



IJRASET

International Journal For Research in
Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IV **Month of publication:** April 2026

DOI: <https://doi.org/10.22214/ijraset.2026.80560>

www.ijraset.com

Call:  08813907089

E-mail ID: ijraset@gmail.com

A Hybrid Ensemble Framework for Cyber bullying Detection using Multi-Model Consensus and Confidence Weighting

Atharva Kadam, Utkarsh Chavan, Ganesh Nagare, Kaushik Moger

Department of Electronics & Telecommunication Engineering, Atharva College of Engineering, Mumbai, India

Abstract: Cyberbullying detection in social media has become a critical research challenge due to the rapid growth of online communication platforms. Traditional approaches rely either on rule-based systems or deep learning models, each with inherent limitations such as poor generalization or lack of interpretability. This paper proposes a hybrid ensemble framework that integrates multiple decision engines, including rule-based logic and transformer-based models, combined with a confidence-aware aggregation mechanism. A novel Multi-Engine Cyberbullying Framework (MECF), along with a Multi-Class Weighted Grading (MCWG) strategy, is introduced to improve detection robustness. The system evaluates predictions from multiple models and aggregates them using confidence-based weighted voting to produce the final classification. Experimental results on a balanced dataset demonstrate that the proposed approach achieves an F1-score of 0.91 and an AUC score of 0.962, outperforming individual models. The results highlight the effectiveness of combining heterogeneous models with confidence-based consensus for robust cyberbullying detection.

Keywords: Cyberbullying Detection, Ensemble Learning, DeBERTa, Hate Speech Detection, MCWG, NLP, Transformer Models.

I. INTRODUCTION

The rapid growth of social media and digital communication platforms has led to a significant rise in cyberbullying, which poses serious psychological and emotional risks to users, especially adolescents. Cyberbullying involves the use of online platforms to harass, threaten, or humiliate individuals, often resulting in anxiety, depression, and reduced self-esteem. Traditional detection methods based on manual moderation and keyword filtering are limited in their ability to identify contextual, sarcastic, and implicit abusive content. The proposed system aims to accurately detect harmful content in real time, provide preventive warnings, and support automated intervention mechanisms to promote safer and more responsible online interactions.[1] With the increasing use of social media platforms, cyberbullying has emerged as a significant challenge affecting users globally. Automated detection systems are crucial for monitoring and mitigating harmful content. Early approaches relied on rule-based systems and traditional machine learning algorithms, which often failed to capture contextual and semantic nuances in language.[3]

Recent advancements in deep learning, particularly transformer-based architectures, have significantly improved text classification tasks. However, single model approaches often struggle with generalization and may produce biased predictions, especially in noisy social media environments.[4]

To address these challenges, this paper proposes a hybrid ensemble framework that combines multiple decision engines with a confidence-based aggregation mechanism. The system integrates rule-based models, transformer-based classifiers, and a novel decision fusion strategy to improve performance and robustness.[2]

The main contributions of this work are:

- A hybrid multi-engine framework combining rule-based and transformer models
- A novel Multi-Class Weighted Grading (MCWG) decision mechanism.
- A confidence-aware detection module for robust prediction aggregation.
- Comprehensive evaluation and comparison across multiple models.

II. LITERATURE REVIEW

Cyberbullying detection has been widely studied using various approaches. Traditional machine learning methods such as Support Vector Machines (SVM) and Logistic Regression were initially used for text classification tasks.

These methods relied heavily on feature engineering and were limited in capturing contextual meaning.

Deep learning models, including Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, improved performance by learning representations directly from data. However, these models still faced challenges in handling long-range dependencies.[3]

Transformer-based models such as BERT, RoBERTa, and DeBERTa have demonstrated state-of-the-art performance in natural language processing tasks. These models leverage attention mechanisms to capture contextual relationships in text effectively. Despite their success, they may still produce inconsistent results when used independently.[14]

Recent research has explored ensemble methods to combine multiple models for improved performance. However, most existing ensemble approaches lack a structured confidence-based aggregation strategy. This work addresses this gap by introducing a hybrid framework with a novel MCWG-based decision mechanism.

III. PROPOSED METHODOLOGY

This research proposes an AI-Based Model for Cyberbullying Prevention, Detection, and Action that leverages Natural Language Processing (NLP) and deep learning techniques to accurately identify and mitigate harmful online content. The proposed framework is designed to operate in three primary stages: data preprocessing, cyberbullying detection, and automated response generation, enabling real-time intervention and improved online safety dataset.

A. Multi-Engine Cyberbullying Framework (MECF)

The MECF integrates outputs from all engines to form a unified decision pipeline. Instead of relying on a single model, the framework leverages diversity in model behaviours to improve robustness.

B. Multi-Class Weighted Grading (MCWG)

The MCWG mechanism assigns weights to predictions based on confidence scores. Each engine contributes to the final decision proportionally to its confidence. The final decision is computed using a weighted aggregation strategy:

- Higher confidence predictions have greater influence.
- Conflicting predictions are resolved through weighted voting.

C. Detection Module

The detection module processes predictions and confidence values from all engines. It evaluates agreement levels and determines the final classification.

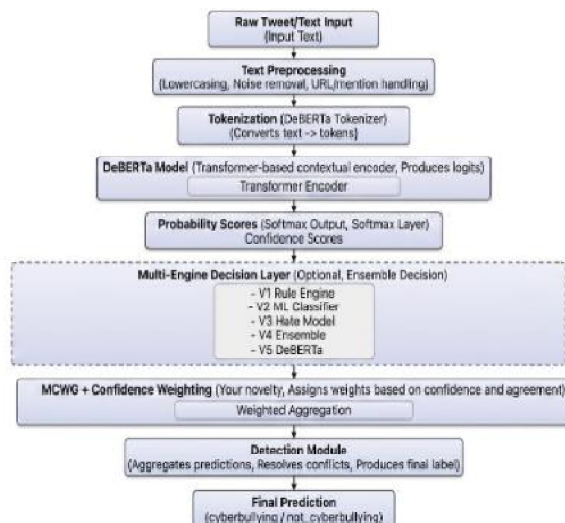


Figure: Block Diagram

1) System Overview

The overall system architecture consists of three functional layers:

- Detection Layer: Identifies whether online text contains cyberbullying using deep learning-based text classification.
- Prevention Layer: Provides early warnings to users before posting potentially harmful content.
- Action Layer: Triggers automated safety actions such as content flagging, reporting, and victim support mechanisms.

The dataset used in this study was compiled from publicly available social media sources, including Twitter, Reddit, YouTube comments, and Kaggle cyberbullying datasets. The collected data contains labelled text samples categorized into classes such as harassment, hate speech, threats, insults, and neutral content. The dataset was divided into training, validation, and testing sets to ensure reliable model evaluation and generalization.[14]

2) Data Preprocessing

To enhance classification performance, raw textual data undergoes several preprocessing steps:

- Removal of URLs, emojis, hashtags, and user mentions
- Conversion of text to lowercase for uniformity
- Elimination of stop words and irrelevant symbols
- Tokenization using the BERT tokenizer
- Handling class imbalance using SMOTE (Synthetic Minority Over-sampling Technique)

These steps help reduce noise and improve the effectiveness of downstream learning processes.

3) Feature Extraction

The system extracts multiple features to capture both semantic and emotional characteristics of text:

- Contextual embeddings generated by BERT, representing each word based on sentence-level context
- Sentiment features to identify emotional tone (positive, negative, neutral)
- Linguistic features, such as word frequency, punctuation intensity, and capitalization patterns

This multifaceted strategy enables more accurate detections of implicit and emotionally driven bullying

IV. MODEL ARCHITECTURE

The proposed system primarily utilizes a transformer-based architecture, specifically the DeBERTa-v3-small model, for cyberbullying classification. DeBERTa enhances contextual understanding using detangled attention mechanisms, allowing an improved representation of the semantic relationships in text. Unlike traditional deep learning approaches such as CNNs or LSTMs, the transformer-based model captures long-range dependencies efficiently without requiring sequential processing. The DeBERTa model is fine-tuned on a binary classification task (cyberbullying vs non-cyberbullying), and its outputs are further integrated into the ensemble framework for final decision-making.[10]

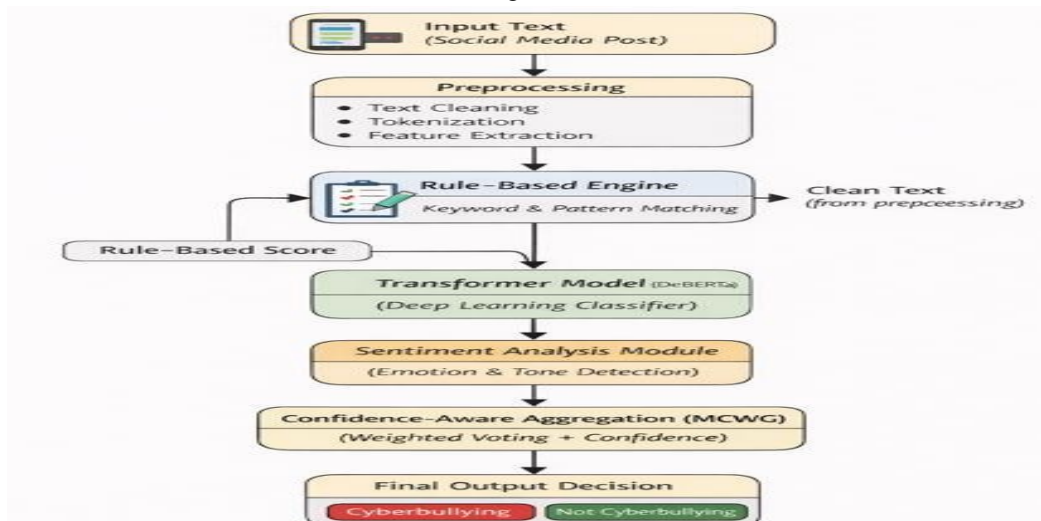


Figure2 :Neuralnetwork[11]

A. Multi-Engine Cyberbullying Framework (MECF)

The MECF integrates outputs from all engines to form a unified decision pipeline. Instead of relying on a single model, the framework leverages diversity in model behavior to improve robustness.

B. Multi-Class Weighted Grading (MCWG)

The MCWG mechanism assigns weights to predictions based on confidence scores. Each engine contributes to the final decision proportionally to its confidence.

- The final decision is computed using a weighted aggregation strategy;
- Higher confidence predictions have greater influence.
- Conflicting predictions are resolved through weighted voting.

V. RESULTS AND DISCUSSION

The model was trained using the PyTorch framework. Cross-entropy loss with class weighting was used to address class imbalance. The AdamW optimizer was employed with a learning rate of 2e-5. Training was conducted for 5 epochs with a batch size of 8. A learning rate scheduler with warm-up steps was applied to improve convergence. Early stopping based on validation of F1-score was used to prevent overfitting.

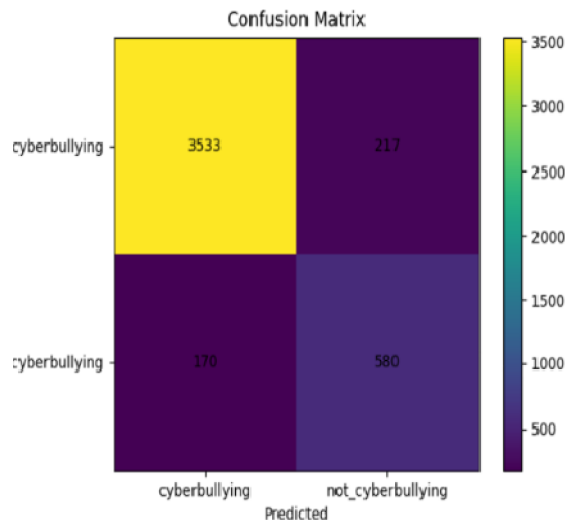


Figure: Confusion Matrix

VI. EXPERIMENTAL SETUP

1) Dataset:

A balanced dataset of social media text samples was used for evaluation. The dataset contains labeled instances of cyberbullying and non-cyberbullying content.

Totalsamples:5000+

Classes: Cyberbullying, NotCyberbullying

2) Data Split:

The dataset was divided as follows:

- Training: 70%
- Validation: 15%
- *Testing: 15%

Stratified sampling was used to maintain class distribution.

3) Model Configuration

The DeBERTa-v3-small model was fine-tuned using:

- Learning rate: $2e-5$
- Batch size: 8
- Epochs: 5
- Optimizer: AdamW

Class weights were applied to handle class imbalance.

4) Evaluation Metrics:

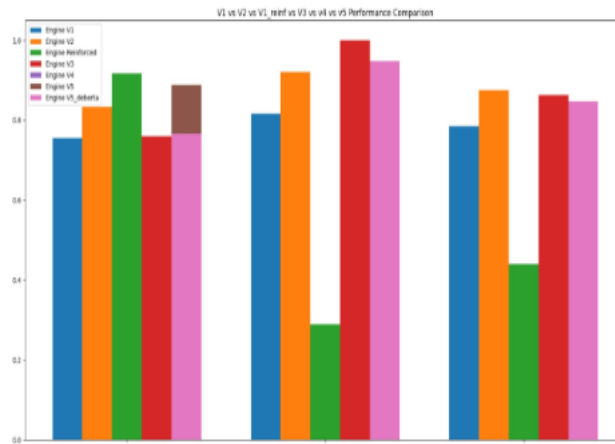
The following metrics were used:

- Precision
- Recall
- F1-score
- ROC-AUC

VII. RESULTS AND DISCUSSION

The model was trained using the PyTorch framework. Cross-entropy loss with class weighting was used to address class imbalance. The AdamW optimizer was employed with a learning rate of $2e-5$.

Training was conducted for 5 epochs with a batch size of 8. A learning rate scheduler with warm-up steps was applied to improve convergence. Early stopping based on validation of F1-score was used to prevent overfitting.



Discussion:

- Engine V1: Rule-based system using predefined patterns.
- Engine V2: Enhanced rule-based system with improved heuristics.
- Engine V3: Transformer-based model (HateBERT).
- Engine V4: Intermediate ensemble combining multiple outputs.
- Engine V5: Fine-tuned DeBERTa model for high-accuracy classification.

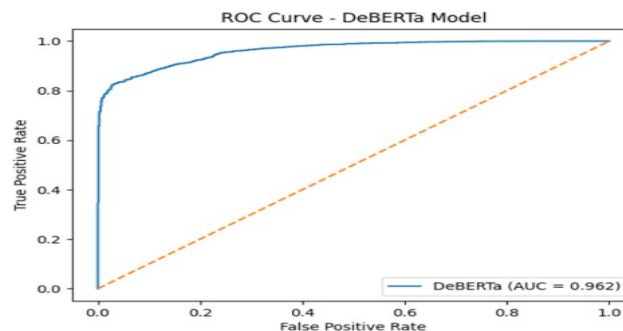


Figure 5: Precision | recall...log

The proposed DeBERTa-based model achieved an AUC score of 0.962, indicating excellent classification capability and strong separability between cyberbullying and non-cyberbullying classes.

Sr. No	Model	Approach	Reported fl
1	SVM	Traditional ML	0.75
2	LSTM	Deep Learning	0.80–0.85
3	BERT	Transformer	0.88–0.90
4	RoBERTa	Transformer	0.90
5	Proposed Model	Hybrid+DeBERTa+MCWG	0.91

TABLE: Performance comparison

VIII. CONCLUSION AND FUTURE SCOPE

This paper presented a hybrid ensemble framework for cyberbullying detection that integrates rule-based and transformer-based models. The proposed MCWG-based decision mechanism enables effective aggregation of multiple predictions using confidence scores.

Experimental results demonstrate that the hybrid approach improves detection performance compared to individual models. The framework provides a balance between accuracy and robustness, making it suitable for real-world applications.

Future work includes incorporating sarcasm detection, improving preprocessing techniques, and extending the model to multilingual datasets.

REFERENCES

- [1] T. Davidson et al., "Automated Hate Speech Detection," ICWSM, 2017.
- [2] Z. Waseem and D. Hovy, "Hateful Symbols or Hateful People," NAACL, 2016.
- [3] J. Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers," NAACL, 2019.
- [4] Y. Liu et al., "RoBERTa: A Robustly Optimized BERT Approach," 2019.
- [5] P. He et al., "DeBERTa: Decoding-enhanced BERT," ICLR, 2021.
- [6] T. Wolf et al., "Transformers: State-of-the-Art NLP," EMNLP, 2020.
- [7] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," JMLR, 2011.
- [8] I. Goodfellow et al., Deep Learning, MIT Press, 2016.
- [9] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, 1997.
- [10] A. Vaswani et al., "Attention is All You Need," NeurIPS, 2017.
- [11] N. Vidgen et al., "Challenges in Hate Speech Detection," 2019.
- [12] H. Zhang et al., "Detecting Offensive Language in Social Media," ACL, 2018.
- [13] K. Chawla et al., "SMOTE: Synthetic Minority Over-sampling," 2002.
- [14] J. Brownlee, "Imbalanced Classification," Machine Learning Mastery, 2020.
- [15] S. Sun et al., "A Survey of Ensemble Methods," Information Fusion, 2020.
- [16] Kaggle Dataset: Cyberbullying Classification Dataset, Available: <https://www.kaggle.com>



10.22214/IJRASET



45.98



IMPACT FACTOR:
7.129



IMPACT FACTOR:
7.429



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089  (24*7 Support on Whatsapp)