Automated Detection of Hate Speech and Toxic Comments Using Machine Learning and Natural Language Processing

Dhruvesh Vaghasiya, Nirma University, India
Aman Deep Singh, Nirma University, India
Dev Detroja, Nirma University, India
Vedant Vaghasiya, Nirma University, India

The Barcelona Conference on Arts, Media & Culture 2025
Official Conference Proceedings

Abstract

The proliferation of hate speech and toxic remarks in online communities presents considerable challenges to individuals, organisations, and society. This research examines the ramifications of detrimental communications and suggests an automated method for their identification utilising sophisticated machine learning (ML) and natural language processing (NLP) techniques. We examine multiple models, including deep learning architectures like BERT and GPT, to improve the precision of hate speech detection from extensive datasets obtained from platforms such as Twitter and Reddit. The study examines the complications involved with identifying hate speech, such as contextual reliance, sarcasm, and dataset biases, which frequently result in false positives and negatives. We provide a systematic review to assess current methodologies and their efficacy, while highlighting ethical considerations and the practical application of our approach. Our findings underscore significant deficiencies in existing research and propose new avenues for creating more efficient algorithms for identifying harmful information, thereby fostering healthier online environments.

Keywords: hate speech, machine learning, natural language processing, BERT, GPT, text classification

iafor
The International Academic Forum
www.iafor.org

Introduction

Hate speech and toxic remarks are an escalating problem in online forums, adversely affecting individuals, organisations, and society at large (Patel et al., 2024). The heightened dependence on digital communication platforms, social media, and online forums has rendered the dissemination of hazardous content a significant issue. Hate speech denotes derogatory, discriminatory, or menacing remarks aimed against individuals or groups based on race, ethnicity, gender, religion, sexual orientation, or other characteristics. Toxic remarks, although typically not legally classified as hate speech, exacerbate online harassment, cyberbullying, and emotional distress (Akshaya et al., 2025). Automated methods for identifying and managing harmful content are essential for sustaining healthy and inclusive online ecosystems. Manual moderation, although advantageous in many instances, is labour-intensive, expensive, and often inconsistent due to individual biases and varying interpretations of objectionable content. Moreover, given the vast volume of user-generated content produced every second, reliance solely on human censors is impractical (Alrehili, 2019; Miran & Yahia, 2023).

Machine learning (ML) and natural language processing (NLP) have arisen as potent solutions for addressing this issue. Automated systems can categorise text as poisonous, offensive, or neutral employing supervised learning, deep learning frameworks (such as BERT and RoBERTa), and sentiment analysis methodologies. These algorithms can be trained to identify patterns of harmful discourse utilising extensive datasets of annotated comments from platforms such as Twitter, Reddit, and Wikipedia (Devlin et al., 2018; Prabhu & Seethalakshmi, 2025). Detecting hate speech is challenging due to contextual reliance, irony, code-switching, and inherent biases in datasets. Current algorithms often demonstrate false positives (misidentifying neutral communication as hate speech) and false negatives (overlooking subtle or implicit toxicity). Resolving these challenges requires sophisticated NLP methodologies, robust datasets, and continuous enhancements in model fairness and interpretability (Mullah & Zainon, 2021).

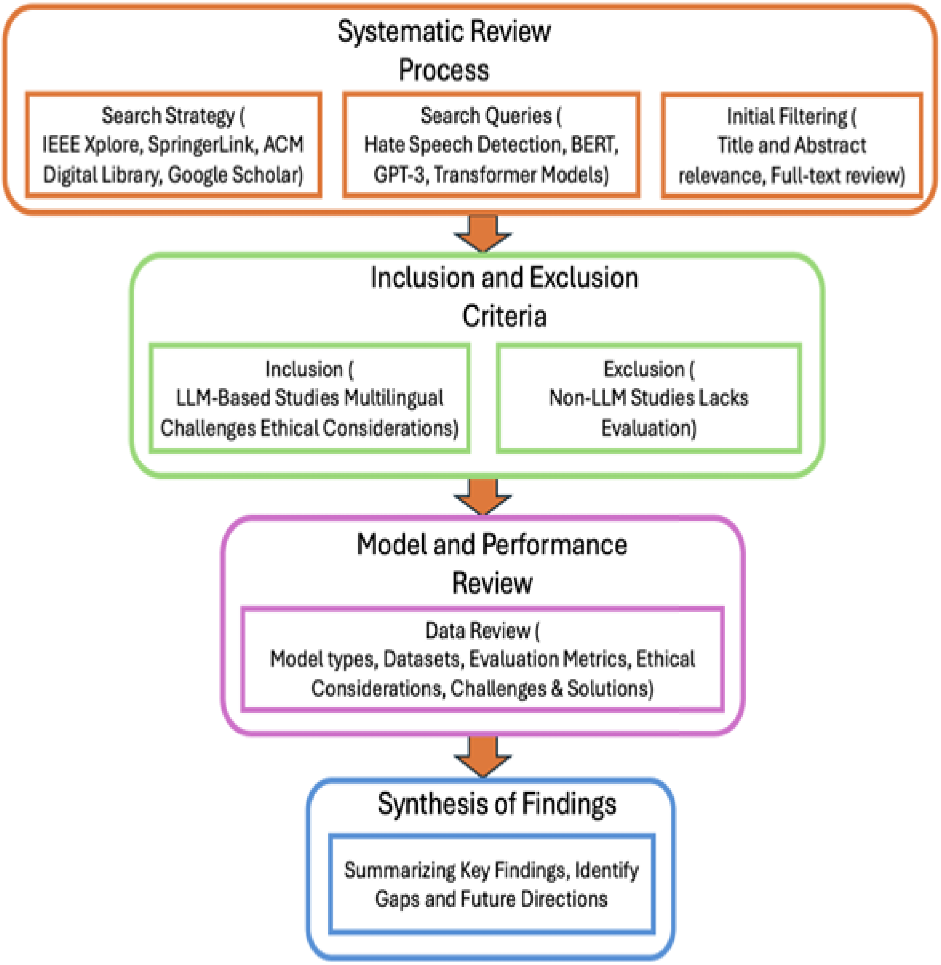
This paper presents an AI-driven approach for the detection and classification of hate speech and toxic comments utilising advanced machine learning methodologies. Our research examines several NLP models, feature extraction techniques, and classification algorithms to enhance the precision and dependability of hate speech detection (Jahan & Oussalah, 2023). Furthermore, we analyse the ethical ramifications, constraints, and prospective applications of our methodology in practical scenarios. Hate speech can be identified by many approaches, including manual evaluation by trained human reviewers who search for certain terms or patterns. Machine learning algorithms can identify hate speech by examining text for prevalent patterns and characteristics. Recurrent Neural Networks (RNNs) transformed machine learning, particularly in simulating sequential data such as language (Binns, 2018). RNNs encounter significant challenges, such as vanishing gradients and the management of long-term dependencies (Irfan et al., 2024; White, 2024). The transformer architecture substantially enhanced large language models (LLMs) by overcoming the limitations of recurrent neural networks (RNNs).

The transformer paradigm, with a self-attention mechanism, transformed the management of long-range dependencies by facilitating parallelisation and efficient processing of sequential data. This design facilitated the creation of sophisticated models, including Google's Bidirectional Encoder Representations from Transformers (BERT) and OpenAI's Generative Pretrained Transformer (GPT) series (G et al., 2021; Guillaume et al., 2022). These models have established new standards in natural language processing, excelling in numerous tasks and showcasing the transformative potential of the LLM architecture. Their success in machine translation and content generation illustrates their versatility and broadens the potential for academic research and commercial technology applications.

Figure 1 illustrates a representation of the manuscript review process (Alrehili, 2019; Mullah & Zainon, 2021). A comprehensive review was performed using databases like IEEE Xplore and Google Scholar, employing search terms such as "Hate Speech Detection" and "GPT-3." Following preliminary evaluation based on title and abstract relevance, inclusion and exclusion criteria were used. We examined research on LLMs that tackles language issues and ethical implications, while omitting non-LLM studies and those lacking evaluation. The model and performance review examined data from selected research, emphasising model types, datasets, evaluation measures, ethical considerations, and persistent challenges and solutions. In conclusion, the synthesis of findings summarised significant lessons that underscore limitations in the research and offer future options for enhancing hate speech detection algorithms.

Figure 1

A Representation of the Manuscript Review Process


Systematic Review Process
Search Strategy (IEEE Xplore, SpringerLink, ACM Digital Library, Google Scholar)
Search Queries (Hate Speech Detection, BERT, GPT-3, Transformer Models)
Initial Filtering (Title and Abstract relevance, Full-text review)
Inclusion and Exclusion Criteria
Inclusion (LLM-Based Studies Multilingual Challenges Ethical Considerations)
Exclusion (Non-LLM Studies Lacks Evaluation)
Model and Performance Review
Data Review (Model types, Datasets, Evaluation Metrics, Ethical Considerations, Challenges & Solutions)
Synthesis of Findings
Summarizing Key Findings, Identify Gaps and Future Directions

Literature Review

Background of LLM

Large language models (LLMs) represent a substantial advancement in natural language processing (NLP), employing deep learning techniques to comprehend and produce human language. Large Language Models have developed in conjunction with machine learning and artificial intelligence over several decades. Large Language Models (LLMs) originated in the 1950s, when pioneering AI systems, such as Alan Turing's Turing Test, sought to evaluate a machine's capacity to replicate human intelligence. The constraints of previous technology necessitated the advancement of superior approaches in later decades. GPT models, especially from GPT-2 onwards, have exhibited remarkable proficiency in producing cohesive and contextually relevant content across several subjects. These models are pre-trained on extensive datasets and subsequently fine-tuned for specific tasks, rendering them useful in many applications such as translation, summarisation, and creative writing (Alkomah & Ma, 2022). BERT introduced bidirectional training, enhancing performance on tasks such as question answering and sentiment analysis by considering both left and right context throughout all layers. Contemporary research emphasises on model bias, the requirement for substantial computational resources, and the difficulties of generalisation across languages and domains.

Brief History of Hate Speech Detection Using LLM

During the 1980s and 1990s, neural networks advanced significantly, particularly due to the backpropagation algorithm. These networks facilitated the creation of intricate models proficient at learning from substantial datasets. These models were constrained in scale and complexity due to time limitations in computation. During the early 2000s, statistical techniques such as Hidden Markov Models (HMMs) were prominent in natural language processing (NLP). These methodologies have been extensively utilised for applications such as part-of-speech tagging, voice recognition, and machine translation. Nonetheless, these models were constrained in their capacity to manage long-range linguistic connections, necessitating the creation of more intricate designs (Nascimento et al., 2023; Parihar et al., 2021).

Recent improvements, including transformer models like GPT and cross-lingual LLMs, have enhanced hate speech identification since 2016. This image underscores the continuous enhancement of methodologies and the growing efficacy of large language models (LLMs) in identifying nuanced and contextually rich hate speech. Comprehending the progression of LLMs is essential for evaluating their constraints in identifying hate speech. Identifying hate speech online is essential due to the widespread presence of harmful content on social media and forums. Conventional hate speech detection systems, dependent on keyword-centric approaches and basic machine learning classifiers, encountered difficulties in addressing linguistic complexities such as irony, context, and evolving vernacular. The utilisation of LLMs has improved the precision and resilience of hate speech detection. Initial endeavours in deep learning addressing this issue employed convolutional neural networks (CNNs) and recurrent neural networks (RNNs), which demonstrated potential but were constrained by their reliance on data type dependence and computational resources.

Refining large language models using hate speech datasets can markedly improve detection precision, as evidenced by studies. Research has utilised datasets such as the hate speech and offensive language dataset and the multilingual hate speech dataset to develop computers proficient in recognising hate speech across many languages and cultural situations. Large language models surpass conventional methods and earlier deep learning architectures in generalisation across diverse datasets.

Methodology

Modules Based Reservation

The methodology section delineated a systematic framework for assessing the efficacy of large language models (LLMs) in identifying hate speech. The Modules Based Reservation technique divides the research process into discrete modules, each accountable for a certain aspect of the study. The following modules are included:

- Literature Review and Search Strategy: Development of targeted search queries based on keywords such as "LLMs and hate speech detection," "NLP models," and "toxic content," alongside the identification of academic resources like IEEE Xplore, ACM Digital Library, Springer, and Google Scholar.
- Screening and Filtering: At first search is conducted, and duplicate entries are eliminated. Titles and abstracts are subsequently examined to exclude irrelevant studies.
- Full-text Review: A thorough evaluation of the selected papers is performed to confirm that each study satisfies the inclusion criteria.

Figure 1 illustrates the operational framework of the proposed system, visually depicting the modules and their interactions during the systematic review process.

Task Arrived Reservation

In the Task Arrived Reservation approach, research tasks are prioritised according to their identification sequence, guaranteeing that each step receives prompt attention. The procedure commences with the identification of pertinent studies using exhaustive search searches. This is succeeded by a sequential methodology wherein duties such as screening, duplicate elimination, and comprehensive text evaluation are executed consecutively. This sequential approach ensures that each element—from query formulation to final study selection—is systematically processed, hence maintaining the rigour and integrity of the research.

Checkpoints Reservation

The reservation of checkpoints is essential for upholding the quality and uniformity of the review process. During the study, multiple checkpoints are instituted to guarantee that the evaluation is thorough and methodologically rigorous:

- Initial Screening Checkpoint: Following the literature review and removal of duplicates, abstracts are assessed to swiftly discard irrelevant studies.
- Full-text Review Checkpoint: A comprehensive evaluation of the selected studies is conducted to ensure adherence to established inclusion criteria.
- Data Extraction and Analysis Checkpoint: Essential data components, such as model designs, datasets, and evaluation criteria, are methodically gathered and examined, guaranteeing that each study provides significant insights.

These checkpoints help to filter out irrelevant data while retaining high-quality, pertinent research for further analysis.

Task Mapper Reservation

The Task Mapper Reservation procedure entails methodically aligning each research task with designated modules in the study's workflow. This technique guarantees clarity in task allocation and smooth integration throughout the research process:

- Mapping Tasks to Modules: Every element of the study, from the preliminary search to the comprehensive data analysis, is distinctly delineated and allocated to certain modules.
- Ensuring Continuity: A seamless transition is upheld between jobs, guaranteeing that no essential procedures are neglected.
- Documentation and Review: Comprehensive documentation of task advancement—encompassing research selection, screening results, and data extraction—is maintained to guarantee transparency and facilitate systematic evaluation.

This organised framework elucidates the responsibilities of each phase while simultaneously improving the overall efficiency and efficacy of the research process.

Feedback-Based Reservation

Feedback-based Reservation employs an iterative methodology that perpetually enhances the research process through insights garnered at each phase. This flexible methodology guarantees that the research is both adaptable and comprehensive:

- Iterative Query Refinement: Initial search terms are optimised according to feedback from preliminary screenings to encompass all pertinent material.
- Methodological and Ethical Feedback: Ongoing assessment of methodological integrity and ethical implications, including model bias and the management of false positives/negatives, guides modifications to the study process.
- Data-Driven Modifications: The quantitative and qualitative evaluations of extracted data inform the enhancement of screening criteria and study selection, guaranteeing the retention of only the most relevant studies.

Future Scope

The domain of hate speech identification utilising large language models (LLMs) is primed for substantial progress as research persists in tackling its existing constraints. The following delineates critical domains for further investigation, highlighting their potential influence on enhancing the efficacy, scalability, and ethical implementation of LLM-based systems.

Advanced Detection of Implicit and Coded Hate Speech

A significant issue in hate speech detection is recognising implicit and coded hate speech. In contrast to explicit hate speech, characterised by direct slurs or abusive words, implicit hate speech utilises subtle indicators such as sarcasm, euphemisms, or cultural references, rendering it more difficult to identify. Future research should prioritise the creation of models that possess enhanced semantic and contextual comprehension, enabling them to decipher intricate verbal structures. Incorporating advancements in natural language processing, such as hierarchical attention mechanisms and sophisticated embeddings, can enhance LLMs' ability to analyse underlying intent. Furthermore, adaptive learning systems that consistently refresh depending on emerging patterns of coded language will be crucial in addressing the swiftly changing landscape of hate speech on internet platforms.

Strengthening Multilingual and Cross-Lingual Capabilities

The international scope of internet communication requires hate speech detection technologies that function efficiently in various languages. Contemporary LLMs demonstrate robust efficacy in resource-abundant languages such as English, although they encounter considerable obstacles in low-resource and under-represented languages owing to insufficient high-quality training data. Future research should focus on the advancement of multilingual and cross-lingual LLMs that can generalise across linguistic and cultural borders. Methods like zero-shot and few-shot learning enable models to excel in low-resource languages with limited labelled data. Furthermore, developing extensive, varied multilingual datasets that integrate cultural subtleties would improve the relevance of these models.

Addressing Bias and Ethical Concerns

Bias in training data and model predictions persists as a significant challenge in hate speech identification, frequently resulting in inequitable consequences, including the disproportionate targeting of certain groups while neglecting others. Future models must integrate comprehensive bias reduction measures throughout the data collecting and training phases (Cortiz & Zubiaga, 2021; Ullmann & Tomalin, 2020; White, 2024). This involves assembling balanced datasets that accurately reflect multiple viewpoints and employing methods such as adversarial training and counterfactual data augmentation to detect and rectify biases. Moreover, interpretability systems that offer transparent explanations for model decisions would enhance user confidence and accountability. Ethical issues must encompass not only justice but also openness, guaranteeing that content filtering systems do not stifle genuine speech.

Enhancing Scalability and Real-Time Moderation

The scalability and efficiency of hate speech detection algorithms are essential for handling the immense volume of content produced every day on platforms such as Twitter, Facebook, and YouTube. Future developments should prioritise the optimisation of LLM designs for real-time applications by enhancing computational efficiency while maintaining performance standards. Methods including model compression, trimming, and distillation can diminish computing demands, facilitating expedited content processing. Furthermore, distributed systems and edge computing offer potential for decentralised processing, enabling local content analysis instead of dependence on centralised servers.

Integrating Multimodal and Context-Aware Systems

Hate speech frequently extends beyond words, manifesting in multimodal formats including visuals, memes, and videos. Future research ought to concentrate on creating multimodal systems that integrate text analysis with visual and aural processing to identify hate speech across various formats (Irfan et al., 2024; Miran & Yahia, 2023). Models that can analyse text superimposed on photos or extract semantic intent from video captions will be crucial for moderating platforms such as Instagram, TikTok, and YouTube. Moreover, context-aware systems that consider user history, conversational dynamics, and cultural contexts might enhance detection accuracy by differentiating between malicious intent and innocuous phrases (Kiritchenko et al., 2021; Kovács et al., 2021).

Reducing Environmental Impact and Resource Requirements

The computing requirements for training and deploying large language models are substantial, prompting worries regarding their environmental impact and accessibility for smaller entities. Subsequent research ought to concentrate on creating energy-efficient designs and training methodologies to reduce the carbon impact of these models. Innovations including sparse attention mechanisms, efficient neural network architectures, and hardware acceleration can diminish energy consumption while maintaining performance integrity. Furthermore, utilising pre-trained models for transfer learning and fine-tuning might reduce the necessity for resource-demanding training cycles, hence enhancing the sustainability and accessibility of these models.

Expanding Applications Beyond Social Media

Although social media platforms are the principal focus of hate speech detection systems, their applicability may extend to numerous other domains. Educational platforms can utilise tools that identify bullying and abusive language, hence promoting safer virtual classrooms. Workplace communication systems may integrate hate speech identification to foster courteous professional relationships. In online gaming communities, where toxic behaviour frequently dominates, LLMs can assist in identifying and mitigating damaging language, fostering a more conducive atmosphere for participants (Yuan et al., 2023). Extending the use of LLMs to these sectors necessitates tailored modifications and partnerships with industry stakeholders to tackle distinct issues in each environment.

Algorithm

1. Initialise the pre-trained BERT model and the LSTM network for textual input processing.
2. Obtain reservation requests from users using the user interface.
3. Prepare each input by cleansing, normalising, and tokenising the text with the BERT tokenizer.
4. Input the tokenised data into the fine-tuned BERT model to obtain contextual embeddings.
5. Input the BERT embeddings into the LSTM network for sequential analysis and enhanced feature extraction.
6. Assess the LSTM outputs to ascertain the authenticity of reservations and their corresponding classifications.
7. Revise the reservation system by documenting the executed request and its result.
8. The amalgamation of fine-tuned BERT with LSTM facilitates the effective and efficient management of user reservation requests through deep contextual analysis and sequential processing (Tarun et al., 2024).

Conclusion

Large Language Models have markedly progressed the domain of hate speech identification, with considerable enhancements in precision and contextual comprehension. Nonetheless, obstacles persist, especially in mitigating biases included in both the training data and the models themselves (Laaksonen et al., 2021). The analysis indicates that although LLMs are useful in numerous contexts, their implementation necessitates ethical concerns and ongoing attempts to reduce prejudice. Future research ought to concentrate on augmenting the interpretability of these models, expanding their efficacy across many languages and circumstances, and establishing frameworks for equitable and transparent hate speech identification. By tackling these problems, we may harness the complete potential of LLMs to foster safer and more inclusive online environments (Albladi et al., 2025).

Nonetheless, despite the notable progress in LLM-based hate speech identification, some obstacles persist. Challenges such as model biases, the necessity for real-time processing optimisation, and cross-lingual generalisation remain inadequately addressed. These constraints underscore the necessity for continuous research aimed at augmenting model transparency, alleviating potential biases, and boosting generalisation across various languages and domains (Reyero Lobo, 2025). Future research should focus on enhancing bias detection and mitigation methodologies to guarantee that LLMs provide equitable and impartial predictions for hate speech identification. Furthermore, enhancing the scalability of LLMs for real-time moderation, especially on resource-constrained platforms, is a vital topic of investigation. Researchers should concentrate on creating more advanced multilingual models capable of functioning efficiently in low-resource language environments (Geetanjali & Kumar, 2025). The future of hate speech detection depends on the development of more resilient, egalitarian, and efficient LLM systems capable of adapting to the changing difficulties of online hate speech, hence fostering a safer digital world.

References

Akshaya, A., Sindhuja, K., Nomula, R., & Yadandla, S. (2025). Multilingual Toxic Comments Classification using Bert. *International Journal of Development Research*, *15*(2), 67737–67742. https://doi.org/10.37118/ijdr.29239.02.2025

Albladi, A., Islam, M., Das, A., Bigonah, M., Zhang, Z., Jamshidi, F., Rahgouy, M., Raychawdhary, N., Marghitu, D., & Seals, C. (2025). Hate Speech Detection Using Large Language Models: A Comprehensive Review. *IEEE Access*, *13*, 20871–20892. https://doi.org/10.1109/ACCESS.2025.3532397

Alkomah, F., & Ma, X. (2022). A Literature Review of Textual Hate Speech Detection Methods and Datasets. *Information*, *13*(6), 1–22. https://doi.org/10.3390/info13060273

Alrehili, A. (2019). Automatic Hate Speech Detection on Social Media: A Brief Survey. *2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA)*, 1–6. https://doi.org/10.1109/AICCSA47632.2019.9035228

Binns, R. (2018). Fairness in Machine Learning: Lessons from Political Philosophy. In S. A. Friedler & C. Wilson (Eds.), *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (Vol. 81, pp. 149–159). PMLR. https://proceedings.mlr.press/v81/binns18a.html

Cortiz, D., & Zubiaga, A. (2021). Ethical and technical challenges of AI in tackling hate speech. *International Review of Information Ethics*, *29*, 1–10. https://doi.org/10.29173/irie416

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, *1*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

G, A., Kumar, H., & D, B. (2021). Toxic Comment Classification using Transformers. *Proceedings of the 11th Annual International Conference on Industrial Engineering and Operations Management*, 1895–1905.

Geetanjali, & Kumar, M. (2025). Exploring hate speech detection: Challenges, resources, current research and future directions. *Multimedia Tools and Applications*, *84*, 1–37. https://doi.org/10.1007/s11042-025-20716-2

Guillaume, P., Duchêne, C., & Dehak, R. (2022, January 25). *Hate Speech and Toxic Comment Detection using Transformers*. Conférence francophone sur l'Extraction et la Gestion des Connaissance, France. https://www.researchgate.net/publication/360085544_Hate_Speech_and_Toxic_Comment_Detection_using_Transformers

Irfan, A., Azeem, D., Narejo, S., & Kumar, N. (2024). Multi-Modal Hate Speech Recognition Through Machine Learning. *2024 IEEE 1st Karachi Section Humanitarian Technology Conference (KHI-HTC)*, 1–6. https://doi.org/10.1109/KHI-HTC60760.2024.10482031

Jahan, M. S., & Oussalah, M. (2023). A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, *546*(126232), 1–30. https://doi.org/10.1016/j.neucom.2023.126232

Kiritchenko, S., Nejadgholi, I., & Fraser, K. C. (2021). Confronting Abusive Language Online: A Survey from the Ethical and Human Rights Perspective. *Journal of Artificial Intelligence Research*, *71*, 431–478. https://doi.org/10.1613/jair.1.12590

Kovács, G., Alonso, P., & Saini, R. (2021). Challenges of Hate Speech Detection in Social Media. *SN Computer Science*, *2*(2), 95. https://doi.org/10.1007/s42979-021-00457-3

Laaksonen, S.-M., Haapoja, J., Kinnunen, T., Nelimarkka, M., & Pöyhtäri, R. (2021). The Datafication of Hate: Expectations and Challenges in Automated Hate Speech Monitoring. *Front. Big Data*, *3*(Article 3), 1–16. https://doi.org/10.3389/fdata.2020.00003

Miran, A. Z., & Yahia, H. S. (2023). Hate Speech Detection in Social Media (Twitter) Using Neural Network. *Journal of Mobile Multimedia*, *19*(3), 765–798. https://doi.org/10.13052/jmm1550-4646.1936

Mullah, N. S., & Zainon, W. M. N. W. (2021). Advances in Machine Learning Algorithms for Hate Speech Detection in Social Media: A Review. *IEEE Access*, *9*, 88364–88376. https://doi.org/10.1109/ACCESS.2021.3089515

Nascimento, F. R. S., Cavalcanti, G. D. C., & Da Costa-Abreu, M. (2023). Exploring Automatic Hate Speech Detection on Social Media: A Focus on Content-Based Analysis. *SAGE Open*, *13*(2), 1–19. https://doi.org/10.1177/21582440231181311

Parihar, A. S., Thapa, S., & Mishra, S. (2021). Hate Speech Detection Using Natural Language Processing: Applications and Challenges. *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)*, 1302–1308. https://doi.org/10.1109/ICOEI51242.2021.9452882

Patel, D., Dutta Pramanik, P., Suryawanshi, C., & Pareek, P. (2024). Detecting toxic comments on social media: An extensive evaluation of machine learning techniques. *Journal of Computational Social Science*, *8*(20). https://doi.org/10.1007/s42001-024-00349-5

Prabhu, R., & Seethalakshmi, V. (2025). A comprehensive framework for multi-modal hate speech detection in social media using deep learning. *Scientific Reports*, *15*(1), 13020. https://doi.org/10.1038/s41598-025-94069-z

Reyero Lobo, P. (2025). *Addressing Bias in Hate Speech Detection: Enhancing Target Group Identification with Semantics* [PhD Thesis, The Open University]. https://doi.org/10.21954/ou.ro.00102419

Tarun, V. G., Sivasakthivel, R., Ramar, G., Rajagopal, M., & Sivaraman, G. (2024). Exploring BERT and Bi-LSTM for Toxic Comment Classification: A Comparative Analysis. *2024 Second International Conference on Data Science and Information System (ICDSIS)*, 1–6. https://doi.org/10.1109/ICDSIS61070.2024.10594466

Ullmann, S., & Tomalin, M. (2020). Quarantining online hate speech: Technical and ethical perspectives. *Ethics and Information Technology*, *22*(1), 69–80. https://doi.org/10.1007/s10676-019-09516-z

White, J. (2024). Advancing Ethical and Accurate Hate Speech Detection with Machine Learning Techniques. *International Journal of Scientific Research & Engineering Trends*, *10*(2), 99–104. https://doi.org/10.61137/ijsret.vol.10.issue2.135

Yuan, L., Wang, T., Ferraro, G., Suominen, H., & Rizoiu, M.-A. (2023). Transfer learning for hate speech detection in social media. *Journal of Computational Social Science*, *6*(2), 1081–1101. https://doi.org/10.1007/s42001-023-00224-9

Contact email: aman_deep.singh@nirmauni.ac.in