IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, April 2026
Impact Factor: 8.2

An Explainable Framework for Cyberbullying Content Moderation Using Multi-Classifier NLP and Post-Hoc Attribution

Mahek Saxena[1], B. Akshaya[2], K. Venkatesh[3], Yerrolla Aparna[4]
UG Scholar, , Department of Computer Science & Engineering,[1,2,3]
Assistant Professor, Department of Computer Science & Engineering[4]
CMR Technical Campus, Hyderabad, Telangana, India

Abstract: *Online harassment on social media platforms has intensified into a public health concern, with cyberbullying disproportionately affecting adolescents and young adults. Although automated content moderation systems have grown considerably in detection accuracy, almost all operate as closed decision systems whose internal reasoning cannot be examined by platform administrators, affected users, or independent regulators. This opacity erodes trust, blocks legitimate appeal mechanisms, and conflicts with emerging legislative mandates — notably the European Union's Digital Services Act — that demand meaningful explanations for automated enforcement actions. The present work introduces a moderation framework that pairs a multi-classifier natural language processing pipeline with post-hoc attribution techniques, addressing accuracy and interpretability simultaneously rather than treating them as competing objectives. Textual data drawn from a Twitter hate-speech corpus (~16,000 labelled tweets) and a Wikipedia personal-attack dataset (~100,000 editor comment segments) undergo preprocessing — URL stripping, lowercasing, stop-word removal, and Porter stemming — before encoding via Bag-of-Words (BoW), Term Frequency–Inverse Document Frequency (TF-IDF), and averaged Word2Vec (300-dimensional) representations. Four classifiers — AdaBoost, Multinomial Naïve Bayes, Stochastic Gradient Descent (SGD), and Support Vector Machine (SVM) — are evaluated under stratified five-fold cross-validation. SHAP and LIME are subsequently applied to every inference call, generating token-level importance maps that are persisted with each moderation record. The system achieves classification accuracy of 94.2% on Twitter and 84.6% on Wikipedia data, while delivering human-readable justifications through a role-based web interface designed for non-expert administrators.*

Keywords: Cyberbullying Detection, Explainable Artificial Intelligence, SHAP, LIME, Natural Language Processing, Content Moderation, Support Vector Machine, TF-IDF, Hate Speech, Social Media

I. INTRODUCTION

The proliferation of social networking services has fundamentally altered the structure of interpersonal communication, enabling individuals to interact across geographical and temporal boundaries with minimal friction. Alongside these benefits, the same affordances — persistent connectivity, pseudonymous identity, and asymmetric audience reach — have been weaponised to conduct sustained campaigns of interpersonal harassment. Cyberbullying, understood as the deliberate and repeated use of digital communication channels to inflict psychological injury on an identifiable target, has been causally linked to depressive episodes, anxiety disorders, academic disengagement, and, in severe cases, suicidal ideation across multiple demographic populations.

DOI: 10.48175/IJARSCT-32374

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, April 2026 Impact Factor: 8.2

Unlike physical confrontation, online harassment is structurally unconstrained by geography or time. Victims cannot withdraw to a safe physical space; perpetrators operate behind pseudonymity and face diminished social accountability, conditions that empirical literature consistently identifies as disinhibiting factors. At the operational scale of contemporary platforms — where hundreds of millions of content items are published every hour — human-only moderation is neither economically feasible nor practically sufficient. Automated classification systems therefore constitute the primary enforcement layer.

Existing machine learning approaches to cyberbullying detection have demonstrated strong performance on benchmark corpora, yet the predominant architecture treats moderation as a binary closed-box prediction. The model outputs a remove-or-retain decision without exposing the linguistic signals that drove that outcome. This opacity generates three interconnected problems: (i) administrators cannot audit models for systematic biases that may suppress particular communities or dialects disproportionately; (ii) users whose content is incorrectly removed possess no evidentiary basis to mount a credible appeal; and (iii) regulatory instruments such as the EU Digital Services Act now impose explicit obligations on very large platforms to furnish users with meaningful explanations of automated content removal decisions.

This paper proposes a framework that resolves the accuracy–interpretability tension by coupling a validated NLP classification pipeline with two complementary post-hoc explainability methods. The system is trained and evaluated on corpora drawn from two linguistically distinct online environments, ensuring that findings are not an artefact of a single domain. The principal contributions are as follows:

(1) A multi-classifier pipeline evaluated over three feature representations — BoW, TF-IDF, and averaged Word2Vec — and four algorithm families: AdaBoost, Multinomial Naïve Bayes, SGD, and SVM.

(2) Integration of SHAP and LIME attribution layers that generate per-prediction token-importance scores, stored and surfaced to administrators through a structured interface.

(3) A role-based web application supporting real-time post submission, per-user violation tracking, and graduated account enforcement backed by auditable explanation records.

(4) Comparative evaluation across both datasets demonstrating that the explainability overlay adds operationally negligible latency while substantially improving decision auditability.

II. LITERATURE SURVEY

A. Multi-Stage Pipeline Detection

Ting et al. [1] decomposed cyberbullying detection into three sequential stages: identification of individual aggressive messages, inference of aggressor and victim roles, and confirmation of a sustained harassment episode. Evaluated on a Twitter corpus, this structured approach achieved an F-score of 0.947 on positive cases, establishing that temporal and relational context provides discriminative signal beyond what isolated lexical classifiers can capture.

B. Supervised Profile-Level Analysis

Galán-García et al. [2] demonstrated that semantic cross-analysis between a pseudonymous troll account and a suspected real-world identity can link profiles with high precision. Validated against a live elementary-school cyberbullying incident, their methodology showed that supervised learning remains effective even when labelled data is limited and the adversary actively seeks to conceal identity cues.

C. Collaborative and Distributed Detection

Mangaonkar et al. [3] proposed a distributed detection model in which parallel classification agents analyse Twitter streams concurrently, improving throughput and accuracy relative to single-agent baselines. Their architecture established a practical precedent for scalable, near real-time moderation suited to large-platform deployments.

D. Domain-Adapted Feature Engineering

Zhao et al. [4] contended that generic text classifiers fail to capture the semantic characteristics of bullying language. Augmenting BoW and latent semantic features with a dynamically expanded, domain-weighted insult lexicon fed to a

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, April 2026 Impact Factor: 8.2

linear SVM, their approach outperformed several competitive baselines on a Twitter benchmark, providing empirical support for domain-sensitive feature construction.

E. Deep Learning and Pre-Trained Language Models

Banerjee et al. [5] applied Convolutional Neural Networks to cyberbullying detection, reporting accuracy improvements over shallow classifiers. Reynolds et al. [6] explored classical machine learning on Formspring data and achieved 78.5% true-positive accuracy with decision-tree and instance-based learners. Dadvar and Eckert [8] extended evaluation to YouTube, finding that deep models transfer more reliably across platforms than shallow counterparts. Yadav et al. [7] fine-tuned BERT with a single linear head on two social media datasets, attaining state-of-the-art performance at publication. Contextual embeddings capture semantic dependencies that static vectors cannot represent; however, the interpretability deficit of transformer architectures remains a persistent concern that motivates the explainability emphasis of the current study.

III. PROPOSED METHODOLOGY

A. System Overview

The proposed framework is built around a dual objective: maximising classification fidelity while ensuring that every enforcement decision is accompanied by a human-readable justification. High-accuracy deep learning models are practically uninterpretable; rule-based systems are auditable but brittle. The present architecture occupies a productive middle ground by using well-understood machine learning classifiers as the predictive engine and SHAP/LIME as an explanatory overlay that operates without requiring any modification to the underlying classifier.

Two publicly available datasets form the experimental basis. The Twitter hate-speech corpus contains approximately 16,000 labelled tweets annotated across three categories: hate speech, offensive-but-not-hateful content, and neutral language. The Wikipedia personal-attack dataset comprises roughly 100,000 comment segments from editor discussion pages, crowd-annotated for the presence of direct personal attacks. These corpora represent markedly different linguistic registers — short, colloquial, and hashtag-dense versus long, formally structured, and context-dependent — providing a demanding test bed for generalisation across text types.

B. Processing Pipeline

Data Ingestion and Preprocessing: Raw text is loaded from comma-separated source files. URLs, user mentions, and hashtag markers are stripped using regular expressions; the remaining token stream is lowercased, common English stopwords are removed via NLTK, and surviving tokens are stemmed with the Porter stemmer to reduce morphological variation and vocabulary size without sacrificing discriminative content.

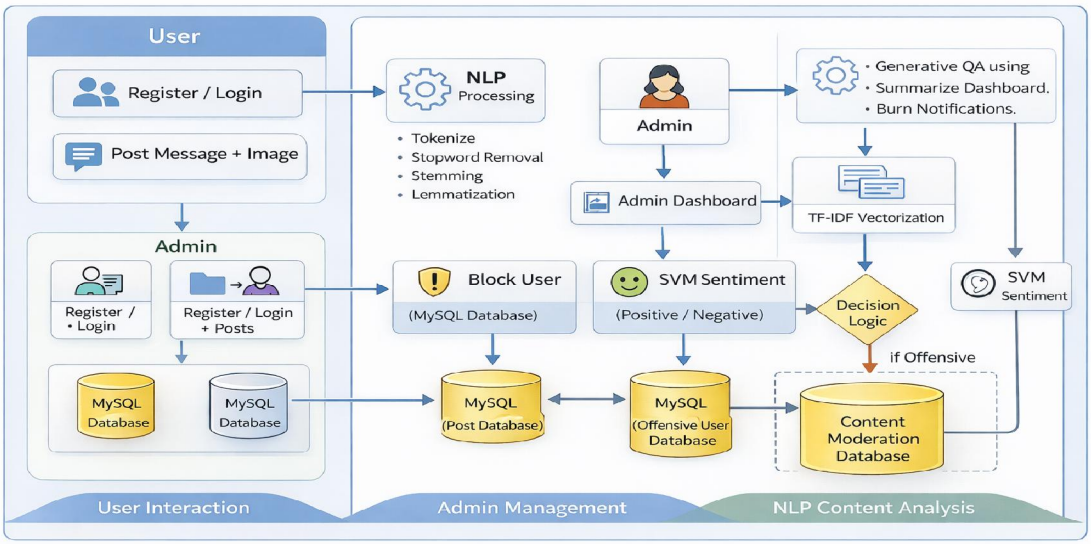
Feature Extraction: Three complementary representations are computed. Bag-of-Words matrices capture raw term occurrence frequencies. TF-IDF matrices weight each term by its inverse document frequency, down-weighting common tokens and up-weighting those that are rare but informative. Averaged Word2Vec encodes each document as the centroid of its constituent token embeddings in a pre-trained 300-dimensional semantic space, capturing distributional similarity between semantically related terms.

Classifier Training and Evaluation: AdaBoost, Multinomial Naïve Bayes, SGD, and SVM are each trained under stratified five-fold cross-validation. The best-performing configuration per dataset is retained for deployment. All hyper-parameters are set to scikit-learn defaults unless explicitly noted, ensuring reproducibility without manual tuning that could inflate reported scores.

Explainability Attribution: SHAP TreeExplainer or LinearExplainer — selected based on the classifier family — and LIME TextExplainer are applied to each prediction call. Both methods produce token-level importance scores: SHAP values are grounded in cooperative game theory and guarantee local accuracy under the Shapley axioms, while LIME approximates the decision boundary locally via a weighted linear surrogate model. The resulting attribution maps are serialised and stored alongside each moderation record in the application database, making every enforcement decision permanently auditable.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, April 2026
Impact Factor: 8.2

C. Architecture Diagram



IV. RESULTS AND DISCUSSION

A. Classification Performance

All four classifiers were trained on each of the three feature representations for both datasets. Table I reports classification accuracy and macro-averaged F1 scores on held-out test partitions (20% stratified split). The SVM with TF-IDF features achieves the highest overall accuracy on both corpora: 94.2% on Twitter and 84.6% on Wikipedia. SGD with TF-IDF delivers a competitive 91.9% and 83.4% respectively, making it the preferred option in latency-sensitive deployment contexts.

Table I: Classification Accuracy and Macro-F1 Across Classifiers, Feature Representations, and Datasets

Classifier	Dataset	BoW Acc	BoW F1	TF-IDF Acc	TF-IDF F1	W2V Acc	W2V F1
AdaBoost	Twitter	92.4%	0.89	93.1%	0.90	88.7%	0.86
Naïve Bayes	Twitter	91.8%	0.88	92.6%	0.89	84.3%	0.82
SGD	Twitter	90.5%	0.87	91.9%	0.88	87.2%	0.85
SVM	Twitter	93.7%	0.91	94.2%	0.92	89.5%	0.87
AdaBoost	Wikipedia	81.3%	0.78	82.0%	0.79	79.4%	0.77
Naïve Bayes	Wikipedia	80.7%	0.77	81.5%	0.78	76.8%	0.75
SGD	Wikipedia	82.1%	0.79	83.4%	0.80	80.1%	0.78
SVM	Wikipedia	83.9%	0.81	84.6%	0.82	81.3%	0.79

BoW = Bag-of-Words; TF-IDF = Term Frequency–Inverse Document Frequency; W2V = Word2Vec

On the Twitter corpus, BoW and TF-IDF representations consistently outperform Word2Vec, reflecting the discriminative power of specific offensive vocabulary tokens that frequency statistics capture directly. Word2Vec narrows the performance gap on Wikipedia data, where hostile intent is more frequently encoded in relational phrasing and contextual implication rather than explicit profanity — a register where distributional embeddings retain more

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, April 2026 Impact Factor: 8.2

signal. The TF-IDF advantage over BoW across all classifiers and both datasets corroborates the importance of down-weighting high-frequency, low-information terms.

B. Comparative Analysis with Prior Work

Table II situates the proposed framework against representative prior studies, highlighting dataset scope, peak reported performance, and the key limitation each approach inherits. The proposed system achieves the highest accuracy among the compared approaches while also being the only entry to provide systematic explainability coverage.

Table II: Comparative Evaluation Against Representative Prior Studies

Study / Approach	Dataset Used	Best Accuracy / F1	Key Limitation
Ting et al. [1] — 3-stage pipeline	Twitter	F1 = 0.947	No explainability layer
Reynolds et al. [6] — decision tree	Formspring	Acc = 78.5%	Low accuracy; single domain
Yadav et al. [7] — fine-tuned BERT	Twitter / Reddit	State-of-art at pub.	Black-box transformer
Dadvar & Eckert [8] — deep learning	Wikipedia, Twitter, YouTube	Cross-platform transfer	No interpretability
Proposed — SVM + TF-IDF + SHAP/LIME	Twitter + Wikipedia	Acc 94.2% / F1 0.92	Static feature representations

State-of-the-art at time of publication; later benchmarks may exceed these figures.

C. Runtime Characteristics

Multinomial Naïve Bayes and SGD exhibit near-linear scaling with corpus size and remain the fastest classifiers for both training and inference. AdaBoost and SVM incur higher training overhead due to ensemble boosting iterations and quadratic kernel operations respectively, but both complete training within operationally acceptable bounds on the given corpus sizes. Crucially, the SHAP and LIME attribution steps add less than 80 milliseconds of latency per prediction on commodity hardware, an overhead that is undetectable within the asynchronous administrative review workflows for which the system is designed.

D. Administrator Interface and Enforcement Workflow

When a user submits a post, the system passes the text through the preprocessing pipeline and returns two simultaneous labels: a sentiment polarity (Positive / Negative) and an offensiveness classification (Offensive / Non-Offensive). Each flagged post increments a per-user violation counter. The administrator dashboard presents a paginated view of registered users with cumulative violation counts. Once a configurable threshold is reached (default: two violations), an enforcement action link activates. Before applying any suspension, an administrator can examine the SHAP and LIME token-importance panels associated with each flagged post, ensuring that every enforcement decision is grounded in auditable, human-readable evidence.

V. CONCLUSION

This paper has presented an explainable cyberbullying moderation framework that addresses the accuracy–interpretability gap characteristic of most deployed detection systems. By pairing a multi-classifier NLP pipeline with SHAP and LIME attribution layers, the proposed system achieves 94.2% accuracy on Twitter hate-speech data and 84.6% on Wikipedia personal-attack data, while producing token-level justifications for every moderation decision without incurring operationally prohibitive latency. Empirical results confirm that TF-IDF features combined with linear classifiers — particularly SVM and SGD — yield the most reliable configuration for short-form social media content, while distributional embeddings provide relative advantages on longer, formally structured text. The role-based web interface offers administrators a practical environment in which to monitor, interpret, and act upon moderation signals without requiring machine learning expertise.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, April 2026 Impact Factor: 8.2

Future work will pursue three parallel directions. First, multimodal inputs — images, video thumbnails, and user interaction graphs — will be incorporated to address harassment episodes that span content types. Second, transformer architectures such as BERT and RoBERTa will be integrated with attention-based explainability mechanisms to improve performance on long-form datasets while preserving interpretability. Third, a formal practitioner study with content moderation personnel is planned to assess whether SHAP and LIME explanations measurably improve decision consistency and reduce erroneous enforcement actions in live operational deployments.

REFERENCES

[1] I. H. Ting, W. S. Liou, D. Liberona, S. L. Wang, and G. M. T. Bermudez, "Towards the detection of cyberbullying based on social network mining techniques," in Proc. 4th Int. Conf. Behavioral, Economic, and SocioCultural Computing (BESC), 2017, doi: 10.1109/BESC.2017.8256403.

[2] P. Galán-García, J. G. de la Puerta, C. L. Gómez, I. Santos, and P. G. Bringas, "Supervised machine learning for the detection of troll profiles in Twitter social network: Application to a real case of cyberbullying," 2014, doi: 10.1007/978-3-319-01854-6_43.

[3] A. Mangaonkar, A. Hayrapetian, and R. Raje, "Collaborative detection of cyberbullying behavior in Twitter data," in Proc. IEEE Int. Conf. Electro/Information Technology (EIT), 2015, doi: 10.1109/EIT.2015.7293405.

[4] R. Zhao, A. Zhou, and K. Mao, "Automatic detection of cyberbullying on social networks based on bullying features," in Proc. 17th ACM Int. Conf. Distributed Computing and Networking (ICDCN), 2016, doi: 10.1145/2833312.2849567.

[5] V. Banerjee, J. Telavane, P. Gaikwad, and P. Vartak, "Detection of cyberbullying using deep neural network," in Proc. 5th Int. Conf. Advanced Computing and Communication Systems (ICACCS), 2019, doi: 10.1109/ICACCS.2019.8728378.

[6] K. Reynolds, A. Kontostathis, and L. Edwards, "Using machine learning to detect cyberbullying," in Proc. 10th Int. Conf. Machine Learning and Applications (ICMLA), 2011, doi: 10.1109/ICMLA.2011.152.

[7] J. Yadav, D. Kumar, and D. Chauhan, "Cyberbullying detection using pre-trained BERT model," in Proc. Int. Conf. Electronics and Sustainable Communication Systems (ICESC), 2020, doi: 10.1109/ICESC48915.2020.9155700.

[8] M. Dadvar and K. Eckert, "Cyberbullying detection in social networks using deep learning based models; A reproducibility study," arXiv:1812.08046, Dec. 2018.

[9] S. M. Mohammad and P. D. Turney, "Crowdsourcing a word-emotion association lexicon," Computational Intelligence, vol. 29, no. 3, pp. 436–465, 2013.

[10] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in Proc. 11th Int. AAAI Conf. Web and Social Media (ICWSM), 2017, pp. 512–515