ISSN 1870-4069

Explainable Artificial Intelligence (XAI) for Hate Speech Detection Using Social Media Discourse

Muhammad Ahmad, Ildar Batyrshin, Grigori Sidorov*

Instituto Politécnico Nacional (IPN),
Centro de Investigación en Computación (CIC), Mexico City, Mexico

sidorov@cic.ipn.mx

Abstract. The proliferation of hate speech on social media platforms has escalated into a critical threat to online safety and societal harmony. Traditional deep learning models used for hate speech detection often operate as black boxes, offering no insight into their decision-making processes. This lack of transparency undermines user trust and hinders real-world deployment, especially in sensitive applications such as content moderation. To address this limitation, this study proposes an explainable deep learning model for hate speech detection in Urdu, a low-resource and morphologically rich language. The proposed framework leverages the UA-HSD-2025 dataset of Urdu tweets, applying tailored preprocessing and data augmentation to improve data quality. A Bidirectional Long Short-Term Memory (BiLSTM) network integrated with FastText embeddings captures subword information and contextual dependencies effectively. Experimental results show that the proposed model achieves 99.3% accuracy, significantly outperforming baseline approaches. To enhance interpretability, Local Interpretable Model-agnostic Explanations (LIME) are incorporated to provide word-level explanations for each prediction, transforming the black-box model into a transparent one. In conclusion, this study presents a highly accurate and explainable deep learning solution for Urdu hate speech detection, promoting accountability and trust in automated moderation systems for low-resource languages.

Keywords: Hate speech detection, explainable AI, social media, NLP, deep learning.

1 Introduction

Hate speech detection has emerged as a critical research area within natural language processing (NLP) due to the rapid growth of user-generated content on social media platforms [12,19,18]. Online spaces such as Twitter, Facebook, and Reddit have become central to communication, but they also facilitate the spread of abusive, offensive, and harmful language targeting individuals or groups

based on attributes such as race, religion, gender, or ethnicity [6,13,8]. The proliferation of such content can lead to serious societal consequences, including the reinforcement of stereotypes, psychological harm, and the escalation of real-world conflicts.

Traditional approaches to detecting hate speech relied heavily on manual moderation and rule-based systems, which are often insufficient due to the scale, diversity, and evolving nature of online language [4,11,16]. To address these challenges, researchers have increasingly turned to machine learning and deep learning techniques. Early methods utilized classical algorithms such as Support Vector Machines (SVM) and Naïve Bayes with handcrafted features like bag-of-words and TF-IDF representations [23,1,5]. However, these approaches often struggle to capture contextual nuances, sarcasm, and implicit forms of hate speech.

Recent advancements in deep learning, particularly transformer-based models such as BERT, RoBERTa, and GPT, have significantly improved the performance of hate speech detection systems [17,7,15]. These models leverage contextual embeddings and large-scale pretraining to better understand linguistic subtleties and semantic relationships in text. Furthermore, the integration of explainable artificial intelligence (XAI) techniques has become increasingly important to ensure transparency and trust in automated decision-making systems.

Despite significant advancements, hate speech detection remains a challenging task due to issues such as data imbalance, domain dependency, multilingual variations, and the ambiguity between hate speech and offensive language. Moreover, many state-of-the-art deep learning models, including BERT, operate as black-box systems, limiting their interpretability and reducing user trust in automated decisions. To address these challenges, this study proposes an explainable hybrid deep learning framework for hate speech detection. The proposed approach combines the strengths of deep learning models with pre-trained word embeddings to enhance classification performance. Furthermore, Explainable Artificial Intelligence (XAI) techniques are integrated to provide transparency in model predictions and improve interpretability. This framework aims to achieve robust, accurate, and interpretable hate speech detection, thereby contributing to safer, more reliable, and trustworthy online environments.

2 Related Work

The rapid expansion of social media platforms has intensified the spread of hate speech, raising critical concerns about digital safety, fairness, and ethical content moderation. This has led to a growing body of research focusing not only on improving detection performance but also on enhancing transparency, robustness, and multilingual capability of hate speech detection systems.

Ribeiro et al. [9] investigated transparency and accountability in hate speech detection, particularly in misogyny-related studies, by analyzing annotator documentation across 25 research papers. They proposed a structured framework consisting of six key dimensions, including annotator demographics, training,

and expertise, and introduced a weighted Annotator Metadata Transparency (AMT) score. Their findings revealed significant transparency gaps, especially in demographic and expertise reporting, and an interesting inverse relationship between model performance and transparency, where higher F1 scores often corresponded to poorer documentation quality.

Building on the need for interpretability, Eilertsen et al. [10] addressed the black-box nature of deep learning models by introducing Supervised Rational Attention (SRA). This framework aligns model attention with human-annotated rationales, enabling the model to focus on meaningful linguistic cues during classification. Evaluations on benchmark datasets demonstrated that SRA improves explainability while maintaining competitive performance and fairness, highlighting the importance of integrating human reasoning into model design.

In contrast, Xu et al. [22] focused on the challenges of detecting hate speech in Chinese social networks, where users often employ cloaking techniques to evade detection. They proposed MMBERT, a multimodal framework based on BERT, integrating text, speech, and visual modalities through a Mixture-of-Experts architecture. With a progressive training strategy and modality-specific routing, MMBERT achieved superior performance over both fine-tuned BERT and large language models, demonstrating robustness against adversarial and multimodal inputs.

Similarly, Mishra et al. [14] conducted a comparative study of traditional and advanced models for hate speech detection. They evaluated CNNs, LSTMs, and transformer-based models such as BERT, alongside hybrid architectures. Additionally, they explored text transformation techniques aimed at converting offensive content into neutral expressions. Their findings suggest that while transformer models achieve higher accuracy, hybrid and transformation-based approaches offer valuable improvements in real-world mitigation scenarios.

From a data-centric perspective, Umansky et al. [20] examined the role of annotation quality in fine-tuning large language models, specifically GPT-4o-mini. Using datasets labeled by annotators with varying expertise, they found that only high-quality annotations—particularly from trained experts—improve model performance. Low-quality annotations, in contrast, may degrade detection effectiveness, emphasizing that data quality is more critical than model complexity alone.

Focusing on ensemble learning, Aksoy et al. [3] addressed hate speech detection against LGBTQ+ individuals in Turkish tweets. They fine-tuned multiple large language models and proposed "Chosen Deep," an ensemble approach combining soft and hard voting strategies. Evaluated against traditional models and GPT-4, their method consistently outperformed baselines, demonstrating the effectiveness of ensemble learning for improving classification accuracy.

Addressing multilingual challenges, Ahmad et al. [2] focused on Arabic and Urdu hate speech detection by introducing the UA-HSD-2025 dataset, enriched with binary and multi-class annotations. They explored both translation-based and joint multilingual strategies, evaluating them using traditional machine learning, deep learning, and transformer models such as XLM-R. Their results

confirmed that multilingual transformer models significantly outperform conventional approaches across both languages.

Extending this direction, Usman et al. [21] developed a trilingual hate speech framework covering English, Spanish, and Urdu. They introduced a manually annotated dataset and conducted extensive experimentation using 41 different model configurations, including machine learning, deep learning, and transformer-based methods. Their study demonstrated that GPT-3.5 achieved the best performance, outperforming strong baselines such as XLM-R, particularly in low-resource languages like Urdu.

3 Study Design

3.1 Dataset Collection

The effectiveness of any hate speech detection system largely depends on the quality and representativeness of the underlying dataset. In this study, we utilize the UA-HSD-2025 dataset introduced by Ahmad et al. [2], which was originally developed for multilingual hate speech detection in Arabic and Urdu social media content. The dataset consists of manually annotated tweets collected from Twitter, ensuring real-world linguistic variability and contextual richness.

From the original dataset, we specifically focus on the Urdu subset to address the challenges of low-resource language processing. This subset contains 1,518 samples, where each instance is labeled into two classes: hate and not hate. This binary classification setup allows for a clear and structured formulation of the hate speech detection task.

The selected Urdu dataset is particularly challenging due to the presence of informal language, cultural expressions, and contextual ambiguity commonly observed in social media text. By leveraging this dataset, our study aims to evaluate the performance of advanced machine learning and deep learning models in a realistic low-resource scenario, contributing to more effective and inclusive hate speech detection systems.

3.2 Data Preprocessing

Data preprocessing plays a crucial role in improving the quality and consistency of textual data before model training. In this study, we applied a comprehensive preprocessing pipeline to the Urdu hate speech dataset. Initially, all text samples were converted to lowercase to ensure uniformity and reduce redundancy caused by case sensitivity. Subsequently, we removed noise such as URLs, user mentions, hashtags, punctuation marks, numbers, and special characters commonly present in social media content from Twitter.

Furthermore, stopwords were eliminated to reduce irrelevant linguistic information, and extra whitespace was normalized. Given the informal nature of social media language, additional cleaning steps were applied to handle elongated words, repeated characters, and informal spellings. These preprocessing steps ensured that the dataset was refined and suitable for effective feature extraction and model training.

3.3 Data Augmentation

To address data sparsity and improve model generalization, data augmentation techniques were employed on the training set. Since the dataset is relatively small and imbalanced, augmentation helps in generating diverse linguistic patterns while preserving semantic meaning.

We applied text-based augmentation strategies such as synonym replacement and paraphrasing to create additional training samples for the minority class. This approach helps the model better learn contextual variations of hate speech expressions. In addition, random perturbation techniques, including word shuffling within limited boundaries, were used to increase robustness against lexical variations.

By incorporating these augmentation methods, the dataset was enriched, leading to improved model stability and better generalization performance, particularly in handling unseen or noisy social media text.

4 Application of Models

In this study, we employ a combination of classical machine learning and deep learning models to effectively detect hate speech in Urdu social media text. Specifically, we utilize three machine learning algorithms—Decision Tree (DT), Random Forest (RF), and Support Vector Machine (SVM)—alongside two deep learning architectures, namely Convolutional Neural Network (CNN) and Bidirectional Long Short-Term Memory (BiLSTM).

The choice of feature representation differs based on the nature of the models. For machine learning models, textual data is transformed using Term Frequency–Inverse Document Frequency (TF-IDF). TF-IDF is selected due to its ability to convert text into sparse, fixed-length feature vectors while effectively capturing the importance of words based on their frequency in a document and rarity across the corpus. This makes it highly suitable for traditional classifiers such as DT, RF, and SVM, which perform well on structured and high-dimensional sparse data.

For deep learning models, we use dense word representations that preserve semantic and contextual relationships between words. Therefore, word embedding techniques such as GloVe and FastText are employed. These embeddings map words into continuous vector spaces where semantically similar words are positioned closer together. Unlike TF-IDF, word embeddings capture contextual meaning, syntactic relations, and semantic similarity, making them more effective for sequential models like CNN and BiLSTM.

The mathematical formulation of TF-IDF is defined as follows:

$$TF(t,d) = \frac{f_{t,d}}{\sum_k f_{k,d}}, \tag{1}$$

where $f_{t,d}$ represents the frequency of term t in document d.

$$IDF(t) = \log\left(\frac{N}{1+n_t}\right), \tag{2}$$

Where N is the total number of documents and n_t is the number of documents containing term t.

The final TF-IDF representation is given by:

$$TF\text{-}IDF(t,d) = TF(t,d) \times IDF(t). \tag{3}$$

For word embeddings, GloVe learns word representations by factorizing the global word co-occurrence matrix. Its objective function is defined as:

$$J = \sum_{i,j=1}^{V} f(X_{ij}) \left(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij}\right)^2, \tag{4}$$

where X_{ij} represents the co-occurrence matrix and w_i, $\tilde{w}_j$ are word vectors.

FastText further enhances word representations by incorporating subword information:

$$w_g = \frac{1}{|G|} \sum_{g \in G} z_g, \tag{5}$$

where g denotes character n-grams and z_g represents subword embeddings.

The Convolutional Neural Network (CNN) is used to extract local feature patterns from embedded sequences. The convolution operation is defined as:

$$c_i = f(W \cdot x_{i:i+k-1} + b), . \tag{6}$$

where $x_{i:i+k-1}$ is the input window, W is the filter matrix, and f is an activation function.

To capture long-range dependencies, we employ a Bidirectional Long Short-Term Memory (BiLSTM) network. The forward and backward hidden states are computed as:

$$\overrightarrow{h_t} = LSTM(x_t, \overrightarrow{h_{t-1}}), \tag{7}$$

$$\overleftarrow{h_t} = LSTM(x_t, \overleftarrow{h_{t+1}}). \tag{8}$$

The final representation is obtained by concatenation:

$$h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}]. \tag{9}$$

Overall, the combination of TF-IDF-based machine learning models and embedding-based deep learning models ensures a comprehensive evaluation of both shallow and deep representations for effective hate speech detection.

4.1 Explainability Using LIME

To enhance the interpretability of the proposed hate speech detection framework, we employ Local Interpretable Model-agnostic Explanations (LIME) for model explanation. Since deep learning models such as CNN and BiLSTM operate as black-box systems, their decision-making process is often not transparent. LIME addresses this limitation by providing local explanations for individual predictions, making the model behavior more interpretable and trustworthy.

LIME works by approximating the complex model locally with an interpretable surrogate model. Given an input text instance, LIME perturbs the input data by generating multiple slightly modified samples and observes the corresponding predictions of the trained model. It then assigns weights to these samples based on their proximity to the original instance and fits a simple linear model to approximate the local decision boundary.

In this study, LIME is applied to identify the most influential words contributing to the classification of a text instance as hate or non-hate. This allows us to highlight key linguistic features that drive model predictions, thereby improving transparency and user trust in the system. By visualizing word-level contributions, LIME helps bridge the gap between high-performance deep learning models and their interpretability requirements in sensitive tasks such as hate speech detection.

5 Overall Strategy

This study follows a structured and systematic pipeline for hate speech detection in Urdu social media text. The overall workflow begins with dataset selection from the UA-HSD-2025 corpus, followed by rigorous preprocessing steps including text cleaning, normalization, and noise removal to ensure data consistency. To enhance model generalization, data augmentation techniques are applied to increase data diversity and address class imbalance.

After preprocessing, the dataset is transformed into numerical representations using TF-IDF for machine learning models and word embeddings (GloVe and FastText) for deep learning models. Subsequently, multiple machine learning algorithms (Decision Tree, Random Forest, and Support Vector Machine) and deep learning architectures (CNN and BiLSTM) are trained and evaluated to capture both shallow and deep semantic patterns in the data.

To improve interpretability, Local Interpretable Model-agnostic Explanations (LIME) is applied to analyze model predictions and identify the most influential features contributing to classification decisions. This ensures transparency and increases trust in the proposed system.

Overall, the proposed framework integrates data preprocessing, augmentation, feature extraction, classification, and explainability into a unified pipeline for robust hate speech detection, where each stage is systematically connected to ensure effective learning and interpretation of the data. Figure [**?**] illustrates the complete methodology architecture, showing the flow and interconnection between all components of the proposed system.

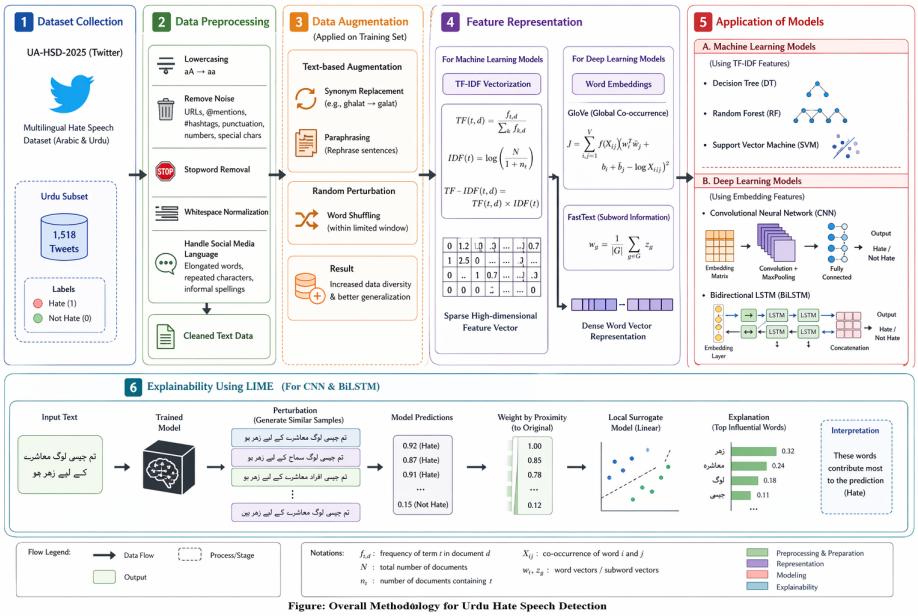

Figure: Overall Methodology for Urdu Hate Speech Detection

Fig. 1. Overall methodology architecture of the proposed hate speech detection framework.

6 Results and Analysis

In this section, we discuss the results based on the proposed methodology and evaluate the performance of both machine learning and deep learning models for hate speech detection in Urdu social media text. The experimental setup follows an 80–20 train-test split, where 80% of the data is used for training and 20% is used for testing to ensure fair evaluation of all models.

For machine learning models, standard configurations of Decision Tree, Random Forest, and Support Vector Machine are applied using TF-IDF features. For deep learning models, the experiments are conducted using the following parameters: maximum vocabulary size of 10,000, maximum sequence length of 100, embedding dimension of 300, 128 filters with a kernel size of 5, pooling size of 5, and 128 LSTM units. The models are trained for 9 epochs with a batch size of 16.

The results demonstrate that deep learning models outperform traditional machine learning approaches due to their ability to capture contextual and sequential information. Overall, this experimental setup provides a comprehensive comparison of different learning paradigms for effective hate speech detection.

Table 1. Performance comparison of machine learning models for binary hate speech detection.

Model	Precision	Recall	F1-score	Accuracy
SVM	0.975	0.974	0.973	0.974
RF	0.984	0.984	0.984	0.984
DT	0.974	0.974	0.974	0.974

6.1 Results for Machine Learning Models

Table 1 shows the performance comparison of machine learning models for binary hate speech detection, where the task is to classify text into *hate* and *not hate* categories. The evaluation is based on precision, recall, F1-score, and accuracy to ensure a comprehensive performance assessment.

The results indicate that all models perform strongly on the binary classification task. Among them, the Random Forest (RF) model achieves the best performance with an accuracy of 98.4%, along with consistently high precision, recall, and F1-score values of 0.984. This demonstrates the effectiveness of ensemble learning in capturing complex patterns within the TF-IDF feature space.

The Support Vector Machine (SVM) also shows competitive performance with an accuracy of 97.4%, highlighting its robustness in handling high-dimensional sparse textual representations. Similarly, the Decision Tree (DT) model achieves comparable results; however, its performance is slightly lower due to its sensitivity to data variations and tendency toward overfitting.

Overall, these findings suggest that ensemble-based approaches such as Random Forest provide better generalization for hate speech detection, while all models demonstrate strong capability in distinguishing between hate and non-hate content in Urdu social media text.

6.2 Results for Deep Learning Models

Table 2 presents the performance comparison of deep learning models for binary hate speech detection using different word embedding techniques, including FastText and GloVe. The evaluation is carried out using precision, recall, F1-score, and accuracy metrics to assess model effectiveness in distinguishing between *hate* and *not hate* classes.

The results show that BiLSTM with FastText embeddings achieves the best overall performance, reaching a remarkable accuracy of 99.3% along with equally high precision, recall, and F1-score values. This indicates that FastText effectively captures subword-level information, which is particularly useful for handling informal and morphologically rich Urdu text commonly found in social media.

In contrast, CNN with FastText embeddings also performs well, achieving an accuracy of 94.7%, demonstrating its ability to extract local semantic patterns from text. However, its performance is slightly lower than BiLSTM due to its limited capability in modeling long-term dependencies.

Table 2. Performance comparison of deep learning models using FastText and GloVe embeddings for binary hate speech detection.

Model	Precision	Recall	F1-score	Accuracy
FastText + CNN	0.947	0.947	0.946	0.947
FastText + BiLSTM	0.993	0.993	0.993	0.993
GloVe + CNN	0.899	0.508	0.460	0.508
GloVe + BiLSTM	0.840	0.799	0.713	0.799

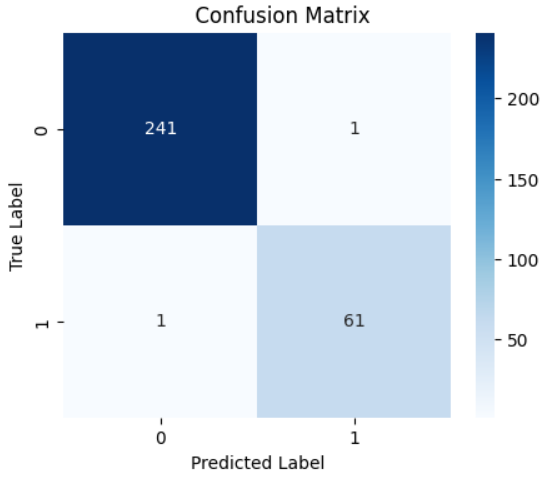

Fig. 2. Confusion matrix of the proposed hate speech detection model.

When using GloVe embeddings, a noticeable performance drop is observed. CNN achieves only 50.8% accuracy, while BiLSTM improves performance to 79.9%. This suggests that GloVe embeddings are less effective for noisy and context-dependent social media text compared to FastText, which benefits from subword information and better generalization.

Overall, the results demonstrate that BiLSTM combined with FastText embeddings provides the most robust performance for hate speech detection, highlighting the importance of both contextual modeling and effective word representation. Figure 2 illustrates the confusion matrix of the proposed model, which shows the distribution of correctly and incorrectly classified instances across the *hate* and *not hate* classes, thereby providing a detailed understanding of the model's classification performance.

6.3 Interpretability with LIME

The interpretability of the proposed model is demonstrated in Figure 3, where a sample Urdu text is analyzed using LIME. The original Urdu sentence shown in

یہ شیطانی سانپوں کا جھنڈ ہیں، ان کا سر قلم کر کے زندہ دفن کر دینا چاہیے۔

the figure is translated into English as ”*These people are extremely hateful and should not be trusted*”.

The proposed BiLSTM model with FastText embeddings classifies this instance as *Hate* with a confidence score of 1.00, indicating a highly confident prediction. The LIME explanation highlights the contribution of individual words toward the final decision, where words shown in blue strongly support the *Hate* class, while those in orange contribute toward the *Not Hate* class.

It can be observed that specific offensive and contextually negative terms in the Urdu text play a dominant role in driving the prediction toward the hate category. This demonstrates that the model effectively captures semantic meaning and contextual dependencies within the text.

Furthermore, the use of FastText embeddings enables the model to understand subword-level information, which is particularly beneficial for handling informal and morphologically rich Urdu language. Overall, the figure confirms that the proposed model is not only highly accurate but also interpretable, as it provides clear insights into which linguistic features influence the classification decision.

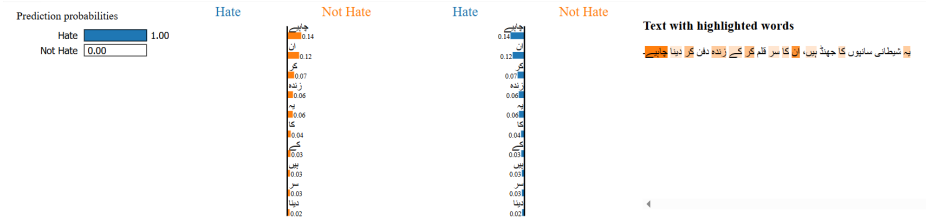

Fig. 3. LIME-based explanation of the proposed BiLSTM model with FastText embeddings for Urdu hate speech classification.

7 Conclusion and Future Work

This study addressed the growing threat of hate speech on social media by proposing an explainable deep learning model for Urdu, a low-resource and morphologically rich language. Unlike traditional black-box approaches, our framework integrates a FastText-based BiLSTM model with LIME explanations, achieving 99.3% accuracy on the UA-HSD-2025 dataset while providing word-level interpretability. The results demonstrate that combining subword-level embeddings with bidirectional contextual modeling effectively captures informal and ambiguous hate speech patterns. By transforming the detection process from an opaque decision system into a transparent one, this work enhances user trust and facilitates real-world deployment in content moderation systems.

Despite these promising outcomes, several limitations remain. The model was evaluated on a single dataset, which may limit generalizability. Future work

should focus on cross-dataset and cross-platform validation to assess robustness. Additionally, the current framework relies on supervised learning with manually annotated data, which is costly and time-consuming for low-resource languages. Exploring semi-supervised or few-shot learning approaches could reduce annotation dependency. Another promising direction is integrating multilingual and code-mixed hate speech detection, as Urdu users frequently blend languages. Finally, extending explainability beyond LIME to counterfactual or concept-based explanations may further improve model transparency and user confidence in real-time moderation systems.

References

1. Abubakar, H.D., Umar, M., Bakale, M.A.: Sentiment classification: Review of text vectorization methods: Bag of words, tf-idf, word2vec and doc2vec. SLU Journal of Science and Technology 4(1), 27–33 (2022)
2. Ahmad, M., Waqas, M., Hamza, A., Usman, S., Batyrshin, I., Sidorov, G.: Ua-hsd-2025: Multi-lingual hate speech detection from tweets using pre-trained transformers. Computers 14(6), 239 (2025)
3. Aksoy, Ç., Demirezen, M.U., Sağıroğlu, Ş.: Hate speech detection in turkish: An ensemble transformer-based deep learning approach. Engineering Applications of Artificial Intelligence 164, 113147 (2026)
4. Albladi, A., Islam, M., Das, A., Bigonah, M., Zhang, Z., Jamshidi, F., Seals, C.: Hate speech detection using large language models: A comprehensive review. IEEE Access 13, 20871–20892 (2025)
5. Alemerien, K., Al-Ghareeb, A., Alksasbeh, M.Z.: Sentiment analysis of online reviews: A machine learning based approach with tf-idf vectorization. Journal of Mobile Multimedia 20(5), 1089–1116 (2024)
6. Castaño-Pulgarín, S.A., Suárez-Betancur, N., Vega, L.M.T., López, H.M.H.: Internet, social media and online hate speech: Systematic review. Aggression and Violent Behavior 58, 101608 (2021)
7. Chapagain, S., Hamdi, S.M.: Advancing hate speech detection with transformers: Insights. In: Advances in Social Networks Analysis and Mining: Proceedings of the 17th International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2025). p. 432. Springer Nature (2025)
8. Chetty, N., Alathur, S.: Hate speech review in the context of online social networks. Aggression and Violent Behavior 40, 108–118 (2018)
9. Costa Ribeiro, L., Paes, A.: Does f1 fool you? a survey on annotator metadata transparency in hate speech detection. Journal of Information, Communication and Ethics in Society pp. 1–21 (2026)
10. Eilertsen, B., Björgfinsdóttir, R., Vargas, F., Ramezani-Kebrya, A.: Aligning attention with human rationales for self-explaining hate speech detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 37369–37378. AAAI Press (March 2026)
11. Geetanjali, Kumar, M.: Exploring hate speech detection: Challenges, resources, current research and future directions. Multimedia Tools and Applications 84(31), 38423–38459 (2025)
12. Kaur, S., Singh, S., Kaushal, S.: Abusive content detection in online user-generated data: A survey. Procedia Computer Science 189, 274–281 (2021)

13. Kiritchenko, S., Nejadgholi, I., Fraser, K.C.: Confronting abusive language online: A survey from the ethical and human rights perspective. Journal of Artificial Intelligence Research 71, 431–478 (2021)
14. Mishra, S., Thakur, S., Mamidi, R.: Enhancing hate speech detection on social media: A comparative analysis of machine learning models and text transformation approaches. arXiv preprint arXiv:2602.20634 (2026)
15. Mukherjee, S., Das, S.: Application of transformer-based language models to detect hate speech in social media. Journal of Computational and Cognitive Engineering 2(4), 278–286 (2023)
16. Qureshi, M.D.M., Qureshi, M.A., Rashwan, W.: Explainable ai for hate speech moderation: A stakeholder-centered and sociotechnical review. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 16(1), e70076 (2026)
17. Ramos, G., Batista, F., Ribeiro, R., Fialho, P., Moro, S., Fonseca, A., Silva, C.: A comprehensive review on automatic hate speech detection in the age of the transformer. Social Network Analysis and Mining 14(1), 204 (2024)
18. Rawat, A., Kumar, S., Samant, S.S.: Hate speech detection in social media: Techniques, recent trends, and future challenges. Wiley Interdisciplinary Reviews: Computational Statistics 16(2), e1648 (2024)
19. Rogers, D., Preece, A., Innes, M., Spasić, I.: Real-time text classification of user-generated content on social media: Systematic review. IEEE Transactions on Computational Social Systems 9(4), 1154–1166 (2021)
20. Umansky, N., Kubli, M., Kotarcic, A., Bronner, L., Kurer, S., Grech, P., Donnay, K.: Improving hate speech detection with large language models. European Journal of Political Research 1, 12 (2026)
21. Usman, M., Ahmad, M., Sidorov, G., Gelbukh, A., Tellez, R.Q.: A large language model-based approach for multilingual hate speech detection on social media. Computers 14(7), 279 (2025)
22. Xue, Q., Dou, Y., Shi, Z.R., Li, X., Gao, W.: Mmbert: Scaled mixture-of-experts multimodal bert for robust chinese hate speech detection under cloaking perturbations. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 34196–34204. AAAI Press (March 2026)
23. Zhang, L.: Feature extraction based on naive bayes algorithm and tf-idf for news classification. PLoS One 20(7), e0327347 (2025)