International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 05, May 2026
Received: 26-03-2026 Accepted: 05-05-2026 Published: 12-05-2026

Cyberbullying Detection Across Social Media Platforms: A Two-Tier Hate Speech Detection System Using SVM and BERT with Confidence-Based Routing.

V. Dadi Naga Siva Sai Pavan
department of IT
Seshadri Rao Gudlavalleru Engineering College
Gudlavalleru,India
dadisaipavan1514@gmail.com

N. Teja
department of IT
Seshadri Rao Gudlavalleru Engineering College
Gudlavalleru,India
tejanadella28@gmail.com

Y. Nithya Sri
department of IT
Seshadri Rao Gudlavalleru Engineering College
Gudlavalleru,India
nithyayeduruvada@gmail.com

T. P. Santhosh
department of IT
Seshadri Rao Gudlavalleru Engineering College
Gudlavalleru,India
santhoshkumartadiboyina@gmail.com

A.Koteswaramma
Assistant Professor
department of IT
Seshadri rao Gudlavalleru Engineering College
Gudlavalleru, India
Koteswari.1201@gmail.com

Abstract—Automatic identification of harmful content on social platforms requires systems that maintain classification precision without sacrificing processing speed. In real-world moderation pipelines, this balance is critical because large-scale user-generated content must be filtered in near real time while preserving contextual understanding. Deep learning architectures such as BERT deliver superior accuracy but demand extensive computing resources, whereas conventional algorithms like SVMs execute rapidly yet fail to capture semantic context adequately. We introduce a dual-stage classification framework employing confidence-based decision routing to optimize the trade-off between computational efficiency and detection performance. Our methodology begins with an SVM utilizing TF-IDF representations for initial assessment, then forwards uncertain predictions to a fine-tuned BERT module for comprehensive semantic evaluation. By restricting intensive processing to ambiguous cases only, this strategy minimizes unnecessary computational expense. Evaluation was conducted on the Davidson et al. benchmark dataset containing three categories: hate speech, offensive but non-hateful language, and neutral messages. Our experimental findings demonstrate that the proposed framework achieves 93.2% classification accuracy with 30 ms mean inference time, approximating BERT's performance while operating approximately five times faster than standalone transformer implementations. Analysis of prediction confidence levels, processing time distributions, and misclassification patterns validates the architectural design. Results indicate that hybrid systems with adaptive routing offer a practical solution for deploying sophisticated hate speech detectors under real-world operational constraints.

Keywords—Hate Speech Detection, Natural Language Processing, Hybrid Classification Systems, Confidence-Aware Routing, Support Vector Machines, BERT, Low-Latency Inference, Content Moderation

I. INTRODUCTION

Online platforms have revolutionized information sharing and interpersonal communication globally. However, this digital expansion has enabled rapid propagation of harmful content including hate speech and abusive language. Automated detection mechanisms are essential, given that such content contributes to real-world harm, reinforces societal biases, and causes psychological distress to affected groups.

Initial content filtering relied on lexicon-based rules and pattern matching. Such techniques proved inadequate for nuanced scenarios involving irony, indirect aggression, or contextual ambiguity. Statistical machine learning methods using feature-based classification offered better performance with modest computational requirements. Nevertheless, these approaches remain limited in semantic comprehension, particularly when differentiating targeted hatred from general profanity.

The advent of transformer architectures like BERT revolutionized text classification through bidirectional attention mechanisms. These models achieve superior accuracy by modeling contextual dependencies effectively. However, their computational intensity creates operational challenges—high memory consumption and slow inference hinder deployment at scale when processing millions of messages per hour.

Our central observation is that uniform application of complex models to all inputs is inefficient. A substantial portion of user-generated content exhibits clear characteristics enabling rapid classification. Ambiguous cases requiring deeper semantic analysis represent a

2250-3676 ijesatj@gmail.com https://www.ijesat.com/

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 05, May 2026
Received: 26-03-2026 Accepted: 05-05-2026 Published: 12-05-2026

minority. This variability motivates an adaptive approach: employing lightweight classification initially, then escalating uncertain instances to comprehensive analysis.

We implement a hierarchical system combining SVM-based initial screening with BERT-based contextual evaluation. The SVM tier handles straightforward cases rapidly using TF-IDF features. Low-confidence predictions trigger escalation to fine-tuned BERT for contextual disambiguation. This strategy achieves near-transformer performance while maintaining practical inference speeds.

Experimental validation on benchmark data confirms substantial latency reduction with minimal accuracy sacrifice—addressing a critical deployment barrier for sophisticated detection systems. Our primary contributions include:

- A confidence-driven framework that adaptively integrates lightweight models with transformer-based classifiers for effective hate speech identification.
- A dynamic routing mechanism that escalates only ambiguous cases to expensive models, eliminating wasted computation.
- Comprehensive evaluation under realistic constraints, analyzing accuracy, latency, threshold sensitivity, and error patterns.
- Demonstration that our system matches BERT accuracy while cutting inference time dramatically-making advanced detection viable for real-time moderation at scale.

II. LITERATURE REVIEW

The automatic identification of abusive language has received growing research interest due to the increasing prevalence of harmful interactions across online platforms. Early work in this area primarily adopted conventional supervised classification strategies based on margin optimization and probabilistic decision frameworks, combined with frequency-driven text representations that emphasize word distribution and relative term importance [1][10]. Although these techniques were computationally efficient and scalable to large datasets, they largely relied on surface-level textual patterns, which limited their ability to reliably separate genuine hate expressions from informal language, sarcasm, or culturally contextual slang.

The rise of transformer-based methods significantly impacted the design and performance of modern natural language processing systems [3][4][5][12]. These models learned latent representations and temporal dependencies with minimal manual feature design. While accuracy improved, fundamental difficulties persisted: detecting sarcasm, recognizing implicit bias, modeling long-distance relationships, and interpreting context-dependent meaning. Dataset imbalance further complicated matters, causing frequent errors on underrepresented hate speech examples.

Several studies have investigated ensemble and hybrid methodologies seeking efficiency-accuracy balance [6][11]. Typical strategies involve combining surface features with deep embeddings or implementing voting schemes for improved robustness. Most designs apply identical processing pipelines to all instances, representing a fundamentally wasteful approach that inadequately addresses latency concerns.

Recent investigations have started considering operational requirements including scalability and response time [7][8]. However, adaptive mechanisms that modulate computational investment based on instance complexity remain uncommon. Our methodology differs by centering confidence estimation as the primary routing criterion. Resource allocation follows prediction certainty directly, enabling efficient processing without accuracy degradation—a practical solution for production environments with stringent resource limitations.

III. EXISTING WORK

Content moderation systems must simultaneously achieve high precision and minimal latency. Social networks handle millions of posts hourly, necessitating instantaneous harmful content identification. Latency increases directly impact system throughput at this operational scale.

We formulate the task as three-way text classification: categorizing inputs as hate speech, offensive-but-not-hateful language, or neutral content. Production datasets demonstrate significant class skew, with hateful content representing a small fraction of total instances. Classification errors carry asymmetric consequences—missed hate speech enables continued harm distribution, while false alarms cause inappropriate censorship damaging platform credibility.

Conventional implementations employ uniform classification strategies. Simple classifiers offer speed advantages but limited semantic understanding, failing on subtle hatred. Deep contextual models provide superior comprehension at substantial computational expense, creating a performance-efficiency tension limiting practical deployment.

Our fundamental premise: computational requirements should match input complexity. Substantial portions of user content exhibit unambiguous characteristics—either clearly innocuous or explicitly abusive—enabling rapid classification. Complex semantic analysis becomes necessary only for ambiguous boundary cases. Input heterogeneity suggests that fixed processing strategies waste computational resources.

This observation motivates our central research objective: designing classification systems that dynamically adjust computational allocation based on input characteristics, applying intensive analysis selectively while preserving overall detection quality.

2250-3676 ijesatj@gmail.com https://www.ijesat.com/

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 05, May 2026
Received: 26-03-2026 Accepted: 05-05-2026 Published: 12-05-2026

IV. PROPOSED WORK

We introduce a confidence-driven two-tier framework that allocates computational resources adaptively according to prediction certainty. Clear-cut instances receive rapid processing; ambiguous cases undergo detailed semantic analysis.

A. Dataset Overview

Our evaluation utilizes the Davidson et al. hate speech corpus [1], which is commonly used in studies on social media content moderation. The dataset contains English-language tweets that have been manually annotated, with each instance labeled to reflect hate-related content, offensive expressions, or neutral language. It captures realistic class imbalance and linguistic ambiguity typically observed in real-world deployments, making it appropriate for evaluating practical system performance.

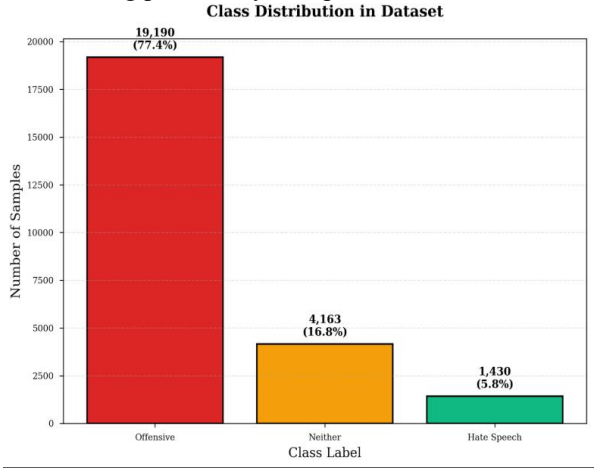

Fig. 1. Frequency of instances per category in the dataset

The dataset contains 24,783 annotated instances demonstrating severe class imbalance—hateful content forms a small minority relative to offensive and neutral categories. This distribution mirrors actual social media environments and tests system capability for accurate minority class identification without excessive false positive rates.

Data partitioning employs stratified splitting to maintain class proportions across training and testing subsets, ensuring evaluation accurately reflects deployment scenarios.

B. System Overview

Our framework processes text through sequential tiers. The majority of inputs receive fast classification; uncertain predictions undergo comprehensive analysis. This architecture maximizes computational efficiency while preserving classification quality.

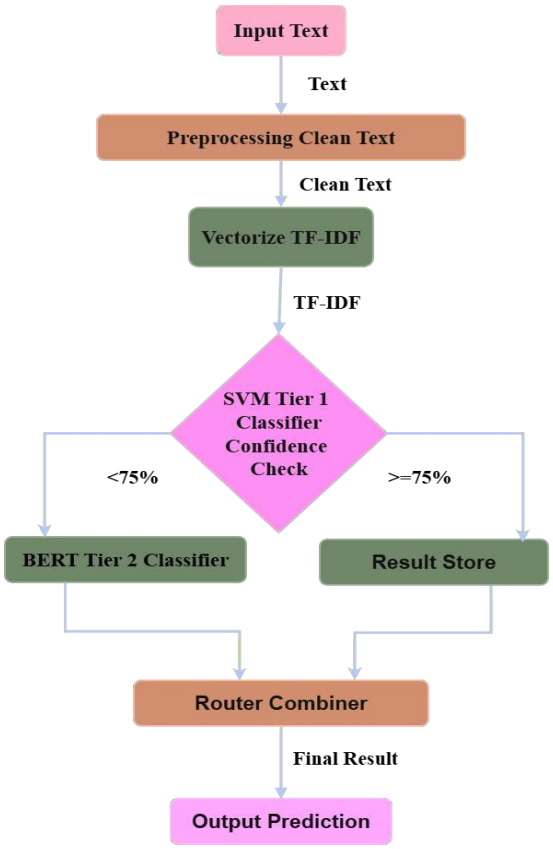

Fig. 2. Architecture of the proposed confidence-aware two-tier hate speech detection system.

The processing pipeline operates as follows: Raw social media text undergoes preprocessing to normalize content while retaining semantic information. Tier-1 (lightweight SVM) generates class predictions with associated confidence scores. High-confidence outputs immediately finalize as system predictions. Low-confidence cases escalate to Tier-2 (fine-tuned BERT) for contextual semantic analysis.

C. Text Processing and Feature Extraction

Social media text contains substantial noise. Our preprocessing workflow normalizes content through lowercase conversion, tokenization, selective stopword elimination, and lemmatization. Non-semantic elements including URLs, user mentions, and special characters are removed while preserving semantically relevant components. Negation terms receive special treatment as they fundamentally alter sentiment and meaning. Word-level tokenization supports SVM feature extraction; BERT employs internal subword tokenization. Lemmatization provides superior base form preservation compared to stemming, enhancing model generalization.

For Tier-1 processing, preprocessed text undergoes numerical transformation via TF-IDF vectorization—an established technique for sparse textual data. Term frequency quantifies term occurrence within documents;

2250-3676 ijesatj@gmail.com https://www.ijesat.com/

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 05, May 2026
Received: 26-03-2026 Accepted: 05-05-2026 Published: 12-05-2026

inverse document frequency weights term significance across the corpus:

$$TF(t,d) = \frac{\text{No. of times term } t \text{ appears in document } d}{\text{Total number of terms in document } d}$$

$$IDF(t) = \log\left(\frac{N}{n_t}\right)$$

where N denotes the total number of documents and represents the number of documents containing term t. The final TF-IDF weight is computed as

$$F - DF(,) = F(,) \times DF()$$

We include unigram through trigram features to capture multi-word expressions like compound insults that single tokens miss. Document frequency thresholds constrain dimensionality, reducing sparsity and improving efficiency.

D. Tier-1: SVM Lightweight Classification

Our first tier uses a linear SVM trained on TF-IDF features [9]. SVMs excel with high-dimensional sparse data, learning maximum-margin decision boundaries efficiently. They deliver strong performance at low computational cost—ideal for fast initial screening.

Beyond class labels, Tier-1 produces confidence scores reflecting prediction reliability. When confidence exceeds threshold τ, we accept the prediction immediately. These cases typically involve explicit lexical signals or clear semantic patterns. Handling them at Tier-1 maximizes throughput while reserving expensive Tier-2 analysis for genuinely ambiguous inputs.

E. Tier-2: BERT Contextual Classification

Low-confidence Tier-1 predictions dynamically escalate to a fine-tuned, domain-specific BERT model. BERT leverages bidirectional self-attention and massive pretraining to grasp deep contextual relationships. This makes it particularly effective at distinguishing hate speech from offensive language in implicit, sarcastic, or context-dependent cases. Selective escalation keeps BERT from dominating system latency. Tier-2 outputs become final predictions for ambiguous inputs, ensuring thorough analysis where it matters most.

F. Confidence-Aware Routing Strategy

Routing logic forms the system's core. Let $f_1(x)$ denote the Tier-1 classifier and c(x) its confidence score for input x. The routing decision:

$$\text{Routing}(x) = \begin{cases} f_1(x), & \text{if } c(x) \geq \tau \\ f_2(x), & \text{if } c(x) < \tau \end{cases}$$

This formulation enables explicit control over the trade-off between accuracy and inference latency. Lower values of favor efficiency, while higher values prioritize accuracy by increasing Tier-2 utilization.

V. RESULTS

We evaluate our system on the Davidson et al. dataset, analyzing both classification performance and system efficiency. Since we target real-world deployment, we measure not just accuracy but also latency, throughput, and routing behavior.

A. SVM Classification Performance

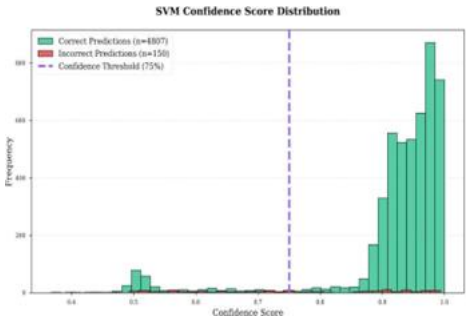

Fig. 3. Distribution of SVM Confidence Scores for Correct and Incorrect Predictions with the Selected Confidence Threshold

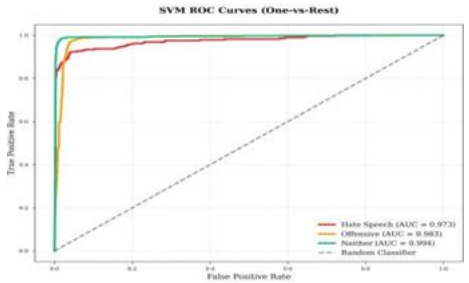

Fig. 4. One-vs-Rest ROC Curves for the Tier-1 SVM Classifier Across Hate Speech, Offensive, and Neutral Classes

Tier-1 serves as our primary decision layer, quickly resolving clear-cut cases and forwarding only uncertain inputs to Tier-2. Its classification and confidence behavior directly impact overall system effectiveness.

The one-vs-rest ROC curves illustrated in Fig. 4 highlight the strong class-separation capability of the SVM classifier. The curves corresponding to the offensive and neither categories rapidly achieve high true positive rates with minimal increases in false alarms, demonstrating that clearly expressed offensive content and neutral language are effectively distinguished using TF-IDF based

ISSN 2250-3676 ijesatj@gmail.com https://www.ijesat.com/

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 05, May 2026
Received: 26-03-2026 Accepted: 05-05-2026 Published: 12-05-2026

representations. In contrast, the hate speech class exhibits comparatively reduced separation, underscoring the challenge of detecting implicit, nuanced, or context-dependent hateful expressions when relying primarily on surface-level textual features. Despite this limitation, the ROC analysis confirms that the SVM model delivers consistent and dependable performance, making it well suited as an efficient first-stage filter in the proposed classification pipeline.

The confidence score distribution illustrated in Fig. 3 further validates the suitability of the SVM classifier for confidence-based routing. Most correct predictions are associated with high confidence values, while incorrect predictions are concentrated in the lower confidence region, indicating that the model's confidence estimates reliably reflect prediction uncertainty.

B. BERT Classification Performance

Threshold	Accuracy	Average Latency	BERT Usage
60%	92.1%	19 ms	12.3%
70%	92.9%	26 ms	22.1%
75%	93.2%	30 ms	26.6%
80%	93.6%	38 ms	34.5%
90%	94.8%	89 ms	72.8%

Table 1: Impact of confidence threshold on accuracy, latency, and BERT utilization

Table 1 reveals the accuracy-latency tradeoff as we vary the confidence threshold. Higher thresholds route more inputs to BERT, improving accuracy but increasing latency. At moderate thresholds (70-80%), we get strong accuracy with reasonable latency—BERT handles ambiguous cases while Tier-1 resolves straightforward ones. Very high thresholds (90%) yield diminishing returns: accuracy improves marginally while latency skyrockets. This validates our adaptive approach—BERT excels on difficult cases, but indiscriminate use destroys efficiency.

C. Two-Tier System Performance Analysis

Fig. 5 illustrates that at τ=75%, Tier-1 processes the majority of inputs while escalating only a small fraction to Tier-2. This selective routing optimizes resource utilization. Fig. 6 confirms the benefit: the two-tier system reduces latency 5-fold versus BERT-only deployment while sacrificing merely 3% accuracy. Compared to SVM-only processing, the system gains 6% accuracy for a 6-fold latency increase—justified by improved detection of subtle hate speech.

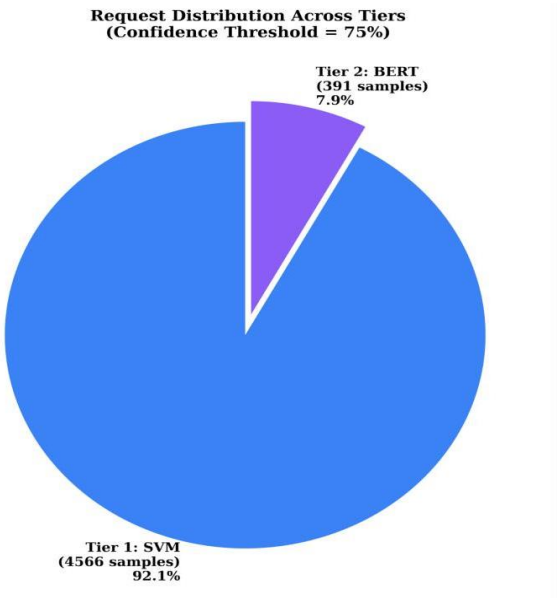

Fig. 5. Distribution of input samples across Tier-1 and Tier-2 classifiers at a confidence threshold of 75%.

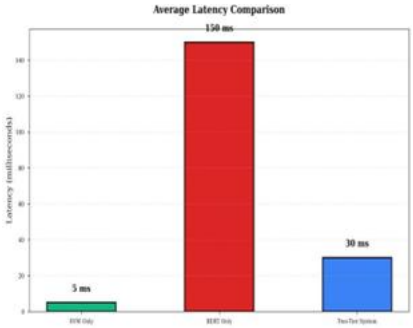

Fig. 6. Average inference latency comparison across SVM-only, BERT-only, and two-tier systems.

System	Accuracy	Average Latency	Throughput
SVM Only	87.2%	5 ms	200 req/s
BERT Only	96.1%	150 ms	6.7 req/s
Two-Tier	93.2%	30 ms	33 req/s

Table 2: System-Level Comparison

Table 2 compares our two-tier system against single-model baselines. SVM-only runs fast (5ms, 200 req/s) but achieves only 87.2% accuracy. BERT-only hits 96.1% accuracy but crawls at 150ms per sample (6.7 req/s)—unacceptable for production. Our two-tier approach achieves 93.2% accuracy at 30ms (33 req/s), splitting the difference intelligently.

2250-3676 ijesatj@gmail.com https://www.ijesat.com/

VI. CONCLUSION

This study proposed a confidence-aware two-tier hate speech detection system designed to reconcile the competing requirements of high classification accuracy and low inference latency in real-time content moderation. The framework combines a lightweight SVM-based classifier with a transformer-based BERT model using a confidence-driven routing strategy, ensuring that computationally expensive contextual analysis is applied only to uncertain inputs. Experimental results demonstrate that the proposed system delivers accuracy close to that of a BERT-only model while substantially reducing inference latency and improving throughput, making it suitable for large-scale deployment. Detailed analysis of confidence behavior and tier utilization shows that most inputs can be reliably classified using the lightweight model, while selective escalation effectively improves detection of ambiguous and minority-class hate speech. By jointly considering predictive performance and system-level efficiency, this work demonstrates that adaptive, confidence-aware hybrid architectures provide a practical and scalable solution for real-time hate speech detection, bridging the gap between high-performing research models and the operational constraints of production moderation systems.

REFERENCES

[1] Davidson T, Warmsley D, Macy M, Weber I. Automated systems for identifying offensive vs hateful language. Proc ICWSM. 2017:512-515.

[2] Waseem J, Hovy D. Identifying hate speech on social platforms: Predictive characteristics. NAACL Student Res Workshop. 2016:88-93.

[3] Zhang Z, Robinson D, Tepper J. Detecting harmful social content via neural networks. Eur Semantic Web Conf. 2018:745-760.

[4] Liu Y, et al. Enhanced optimization for bidirectional encoder representations. arXiv:1907.11692. 2019.

[5] Devlin J, et al. Bidirectional transformer pretraining for language comprehension. Proc NAACL-HLT. 2019:4171-4186.

[6] Schmidt A, Wiegand M. Computational approaches for abusive language: A review. Int Workshop NLP Social Media. 2017:1-10.

[7] Fortuna P, Nunes S. Computational methods for detecting harmful textual content: A survey. ACM Comput Surv. 2018;51(4):1-30.

[8] Vidgen N, Derczynski T. Dataset quality for training abusive language classifiers. PLoS ONE. 2020;15(12):1-26.

[9] Cortes C, Vapnik V. Support-vector networks for pattern recognition. MLJ. 1995;20(3):273-297.

[10] Salton G, Buckley C. Term-weighting strategies in automatic text retrieval. IP&M. 1988;24(5):513-523.

[11] Qian K, et al. User-level representation learning for harmful content. Proc NAACL-HLT. 2018:118-123.

[12] Ruder S. Transfer learning in computational linguistics [Dissertation]. National Univ Ireland; 2019.

[13] Ribeiro M, Singh S, Guestrin C. Interpretability of machine learning classification. Proc ACM SIGKDD. 2016:1135-1144.

[14] Guo A, et al. User requirements for interpretable AI. CHI Conf Human Factors. 2019:1-12.

[15] Hochreiter S, Schmidhuber J. Long Short-Term Memory architectures. Neural Comput. 1997;9(8):1735-1780.

[16] Kim Y. Sentence classification using convolutional neural networks. EMNLP. 2014:1746-1751.