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

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.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 becomea critical research challengedueto therapidgrowth ofonlinecommunicationplatforms.Traditionalapproachesrelyeitheronrule-basedsystemsordeeplearningmodels, 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,combinedwithaconfidence-awareaggregationmechanism.AnovelMulti-EngineCyberbullyingFramework (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 anF1-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 communicationplatformshasledtoasignificantrisein 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. Traditionaldetectionmethodsbasedonmanual 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. Earlyapproaches relied onrule-basedsystems and traditional machine learning algorithms, which oftenfailedtocapturecontextualandsemanticnuances 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 andmayproducebiasedpredictions,especiallyinnoisy 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]

Themaincontributionsofthisworkare:

- 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.
- Comprehensiveevaluationandcomparison across multiple models.

II. LITERATURE REVIEW

Cyberbullyingdetection has been widelystudiedusing 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.
Deeplearningmodels, includingConvolutionalNeural Networks(CNNs)andLongShort-TermMemory(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. Thesemodelsleverageattentionmechanismstocapture contextualrelationshipsintexteffectively.Despitetheir success, they may still produce inconsistent results when used independently.[14]
Recentresearchhasexploredensemblemethodsto combine multiple models for improved performance. However,mostexistingensembleapproacheslacka structured confidence-based aggregation strategy. This work addresses this gap by introducing a hybrid framework with a novel MCWG-based decision mechanism.

III. PROPOSEDMETHODOLOGY

This research proposes an AI-Based Model for Cyberbullying Prevention, Detection, and Action that leveragesNaturalLanguageProcessing(NLP)anddeep 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-EngineCyberbullyingFramework(MECF)

TheMECF integrates outputs fromall 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-ClassWeightedGrading(MCWG)

TheMCWG mechanismassigns weightstopredictions basedonconfidencescores. Eachenginecontributesto the final decision proportionally to its confidence. The final decisioniscomputedusinga weightedaggregation strategy:

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

C. DetectionModule

The detection module processes predictions and confidence values from all engines. It evaluates agreementlevelsanddeterminesthefinalclassification.

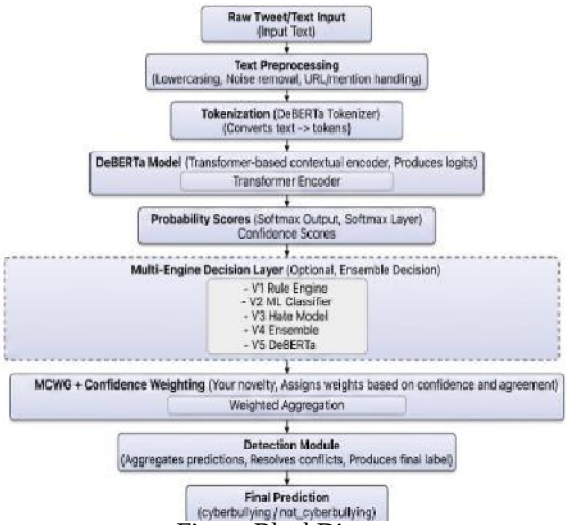


Figure:BlockDiagram

1) System Overview

The overall system architecture consists of three functional layers:

- DetectionLayer:Identifies whether onlinetext 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 actionssuchascontentflagging,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

Toenhanceclassificationperformance,rawtextualdata undergoes several preprocessing steps:

- Removal of URLs, emojis, hashtags, and user mentions
- Conversionoftexttolowercaseforuniformity
- Elimination of stop words and irrelevant symbols
- TokenizationusingtheBERTtokenizer
- Handling class imbalance using SMOTE (SyntheticMinorityOver-samplingTechnique)

Thesestepshelpreducenoiseandimprovethe effectiveness of downstream learning processes.

3) Feature Extraction

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

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

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

IV. MODEL ARCHIETECTURE

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 CNNsor LSTMs,thetransformer-basedmodel 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]

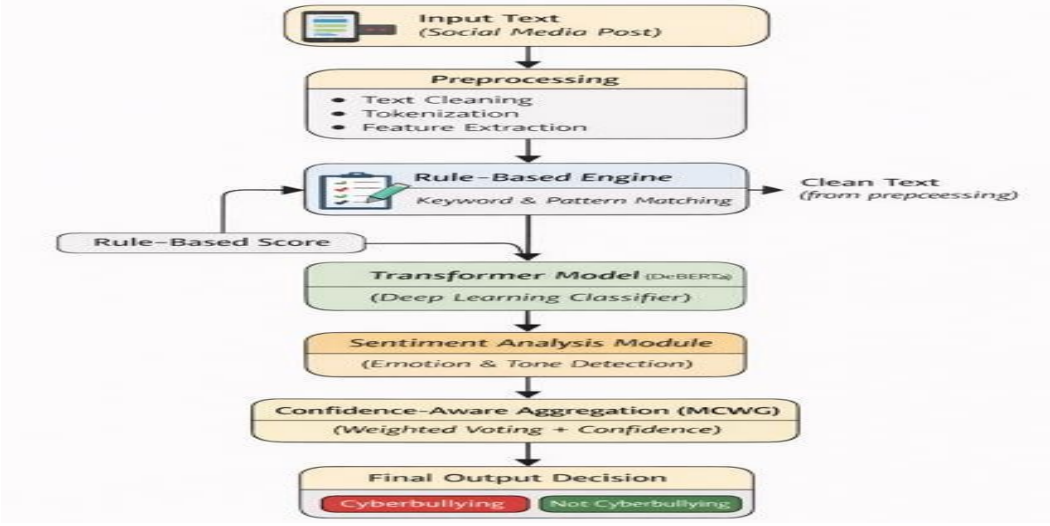


Figure2 :Neuralnetwork[11]

A. Multi-EngineCyberbullyingFramework(MECF)

The MECF integrates outputs fromall 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-ClassWeightedGrading(MCWG)

TheMCWG mechanismassigns weightstopredictions basedonconfidencescores. Eachenginecontributesto the final decision proportionally to its confidence.

- The final decision is computed using a weighted aggregation strategy:
- Higherconfidencepredictionshavegreater 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 learningrateschedulerwithwarm-upstepswasapplied to improve convergence. Early stopping based on validation of F1-score was used to prevent overfitting.

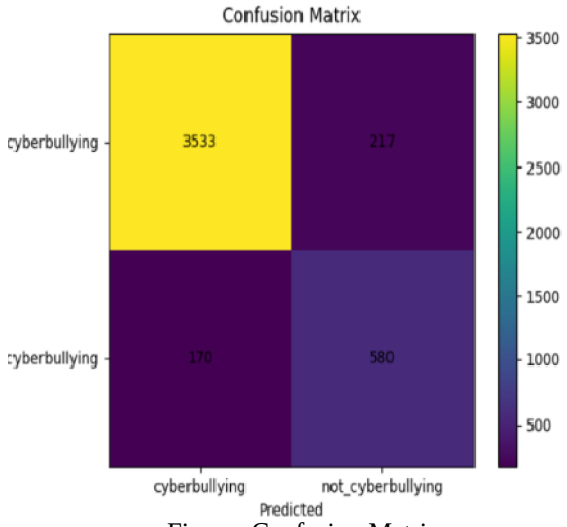


Figure: Confusion Matrix

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) DataSplit:

Thedatasetwasdividedas follows:

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

Stratifiedsamplingwasusedtomaintainclassdistribution.

3) Model Configuration

TheDeBERTa-v3-smallmodelwasfine-tunedusing:

- Learningrate:2e-5
- Batchsize:8
- Epochs: 5
- Optimizer:AdamW

Classweightswereappliedtohandleclassimbalance.

4) Evaluation Metrics:

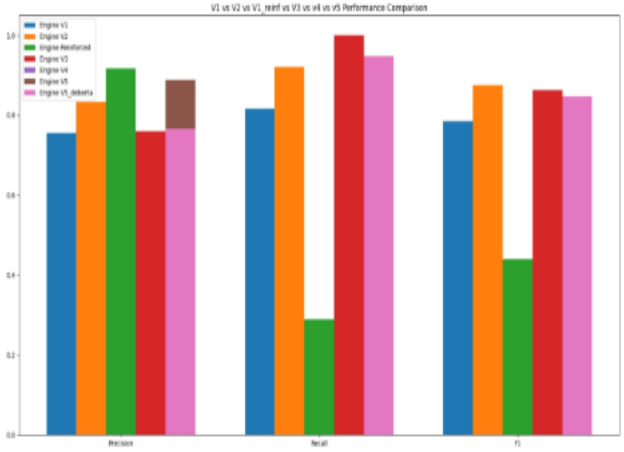
Thefollowingmetricswereused:

- 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 of8.Alearningrateschedulerwithwarm-upsteps 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.
- EngineV5:Fine-tunedDeBERTamodelfor high-accuracy classification.

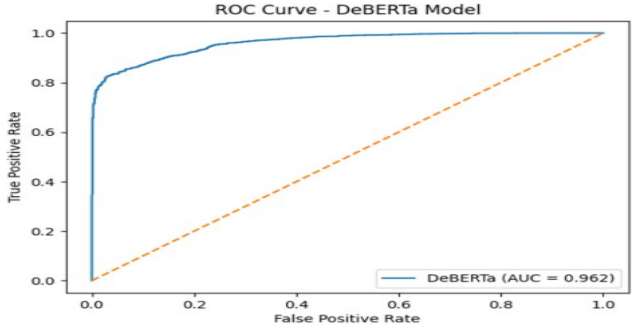


Figure5:Precisionf1recall….log

The proposed DeBERTa-based model achieved an AUC scoreof0.962, indicating excellent classification capability and strong separability between cyberbullying and non-cyberbullying classes.

Sr. No	Model	Approach	Reported f1
1	SVM	TraditionalML	0.75
2	LSTM	DeepLearning	0.80–0.85
3	BERT	Transformer	0.88–0.90
4	RoBERTa	Transformer	0.90
5	Proposed Moodel	Hybrid+DeBERTa +MCWG	0.91

TABLE:Performancecomparison

VIII. CONCLUSION AND FUTURESCOPE

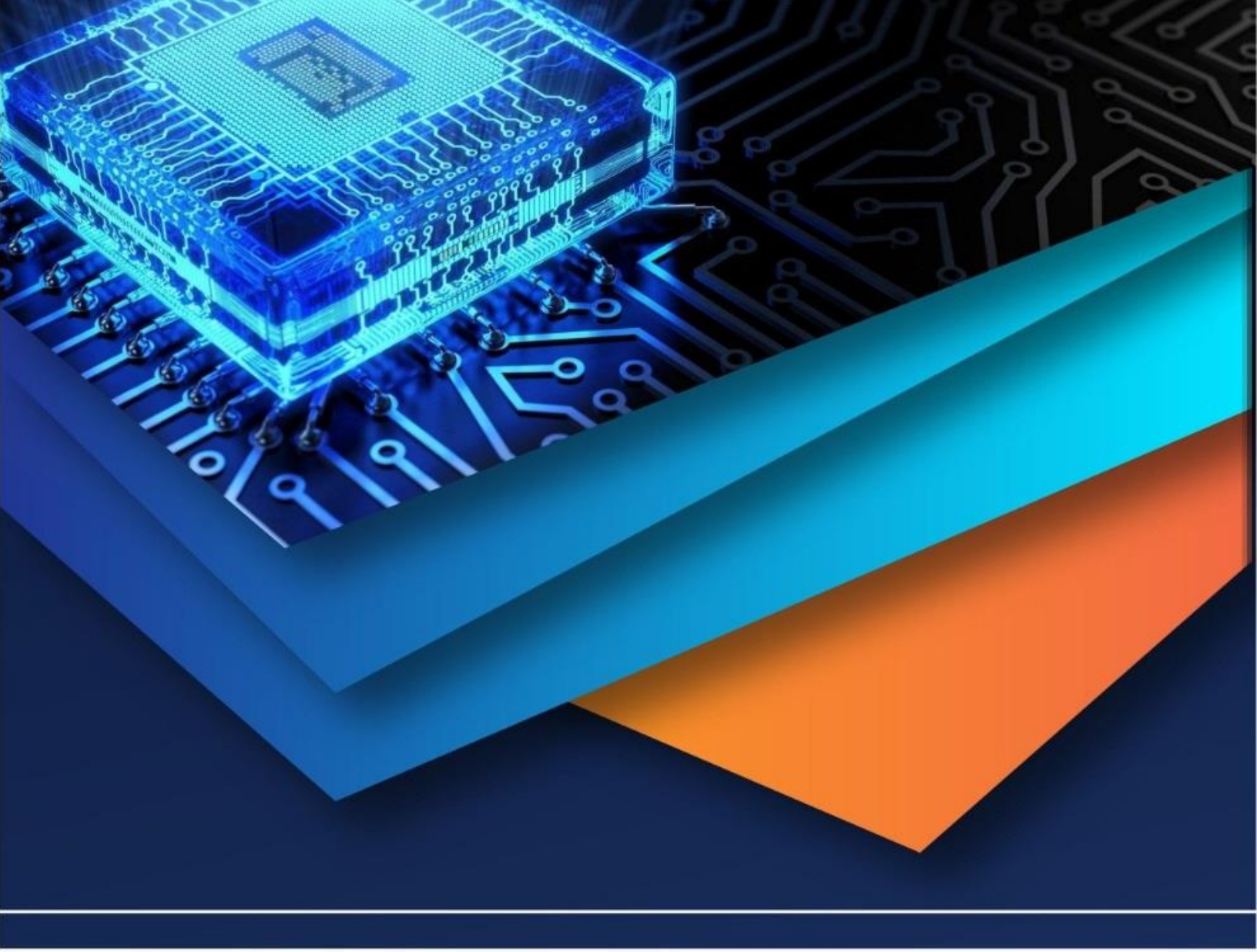
This paper presenteda hybrid ensembleframeworkfor cyberbullying detection that integrates rule-based and transformer-based models. The proposed MCWG- based decision mechanism enables effective aggregation of multiple predictions using confidence scores.

Experimental results demonstrate that the hybrid approachimproves detectionperformancecomparedto individual models. The framework provides a balance betweenaccuracyandrobustness,makingitsuitablefor real-world applications.

Future work includes incorporating sarcasm detection, improvingpreprocessingtechniques,andextendingthe model to multilingual datasets.

REFERENCES

[1] T.Davidsonetal.,“AutomatedHateSpeech Detection,” ICWSM, 2017.
[2] Z.WaseemandD.Hovy,“HatefulSymbolsor Hateful People,” NAACL, 2016.
[3] J.Devlinetal.,“BERT:Pre-trainingofDeep Bidirectional Transformers,” NAACL, 2019.
[4] Y.Liuetal., “RoBERTa:ARobustlyOptimized BERT Approach,” 2019.
[5] P.Heetal.,“DeBERTa:Decoding-enhanced BERT,” ICLR, 2021.
[6] T.Wolfetal.,“Transformers:State-of-the-Art NLP,” EMNLP, 2020.
[7] F.Pedregosaetal.,“Scikit-learn:MachineLearning in Python,” JMLR, 2011.
[8] I.Goodfellowetal.,DeepLearning,MITPress, 2016.
[9] S.HochreiterandJ.Schmidhuber,“LongShort- Term Memory,” Neural Computation, 1997.
[10] A.Vaswani etal., “Attention isAll YouNeed,” NeurIPS, 2017.
[11] N.Vidgenetal.,“ChallengesinHateSpeech Detection,” 2019.
[12] H.Zhangetal.,“DetectingOffensiveLanguagein Social Media,” ACL, 2018.
[13] K.Chawlaetal.,“SMOTE:SyntheticMinority 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] KaggleDataset:CyberbullyingClassification 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