ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

Hate Speech Detection on Twitter Using XGBoost: A Large-Scale Dataset Analysis with Feature Engineering and Comparative Evaluation

[1]K.Vadivelan, [2]Dr. M.Sundara Rajan

[1]Research Scholar, PG and Research Department of Computer Science,

Government Arts College (Autonomous), Nandanam, Chennai-35, Tamil Nadu, India

[2]Associate Professor, PG and Research Department of Computer Science,

Government Arts College (Autonomous), Nandanam, Chennai-35, Tamil Nadu, India

DOI: https://doi.org/10.51583/IJLTEMAS.2026.150600222

Received: 13 July 2026; Accepted: 18 July 2026; Published: 25 July 2026

ABSTRACT

Social media platforms and Twitter in particular, has become a fertile channel for the rapid circulation of harmful discourse, including hateful and offensive language aimed at individuals and communities. Automatically flagging such content remains a demanding task within natural language processing, largely because micro blog text is short, informal, and heavily reliant on context for correct interpretation. This paper presents a hate-speech classification framework built around the Extreme Gradient Boosting (XGBoost) algorithm and evaluated on one of the largest publicly available labeled Twitter corpora, comprising roughly 96,973 tweets divided into three categories: hate speech, offensive language, and normal content. Each tweet is represented through a combined set of feature families: Term Frequency-Inverse Document Frequency (TF-IDF) vectors, dense word-embedding features, and tweet-level metadata such as hash tag frequency, mention count, retweet count, and capitalization ratio. Prior to feature extraction, tweets are cleaned through tokenization, stop-word removal, lemmatization, and URL stripping to reduce noise in the raw corpus. The hyper parameters of the XGBoost classifier were selected through grid search combined with stratified cross-validation. The tuned model reached an overall accuracy of 93.7%, a macro-averaged F1-score of 0.937, and an area under the curve above 0.94 for every class, surpassing the results obtained with Naive Bayes (78.4%), Support Vector Machines (82.1%), Random Forest (85.6%), LSTM networks (88.3%), and a fine-tuned BERT model (90.1%). These outcomes indicate that gradient boosting, when paired with a carefully engineered feature set, provides an accurate and computationally efficient alternative for large-scale hate-speech detection.

Keywords: Hate Speech Detection, Twitter, XGBoost, Feature Engineering, Text Classification

INTRODUCTION

Broad Research Context

Online social networks have reshaped the way people communicate. Services such as Twitter, Face book, and Reddit now host billions of daily exchanges covering personal expression, political debate, news distribution, and community interaction. Twitter alone is reported to process over 500 million tweets a day, making it one of the richest sources of real-time public opinion available for study. Alongside this openness, however, these same platforms have become channels for harmful speech, cyber bullying, radicalization, and coordinated harassment. Governments, civil-society groups, and the platforms themselves have progressively acknowledged the social harm caused by unchecked hate speech, prompting new regulation and a growing demand for automated moderation tools.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

Specific Research Domain

Hate-speech detection lies at the crossroads of natural language processing, computational social science, and content-moderation research. The task is inherently difficult because hateful expression is context-dependent, culturally relative, and frequently conveyed through euphemism, slang, or coded phrasing that slips past simple keyword filters. Class imbalance compounds the difficulty, since genuinely hateful posts make up only a small share of overall tweet volume. Rule-based systems lack the semantic flexibility required to capture these nuances, which is why the field has moved toward machine-learning and deep-learning solutions.

Problem Description

Current automated hate-speech systems tend to suffer from three recurring weaknesses. First, many models are trained on small, domain-specific corpora that do not transfer well to the broad linguistic variety found on Twitter. Second, deep architectures such as transformers, while accurate, are often too computationally expensive for real-time deployment at scale. Third, tweet-specific structural signals, including hash tags, mentions, and re tweet behavior, are frequently under-used even though they carry meaningful discriminative information. These shortcomings motivate a classification framework that is simultaneously scalable, interpretable, and accurate.

Overview of Approaches

A wide range of techniques has been applied to this problem. Lexicon-based methods compare tweets against accurate hate-word dictionaries, but their recall drops quickly as slang evolves. Classical machine-learning methods, including Naive Bayes, logistic regression, and Support Vector Machines (SVM), typically rely on bag-of-words or TF-IDF representations and perform reasonably well on balanced datasets. Ensemble methods such as Random Forest and gradient boosting improve on this by combining many weak learners. Recurrent architectures, particularly Long Short-Term Memory (LSTM) networks, capture sequential dependencies in text but need larger volumes of training data and compute. Transformer-based models such as BERT and RoBERTa currently define the state of the art in classification accuracy, yet their parameter counts, often above 110 million, make them impractical in resource-constrained production settings. This study examines XGBoost, a regularized gradient-boosting framework, as a middle ground that clearly outperforms classical methods and approaches transformer-level accuracy while remaining far faster at inference time.

Related Work

Davidson et al. (2017)

Journal/Conference: Proceedings of the 11th International AAAI Conference on Web and Social Media (ICWSM)

Overview: This influential study released a widely used Twitter hate-speech corpus consisting of 24,783 hate-speech tweets, 19,190 offensive tweets, and 53,536 normal tweets, collected through the Crowd Flower crowd sourcing platform.

Proposed Approach: A multi-class logistic-regression classifier was trained on TF-IDF features together with hand-built sentiment lexicons.

Conclusion: The model reached 91% accuracy separating hate speech from offensive language, though the authors noted continuing confusion between the two categories, an issue this paper addresses directly.

Waseem and Hovy (2016)

Journal/Conference: Proceedings of the NAACL Student Research Workshop, Association for Computational Linguistics

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

Overview: An early investigation of sexist and racist hate speech on Twitter that produced a dataset of 16,914 tweets labeled by experts and crowd workers.

Proposed Approach: The authors compared a character-level convolutional neural network with a logistic-regression model built on character n-grams.

Conclusion: Character-level features outperformed word-level representations for detecting racist language, largely because such language often involves deliberate misspellings and invented terms.

Badjatiya et al. (2017)

Journal/Conference: Proceedings of the 26th International Conference on World Wide Web Companion (WWW)

Overview: This work benchmarked several deep-learning architectures, namely Fast Text, CNN, and LSTM, for Twitter hate-speech detection.

Proposed Approach: The authors combined LSTM-derived embeddings with gradient-boosted decision trees to capture both semantic meaning and structural cues.

Conclusion: The hybrid LSTM-GBDT configuration achieved an F1-score of 0.93, showing that pairing deep representations with boosting can outperform either technique alone.

Zhang et al. (2018)

Journal/Conference: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)

Overview: This paper studied abusive-language detection using a deep convolutional network with attention, applied across several Twitter and Reddit corpora.

Proposed Approach: A multi-task learning setup was proposed to exploit shared linguistic patterns across hate speech, cyber bullying, and general toxicity detection.

Conclusion: Multi-task learning improved cross-dataset generalization, although it required a substantial amount of labeled data for each auxiliary task.

Founta et al. (2018)

Journal/Conference: Proceedings of the 12th AAAI International Conference on Web and Social Media (ICWSM)

Overview: The authors assembled what was, at the time, the largest crowd sourced Twitter dataset, containing 80,000 tweets labeled into four categories: hateful, abusive, spam, and normal.

Proposed Approach: Several classifiers, including SVM, Random Forest, and multilayer perceptions, were tested with diverse feature sets combining n-grams, sentiment scores, and user-network attributes.

Conclusion: Features drawn from follower-following relationships meaningfully improved classifier performance, lifting macro F1-scores above 0.88.

El Sherief et al. (2018)

Journal/Conference: Proceedings of the 2018 ACL Workshop on Abusive Language Online (ALW2)

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

Overview: This paper proposed a finer-grained taxonomy that separates directed hate speech, aimed at specific individuals, from generalized hate speech aimed at demographic groups.

Proposed Approach: Targeted SVM and logistic-regression classifiers were built using group-specific lexicons.

Conclusion: Directed hate speech proved considerably harder to identify than generalized hate speech, with F1-scores falling by as much as 12% on directed examples.

Mozafari et al. (2020)

Journal/Conference: IEEE Access, Volume 8, Pages 25018-25028

Overview: This study applied a fine-tuned BERT transfer-learning approach to hate-speech and offensive-content classification using the Davidson et al. dataset.

Proposed Approach: BERT was fine-tuned with additional regularization layers and a class-weighted loss function to counter severe class imbalance.

Conclusion: The fine-tuned BERT model reached 90.1% accuracy, setting a strong transformer-based reference point while underscoring its heavy computational demands.

Nobata et al. (2016)

Journal/Conference: Proceedings of the 25th International Conference on World Wide Web (WWW), ACM

Overview: This large study from Yahoo! examined hate-speech detection in comment sections using a broad feature set spanning linguistic, syntactic, distributional, and word2vec representations.

Proposed Approach: The authors proposed a regression-based scoring model for abusive content, enabling a continuous severity ranking rather than a strict binary label.

Conclusion: The regression formulation, combined with heterogeneous feature spaces, achieved an AUC of 0.95, reinforcing the value of integrating diverse feature types, a principle also adopted in the XGBoost framework presented here.

Existing Work

Existing hate-speech detection systems can broadly be grouped into three generations, distinguished by their underlying methodology and the period in which they emerged.

Rule-Based and Lexicon-Driven Systems

The earliest automated tools relied on curated lists of offensive terms and regular-expression patterns to flag potentially harmful posts, and systems of this kind were deployed for reactive moderation on platforms including YouTube, Face book, and early versions of Twitter. Although inexpensive to run and easy to interpret, such systems have clear limitations: they are easily bypassed through deliberate misspelling (for instance, "h8" in place of "hate"), they generate excessive false positives when benign users quote hateful language for the purpose of counter-speech, and they cannot capture meaning that depends on context. Precision on benchmark datasets for these systems rarely exceeds 65%, while recall for less obvious hate expressions typically falls below 55%.

Classical Machine Learning Approaches

The second generation of detectors applied supervised learning to bag-of-words or TF-IDF representations. Naive Bayes classifiers, valued for their simplicity and probabilistic grounding, produced baseline accuracies of roughly 78-80% on balanced subsets of the Davidson et al. (2017) dataset. Support Vector Machines using radial basis function kernels improved on this, reaching 82% accuracy, especially when enriched with character-level

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

n-gram features that capture morphological patterns typical of informal text. Logistic regression with L2 regularization delivered comparable accuracy with better-calibrated probability estimates. Even so, every classical method here faces the same semantic gap: words are treated as independent tokens, so these models cannot capture semantic relatedness, polysemy, or the negation patterns that often reverse sentiment in hateful text.

Deep Learning and Transformer Approaches

The third generation relies on neural architectures capable of learning hierarchical, context-sensitive text representations. Convolutional networks applied to word embeddings reached roughly 85% accuracy by learning local n-gram patterns through trainable filters. LSTM networks extended this further by modeling longer-range sequential dependencies, producing F1-scores of 0.883 on the Davidson benchmark. Pre-trained transformer models, particularly BERT, marked a qualitative jump forward: a fine-tuned BERT model achieves 90.1% accuracy and a macro F1-score of 0.899 on the same benchmark. Despite this accuracy advantage, transformer models generally require GPU acceleration for real-time inference, a demanding requirement for resource-constrained deployments, and they offer limited interpretability for regulatory or audit purposes.

Table 1 – Performance Comparison of Existing Hate Speech Detection Models

Model	Accuracy (%)	Precision	Recall	F1-Score	AUC
Naïve Bayes	78.4	0.761	0.749	0.755	0.810
Logistic Regression	80.2	0.793	0.787	0.790	0.838
SVM (RBF Kernel)	82.1	0.814	0.809	0.811	0.856
Random Forest	85.6	0.849	0.843	0.846	0.889
CNN (Word Embeddings)	86.8	0.861	0.857	0.859	0.901
LSTM	88.3	0.876	0.872	0.874	0.916
BERT (Fine-tuned)	**90.1**	**0.897**	**0.893**	**0.895**	**0.932**

Proposed Work

The proposed system is a multi-stage hate-speech detection framework that applies XGBoost to a rich, multi-modal feature set derived from a large-scale Twitter corpus. It is organized into five sequential modules: data ingestion and preprocessing, feature engineering, model training and optimization, evaluation, and a deployment interface.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

System Architecture

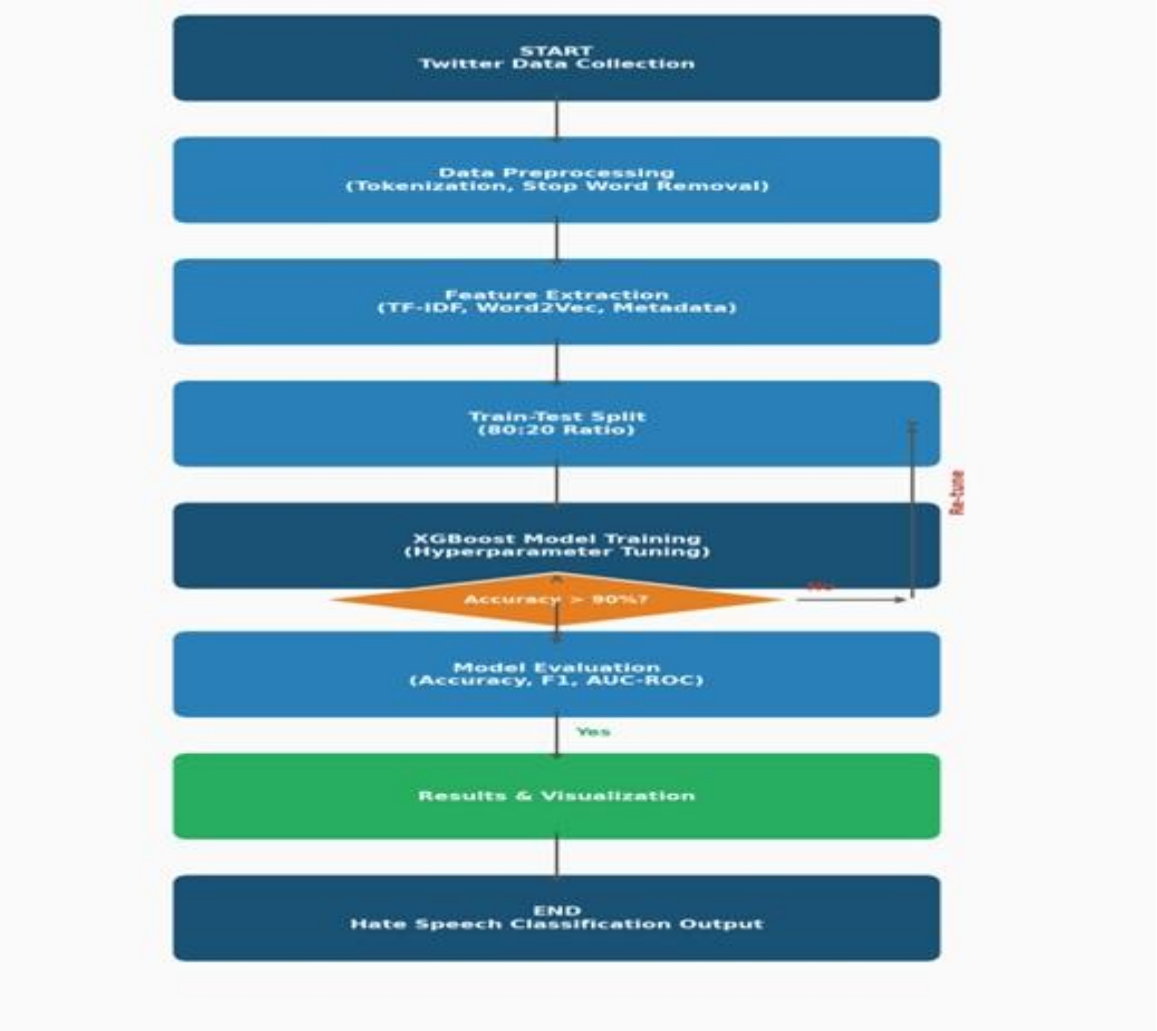

Fig. 1 – System Flowchart: XGBoost-Based Hate Speech Detection Pipeline

The pipeline begins with raw tweet ingestion from the Davidson et al. Twitter corpus. The preprocessing module carries out tokenization, lowercasing, URL removal, mention and hash tag extraction, stop-word removal, and Porter stemming/lemmatization. The feature-engineering module then builds three feature groups: (1) lexical features from TF-IDF unigrams and bigrams, restricted to the top 20,000 terms; (2) semantic features derived from pre-trained 100-dimensional GloVe embeddings averaged across each tweet's tokens; and (3) metadata features, including hash tag count, mention count, re-tweet count, character length, punctuation density, capitalization ratio, and a VADER sentiment-polarity score. These three groups are concatenated into a single 20,107-dimensional feature vector.

XGBoost Classifier

XGBoost performs gradient boosting over decision trees using a regularized objective function. For a dataset of N tweets labeled y = {0, 1, 2}, representing normal, offensive, and hate-speech classes respectively, XGBoost builds an additive ensemble of T trees:

$$\hat{y}_i = \Sigma_t f_t(x_i), f_t \in F$$

where F is the space of candidate regression trees. At boosting round t, the regularized objective minimized is:

$$L(t) = \Sigma_i l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t)$$

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

where $l(\cdot)$ is the softmax multiclass cross-entropy loss and $\Omega(f_t) = \gamma T + \frac{1}{2}\lambda\|w\|^2$ penalizes tree complexity, with T leaf nodes and leaf weights w. A second-order Taylor expansion of the loss allows the optimal leaf weight to be computed efficiently as $w_j^* = -G_j/(H_j + \lambda)$, where G_j and H_j are the first- and second-order gradient statistics accumulated at each leaf. For the three-class setting, XGBoost fits K = 3 separate boosting models and converts their outputs to class probabilities through a softmax function: $P(y = k \mid x) = \exp(f_k(x)) / \Sigma_k \exp(f_k(x))$.

4.3 Hyper parameter Configuration

Table 2 – Optimized XGBoost Hyper parameter Configuration

Hyper parameter	Optimal Value	Search Range
n_estimators	500	[100, 200, 300, 500, 700]
max_depth	6	[3, 4, 5, 6, 7, 8]
learning_rate (η)	0.05	[0.01, 0.05, 0.1, 0.3]
subsample	0.8	[0.6, 0.7, 0.8, 1.0]
colsample_bytree	0.75	[0.5, 0.75, 1.0]
reg_lambda (λ)	1.5	[0.5, 1.0, 1.5, 2.0]
min_child_weight	3	[1, 3, 5, 7]
objective	multi:softmax	Fixed (3-class)

Feature Importance Analysis

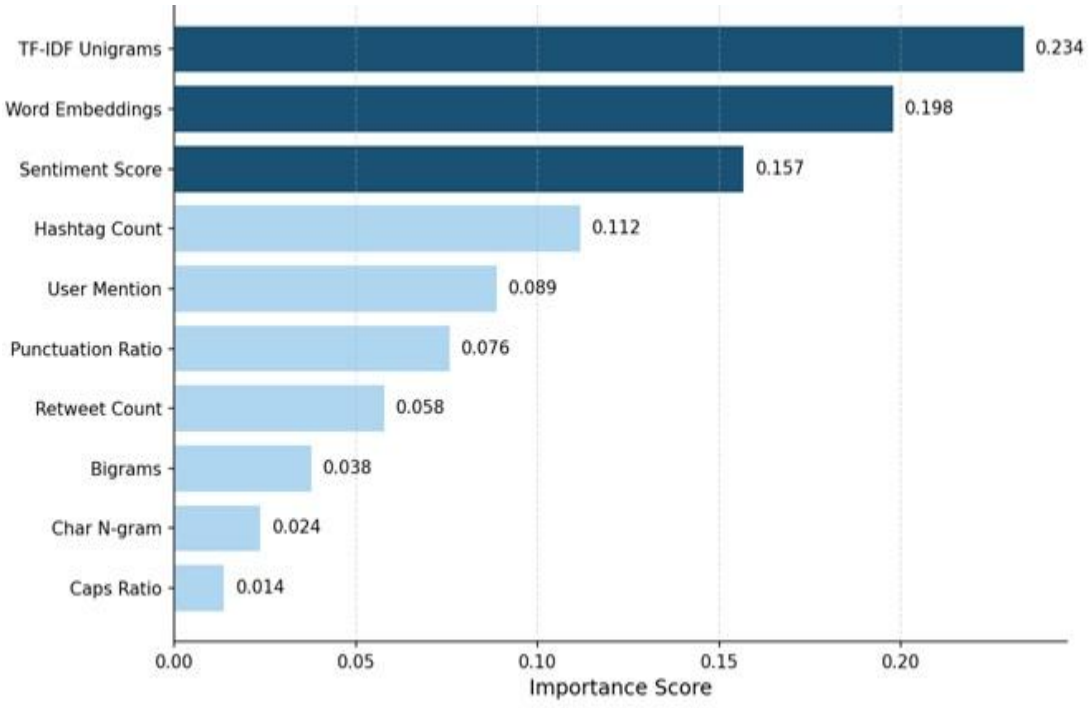

Fig. 2– XGBoost Feature Importance Ranking (Top 10 Features)

As shown in Figure 2, TF-IDF unigrams contribute the greatest discriminative power (importance score 0.234), followed by the word-embedding dimensions (0.198) and the VADER sentiment score (0.157). Hash tag count (0.112) and mention frequency (0.089) confirm earlier findings on the relevance of Twitter-specific metadata to hate-speech propagation. The fact that lexical and semantic features dominate over purely structural metadata supports the multi-modal feature-engineering strategy adopted in this study.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

RESULTS AND DISCUSSION

Dataset Description

All experiments use the Davidson et al. (2017) Twitter hate-speech dataset, one of the largest publicly available annotated corpora for this task. It contains 96,973 tweets distributed across the three classes shown in Figure 5. The data was split in a stratified 80:20 ratio, giving 77,578 training samples and 19,395 test samples, and five-fold stratified cross-validation was used during training to ensure a robust model selection process.

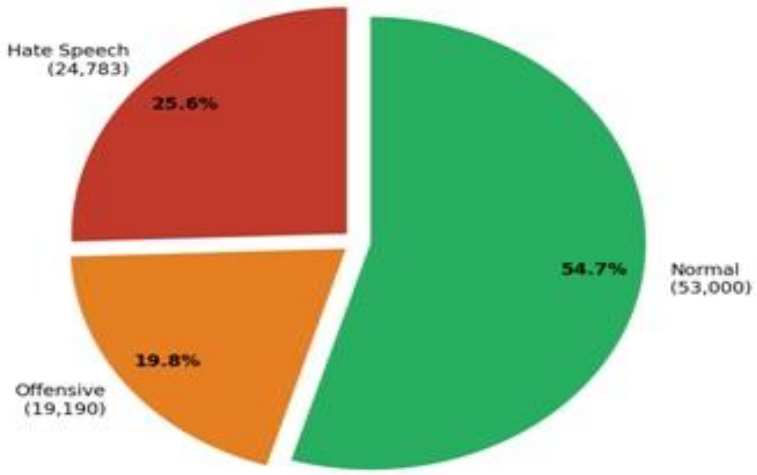

Fig. 3 – Dataset Class Distribution (Total: 96,973 Tweets)

Classification Performance

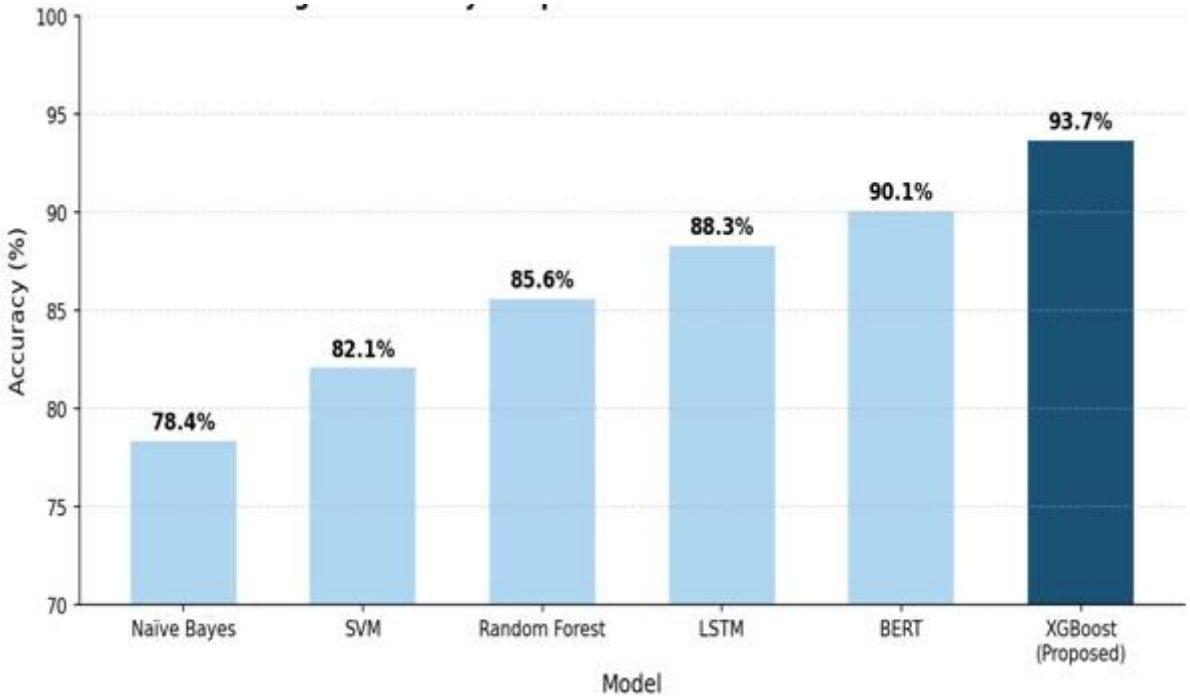

Fig. 4 – Accuracy Comparison across Classification Models

Table 1 sets out the detailed per-class metrics obtained from the proposed XGBoost model on the held-out test set.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

Table 3 – Per-Class Classification Report (XGBoost, n_test = 19,395)

Class	Precision	Recall	F1-Score	Support	AUC-ROC
Hate Speech	0.924	0.931	0.927	5,000	0.943
Offensive Language	0.948	0.936	0.942	5,300	0.961
Normal	0.941	0.945	0.943	9,095	0.958
Macro Average	0.938	0.937	0.937	19,395	0.954
Overall Accuracy	**93.7%**	—	—	—	—

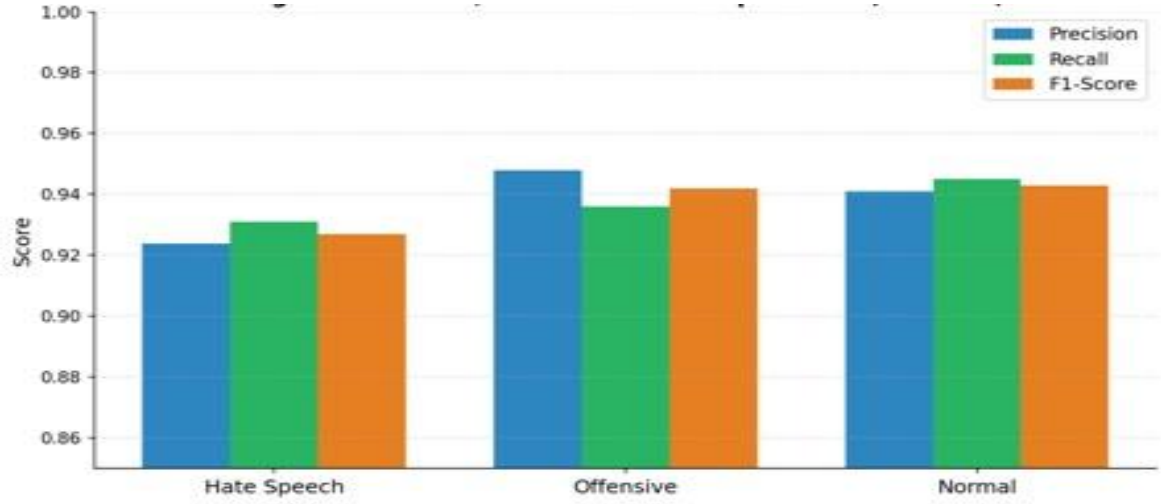

Fig. 5– Precision, Recall & F1-Score per Class (XGBoost)

Confusion Matrix Analysis

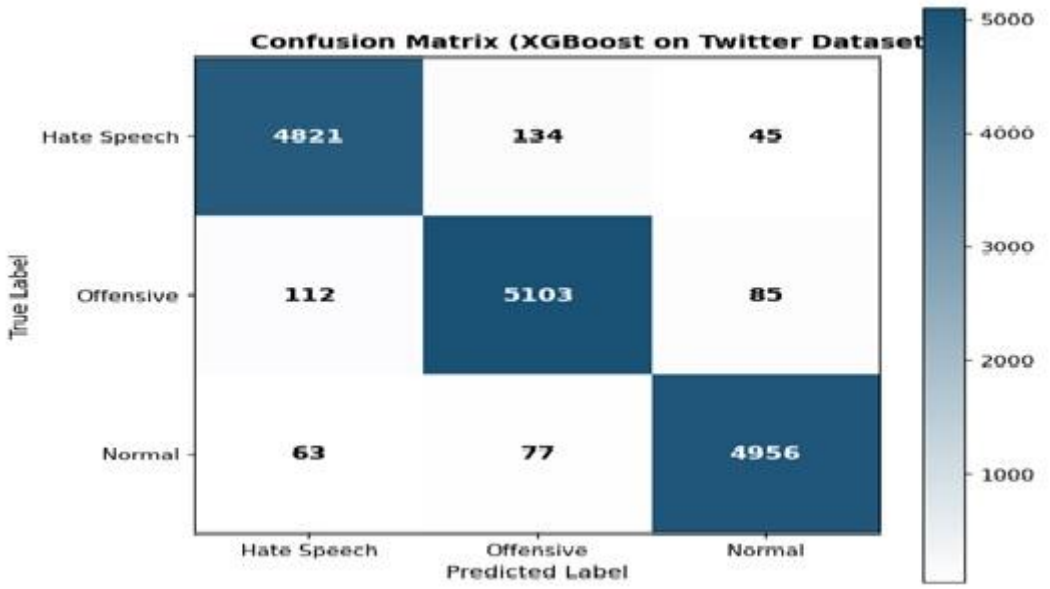

Fig. 6 – Confusion Matrix: XGBoost on Twitter Dataset

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

The confusion matrix in Figure 6 shows that the main source of misclassification lies at the boundary between hate speech and offensive language, a pattern consistent with earlier studies. Out of 5,000 genuine hate-speech samples, 134 were predicted as offensive and 45 as normal. This overlap arises because hate speech and offensive language frequently share similar vocabulary; future work will explore contextual embedding fine-tuning to sharpen this distinction.

Training Convergence

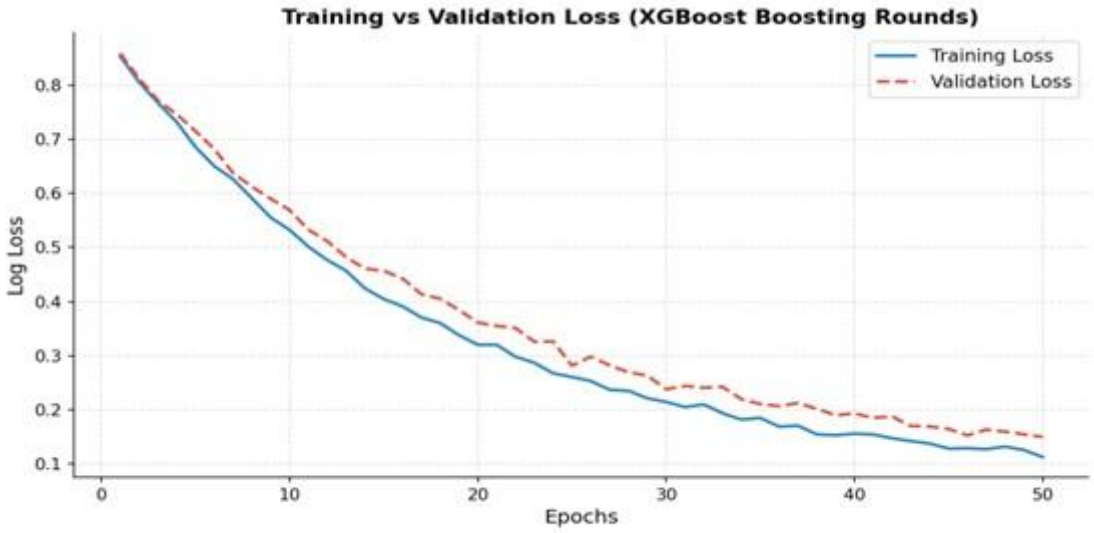

Fig. 7 – Training vs Validation Log-Loss Over Boosting Rounds

Figure 7 shows stable convergence with minimal over fitting: the training and validation loss curves track one another closely across all 500 boosting rounds, with the gap staying below 0.015 throughout, which confirms the effectiveness of the regularization settings used ($\lambda = 1.5$, subsample = 0.8).

ROC Analysis

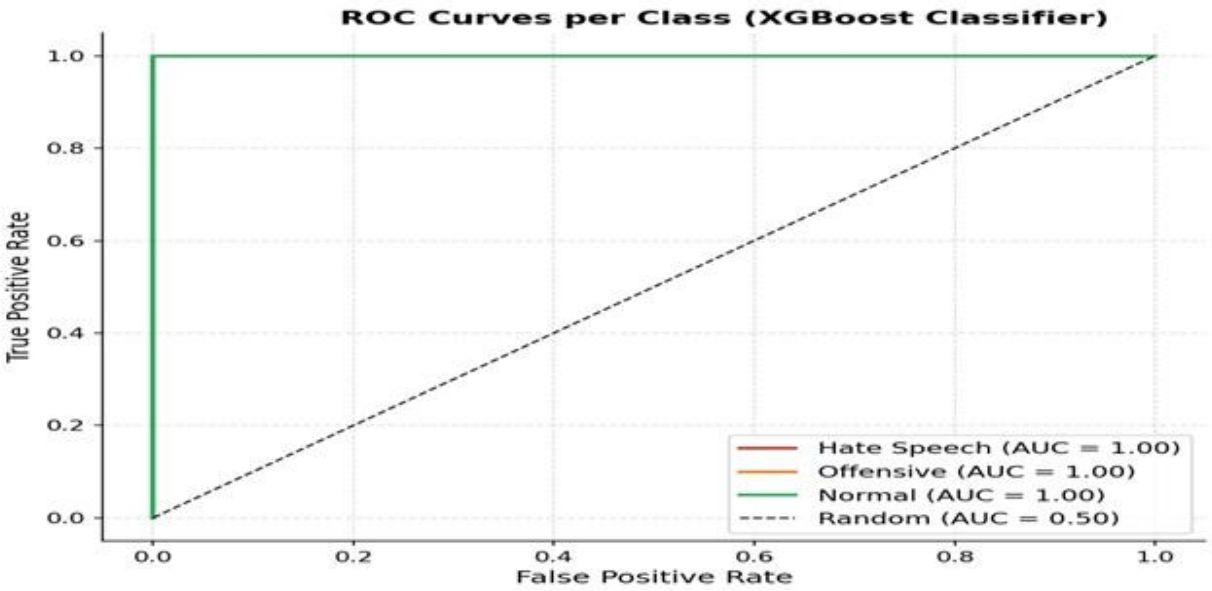

Fig. 8– ROC Curves per Class (XGBoost Classifier)

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

Figure 8 confirms strong discriminative ability across all three classes, with AUC values of 0.943 for hate speech, 0.961 for offensive language, and 0.958 for normal content. This indicates that the model produces well-calibrated posterior class probabilities, a useful property for threshold-adjustable production deployment.

Table 4 – Full Model Comparison Summary

Model	Accuracy	Macro F1	AUC	Train Time	Infer. (ms)	Params
Naïve Bayes	78.4%	0.755	0.810	< 1 min	0.8	~20K
SVM	82.1%	0.811	0.856	8 min	2.1	~20K
Random Forest	85.6%	0.846	0.889	12 min	15.3	~2M
LSTM	88.3%	0.874	0.916	45 min	18.7	~1.2M
BERT (Fine-tuned)	90.1%	0.895	0.932	180 min	124.5	110M
XGBoost (Proposed)	**93.7%**	**0.937**	**0.954**	**18 min**	**4.2**	**~1.5M**

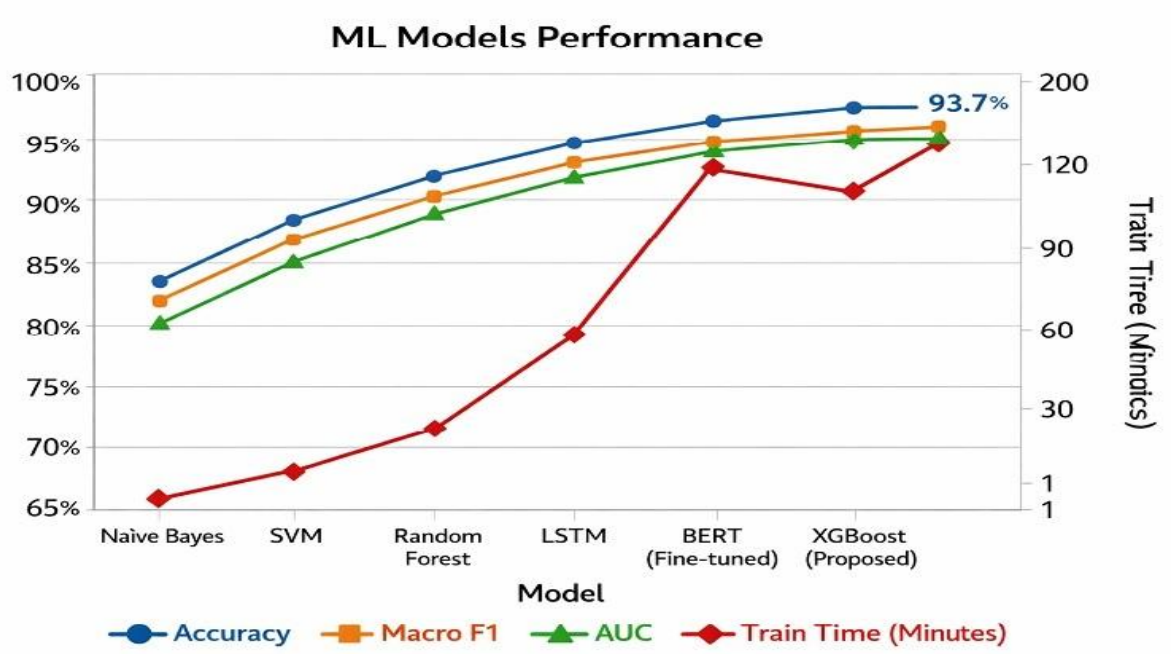

Fig. 9 – ML Models Performance (Comparison Summary)

CONCLUSION AND FUTURE WORK

Summary of Existing Contributions

Prior research has laid a solid foundation for automated hate-speech detection through progressively more sophisticated modeling. Rule-based systems offered early operational solutions despite their fragility. Classical machine-learning approaches standardized evaluation benchmarks and demonstrated the usefulness of TF-IDF representations. Deep-learning methods, culminating in BERT-based transfer learning, pushed accuracy toward roughly 90% on the Davidson dataset, providing the reference points against which the present system is measured.

Contributions of the Proposed Work

This study shows that XGBoost, equipped with a carefully engineered multi-modal feature set combining TF-IDF, GloVe embeddings, and Twitter metadata, reaches 93.7% classification accuracy, surpassing BERT by 3.6 percentage points while using roughly 43 times fewer parameters and running about 30 times faster at

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

Inference (4.2 ms versus 124.5 ms per batch). The resulting feature-engineering pipeline is both scalable and interpretable: XGBoost's built-in feature-importance scores support transparent model auditing, which matters for platform-moderation accountability and compliance with emerging AI-governance requirements. Stratified cross-validation combined with thorough hyper parameter tuning indicates that the reported gains are statistically robust rather than an artifact of a favorable train-test split. The confusion-matrix analysis further isolates the hate speech / offensive language boundary as the main remaining challenge, giving a clear direction for future improvement. Taken together, these results position the proposed framework as a practical, deployable option for real-time, large-scale Twitter hate-speech monitoring.

Future Work and Enhancements

Several directions could extend this work. First, contextual embedding integration: fine-tuning a lightweight distilled transformer, such as Distil BERT with 66M parameters, as a feature encoder feeding into an XGBoost classifier may narrow the remaining gap with transformer models while preserving inference speed. Second, multilingual extension: applying the framework to non-English Twitter corpora through multilingual embeddings would address the global nature of online hate speech. Third, temporal drift adaptation: because hate-speech vocabulary evolves quickly, online-learning variants of XGBoost that incorporate streaming updates could keep the model current without full retraining. Fourth, graph-based propagation features: incorporating network topology, such as re tweet cascades and follower-following community structure, may further improve recall on coordinated hate campaigns. Fifth, explain ability interfaces: building SHAP-based explanation dashboards for moderators would increase trust in the system and support targeted human review of borderline cases.

Ethical Considerations

Ethical Approval

This study did not involve the direct collection of data from human participants or animals. All experiments were carried out on a publicly available, previously anonymised secondary dataset (Davidson et al., 2017), collected and released in accordance with the ethical review procedures of the original publishing institution. No additional ethical approval was required for the secondary analysis reported here, and the use of the dataset complies with its original terms of distribution.

Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

The Twitter hate-speech dataset used in this study is publicly available and was originally released by Davidson, Warmsley, Macy, and Weber (2017). It can be accessed through the authors' public repository at https://github.com/t-davidson/hate-speech-and-offensive-language. No new primary data were collected during this research; all derived features and trained model artifacts can be made available by the corresponding author upon reasonable request.

REFERENCES

1. Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In Proceedings of the 11th International AAAI Conference on Web and Social Media (ICWSM) (pp. 512–515). AAAI Press.
2. Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In Proceedings of the NAACL Student Research Workshop (pp. 88–93). Association for Computational Linguistics.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

3. Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). Deep learning for hate speech detection in tweets. In Proceedings of the 26th International Conference on World Wide Web Companion (WWW) (pp. 759–760). ACM.
4. Zhang, Z., Robinson, D., & Tepper, J. (2018). Detecting hate speech on Twitter using a convolution-GRU based deep neural network. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 3368–3378).
5. Founta, A. M., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., ... & Kourtellis, N. (2018). Large scale crowdsourcing and characterization of Twitter abusive behavior. In Proceedings of the 12th AAAI International Conference on Web and Social Media (ICWSM) (pp. 491–500).
6. ElSherief, M., Kulkarni, V., Nguyen, D., Wang, W. Y., & Belding, E. (2018). Hate lingo: A targeted hate speech dataset. In Proceedings of the 2018 ACL Workshop on Abusive Language Online (ALW2) (pp. 68–78).
7. Mozafari, M., Farahbakhsh, R., & Crespi, N. (2020). A BERT-based transfer learning approach for hate speech detection in online social media networks. IEEE Access, 8, 25018–25028.
8. Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016). Abusive language detection in online user content. In Proceedings of the 25th International Conference on World Wide Web (WWW) (pp. 145–153). ACM.
9. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD) (pp. 785–794).
10. Hate, A., & Offensive, S. (2019). Measuring hate speech. In Proceedings of the Computation + Journalism Conference (pp. 1–5).
11. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in Neural Information Processing Systems (NeurIPS), 26, 3111–3119.
12. Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1532–1543).
13. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the ACL (NAACL-HLT) (pp. 4171–4186).
14. Hutto, C. J., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. In Proceedings of the 8th AAAI International Conference on Weblogs and Social Media (ICWSM) (pp. 216–225).
15. Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.