Sakshi Dinesh Pandit, 2026, 14:3
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

Social Media Forensic: Neural Network Techniques for Cyber Bullying Hate Speech Detection Under Uncertainty

Sakshi Dinesh Pandit[1], V. D. Dabhade[2]
[1]PG Student, [2]Associate Professor, Department of Artificial Intelligence & Data Science
MET Institute of Engineering, Nashik, India

Abstract— The rapid growth in online hostile behavior, in- cluding hate speech and cyberbullying, has created significant challenges for users and platform moderators. While early detection methods relied on traditional machine learning ap- proaches like SVM, Naïve Bayes, and Random Forests using hand-crafted features [3], [4], these techniques struggled with noisy, multilingual, and imbalanced social media data. Recent advances in deep learning, particularly transformer-based models such as BERT, RoBERTa, and architectures like CNN, LSTM, and BiLSTM, have substantially improved detection capabilities by automatically learning rich textual representations [5]–[8]. However, most existing systems generate deterministic, overcon- fident predictions even in ambiguous contexts, posing risks in high-stakes moderation scenarios. This review comprehensively examines hate speech and cyberbullying detection techniques with emphasis on uncertainty-aware learning approaches, in- cluding Monte Carlo dropout, Bayesian modeling, and ensemble methods. We identify critical research gaps related to dataset bias, annotation inconsistency, limited multilingual coverage, and insufficient robustness. Our findings highlight the necessity of integrating uncertainty quantification into detection systems to enhance reliability, interpretability, and real-world applicability, guiding future development of trustworthy and ethically respon- sible content moderation frameworks.

Keywords: Hate speech detection, cyberbullying, social me- dia forensics, uncertainty estimation, Bayesian neural net- works, Monte Carlo dropout, transformer models, deep learn- ing, multimodal analysis, content moderation

I. INTRODUCTION

Digital communication platforms including X (previously known as Twitter), Facebook, Instagram, YouTube, and TikTok have fundamentally reshaped human interaction, opinion shar- ing, and engagement in public conversations. However, these technological advances have been accompanied by a concern- ing escalation in detrimental online behaviors, particularly hate speech, harassment, and cyberbullying. Such hostile conduct typically manifests through brief, casual, and emotionally intense messages targeting individuals or communities based on characteristics including race, gender, religion, nationality, or similar attributes. Continuous exposure to this harmful content has been associated with mental health issues such as anxiety, depression, social isolation, and in severe instances, self-harm or suicidal thoughts, particularly among adolescents and vulnerable Populations [1,2]. Consequently, digital foren- sics and automated content moderation have emerged as vital Research domains aimed at identifying toxic material before it inflicts significant harm.

Initial computational methodologies for identifying hate speech and cyberbullying primarily employed conventional machine learning (ML) techniques including Support Vector Machines (SVM), Naïve Bayes (NB), Logistic Regression (LR), k-Nearest Neighbours (k-NN), Decision Trees, and Random Forests [3], [4]. These algorithms generally relied on manually engineered features such as n-grams, TF–IDF representations, sentiment analysis metrics, and dictionaries containing offensive or abusive language. Although these methods offered simplicity and computational efficiency, they exhibited declining performance when confronted with noisy, multilingual, code-switched, or class-imbalanced datasets—all characteristic of social media environments.

Additionally, manually designed features proved insufficient for capturing nuanced semantic and contextual information, including sarcasm or subtle forms of bullying.

With deep learning's success in NLP, recent studies have adopted architectures like CNN, RNN, LSTM, BiL- STM, and transformer models (BERT, RoBERTa, XLM-R) for hate speech and cyberbullying detection [5]–[8]. These models automatically learn rich text representations, handle large datasets, and capture long-range dependencies. Hybrid CNN/LSTM approaches combined with pre-trained embed- dings have achieved strong performance on Twitter and Face- book data [6,7], while other work has adapted these frame- works for low-resource and dialectal languages, addressing challenges like code-switching and class imbalance [9,10].

Despite these technological advancements, most current hate speech and cyberbullying detection systems function deterministically, producing single classification labels with probability scores often misinterpreted as confidence mea- sures, even when facing uncertain or out-of-distribution inputs. In high- stakes automated moderation contexts, false negatives enable harmful content spread while false positives cause unjust censorship, raising ethical and legal concerns. Therefore, recent trustworthy NLP research emphasizes that models must not only predict labels but also quantify prediction uncertainty [11], [12].

II. LITERATURE REVIEW

Methodologies for identifying hate speech and cyberbully- ing have undergone significant transformation over the last ten years, driven by the proliferation of social networking sites, rising incidents of digital harassment, and heightened need for automated moderation systems. This section pro- vides a systematic, thematically organized examination of current research, encompassing conventional machine learning techniques, deep learning models, transformer-based systems, cross-lingual challenges, multimodal approaches, uncertainty quantification methods, and dataset issues. The analysis fur- ther identifies critical shortcomings in existing literature and articulates research deficiencies that justify the development of uncertainty-informed detection frameworks.

Traditional Machine Learning Approaches

Early work on hate speech and abusive content detection relied heavily on classical machine learning classifiers and hand-engineered text features. Davidson et al. [13] used lo- gistic regression with TF–IDF representations to differentiate between "hate speech," "offensive language," and neither observing considerable ambiguity in language usage. Sim- ilarly, Nobata et al. demonstrated that combining lexical, syntactic, and sentiment-level features improves abusive lan- guage identification, though performance deteriorates when encountering informal slang, sarcasm, or unseen linguistic patterns. These models offered simplicity and interpretability but lacked the capacity to model semantic dependencies or contextual meaning.

Deep Learning-Based Methods

Deep learning introduced a major shift in this domain by enabling automatic feature extraction and improved contex- tual understanding. Badjatiya et al. [14] combined word em- beddings with CNN and LSTM architectures, demonstrating notable performance gains over traditional methods. Rizos et al .developed an ensemble of deep neural networks that cap- tured both local and long-range dependencies, which proved useful in cyberbullying detection where harmful content may span multiple messages. Studies such as Saha and Chandra incorporated attention mechanisms to emphasize critical of- fensive expressions, improving interpretability and robustness.

Transformer-Based Architectures

With the success of models such as BERT, RoBERTa, and their derivatives, transformer architectures have become the dominant approach for hate speech and cyberbullying detec- tion. Their ability to capture bidirectional context enables them to model nuanced and ambiguous expressions more effectively than

CNNs or RNNs. Mandl et al. [15] demonstrated signif- icant performance improvements in multilingual hate speech benchmarks when using transformer-based models. Muneer et al. [1] combined a fine-tuned BERT model with traditional machine learning classifiers in a stacking ensemble, offering improved stability on imbalanced datasets.

However, transformer models exhibit a critical drawback: they tend to produce highly confident predictions, even when incorrect. This overconfidence becomes problematic in real world moderation systems that rely on trustworthy clas- sification. High computational cost and the need for large GPU resources further complicate deployment for continuous monitoring in fast-moving social media environments.

Multilingual and Cross-Cultural Challenges

A significant proportion of hate speech research focuses on the English language, leaving a notable gap for resource- poor languages. Mnassri et al. [6] highlighted that offensive expressions differ greatly across cultures, and many datasets lack proper representation of dialectal variations, code-mixed text, and minority languages. Alzaqebah et al. [3] focused on Arabic cyberbullying detection and demonstrated that models trained solely on English perform poorly when transferred to morphologically different languages. Cultural context further complicates annotation, as phrases considered hateful in one region may be benign in another. These multilingual chal- lenges underscore the need for adaptable and uncertainty- aware models capable of recognizing ambiguity and mitigating bias.

Multimodal Hate Speech Detection

With the rise of memes, edited images, and short videos, hate speech increasingly appears in multimodal form. Text- only models fall short in such settings. The TrusV-HSD model [10] introduced multimodal classification by combin- ing visual augmentation with transformers to detect hateful memes. Other works incorporate facial expressions, scene con- text, or image–text alignment. However, multimodal datasets are scarce and often noisy, making training unstable. This complexity reinforces the value of uncertainty modeling to assess confidence, especially when handling ambiguous visual content.

Uncertainty-Aware Learning Approaches

Although widely adopted in other AI domains, uncertainty modeling has only recently gained attention in hate speech detection. Miok et al. [8] introduced Bayesian Attention Net- works, integrating Bayesian inference into attention layers to capture epistemic uncertainty. Basiri [9] applied Monte Carlo dropout to estimate predictive variance, enabling systems to identify potentially misclassified samples. More recent work by Yang et al. [7] examined uncertainty-aware cross-modal alignment and demonstrated its value in addressing noisy user-generated content. Despite promising results, the integration of uncertainty into transformer models for cyberbullying remains limited, highlighting a key research opportunity.

Comparative Summary of Key Studies

Table I summarises representative studies across different methodological categories.

TABLE I: Comparison of Major Studies in Hate Speech and Cyberbullying Detection

Study	Model	Dataset	Key Insights
Davidson et al. [13]	Logistic Regression	Twitter	Early lexical detection;struggles with ambiguity.
Badjatiya et al. [14]	CNN/LSTM	Twitter	DL improves feature learning; sensitive to noisy labels.
Muneer et al. [1]	BERT Ensemble	Custom	Improved stability on imbalanced data.
Miok et al. [8]	Bayesian Atten- tion	Multiple	Introduces epis- temic uncertainty modelling.
Yang et al. [7]	Cross-modal + Uncertainty	LREC	Enhances reliability for multimodal content.

III. METHODOLOGY

This review follows a structured and systematic methodol- ogy designed to ensure comprehensive coverage of the existing research on hate speech and cyberbullying detection, with a particular emphasis on uncertainty-aware approaches. The methodology includes four major stages:

1. identification of relevant literature,
2. screening and selection of studies,
3. thematic organization and analysis, and
4. Synthesis of findings.

Literature Search Strategy

To identify relevant academic work, a systematic and struc- tured review was executed across prominent digital reposito- ries and indexing services, including IEEE Xplore, ACM Dig- ital Library, SpringerLink, ScienceDirect, MDPI, and arXiv. This search was supplemented by examining proceedings from leading conferences such as AAAI, ACL, COLING, ICWSM, FIRE, and LREC.

The search strategy integrated both general and highly targeted terms to achieve comprehensive coverage. Principal keywords included "hate speech identification", "cyberbully- ing classification", "toxic language analysis", "offensive con- tent detection", "uncertainty quantification", "Bayesian neu- ral architectures", "Monte Carlo dropout", "transformer-based frameworks", "multimodal harassment", and "digital foren- sics". These expressions were combined through Boolean operators (AND, OR) and exact-phrase searches to refine the results and maintain precision across the chosen scholarly databases.

Inclusion and Exclusion Criteria

To maintain scientific rigor, the retrieved studies were screened based on the following criteria:

Inclusion Criteria:

- **Linguistic Scope:** Research on abusive language detection in English, Arabic (including dialects), and low-resource languages, with emphasis on multilingual, cross-lingual, and code-switched settings.
- **Thematic Coverage:** Studies addressing hate speech, cyber-harassment, offensive discourse, or toxicity, includ- ing fine-grained distinctions such as overt/covert and targeted/untargeted abuse.
- **Methodological Frameworks:** Implementation of ma- chine learning, deep learning, transformer-based, or uncertainty-aware models, particularly those handling di- alectal variation, class imbalance, adversarial robustness, and multilingual transfer.
- **Empirical Validation:** Use of benchmark or well-documented datasets with standardized evaluation metrics (precision, recall, F1-score, AUC) to ensure rigor and reproducibility.

Exclusion Criteria

- Non-English papers lacking reliable translations.
- Studies not directly addressing automated detection (e.g., purely sociological or psychological analyses).
- Duplicate records or low-quality preprints without sufficient methodological detail.
- Opinion pieces, blogs, and non-academic sources.

After applying these criteria, 82 papers were shortlisted from an initial set of more than 250 retrieved documents. From these, 45 highly relevant studies were selected for in- depth analysis, prioritizing novelty, methodological clarity, and contributions related to uncertainty estimation.

Data Extraction and Classification

Each selected study was examined in detail to extract key information regarding:

- Model type (traditional ML, CNN, LSTM, transformer, ensemble, Bayesian, etc.)
- Dataset used, including language, domain, and size
- Feature engineering and preprocessing techniques
- Evaluation metrics (accuracy, F1-score, precision, recall, AUC, etc.)
- Presence of uncertainty estimation or reliability assess- ment

- Identified limitations, challenges, and assumptions

The extracted information was then grouped into thematic categories, enabling structured comparison across methodologies.

Thematic Categorization and Analysis

To improve clarity and coherence, the literature was organized into major themes:

- Traditional machine learning approaches
- Deep learning-based detection frameworks
- Transformer-based and pre-trained language models
- Multimodal hate speech detection
- Multilingual and cross-cultural challenges
- Uncertainty-aware learning approaches
- Dataset characteristics and annotation difficulties

This thematic framework facilitated the identification of trends, strengths, and weaknesses across different research directions. It also supported a deeper understanding of how uncertainty modeling has emerged as a critical requirement in modern hate speech detection systems.

Synthesis and Identification of Research Gaps

Finally, a synthesis of findings was conducted to evaluate recurring themes, methodological limitations, and assumptions across the reviewed studies. Particular attention was given to the integration of uncertainty modeling within detection frameworks.

This analysis revealed several key research gaps, including the limited incorporation of uncertainty-aware transformer architectures, insufficient robustness in multilingual settings, persistent challenges in multimodal content interpretation, and inconsistencies in dataset annotation practices.

IV. RESEARCH GAPS

Despite substantial progress in automated hate speech and abusive content detection, several limitations continue to re- strict the development of robust and deployable systems.

First, the reliability of annotated datasets remains a significant concern. Many publicly available datasets rely on loosely defined labeling criteria, leading to subjective interpretations and inconsistencies. This issue is particularly evident in borderline cases where distinguishing between offensive language and explicit hate speech is nuanced. Consequently, models trained on such datasets may inherit annotation biases, affecting their generalization capability.

Second, the dynamic and evolving nature of online communication poses an ongoing challenge. Digital platforms frequently introduce new slang, coded expressions, emojis, and multimodal elements. Existing datasets often fail to capture these rapidly changing linguistic patterns, resulting in performance degradation when models are applied in real- world scenarios. This highlights the necessity for continuously updated datasets and adaptive learning mechanisms.

Third, current research predominantly emphasizes English- language corpora, whereas social media environments are in- herently multilingual. Models commonly struggle with translit- erated text, regional dialects, and code-mixed language, lim- iting their effectiveness across diverse linguistic communities. The lack of sufficient resources for low-resource and regional languages further exacerbates this limitation.

Additionally, adversarial manipulation represents a grow- ing concern. Users may intentionally modify text to evade detection systems. However, many existing approaches lack robustness against such adversarial inputs, raising questions regarding system reliability in practical deployments.

Finally, scalability and real-time implementation remain underexplored. While numerous models demonstrate strong performance in experimental settings, fewer are optimized for processing large-scale streaming data with minimal latency. Moreover, the integration of uncertainty estimation within real-time detection frameworks has received limited attention in existing literature.

Overall, these gaps underscore the need for improved dataset quality, multilingual adaptability, adversarial robustness, interpretability, and scalable real-time architectures to ensure effective real-world deployment.

V. SIGNIFICANCE OF THE STUDY

The prevalence of digital hostility on social platforms neces- sitates robust detection systems to safeguard mental health and maintain social stability. This review evaluates uncertainty- aware strategies as a vital advancement for modern, techno- logically and socially responsible content moderation.

A primary contribution of this work is its focus on un- certainty quantification—a frequently neglected research area. Unlike deterministic models that provide fixed, often overcon- fident outputs, uncertainty-aware architectures offer transpar- ent and cautious predictions. This transparency is essential for high-stakes moderation decisions, such as account suspensions or content removals, ensuring they are both interpretable and trustworthy.

By addressing inherent flaws in current datasets and anno- tation practices, this study advocates for learning frameworks capable of navigating complex linguistic nuances, including sarcasm, evolving slang, and adversarial text manipulations. Such adaptability is critical for platforms serving diverse and linguistically rich global communities.

Furthermore, incorporating confidence metrics enhances overall system reliability and assists human moderators in prioritizing high-risk content. This integration reduces manual workloads and minimizes wrongful censorship, fostering fairer and more accountable moderation pipelines. Ultimately, this work synthesizes technical, ethical, and practical insights to guide the development of next-generation, context-aware, and responsible AI detection systems.

The proposed architecture for hate speech and cyberbullying detection is organized as a sequential pipeline, beginning with data collection and ending with an uncertainty-aware decision layer. Fig. 1 illustrates the overall workflow.

VI. PROPOSED ARCHITECTURE

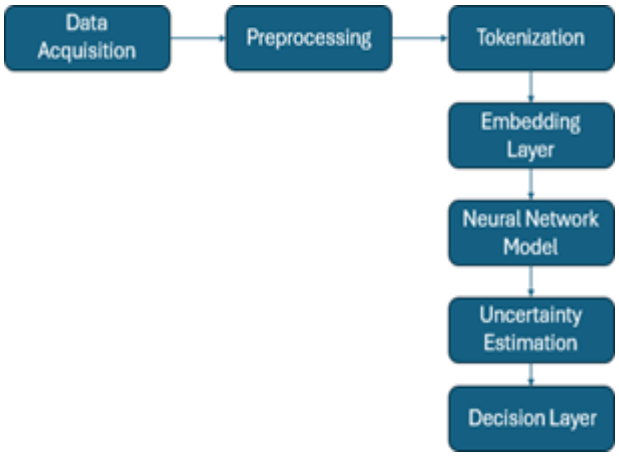

Fig. 1: Proposed neural network-based architecture for hate speech and cyberbullying detection.

The process starts with Data Acquisition, where raw posts, comments, and messages are collected from social media platforms. These inputs often contain noise, informal language, emojis, and platform-specific symbols, necessitating an initial Preprocessing stage. This stage handles cleaning operations such as lowercasing, removal of URLs and punctuation, normalization of slang terms, and elimination of irrelevant characters.

After preprocessing, the cleaned text undergoes Tokenization, where each sentence is divided into meaningful units (tokens). These tokens serve as the input for the Embedding Layer, which transforms discrete text into dense numerical vectors. Depending on the model, this layer may use Word2Vec, GloVe, fastText, or contextual embeddings such as BERT.

The vectorized representation is then passed to the Neural Network Model, which forms the core of the architecture. In this component, deep learning models such as CNNs, LSTMs, BiLSTMs, GRUs, or Transformer-based architectures learn hierarchical patterns associated with hateful, abusive, or cyberbullying behavior. The model generates a predicted class label along with an intermediate confidence score.

To incorporate reliability into the prediction, an additional Uncertainty Estimation module is included. Techniques such as Monte Carlo dropout, deep ensembles, or Bayesian neural layers quantify the epistemic and aleatoric uncertainties present in the model's output. This ensures that ambiguous or borderline cases are properly flagged.

Finally, the Decision Layer integrates both the predicted class and the estimated uncertainty score. High-confidence predictions are automatically labeled, while low-confidence or high-uncertainty cases are routed for manual review or further analysis. This design significantly improves robustness, especially in real-world applications where harmful content may exhibit ambiguous linguistic cues.

VII. RESULTS AND DISCUSSION

As this study is presented in the form of a comprehensive review, the discussion focuses on the analysis of findings reported across existing research works rather than on newly generated experimental results. The objective is to highlight common trends, methodological advantages, and identified limitations within the domain of hate speech and cyberbullying detection, particularly in relation to uncertainty-aware neural models.

Performance Trends in Existing Literature

Across the reviewed papers, deep learning models such as CNNs, LSTMs, BiLSTMs, GRUs, and Transformer-based architectures consistently outperform traditional machine learning methods. Studies show that contextual embeddings (e.g., BERT, RoBERTa, ALBERT) significantly improve classification accuracy, especially for complex and implicit forms of hate speech.

Research also indicates that uncertainty-aware models pro- vide more reliable predictions compared to deterministic neural networks. Works employing Monte Carlo dropout, Bayesian neural networks, and ensemble methods report improved calibration and robustness, particularly for noisy or ambiguous samples.

Effectiveness of Uncertainty Estimation

A major theme observed in the literature is the importance of measuring prediction reliability in sensitive tasks. Uncer- tainty estimation methods help identifies:

- linguistically ambiguous or sarcastic posts,
- multilingual and code-mixed expressions,
- out-of-vocabulary or unseen toxic keywords,
- Samples lacking contextual clarity.

Studies demonstrate that high-uncertainty samples often correlate with incorrect classifications. By incorporating un- certainty scores, researchers were able to improve human–AI collaboration, reduce false positives, and enhance moderation workflows on social platforms.

Comparative Insights From Reviewed Studies

Several reviewed works compare uncertainty-aware models with standard deep learning baselines. Common observations include:

- uncertainty-aware models provide more stable confidence scores,
- they offer better generalization in low-resource datasets,
- they reduce model overconfidence, a frequent issue in neural networks,
- They perform better under domain shift and noisy text conditions.

While Transformer-based models achieve high accuracy, uncertainty-aware versions of these models demonstrate im- proved interpretability and decision reliability, making them more suitable for real-world deployment.

Identified Limitations in Current Approaches

Despite their advantages, existing studies still face several challenges. Many datasets suffer from class imbalance, incon- sistent annotation guidelines, and cultural bias. In addition, most current models focus solely on text, overlooking multi- modal cues such as images, emojis, or metadata.

Uncertainty estimation techniques also introduce higher computational cost, particularly Bayesian neural networks and ensemble-based methods. Some studies report difficulty inte- grating uncertainty into real-time content moderation pipelines due to inference overhead.

Synthesis of Findings

Overall, the findings across the reviewed literature con- firm that the integration of deep learning with uncertainty estimation enhances both model performance and reliability. The combined approach is especially beneficial for ambiguous or borderline hate speech cases, where deterministic models often struggle. These insights underscore the importance of developing scalable, uncertainty-aware systems for safer and more effective content moderation on social media platforms.

VIII. DATASETS

In hate speech and cyberbullying detection, dataset selection is fundamental to model effectiveness, generalizability, and evaluation reliability. This review focuses on established public datasets commonly cited in previous work, highlighting their variations in size, linguistic diversity, and annotation frame- works.

Twitter-Based Data: Twitter remains a primary source for hate speech corpora, largely due to its accessible API and high frequency of user-generated toxic content. Frequently ref- erenced datasets classify posts into hate, offensive, or neutral categories. While these often rely on expert or crowdsourced annotations, they present inherent challenges such as brief text sparsity, subjective labels, and the rapid evolution of digital slang, all of which can hinder model robustness.

Cyberbullying and Forum-Based Data: Research in cyber- bullying often utilizes datasets from platforms like YouTube, Instagram, and online forums, reflecting real-world harassment patterns. These datasets frequently suffer from severe class imbalance, with harmful content being a minority. As a result, systems developed using this data must be specifically de- signed to handle skewed distributions and linguistic ambiguity to ensure reliable classification outcomes.

Future Work

Future research on hate speech and cyberbullying detection should focus on developing more comprehensive and contin- uously updated datasets that capture the evolving nature of online language, slang, and multimodal content. Uncertainty- aware deep learning models should be further explored to enhance the reliability of predictions, especially in ambiguous or borderline cases. There is also a strong need for multilingual and code-mixed datasets, as current systems struggle to gener- alize beyond English. Improving robustness against adversarial text manipulation, such as intentional misspellings or symbol insertions, remains another important direction. Finally, future work should aim to integrate explainability and uncertainty estimation into real-time moderation pipelines, enabling more transparent, fair, and scalable content moderation systems in practical social media environments.

IX. CONCLUSION

This review highlights the growing importance of uncertainty-aware computational approaches for detecting hate speech and cyberbullying on social media. While deep learning models have significantly improved the ability to identify abusive and harmful content, their deterministic and uninter- pretable nature limits their reliability in real-world moderation scenarios.

The literature reviewed shows that ambiguity in language, subjective annotations, evolving online expressions, and multilingual usage continue to challenge existing sys- tems. Incorporating uncertainty estimation into neural models offers a promising direction for overcoming these limita- tions by providing calibrated confidence scores, improving robustness, and enabling safer automated decision-making. However, challenges remain in building large, diverse datasets, handling code-mixed languages, mitigating adversarial inputs, and ensuring fairness and transparency. Overall, this study emphasizes the

need for next-generation moderation systems that combine uncertainty modeling, linguistic adaptability, and explainability to support trustworthy and effective online safety solutions.

REFERENCES

1. A. Muneer, S. A. Alzaqebah, and M. Alsharoa, "Cyberbullying detection on social media using stacking ensemble learning and Enhanced BERT model," Information, vol. 14, no. 8, p. 467, 2023. [Online]. Available: https: //w ww.mdpi.com/2078-2489/14/8/467
2. D. Sultan, T. A. Khan, and A. Shahid, "Cyberbullying-related hate speech detection using shallow-to-deep learning models," Computers, Materials & Continua, vol. 75, no. 3, pp. 5481–5498, 2023.
3. S. A. Alzaqebah et al., "Cyberbullying detection framework for short and imbalanced Arabic datasets," Journal of King Saud University – Computer and Information Sciences, 2023.
4. L. J. Thun, V. R. Vijayalakshmi, and K. V. Mithuna, "CyberAid: Are your children safe from cyberbullying? A mobile application for child safety," Journal of King Saud University – Computer and Information Sciences, 2022.
5. M. S. Jahan, M. P. Dey, and T. Chakraborty, "A systematic review on hate speech automatic detection," Neurocomputing, vol. 537, pp. 27–49, Jul. 2023. [Online]. Available: https:/ /doi.org/10.1016/j.neucom.2023.04.040
6. K. Mnassri, M. Benamara, and Y. Mathieu, "A survey on multilingual offensive language and hate speech detection," Journal of Big Data, vol. 11, 2024. [Online]. Available: https: //pmc.ncbi.nlm.nih.gov/articles/PMC11042037/
7. C. Yang, Z. Sun, and J. Li, "Uncertainty-aware cross-modal alignment for hate speech detection," in Proceedings of LREC–COLING 2024, pp. 2157–2168. [Online]. Available: https ://aclanthology.org/2024.lrec- main.1475.pdf
8. K. Miok, A. Mathur, and A. Karan, "Bayesian Attention Networks for reliable hate speech detection," Cognitive Computation, 2022. [Online].Available: https:// arxiv.org/abs/2007.05304
9. M. E. Basiri, "Prediction uncertainty estimation for hate speech and abusive language classification," 2019. [Online]. Available: https:/ /www.researchgate.net/publication/336087067
10. M. Singh, P. Gupta, and D. Kumar, "Trustworthy Hateful Meme De- tection via Visual Augmentation (TrusV-HSD)," arXiv preprint, 2024. [Online]. Available: https: //arxiv.org/abs/2409.13557
11. S. Fortuna and J. Nunes, "A survey on automatic detection of hate speech in text," ACM Computing Surveys, vol. 53, no. 3, pp. 1–30, 2020.
12. Z. Zhang and R. Luo, "Deep neural networks for cyberbullying detection on Twitter," IEEE Access, vol. 8, pp. 187–198, 2020.
13. T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in Proceedings of ICWSM, 2017.
14. P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep learning for hate speech detection in tweets," in Proceedings of WWW Companion, 2017.
15. A. Mandl et al., "Overview of the HASOC track: Hate speech and offensive content detection," in FIRE, 2019.