety," "large language models in social media analysis," "online threat detection," "AI-based content moderation," "social media monitoring for threats."
  - Boolean Operators: Keywords were combined using Boolean operators (AND, OR) to refine search results. For instance, a search might include terms like: *"AI AND threat detection" OR "machine learning AND social media threats."*
- **Timeframe**:
  - The initial search covered publications from the last ten years, with a primary focus on the most recent five years to capture the latest advancements in AI, ML, and LLMs.
- **Screening Process**:
  - Abstracts of identified articles were reviewed to ensure relevance and alignment with the study's purpose.
  - Selected articles were reviewed in full, applying the inclusion and exclusion criteria to finalize the sources.

C. Data Analysis Methods

A systematic approach was taken to analyze and categorize the selected literature. The primary data analysis methods include **thematic analysis** and **content analysis**, ensuring a detailed and organized review of findings.

- **Thematic Analysis**:
  - Relevant themes were identified across literature, such as AI algorithms used in threat detection, challenges with LLMs, ethical considerations, and platform-specific performance.
  - Articles wrere grouped to allow for a detailed comparison and synthesis of approaches, findings, and limitations within each theme.
  - Key themes will include:
    - **Algorithmic Approaches**: Techniques like natural language processing, sentiment analysis, and neural networks.
    - **Ethical and Privacy Concerns**: Issues related to surveillance, bias, privacy, and civil liberties.
    - **Effectiveness Across Platforms**: Differences in AI performance across social media and online platforms.
- **Content Analysis**:
  - Detailed content analysis will identify commonalities in methodological approaches, such as the types of algorithms used, and data sources analyzed.
  - A comparative table was created to summarize findings across studies, highlighting areas such as data sources, models, and performance metrics.
- **Data Extraction and Coding**:
  - Key data points, such as study objectives, methodologies, findings, limitations, and implications, were extracted from each source and coded into a spreadsheet or software (e.g., NVivo or Excel) for easy reference.

- Coded data were grouped according to thematic relevance and analyzed to draw out common patterns, differences, and critical insights.

The literature review through this methodology will present a cohesive analysis of the research landscape to reveal the effectiveness, limitations, and ethical concerns of using AI, ML, and LLMs in online threat detection. This approach will, in the end, help to better understand the current research directions and gaps and gain critical insights for future work in this evolving field. Hence, the literature review through this methodology will present a cohesive analysis of the research landscape to reveal the effectiveness, limitations, and ethical concerns of using AI, ML, and LLMs in online threat detection. This approach will, in the end, help to better understand the current research directions and gaps and gain important insights into future work in this evolving field.

4. Body of the Literature Review

This review encompassed 19 peer-reviewed studies, each selected based on relevance, quality, and focus on machine learning applications for detecting hate speech in social media contexts. The studies were published between 2018 and 2024, a period marked by rapid advancements in AI and machine learning, particularly in NLP and model architectures. Researchers employed various data collection methods, including web scraping and the analysis of publicly available datasets. These sources of data span a diverse range of social media platforms, providing a robust foundation for generalizing findings across different online environments.

The studies also demonstrated international scope, with some focusing on social media data from non-Western countries, such as China and Korea, to capture cultural and linguistic diversity. This geographic and linguistic diversity offers valuable insights into the models' adaptability to various languages, dialects, and online behaviors. Additionally, cross-national data sources highlighted distinct hate speech patterns and slang unique to particular regions, underscoring the importance of context in model training and validation.

In terms of methodology, machine learning techniques were central to all studies, with BERT emerging as the most frequently used model. BERT’s pre-trained transformer-based architecture, which can be fine-tuned for specific tasks, made it particularly effective in capturing the subtleties of language necessary for detecting hate speech across diverse social media platforms. In some studies, BERT was integrated with other machine learning models, such as CNNs or LSTMs, to enhance accuracy and manage the complex datasets typical of social media environments. Other models, including support vector machines (SVMs) and logistic regression, were utilized as benchmarks, enabling researchers to compare traditional algorithms with newer, more advanced machine learning techniques.

Table 1 provides an overview of each study's key characteristics, including publication date, data source, language focus, and model type. This table summarizes the heterogeneity among the studies, reflecting a broad array of machine-learning strategies and social media contexts. Through this detailed comparison, Table 1 reveals trends in model selection and adaptation, shedding light on the evolving methodologies in online hate speech detection.

Table 1: Characteristics of Included Studies

Author and Year	*Methodology*	*Results*
Modha et al. (2022)	Deep learning-based classifiers (e.g., Bag of Words) and models (e.g., BERT)	Deep neural networks outperform classifiers by 6% to 10%
Faraj & Utku (2024)	Word embeddings (e.g., GloVE) and models (e.g., BERT)	BERT yielded the highest accuracy of 85%
Dehingia et al. (2023)	Naïve Bayes, BERT, SVM, Ridge logistic regression, multi-layer perceptron neural network	BERT performed the best (AUC F1 of 91)
Kwon & Kim (2023)	LLM (GPT-4)	97.9% accuracy for non-threats and 100% for threats
Husain (2021)	Ensemble machine learning (e.g., bagging)	Bagging performed with most accurately (offensive language detection F1 score of 88%)
McKeever et al. (2020)	Word and sentence embeddings (e.g., Word2Vec, SentEncoder)	Sentence embeddings were most effective (accuracy of 96%)
Mossie & Wang (2020)	Deep learning (e.g., LSTM)	Feature extraction with word embedding techniques performed most effectively in hate speech detection
Coppersmith et al. (2018)	Word embeddings (e.g., GloVE)	70% to 85% true accuracy rate
Elzayady et al. (2022)	Deep learning (e.g., LSTM)	97.1% accuracy using SVM and ensemble methods
Kulkarni et al. (2023)	NLP and deep learning (e.g., CNN)	Unknown (proposed article and methodology)
Chakraborty & Seddiqui (2019)	Machine learning (e.g., SVM)	SVM with a linear kernel had the highest accuracy (78%)
Monnar et al. (2024)	NLP (e.g., word embeddings) and machine learning models (e.g., BERT)	Monolingual – BERT performed best (F1 score of 74.51); cross-lingual – LSTM and CNN performed best in English (F1 of 63.91 and 58.04 respectively); performance depended on language in cross-lingual tests
Mohamed et al.	Machine learning	Ensemble-learning

(2023)	(e.g., BERT and GPT)	BERT-based approach – F1 score of approximately 92%
Paraschiv et al. (2023)	LLMs (e.g., GPT-4) and machine learning (e.g., BERT)	BERT-based model performed best (F1 score of 75.72)
Mnassri et al. (2024)	Machine learning (e.g., GAN)	GAN performed best in all languages (e.g., accuracy of 0.753 in English)
Huang et al. (2022)	Machine learning (e.g., BERT) and deep learning (e.g., CNN)	MFAE (ensemble learning method) performed best (e.g., 96% accuracy in one dataset)
Adewumi et al. (2023)	Machine learning (e.g., T5)	T5 performed best (e.g., 91.73% macro F1 score on one subtask)
Manias et al. (2023)	Machine learning (e.g., BERT); comparison between models	Multilingual BERT-based models yielded the highest accuracy out of compared models; zero shot classifiers are more scalable and faster than BERT-based models
Muneer et al. (2023)	Deep learning (e.g., LSTM)	Stacking ensemble deep learning model performs the best (97.4% accuracy and 0.964 F1 score on Twitter dataset; 90.97% accuracy on combined Twitter and Facebook dataset

5. Synthesis of Findings

Most studies in this review show that BERT-based models achieve the highest accuracy, efficiency, and speed in detecting hate speech across different social media sites and languages. The ability of BERT-based models to generalize effectively across websites and languages is particularly useful in the context of multilingual and culturally diverse online environments (Dehingia et al., 2023; Manias et al., 2023; Muneer et al., 2023). This is because BERT-based models are founded on the transformer architecture that is able to capture the linguistic features and help in correct hate speech classification even in the datasets that are diverse in language. While many studies are focused on a particular language, BERT has been found to be adaptable for detecting hate speech across multiple sites and many languages with minimal retraining. This flexibility across platforms and languages is a critical advantage over other models which may require more iterative tuning to achieve similar levels of accuracy across languages and regions.

In addition to performance, BERT-based models improve the explainability and interpretability of the models, a feature that sets them apart from the deep neural networks with the black-box like architectures. In traditional neural networks, for example, hate speech can be detected through the hidden layers, which makes it hard to understand how the decision was made, but BERT has a transformer-based structure which makes it easier to understand the decision-making process in the model output (Paraschiv et al., 2023). However, for the best interpretability and accuracy, BERT is opposed to other models, including CNNs and LSTM networks, which are more effective in particular hate speech detection tasks. These models are generally better than the traditional classifiers like support vector machines and logistic regression, especially when the data is sequential and context is crucial in the interpretation process (Elzayady et al., 2022; Modha et al., 2022).

In applying these machine learning approaches, the researchers have consistently used parameters such as AUC (area under the curve) and F1 scores to compare the model performance in hate speech detection and other related areas, for instance, cyberbullying (Dehingia et al., 2023; Faraj & Utku, 2024). Other approaches which include semi-supervised models, text categorization, and sentiment analysis were also efficient in detecting hate speech across different languages especially in less popular languages such as Bengali and Arabic and on different social media platforms including Facebook and X (formerly Twitter) (Chakraborty & Seddiqui, 2019; Manias et al., 2023; Mnassri et al., 2024; Mohamed et al., 2023).

Out of the models above, these models were also useful in identifying hate speech, but they were most useful in identifying threats to the public in the online world, which can be used for public safety and law enforcement purposes. For instance, it was determined that GPT-4 could detect potential threats with 97% accuracy, thus assisting the police in identifying people who may pose a threat to the public and causing violence, particularly in particular communities and areas (Kulkarni et al., 2023; Mossie & Wang, 2020). Furthermore, since hate speech is related to some suicide risk factors, machine learning models can identify those at high risk and help in preventing mental health crises (Coppersmith et al., 2018).

A major problem in all the studies was the complexity of the social media data especially from platforms such as X. This data is often characterized by noise, linguistic diversity, and terms only found in specific communities, for example, the hacker community, which decreases the accuracy of the model (Husain, 2021; McKeever et al., 2020). But the use of CNNs and sentence embedding in some studies helped to reduce these problems and increase the model's efficiency in identifying hate speech, threats, and abuse in social media text. Furthermore, the majority of researchers applied NLP tools to improve the model's sensitivity to hate and abusive language (Chakraborty & Seddiqui, 2019; Monnar et al., 2024). As compared to the deep neural networks, the NLP approaches based on the domain specific word embedding also addressed the problems of noise and jargon and significantly enhanced the detection rate. For instance, researchers who worked on the problem of imbalanced datasets, a typical challenge in hate speech detection, discovered that NLP methods decreased the noise and prejudice, which increased the model's accuracy (Mohamed et al., 2023).

The integration of BERT-based models with other approaches such as TextCNN and BiLSTM shows the strength of the ensemble models in hate speech detection (Huang et al., 2022). In this way, BERT has been used for encoding pre-trained models and TextCNN or BiLSTM for decoding to enhance the accuracy and confidence of hate speech detection. This new approach provided a better emotion recognition, which in turn improved the sensitivity and efficiency of the model. Furthermore, the application of ensemble models to various datasets including augmented data – highlights the possibility of hybrid strategies for improving hate speech detection across languages and platforms (Adewumi et al., 2023).

6. Conclusion

This literature review shows that BERT-based models are very accurate and fast in identifying hate speech across various languages and social media platforms; they also have better interpretability than other neural network-based models. The ability of BERT to increase the level of explainability is an significant advantage in a field that relies heavily on these qualities for practical use. However, other types of models, including machine learning and deep learning models such as CNNs and LSTM networks, have also performed well in equal measure, especially in multilingual and complex data sets.

Hate speech detection is not the only area where models like GPT-4 can help identify online threats; it has also performed very well in identifying cyber threats and abusive language, which has implications for public safety and mental health. This wider relevance shows the potential of AI-based solutions in assisting efforts to ensure public safety, for instance, in the fight against cyberbullying, suicide cues, and other forms of online aggression.

However, the field has still inherited some conventional problems, for instance, the problem of noisy, diverse, and imbalanced data that is typical for social media platforms. The research conducted with the help of NLP methods, including the domain-specific word embeddings, presents the potential solutions to these problems and how NLP can help to overcome the problems of language and community-specific jargon. Model ensembles such as the use of BERT encoding with TextCNN and BiLSTM decoding have also been found to improve the detection rates and are particularly useful in challenging or augmented data sets.

This review indicates that although BERT-based and other deep learning models are very efficient, there is a strong rationale to keep examining ensemble and hybrid strategies to solve the data issues in hate speech detection. Future work should also aim at enhancing the model's transparency and coping with ethical concerns such as data privacy as these are crucial in developing trustworthy and responsible AI systems for online threat analysis and hate speech.

References

[1] Adewumi, T., Sabry, S. S., Abid, N., Liwicki, F., & Liwicki, M., T5 for Hate Speech, Augmented Data, and Ensemble. Sci, Vol.**5**, Issue.**4**, pp.**3-7, 2023.** https://doi.org/10.3390/sci5040037

[2] Chakraborty, P., & Seddiqui, Md. H., Threat and Abusive Language Detection on Social Media in Bengali Language. 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT), pp.**1–6, 2019.** https://doi.org/10.1109/ICASERT.2019.8934609

[3] Coppersmith, G., Leary, R., Crutchley, P., & Fine, A., Natural Language Processing of Social Media as Screening for Suicide Risk. Biomedical Informatics Insights, 10, **2018.** https://doi.org/10.1177/1178222618792860

[4] Dehingia, N., McAuley, J., McDougal, L., Reed, E., Silverman, J. G., Urada, L., & Raj, A., Violence against women on Twitter in India: Testing a taxonomy for online misogyny and measuring its prevalence during COVID-19. PLOS ONE, Vol.**18**, Issue.**10**, pp.**e0292121, 2023.** https://doi.org/10.1371/journal.pone.0292121

[5] Elzayady, H., S. Mohamed, M., M. Badran, K., & I. Salama, G., Detecting Arabic textual threats in social media using artificial intelligence: An overview. Indonesian Journal of Electrical Engineering and Computer Science, Vol.**25**, Issue.**3**, pp.**1712, 2022.** https://doi.org/10.11591/ijeecs.v25.i3.pp1712-1722

[6] Faraj, A., & Utku, S., Comparative Analysis of Word Embeddings for Multiclass Cyberbullying Detection. UHD Journal of Science and Technology, Vol.**8**, Issue.**1**, pp.**55–63, 2024.** https://doi.org/10.21928/uhdjst.v8n1y2024.pp55-63

[7] Huang, Y., Song, R., Giunchiglia, F., & Xu, H., A Multitask Learning Framework for Abuse Detection and Emotion Classification. Algorithms, Vol.**15**, Issue.**4**, pp.**1-16, 2022.** https://doi.org/10.3390/a15040116

[8] Husain, F. A., Arabic Offensive Language Detection in Social Media. George Mason University, **2021.**

[9] Kulkarni, V., Baghwat, V., Patli, A., Kumari, S., & Deep Kour, S., A System to Identify Threats on Social Media Conversations and Providing Preliminary Legal Actions. International Journal of Research and Analytical Reviews, Vol.**10**, Issue.**4**, pp.**338–342, 2023.**

[10] Kwon, T., & Kim, C., Efficacy of Utilizing Large Language Models to Detect Public Threat Posted Online, **2023.**

[11] Manias, G., Mavrogiorgou, A., Kiourtis, A., Symvoulidis, C., & Kyriazis, D., Multilingual text categorization and sentiment analysis: a comparative analysis of the utilization of multilingual approaches for classifying twitter data. Neural Computing and Applications, Vol.**35**, Issue.**29**, pp.**21415–21431, 2023.** https://doi.org/10.1007/s00521-023-08629-3

[12] McKeever, S., Keegan, B., & Quieroz, A., Detecting Hacker Threats: Performance of Word and Sentence Embedding Models in Identifying Hacker Communications. AICS 2019 27th AIAI Irish Conference on Artificial Intelligence and Cognitive Science, **2020.**

[13] Mnassri, K., Farahbakhsh, R., & Crespi, N., Multilingual Hate Speech Detection: A Semi-Supervised Generative Adversarial Approach. Entropy, Vol.**26**, Issue.**4**, p.**344**, **2024.** https://doi.org/10.3390/e26040344

[14] Modha, S., Majumder, P., & Mandl, T., An empirical evaluation of text representation schemes to filter the social media stream. Journal of Experimental & Theoretical Artificial Intelligence, Vol.**34**, Issue.**3**, pp.**499–525, 2022.** https://doi.org/10.1080/0952813X.2021.1907792

[15] Mohamed, M. S., Elzayady, H., Badran, K. M., & Salama, G. I., An efficient approach for data-imbalanced hate speech detection in Arabic social media. Journal of Intelligent & Fuzzy Systems, Vol.**45**, Issue.**4**, pp.**6381–6390, 2023.** https://doi.org/10.3233/JIFS-231151

[16] Monnar, A. A., Perez Rojas, J., & Labra, B. P., Cross-lingual hate speech detection using domain-specific word embeddings. PLOS ONE, Vol.**19**, Issue.**7**, pp.**e0306521, 2024.** https://doi.org/10.1371/journal.pone.0306521

[17] Mossie, Z., & Wang, J.-H., Vulnerable community identification using hate speech detection on social media. Information Processing & Management, Vol.**57**, Issue.**3**, pp.**102087, 2020.** https://doi.org/10.1016/j.ipm.2019.102087

[18] Muneer, A., Alwadain, A., Ragab, M. G., & Alqushaibi, A., Cyberbullying Detection on Social Media Using Stacking Ensemble Learning and Enhanced BERT. Information, Vol.**14**, Issue.**8**, pp.**467, 2023.** https://doi.org/10.3390/info14080467

[19] Paraschiv, A., Ion, T. A., & Dascalu, M., Offensive Text Span Detection in Romanian Comments Using Large Language Models. Information, Vol.**15**, Issue.**1**, **8, 2023.** https://doi.org/10.3390/info15010008

AUTHORS PROFILES

Dr. Sameh Shamroukh, Ph.D. in Data Analytics, brings over 27 years of expertise in supply chain management, data analytics, technology, and academia. His extensive experience with global companies has given him deep insights into the complexities of data-driven decision-making. Throughout his career, he has leveraged advanced technologies to optimize processes and develop innovative strategies to address evolving industry challenges. Passionate about education, Dr. Shamroukh simplifies complex concepts, making them accessible and actionable for business professionals, students, and readers. He has published multiple peer-reviewed articles across various fields and authored several professional books supporting the supply chain and healthcare industries.

Dr. Teray Johnson is the Director of Data Automation and Transformation at Lifepoint Health, overseeing data analytics for operations and performance improvement. She holds a PhD in Data Science and has 9 years of healthcare experience, including managing patient transportation, leading nursing administration, and improving population health. She also serves on the board of the American College of Healthcare Executives, co-chairing the Career Development Committee. She has published several peer-reviewed articles on improving organizational culture and reducing employee burnout. In addition, Dr. Johnson is a member of the Decision Sciences Institute.