International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(10s) 2026 pp. 82-90
https://doi.org/10.70917/ijcisim-2026-3568
Article

SOCIAL MEDIA HATEFUL CONTENT DETECTION USING DEEP LEARNING RESOURCES

G. T. Prabavathi[1], N. Premalatha[2]

[1]Department of Computer Science, Gobi Arts and Science College, Gobichettipalayam.
Email: gtpraba@gmail.com
[2]Department of Computer Science, Gobi Arts and Science College, Gobichettipalayam.
Email: premanatesan89@gmail.com

Abstract: Social media platforms enable large-scale communication where users express a wide range of emotions, including anger, frustration, solidarity, satire and political opinions. While these platforms promote interaction and information sharing, it also facilitates the rapid spread of hate speech, offensive language and identity-based hostility. Detecting such harmful content is essential to ensure safe and inclusive digital environments. Existing machine learning approaches rely on hand-crafted textual features, whereas deep learning models automatically learn semantic and contextual representations from data. More recently, transformer-based natural language processing techniques have significantly improved the ability to capture context in social media and code-mixed text. This survey presents a structured and comparative analysis of machine learning; deep learning, transformer-based, hybrid and context-aware approaches for hate speech detection. It examines feature engineering strategies, data characteristics, social media challenges, evaluation methods and interpretability techniques. Furthermore, the survey highlights the importance of multimodal integration, real-time deployment considerations and responsible content moderation systems. Overall, this survey provides a comprehensive review of existing approaches identifies key challenges and outlines emerging research directions in hate speech detection.

Keywords: Social Media, Hate Speech Detection, Machine Learning, Deep Learning, Transformer Models, Multimodal systems.

1. Introduction

Social media platforms such as Twitter/X, Facebook, Instagram and YouTube enable individuals to communicate with a large global audience with minimal effort. While these platforms facilitate rapid information sharing and public interaction, it also weakens existing mechanisms that once ensured accountability in public discourse. Social media provides opportunities for immediate worldwide communication, allowing users to share opinions, emotions and ideological perspectives without geographical restrictions. The continuously evolving nature of social media, where posts are constantly updated, re-shared and amplified, contributes to the rapid spread and reinforcement of various emotional expressions across online communities. Users on social media platforms express a wide range of emotions including happiness, excitement, anger, frustration, empathy and political opinions. While these emotional exchanges often foster communication and social bonding, it can also lead to hostile interactions when negative sentiments are expressed in aggressive or abusive ways.

The detection of hate speech on social media has become an essential research area due to the widespread dissemination of offensive, abusive and identity-targeted content online. Early machine learning techniques provided the foundation for automated text classification tasks. Support Vector Machines (SVMs), introduced by Cortes and Vapnik [1], offered a robust approach for separating data with maximal margins, proving highly effective in text classification tasks. Similarly, Naïve Bayes classifiers, as analyzed by McCallum [2], provided probabilistic models for text-based applications, particularly suitable for high-dimensional datasets typical in natural language processing.

Random Forests developed by Breiman [3], introduced ensemble learning approaches that combined multiple decision trees to improve predictive performance and robustness against overfitting. These classical methods established the methodological basis for detecting hate speech and other forms of abusive content.

The evolution of machine learning into deep learning enabled more sophisticated representations of textual data. Deep learning is a subset of machine learning that uses multi-layered artificial neural networks inspired by the human brain to analyze large datasets, recognize complex patterns and make autonomous decisions. Early machine learning research, such as Samuel's work on self-learning systems [4], laid the foundation for the development of intelligent models which later evolved into advanced techniques for applications such as text classification and hate speech detection.

Building on this, Hinton et al. [5] introduced Deep Belief Networks, capable of hierarchical feature extraction, which inspired subsequent neural architectures for natural language understanding. Long Short-Term Memory (LSTM) networks [6] addressed challenges in learning long-term dependencies in sequences, critical for capturing context in text data, while Convolutional Neural Networks (CNNs) [7] demonstrated efficacy in document recognition tasks by automatically learning spatial hierarchies in data. The Recurrent Neural Network (RNN) encoder–decoder framework proposed by Cho et al. [8] further advanced sequence-to-sequence learning, facilitating applications such as machine translation and contextual text understanding, which later influenced hate speech detection models. These innovations enabled the development of more accurate, context-aware models that could handle the complexity of human language.

Transformer-based models are deep learning architectures that use self-attention mechanisms to capture contextual relationships between words in a sequence. Multimodal hate speech detection refers to the integration of multiple data types, such as text, images and videos, to improve classification performance. Various researchers have focused on hate speech detection on social media by incorporating temporal, linguistic and contextual factors. Florio et al. [9] highlighted the role of time in hate speech propagation, emphasizing that temporal patterns influence detection performance. Explainable AI approaches, as investigated by Mehta and Passi [10], provide transparency and interpretability, enabling users to understand why certain content is flagged as offensive. The socio-political dimensions of online hate speech were analyzed by Chekol et al. [11], showing that context and cultural factors play a critical role in understanding severity and prevalence. Transfer learning approaches Priyadarshini et al. [12] have been used to fine-tune pre-trained models for offensive content detection. Multimodal approaches enhance hate speech detection by combining diverse data types, including text, images and videos. By integrating textual and visual information, these methods Perifanos, K., & Goutsos, D. [13] improve classification accuracy and capture complex patterns across social media platforms

BERT (Bidirectional Encoder Representations from Transformers) is a pre-trained transformer-based model that learns bidirectional context by analyzing words based on both the preceding and succeeding terms, making it highly effective for understanding semantic meaning in natural language. Key challenges in the field include data scarcity, annotation quality and the effective use of external resources Kovács et al. [14]. Comprehensive reviews of machine learning approaches have summarized advancements and limitations in hate speech detection Mullah & Zainon [15]. Comparative studies, such as Bag-of-Words and Term Frequency–Inverse Document Frequency (TF-IDF) feature representations Akuma et al. [16], help optimize model input for improved classification performance. Recent works have focused on multi-model frameworks that combine machine learning and deep learning methods to enhance detection accuracy Naseeb et al. [17]. Furthermore, language-specific adaptations Almaliki et al.[18] and ensemble learning with transformer-based architectures Mazari et al. [19] demonstrate the importance of contextualized embeddings for handling diverse social media data. Dual contrastive learning approaches Chavinda, K., & Thayasivam [20] and curated datasets Mody et al. [21] have further enabled robust model training in low-resource scenarios. Additionally, data augmentation techniques, such as back-translation and paraphrasing Beddiar et al. [22], improve model generalization, while ensemble learning and visualization-based methods Mubeen et al. [23], Turki & Roy [24] enhance interpretability and detection performance. Finally, comprehensive multi-label datasets, such as ETHOS (Explainable Hate Speech Detection Dataset) Mollas et al. [25], provide standardized benchmarks for evaluating models across multiple hate speech categories. The recent advancements in hate speech detection have increasingly focused on hybrid and transformer-enhanced architectures that combine multiple learning paradigms to improve robustness and generalization. Iftikhar et al. [26] proposed a hybrid transformer–ensemble framework that integrates deep contextual representations with ensemble learning strategies to effectively detect both hate speech and offensive language in social media environments. Similarly, Azhar et al. [27] introduced HateBertBN, a transformer-based hybrid model designed for Bangla hate speech detection, demonstrating strong performance across diverse social contexts and linguistic variations. In addition, Kumar et al. [28] developed a hybrid deep BiLSTM–CNN

architecture that combines sequential modeling and local feature extraction to improve detection performance in multi-social media datasets. Furthermore, Siddiqui et al. [29] proposed a fine-grained multilingual hate speech detection framework using transformers integrated with explainable AI techniques, enhancing both classification accuracy and interpretability across multiple languages.

Overall, the literature reflects a progression from classical machine learning algorithms to deep learning, transfer learning and ensemble-based methods, addressing the complex nature of social media hate speech. These advancements further extend to multimodal approach, enabling more comprehensive and robust detection across diverse data types and languages. The above mentioned research works overall emphasize the necessity of context-aware, explainable and adaptable models to accurately detect and mitigate harmful content across diverse platforms and languages.

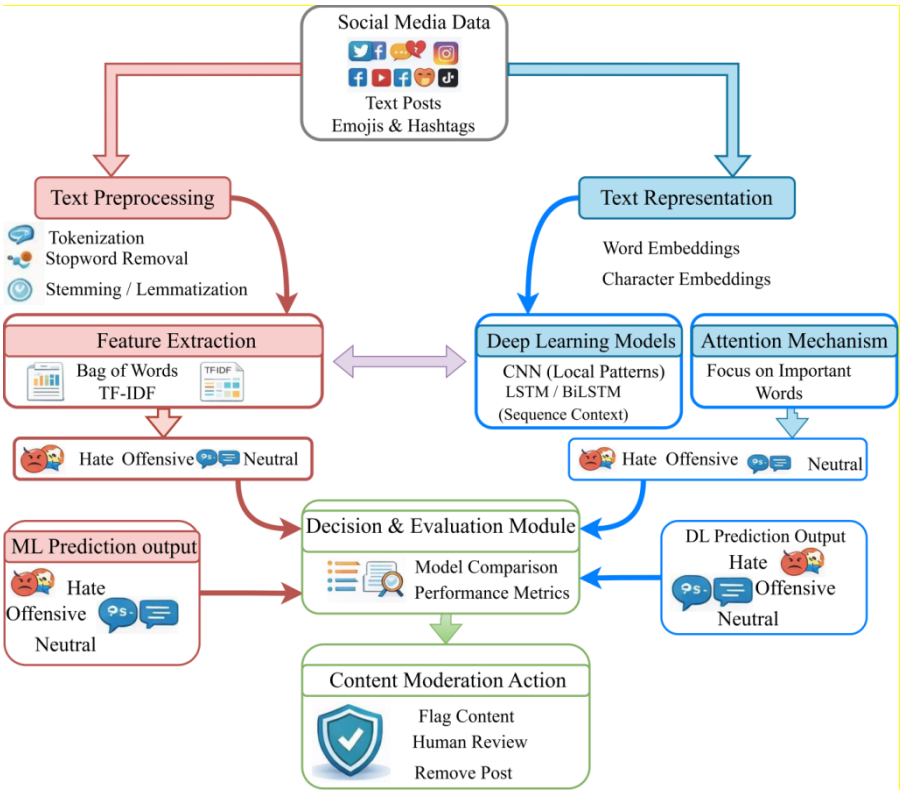

Figure 1: Integrated Machine Learning and Deep Learning Framework for Hate Speech Detection

Figure 1 presents an integrated framework approach for hate speech detection where social media data is processed through text preprocessing and feature extraction for machine learning models, alongside deep learning models that capture contextual and semantic features. The architecture incorporates attention mechanisms to enhance important textual patterns and combines outputs from both ML and DL models for improved classification accuracy. The final stage includes decision evaluation and content moderation actions such as flagging, human review and removal of harmful content.

2. Machine Learning-Based Hate Speech Detection

McCallum [2] focused on Naïve Bayes classifiers for text classification. The research compared different event models to determine which probabilistic assumptions lead to better classification accuracy on high-dimensional textual data. In the context of hate speech detection, this concept is implemented by treating each word in a social media post as a feature and calculating the probability of the post being hateful based on the occurrence of words. This approach allows fast and scalable classification of large volumes of text, forming a foundation for early ML-based hate speech detection systems.

Breiman, [3] introduced Random Forests, an ensemble of decision trees that improves predictive performance by combining multiple weak classifiers. The key concept implemented here is bagging and random feature selection

for each tree, which reduces overfitting and increases robustness. In hate speech detection, Random Forests are applied to textual feature vectors (like TF-IDF or Bag-of-Words) to classify posts as hateful or non-hateful, utilizing the ensemble's ability to handle noisy and high-dimensional social media text.

The research work by Florio et al., [9] investigated the temporal dimension of hate speech and implemented machine learning classifiers to detect how hate speech varies over time and spreads across social media platforms. By modeling the timing of posts and incorporating temporal features, it improved detection accuracy for posts that may be contextually hateful at certain times. The research emphasizes that understanding the timing of messages can significantly enhance ML-based hate speech detection.

Mehta & Passi, [10] Combined classical ML classifiers with explainable AI (XAI) techniques. It implemented ML models (such as Random Forest and SVM) to classify hate speech and used XAI methods to visualize feature importance, showing which words or phrases contributed most to a classification. This allows better understanding and transparency of ML predictions, helping developers and moderators trust automated hate speech detection systems.

Kovács et al., [14] Focused on practical challenges in applying ML to hate speech detection, especially data scarcity and annotation limitations. It implemented strategies to enhance performance with limited labeled data, such as using external resources and feature engineering. The concept here is that even with smaller datasets, ML models can still effectively detect hate speech if augmented with additional features and smart preprocessing.

A comprehensive review of machine learning algorithms for hate speech detection was done by Mullah & Zainon [15] . It examined implementations of SVM, Naïve Bayes, Random Forest and ensemble methods, discussing which models work best for different social media platforms and datasets. The concept implemented is model comparison and evaluation, which helps researchers to select appropriate ML methods for hate speech detection research.

Akuma et al.[16] compared Bag-of-Words and TF-IDF feature representations with classical ML classifiers (SVM, Random Forest) for detecting hate speech in live tweets. Different combinations of features and classifiers were evaluated to identify the setup that yielded the highest accuracy. This research demonstrates how feature engineering directly impacts the performance of ML-based hate speech detection systems.

Mubeen et al., [23] implemented a stacked ensemble classification method for detecting cyberbullying-related hate speech. The concept involves combining predictions from multiple base classifiers to create a more accurate meta-classifier. By stacking models, the ensemble classification improved detection accuracy and robustness compared to single ML models, illustrating the power of ensemble strategies in handling diverse social media text.

Turki & Roy [24] developed a framework combining count vectorizer features, word cloud visualization and ensemble machine learning models (Random Forest, XGBoost) for hate speech detection. Feature extraction and visualization techniques were used to highlight key terms contributing to hatefulness, followed by the application of ensemble classifiers for final detection. This approach demonstrated a practical machine learning pipeline that integrates feature representation, interpretability and ensemble learning for effective hate speech detection.

3. Deep Learning-Based Hate Speech Detection

Detecting hate speech on social media is challenging due to the subtlety and hierarchical nature of offensive content. Hinton et al. [5] proposed Deep Belief Networks (DBNs), composed of stacked restricted Boltzmann machines (RBMs), which automatically learn hierarchical feature representations from raw data. In the context of hate speech detection, DBNs capture latent patterns and semantic relationships in posts, enabling the model to distinguish between subtle, implicit hateful content and benign language. This research demonstrated that hierarchical feature learning improves classification accuracy compared to shallow or manually engineered feature models, making it effective for nuanced social media content; nevertheless, its performance is limited by high computational complexity and difficulty in training deep architectures efficiently.

Sequential dependencies in social media posts make detecting hate speech difficult, especially for context-dependent offensive expressions. Hochreiter & Schmidhuber [6] introduced Long Short-Term Memory (LSTM) networks, which solve the vanishing gradient problem of standard RNNs using gated memory cells. In hate speech detection, LSTMs process entire sentences and capture long-range dependencies between words, enabling detection of nuanced and context-sensitive offensive content.Experimental analysis demonstrated that LSTMs outperform

traditional RNNs in sequential text classification tasks; while effective, the model suffers from longer training time and higher computational requirements.

Hate speech often appears in localized textual patterns, such as repeated offensive phrases, slang or character-level variations. LeCun et al. [7] proposed Convolutional Neural Networks (CNNs) for document recognition, using convolutional filters and pooling layers to extract local features. Applied to hate speech detection, CNNs capture these linguistic patterns, enabling identification of abusive content in noisy social media text. It demonstrates high classification performance and reduced reliance on handcrafted features; despite its strengths, CNNs are limited in capturing long-range dependencies and contextual relationships in text.

Semantic understanding of sequences is critical for accurate hate speech detection. Cho et al. [8] introduced the RNN encoder–decoder architecture, which encodes a sequence into a fixed-length vector capturing contextual information. In social media applications, this enables holistic representation of posts and improved understanding of implicit hate speech. The architecture improved classification of context-dependent offensive content; nonetheless, its effectiveness is constrained by information loss in fixed-length representations and scalability issues for long sequences.

Limited labeled datasets make hate speech detection challenging, particularly on emerging social media platforms. Priyadarshini et al. [12] proposed a transfer learning approach by fine-tuning pre-trained deep learning models on hate speech datasets. This approach enables models to leverage semantic knowledge from large corpora and adapt effectively to domain-specific data. Experimental results demonstrated improved accuracy compared to traditional machine learning methods; Even so, the approach depends heavily on the quality of pre-trained models and may suffer from domain mismatch issues.

Facebook posts often contain noisy and ambiguous content. Naseeb et al. [17] proposed a multi-model hybrid framework combining machine learning classifiers with deep learning networks such as LSTM and CNNs. TF-IDF represents explicit lexical features, while deep learning models capture semantic patterns. A fusion strategy combines predictions from both approaches for final classification. This framework achieved improved precision, recall and F1-score compared to individual models; despite these improvements, increased model complexity and computational cost remain significant limitations.

Arabic social media is challenging due to dialectal variations and code-mixing. Almaliki et al. [18] introduced the BERT-mini model (ABMM), fine-tuned for social media hate speech detection. The model captures contextual embeddings that identify offensive and subtle abusive content effectively. ABMM demonstrated superior performance and reduced error rates compared to generic models; although efficient, its applicability is limited to specific languages and require adaptation for cross-lingual scenarios.

Hate speech is multi-dimensional, including insults and identity-based threats. Mazari et al. [19] proposed a BERT-based ensemble framework integrated with Bi-LSTM and Bi-GRU networks. The ensemble combines contextual embeddings with sequential modeling for multi-label classification. This approach improved detection accuracy across diverse datasets and enabled better identification of overlapping hate categories; In spite of its benefits, the model complexity and training cost are significantly higher compared to standalone models.

Detecting hate speech in low-resource languages is challenging due to limited data. Chavinda and Thayasivam [20] proposed a dual contrastive learning framework that enhances feature representation by aligning similar samples and separating dissimilar ones. This model improved detection of fine-grained hate speech in sparse data conditions. Experimental results demonstrated superior performance compared to conventional models; despite its superior performance, the framework requires careful tuning and increased training complexity.

High-quality datasets are essential for training robust models. Mody et al. [21] developed a curated dataset for social media hate speech detection with accurate annotations. This dataset enables models to learn reliable representations and improves classification performance. Models trained on this dataset achieved higher accuracy and generalization; still, dataset creation is time-consuming and may not fully capture real-world diversity.

Limited labeled data is a common challenge in hate speech detection. Beddiar et al. [22] proposed data augmentation techniques such as back-translation and paraphrasing to increase training data diversity. These methods improve model generalization and reduce overfitting. Experimental results showed enhanced performance in detecting subtle and noisy hate speech; yet, augmented data may introduce noise and semantic distortion.

Hate speech often requires multi-label classification. Mollas et al. [25] introduced the ETHOS dataset, enabling models to classify multiple categories of hate speech simultaneously. This improves detection of complex and

overlapping offensive content. Models trained on ETHOS achieved better multi-label classification performance; however, challenges remain in handling class imbalance and subjective annotation.

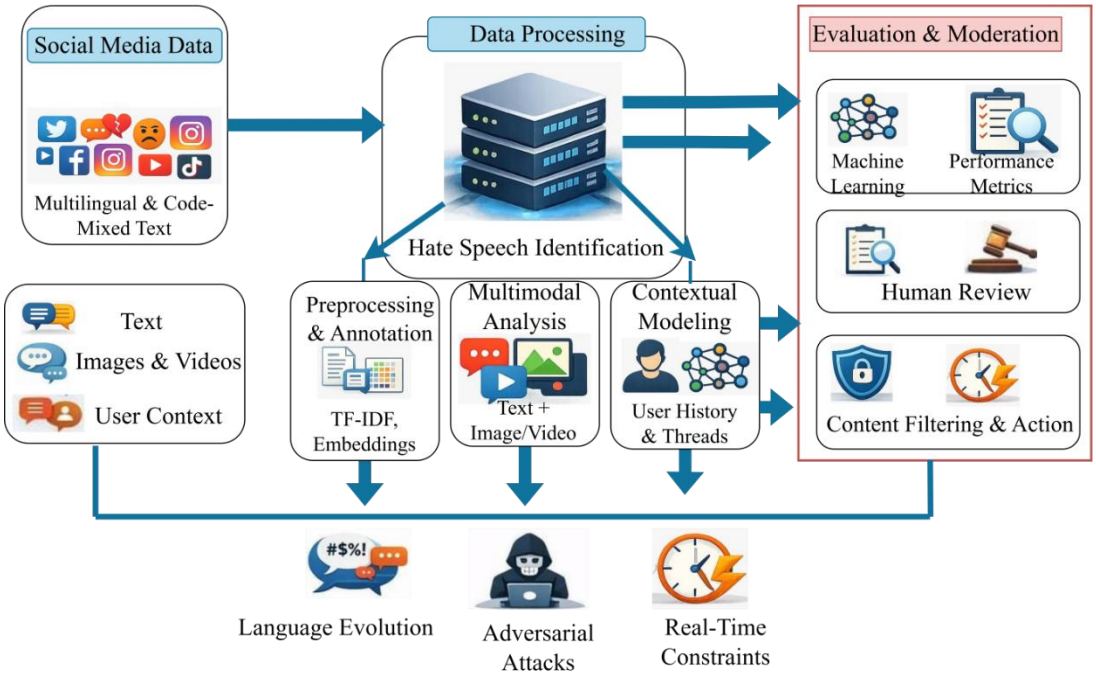

Figure 2: Multimodal and Context-Aware Hate Speech Detection Architecture

Figure 2 illustrates a multimodal hate speech detection system that integrates text, images, videos and user context to improve identification accuracy. The framework includes preprocessing, embedding and contextual modeling techniques such as user history and conversation threads to capture complex patterns in social media data. It further incorporates evaluation metrics and human moderation to ensure effective, real-time and responsible content filtering.

4. Transformer-based Methods

Transformer-based methods are advanced deep learning architectures that utilize self-attention mechanisms to capture contextual relationships between words in a sequence without relying on recurrent structures. These models process entire input sequences in parallel, enabling efficient handling of long-range dependencies and improving semantic understanding in natural language processing tasks. A widely used transformer model is BERT, which learns bidirectional context by analyzing both preceding and succeeding words, making it highly effective for understanding complex and context-dependent textual data.

Detecting hate speech on social media is challenging due to code-mixed content, dialectal variations and subtle contextual nuances. Transformer-based architectures have become effective tools for this task because it captures long-range dependencies and contextual semantics in textual data. Priyadarshini et al. [12] proposed a transfer learning approach using pre-trained BERT models fine-tuned on hate speech datasets. This framework utilizes semantic knowledge encoded in BERT embeddings to detect implicit and subtle offensive content. The fine-tuning process enables adaptation to specific social media platforms while retaining general language understanding. The approach demonstrated improved classification accuracy and better generalization in low-resource settings; yet, its performance depends on large pre-trained models and may suffer from domain adaptation challenges.

To improve multi-label detection and handle sequential dependencies, Mazari et al. [19] proposed a BERT-based ensemble framework integrated with Bi-LSTM and Bi-GRU networks. The framework combines contextual embeddings with sequential modeling to classify overlapping categories of hate speech. The Framework demonstrated superior performance compared to standalone BERT and classical models in terms of accuracy and F1-score; even so, the model introduces increased computational complexity and higher training cost.

For low-resource and underrepresented languages, Chavinda and Thayasivam [20] proposed a dual contrastive learning framework utilizing transformer embeddings. The approach aligns semantically similar hateful content while

separating dissimilar samples in embedding space, improving detection in limited data scenarios. The Framework improved classification accuracy and robust feature representation; though effective, the framework requires careful parameter tuning and increases training complexity.

Overall, these studies demonstrate that transformer-based models, including pre-trained BERT, compact variants, ensemble architectures and contrastive learning frameworks, provide strong contextual understanding for hate speech detection. These models achieved high accuracy and improved performance in handling complex, implicit and multi-dimensional content; Even so, challenges remain in terms of computational cost, data dependency and generalization across diverse domains.

4a. Hybrid and Context-Aware Approaches

Detecting hate speech on social media is challenging due to informal language, code-mixing and context-dependent expressions. To overcome these challenges, hybrid and context-aware approaches integrate machine learning, deep learning and transformer-based models to improve detection performance and robustness.

Iftikhar et al. (2025) [26] proposed a hybrid transformer–ensemble model to identify hate speech and offensive text in social media. To boost the classification robustness with respect to a wide range of data sets, the model is designed to fuse transformer-based contextual embeddings and ensemble learning techniques. Overall, the method performs well in recognizing semantic relationships and dealing with the complex social media text, particularly when the expression is implicit or context-dependent. But, it is expensive to compute since more than one model is involved and the training effort is enormous, which makes deployment in real time difficult.

Azhar et al. (2026) [27] created a Bangla contextual hate speech detection model named “HateBERTBN”, which is based on the BERT transformer. The model can accurately reflect dialectal differences and context in low-resource language domains, resulting in a better performance in detecting offensive nuances than the baseline models. Despite its good performance, the model is very language specific and needs to be adapted to be used for multilingual or cross domain applications, so that it can be generalized.

Kumar et al. (2024) [28] proposed a hybrid deep learning model that integrates BiLSTM with CNN for hate speech detection across multi-social media platforms. The BiLSTM component is used to capture the sequential dependencies of the text data, and the CNN component is used to extract the local feature patterns of the text data, such as offensive text phrases and word groups based on context. The integrated architecture enhances the performance of classification and gives a better accuracy than traditional machine learning techniques. The model, however, comes with increased complexity and training duration because it consists of several deep learning components.

Siddiqui et al. (2024) [29] proposed a fine-grained multilingual hate speech detection framework by employing a transformer-based model and explainable AI techniques. The methodology combines the strengths of transformer embeddings, which capture contextual semantics across multiple languages, with XAI techniques that increase the interpretability of the prediction by highlighting the important features responsible for the prediction. The model has shown to be more accurate than the traditional methods in the multilingual context, although in this context it consumes a lot of computational resources and needs to be carefully tuned to achieve the best results.

Overall, hybrid and context-aware approaches provide improved performance by integrating multiple modeling techniques to capture lexical, contextual and sequential features. Despite the effectiveness, these models face challenges related to computational cost, scalability and domain adaptability.

5. Conclusion

The problem of hate speech on social media remains a complex and evolving phenomenon due to informal language, code-mixing, social media information and the expressions that are related to the context. Classical machine learning systems, which rely on those attributes as TF IDF, n-grams and lexical patterns, provide easy to understand and efficient results, yet it is incapable of dealing with implicit, subtle and multi-aspect hate speech. The deep learning approaches including BERT, LSTM, GRU and transformer-based models have been found to perform better because of automatic learning of the semantic and contextual representations and the capacity to work within a social media or low-resource environment. Hybrid and context-aware methods take an extra step in enabling prediction with the usage of user-level data, social context and ensemble classifier to come up with powerful and socially sensitive forecasts. Fine-tuning, contrastive learning and transfer learning have been found to be helpful in cross-domain and cross-linguistic generalization in low resource situations. Combination architectures (also referred to as ensemble architectures) which are a mixture of several deep learning or ML models reduces misclassification and is better at

dealing with noisy text in social media. The attention algorithms and contextual embeddings teach the models the subtleties and implicit hate speech. The explainable AI models increase the readability, faith and interpretability which is critical in real life. Altogether, the discussion of feature engineering, deep learning and context-aware solutions develops a balanced system of an effective solution to hate speech on social media in the current times. Despite this progress, it is not an easy task to detect sarcasm; slang is constantly evolving and culturally inclined expressions.

A comprehensive review of hate speech detection techniques on social media has been presented, covering machine learning, deep learning, transformer-based and hybrid approaches. This survey analyzed various models in terms of the effectiveness, feature representation and applicability to real-world social media data. Future research directions should focus on developing cross-lingual and multilingual models to handle diverse languages, integrating multimodal data such as text, images and videos for improved detection accuracy and designing lightweight architectures that reduce computational cost while enabling real-time and scalable deployment in practical systems.

Reference

1. Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297. https://doi.org/10.1007/BF00994018
2. McCallum, A. (1998). A comparison of event models for naive Bayes text classification. *AAAI-98 Workshop on Learning for Text Categorization*.
3. Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324
4. Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development, 3*(3), 210–229. https://doi.org/10.1147/rd.33.0210
5. Hinton, G. E., Osindero, S., & Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation, 18*(7), 1527–1554. https://doi.org/10.1162/neco.2006.18.7.1527
6. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
7. LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE, 86*(11), 2278–2324. https://doi.org/10.1109/5.726791
8. Cho, K., van Merriënboer, B., Bahdanau, D., & Bengio, Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*. https://arxiv.org/abs/1406.1078
9. Florio, K., Basile, V., Polignano, M., Basile, P., & Patti, V. (2020). Time of your hate: The challenge of time in hate speech detection on social media. *Applied Sciences, 10*(12), 4180.
10. Mehta, H., & Passi, K. (2022). Social media hate speech detection using explainable artificial intelligence (XAI). *Algorithms, 15*(8), 291.
11. Chekol, M. A., Moges, M. A., & Nigatu, B. A. (2023). Social media hate speech in the wake of Ethiopian political reform: Analysis of prevalence, severity, and nature. *Information, Communication & Society, 26*(1), 218–237.
12. Priyadarshini, I., Sahu, S., & Kumar, R. (2023). A transfer learning approach for detecting offensive and hate speech on social media platforms. *Multimedia Tools and Applications, 82*(18), 27473–27499.
13. Perifanos, K., & Goutsos, D. (2021). Multimodal hate speech detection in Greek social media. *Multimodal Technologies and Interaction, 5*(7), 34.
14. Kovács, G., Alonso, P., & Saini, R. (2021). Challenges of hate speech detection in social media: Data scarcity and leveraging external resources. *SN Computer Science, 2*(2), 95.
15. 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.
16. Akuma, S., Lubem, T., & Adom, I. T. (2022). Comparing bag of words and TF-IDF with different models for hate speech detection from live tweets. *International Journal of Information Technology, 14*(7), 3629–3635.
17. Naseeb, A., Zain, M., Hussain, N., Qasim, A., Ahmad, F., Sidorov, G., & Gelbukh, A. (2025). Machine learning- and deep learning-based multi-model system for hate speech detection on Facebook. *Algorithms, 18*(6), 331.
18. Almaliki, M., Almars, A. M., Gad, I., & Atlam, E. S. (2023). ABMM: Arabic BERT-mini model for hate-speech detection on social media. *Electronics, 12*(4), 1048.
19. Mazari, A. C., Boudoukhani, N., & Djeffal, A. (2024). BERT-based ensemble learning for multi-aspect hate speech detection. *Cluster Computing, 27*(1), 325–339.
20. Chavinda, K., & Thayasivam, U. (2025). A dual contrastive learning framework for enhanced hate speech detection in low-resource languages. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)* (pp. 115–123).
21. Mody, D., Huang, Y., & De Oliveira, T. E. A. (2023). A curated dataset for hate speech detection on social media text. *Data in Brief, 46*, 108832.
22. Beddiar, D. R., Jahan, M. S., & Oussalah, M. (2021). Data expansion using back translation and paraphrasing for hate speech detection. *Online Social Networks and Media, 24*, 100153.
23. Mubeen, M., Muskan, A., Akram, A., Rashid, J., Alshalali, T. A. N., & Sarwar, N. (2025). Cyberbullying-related automated hate speech detection on social media platforms using stack ensemble classification method. *International Journal of Computational Intelligence Systems, 18*(1), 174.

24. Turki, T., & Roy, S. S. (2022). Novel hate speech detection using word cloud visualization and ensemble learning coupled with count vectorizer. *Applied Sciences, 12*(13), 6611.
25. Mollas, I., Chrysopoulou, Z., Karlos, S., & Tsoumakas, G. (2022). ETHOS: A multi-label hate speech detection dataset. *Complex & Intelligent Systems, 8*(6), 4663–4678.
26. Iftikhar, U., Ali, S. F., Mustafa, G., Bahar, N., & Ishaq, K. (2025). Beyond words: A hybrid transformer-ensemble approach for detecting hate speech and offensive language on social media. *PeerJ Computer Science, 11*, e3214. https://doi.org/10.7717/peerj-cs.3214
27. Azhar, T., Mahmud, T., Hasan, M. A., et al. (2026). HateBERTBN: A hybrid transformer-based model for Bangla hate speech detection across various social contexts. *Discover Computing, 29*, 15. https://doi.org/10.1007/s10791-025-09804-x
28. Kumar, A., Kumar, S., Passi, K., & Mahanti, A. (2024). A hybrid deep BiLSTM-CNN for hate speech detection in multi-social media. *ACM Transactions on Asian and Low-Resource Language Information Processing, 23*(8), 127. https://doi.org/10.1145/3657635
29. Siddiqui, J. A., Yuhaniz, S. S., Shaikh, G. M., Soomro, S. A., & Mahar, Z. A. (2024). Fine-grained multilingual hate speech detection using explainable AI and transformers. *IEEE Access, 12*, 143177–143192. https://doi.org/10.1109/ACCESS.2024.3470901