iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** IV **Month of publication:** April 2025

DOI: https://doi.org/10.22214/ijraset.2025.69732

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 13 Issue IV Apr 2025- Available at www.ijraset.com

Cyber-Bullying and Harassment Detection using ML

Omkar Horate[1], Paras Mohite[2], Ashish Padhy[3], Kalpesh Sonar[4], Anas Dange[5]
Department of Artificial Intelligence and Machine Learning, Mumbai university, India

Abstract: ***Cyber-Bullying and Harassment Detection using ML presents a hybrid approach for detecting and hiding cyberbullying content on social media by integrating a machine learning-based detection system into the browser extension. AmultilingualdatasetcontainingEnglish,Hindi,andHinglishtextwaspreprocessedandusedtotrainvariousmachinelearningmodels.Thebestperformingmodel,LinearSVCwithTFIDFfeatures,achieved92.1%accuracyandwasembeddedintotheextensionforrealtimemoderation.The system first applies rule-based filtering, followed by ML classification for ambiguous content. Successfully tested on Facebook and Instagram, the solution enhances online safety by automating cyberbullying detection without affecting user experience. The project demonstrates a scalable and adaptive method for content moderation across dynamic and diverse digital environments.***

Keywords: Cyberbullying Detection, Machine Learning, Browser Extension, Natural Language Processing, Social Media Moderation, Multilingual Text Classification

I. INTRODUCTION

With the exponential growth of social networking services such as Facebook and Instagram, the propagation of hate speech and abusivecontenthasbecomeincreasinglypervasive,affectingusersofallagesandbackgrounds.Browserofferafirstlineofdefensebyemployingpredefinedquerylogic,regularexpressions,andheuristicrulestohideoffensivecommentsinrealtime.However,theserulebasedapproaches often struggle to keep pace with evolving slang, context-dependent expressions, and new forms of coded language, leading to both false positives (benign content hidden) and false negatives (harmful content missed).

To overcome these challenges, this project proposes the seamless integration of machine learning models—specifically, the Linear Support Vector Classifier (LinearSVC) pipeline developed in the Cyberbullying Detection framework—into the extension. The LinearSVCmodelisencapsulatedwithinascikit-learnPipelinethatchainsTF-IDFvectorizationwithclassification,andisserialized via joblib for deployment. At runtime, the extension continues to apply its legacy query filters for rapid screening; comments that pass or trigger these filters are then passed to the ML pipeline. This hybrid filtering architecture preserves backward compatibility and performance, while leveraging data-driven insights to detect nuanced or emergent patterns of abusive language.

Theprimaryobjectiveistodeliverarealtime,automatedmoderationtoolthatenhancestheoriginalextension'scapabilitieswithoutalteringitsuserfacingbehavior.Bycombininglightweight,rulebasedcheckswithrobustMLbasedpredictions,thesystemaimstosignificantlyreducebothfalsepositivesandfalsenegatives,ensuringthatharmfulremarksarereliablyhiddenandrespectfuldiscourse remains uninterrupted. Additionally, the modular design allows for periodic model retraining and updates, enabling the system to adapt to shifting linguistic trends, emerging slang, and new vectors of online harassment.

Ultimately,thisintegratedsolutionservesasaprototypefornext-generationcontentmoderation,demonstratinghowexistingbrowser- based toolscan beaugmentedwith machinelearning to createasafer,morerespectfulonlineenvironment.Itexemplifiesapractical pathway for extending legacy applications with data-driven techniques—maintaining core functionality while introducing scalable, adaptive hate-speech mitigation.

II. AIM & MOTIVATION

The primary aim of this project is to develop and integrate a machine learning-based cyberbullying detection framework into an existing browser extension, which is designed to hide abusive or hateful comments on platforms like Facebook and Instagram. The coreofthissystemisbuiltontraditionalmachinelearningalgorithmssuchasSupportVectorMachines(SVM),LogisticRegression, Random Forest, Multinomial Naive Bayes, and ensemble methods like AdaBoost and Bagging. These models are trained on a multilingual dataset comprising Hindi, English, and Hinglish content, allowing the system to recognize offensive language across diverse linguistic patterns, including code-switching scenarios.

Theintegration involvesreplacing or enhancing theextension'sexisting rule-based commentfilteringmechanismwith atrained ML model(e.g.,LinearSVCpipeline)capableofidentifyingsubtleorevolvingformsofcyberbullying.Thishybridapproach—combining both heuristic and ML-based filtering—ensures backward compatibility while significantly improving detection accuracy. The ML modelisserializedandloadedbytheextensioninrealtimetoanalyzeuser-generatedcontent,withoutalteringtheextension'soriginal user interface or operational workflow.

The motivation behind this work stems from the growing concern around cyberbullying on social media, which affects millions of users,particularlyteenagersandyoungadults.Manualmoderationandbasickeyword-matchingsystemsarenolongersufficientdue to the scale and complexity of online interactions. Harmful content is often disguised using sarcasm, slang, or indirect language, making it difficult for rule-based systems to detect effectively.

Thisprojectaddressesthatgapbyintroducinganautomated,adaptive,andlanguage-awaresolutionthatcanbedeployedinreal-time.

Additionally, by enhancing the existing tool with machine learning capabilities, this work provides a practical and scalable step forward in ensuring safer digital spaces without relying solely on human intervention.

III. PROPOSED METHODOLOGY

Thissectionoutlinesthemethodologyadoptedtodevelopandintegrateamachinelearningbasedcyberbullyingdetectionmodelintotheextension. The combined system aims to identify and hide cyberbullying-related comments on Facebook and Instagram by augmenting rule-based filtering with intelligent, real-time classification.

A. Data Collection and Integration

1) Sources: The dataset is curated from various open-source platforms, including Twitter, Facebook, Kaggle, and academicdatasets containing labeled instances of cyberbullying and non-cyberbullying text in English, Hindi, and Hinglish.

2) Multilingual Focus: The dataset emphasizes multilingual communication to effectively detect abusive language in code-switched and regional dialects commonly found in Indian digital spaces.

B. Data Preprocessing

Preprocessingensuresthedataisclean,consistent,andsuitablefortrainingMLmodels:

1) DataCleaning:Removalofnullvalues,duplicateentries,andirrelevantattributes.

2) DataTransformation:Standardizedlabelingusingbinaryclassification(0=non-cyberbullying,1=cyberbullying).

3) TextNormalization:

4) Lowercasing

5) Removalofstopwords,URLs,emojis,andspecialcharacters

6) Stemmingandlemmatization

7) Tokenizationand Vectorization:

8) Tokenizationsplitssentencesintoindividualwords.

9) TF-IDFvectorizerconvertstextintonumericalformwhilecapturingtermimportance

C. Feature Engineering

1) UnigramFeatureExtraction:Emphasizesindividualkeywordsthatarehighlyindicativeofbullyingbehavior.

2) LinguisticFeatures:Includesprofanitycounts,sentimentscores,andlengthofmessages.

3) These features arecritical formodels todistinguishbetweenbenignandharmful comments,especiallywhenoffensive language is implicit or sarcastic.

D. Model Selection and Training

Multipletraditionalmachinelearningmodelsweretrainedandevaluated:

1) SupportVectorMachine(LinearSVC):

- Accuracy:92.1%
- Strongperformerinhigh-dimensional,sparsedatascenarios.

2) Logistic Regression:
- Accuracy:90.4%
- Efficientandinterpretablemodel,well-suitedforbinaryclassification.

3) RandomForestClassifier:
- Accuracy:89.7%
- Providesrobustnessthroughensembledecisiontrees.

4) MultinomialNaiveBayes:
- Accuracy:86.9%
- Fastandefficientfortextclassificationbutstruggleswithcomplexlanguage.

5) AdaBoostClassifier:
- Accuracy:88.2%
- Boostsweaklearnersbyfocusingonmisclassifiedexamples.

6) BaggingClassifier:
- Accuracy:87.8%
- Reducesvarianceandoverfittingthroughbootstrapaggregation.
- Eachmodelwasevaluatedusingconfusionmatrix-basedmetrics:
- Precision:Correctlypredictedbullyinginstances/Totalpredictedbullying
- Recall:Correctlypredictedbullying/Totalactualbullying
- F1-score:Harmonicmeanofprecisionandrecall
- Thesehelpassessthemodel'srobustness,especiallyonimbalanceddatasets.

E. ModelExportandIntegration

1) Thebest-performingmodel,LinearSVCwithTF-IDFpipeline,isserializedusingjoblibandexportedasa.pklfile.

2) Thistrainedpipelineincludespreprocessingsteps(TF-IDFvectorization)andclassificationinonepackage,ensuringconsistency during inference.

F. ExtensionIntegration

1) TheextensionscansallvisiblecommentsonFacebookandInstagramposts.

2) Originally,itusedregexandkeyword-basedfilteringtohidehatefulorabusivecomments.

3) Intheupdatedsystem:

4) TheextensionloadstheLinearSVCMLmodelduringruntime.

5) Foreverycomment,itfirstappliesexistingrule-basedlogic.

6) Ifacommentisambiguousorpassesinitialrules,itispassedtotheMLmodelforclassification.

7) Ifclassifiedascyberbullying(label=1),thecommentisautomaticallyhiddenusingDOMmanipulation.

8) Thishybridapproachmaintainsbackwardcompatibilitywhileaddingintelligenceandadaptabilitytodetectsubtle,evolvingcyberbullying patterns.

IV. SYSTEM ARCHITECTURE

Thesystemarchitectureofthecombinedprojectseamlesslyintegratesamachinelearningbasedcyberbullyingdetectionmodelwiththebrowserextensiontoenablereal-timemoderationofonlinecontent.Itbeginswithpreprocessingmultilingualtextualdata—such as English, Hindi, and Hinglish—using techniques like tokenization, stopword removal, stemming, and TF-IDF vectorization. The processeddataisthenpassedtoatrainedLinearSVCmodel,whichhasbeenserializedusingjoblibforefficientdeployment. Within the browser extension, user comments on Facebook and Instagram are dynamically captured. While existing rule-based filters are applied first, ambiguous cases are routed through the ML model. If flagged as cyberbullying, comments are hidden instantly using DOM manipulation. The architecture ensures privacy, scalability, and adaptability to evolving language patterns.

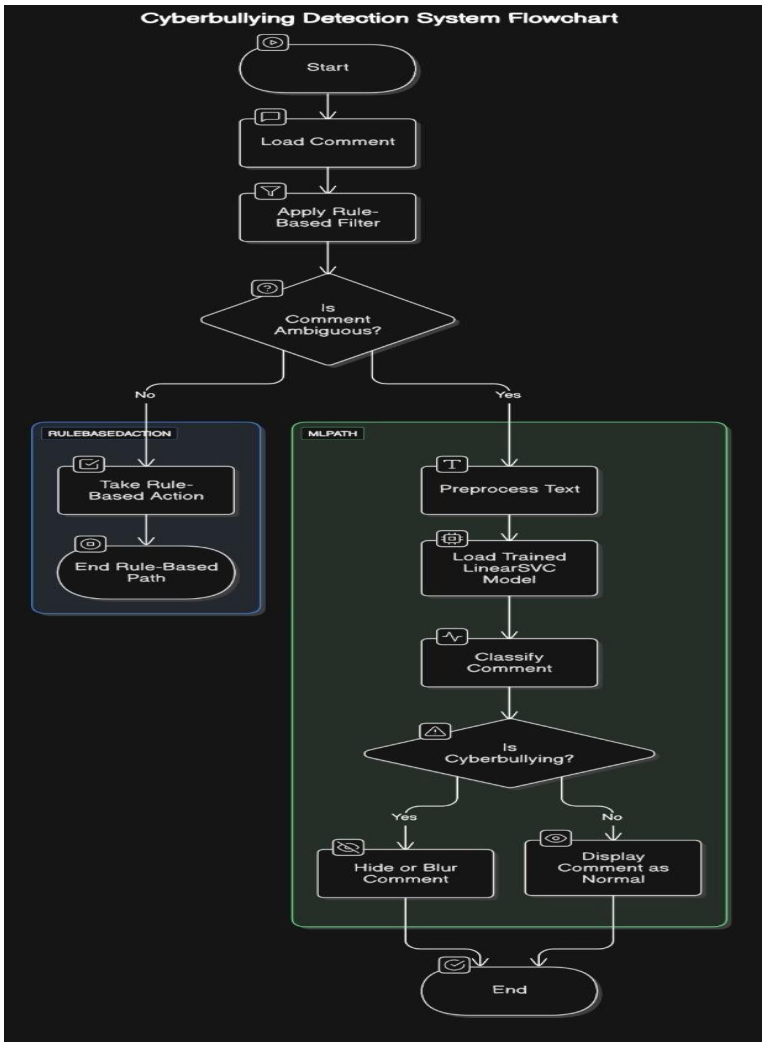


Fig.5.1System Architecture

V. EXPERIMENTATION & RESULTS

FIGURE5.1DETECTIONONFACEBOOK

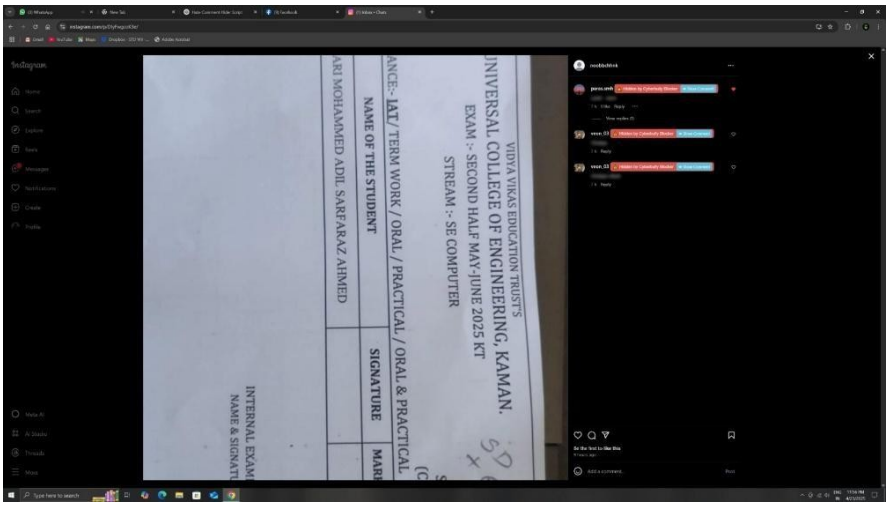


FIGURE5.2DETECTIONONINSTAGRAM

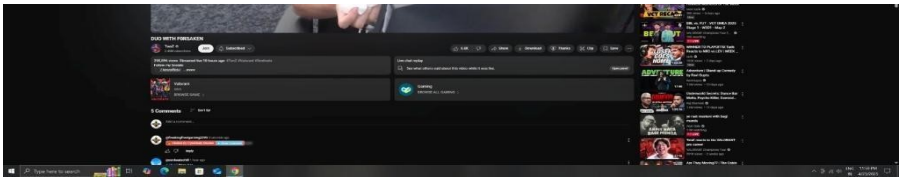


FIGURE5.3DETECTIONONYOUTUBE

The experimentation phase involved training and evaluating multiple machine learning models on a multilingual dataset containing labeled instances of cyberbullying and non-cyberbullying content in English, Hindi, and Hinglish. The dataset underwentextensivepreprocessing,includingtextnormalization,stemming,lemmatization,andTF-IDFvectorization.Models suchasSupportVectorMachine(LinearSVC),LogisticRegression,RandomForest,MultinomialNaiveBayes,AdaBoost,and Bagging were trained and tested using a standard 80-20 train-test split.

Amongthese,theLinearSVCmodelachievedthehighestperformancewithanaccuracyof92.1%,followedcloselybyLogistic Regression at90.4%.Precision,recall,and F1-scoreswerealso calculated to assesstherobustnessof eachmodel,especially in handlingimbalanced classes.

TheLinearSVCmodeldemonstratedstronggeneralizationandlowfalsepositiverates,makingit suitable for real-time implementation.

Post-training, the model was integrated into the browser extension. The extension was modified to include a pipeline that first appliesexistingrule-basedfiltersandthenroutesuncertaincommentstotheMLmodelforclassification.Thefinalsystemwas testedliveonFacebook,InstagramandYoutubeposts.

Visualresultsand screenshotscaptured fromtheworking extension confirmthattheintegrated system successfully detectsand hides abusive comments in real time. The hidden comments are replaced with placeholder messages, maintaining a clean interface forusers.Theseoutputimagesprovidetangibleevidence of thesystem'seffectivenessand demonstrate howmachine learning enhances the hate hider's original functionality without affecting user experience or platform interaction.

VI. CONCLUSION

Inthispaper,weproposedamachinelearningbasedcyberbullyingdetectionsystemintegratedwiththebrowserextensionforrealtimemoderation onsocialmediaplatforms.OurapproachcombinestraditionalMLmodels,particularlyLinearSVC,with TF-IDF feature extraction to accurately classify multilingual comments. The system preserves the extension's original workflowwhile significantlyimprovingdetectionaccuracy.Future work will explore multilingual model expansion, adaptive retraining, and integration with explainable AI techniques for greater transparency.

VII. ACKNOWLEDGEMENT

We take this opportunity to express our heartfelt gratitude to our project guide and coordinator, Mr. Anas Dange, for his invaluable guidance, support, and continuous encouragement throughout the duration of our project. His expertise, insightful suggestions, and patient mentoring played a vital role in the successful completion of our work within the given timeframe.

Our deep appreciation also extends to Dr. J.B. Patil, Principal of Universal College of Engineering, Vasai, Mumbai, and the collegemanagementforprovidinguswith thenecessary infrastructure, facilities,and aconduciveatmospheretocarryoutour project efficiently.

We would like to thank the departmental staff, as well as the library and lab assistants, for their assistance and cooperation throughout the project. Their support contributed significantly to the smooth execution of our work.

REFERENCES

[1] .K.Dinakar,R.Reichart,andH.Lieberman,"ModelingtheDetectionofTextualCyberbullying,"inProc.Int.Conf. Weblogs and Social Media (ICWSM), 2011.

[2] R. Zhao, A. Zhou, and K. Mao, "Automatic Detection of Cyberbullying on Social Networks Based on Bullying Features," in Proc. 17th Int. Conf. Distributed Computing and Networking (ICDCN), 2016.

[3] S.Salawu,Y.He,andJ.Lumsden,"ApproachestoAutomatedDetectionofCyberbullying:ASurvey,"IEEETrans. Affective Computing, 2020.

[4] D. Chatzakou, N. Kourtellis, J. Blackburn et al., "Mean Birds: Detecting Aggression and Bullying on Twitter," in Proc.ACM Conf. Web Science (WebSci), 2017.

[5] N.SoniandA.Singh,"AMachineLearningApproachforDetectionofCyberbullyinginTwitter,"inProc.Int.Conf. Intelligent Computing and Communication (ICICC), 2018.

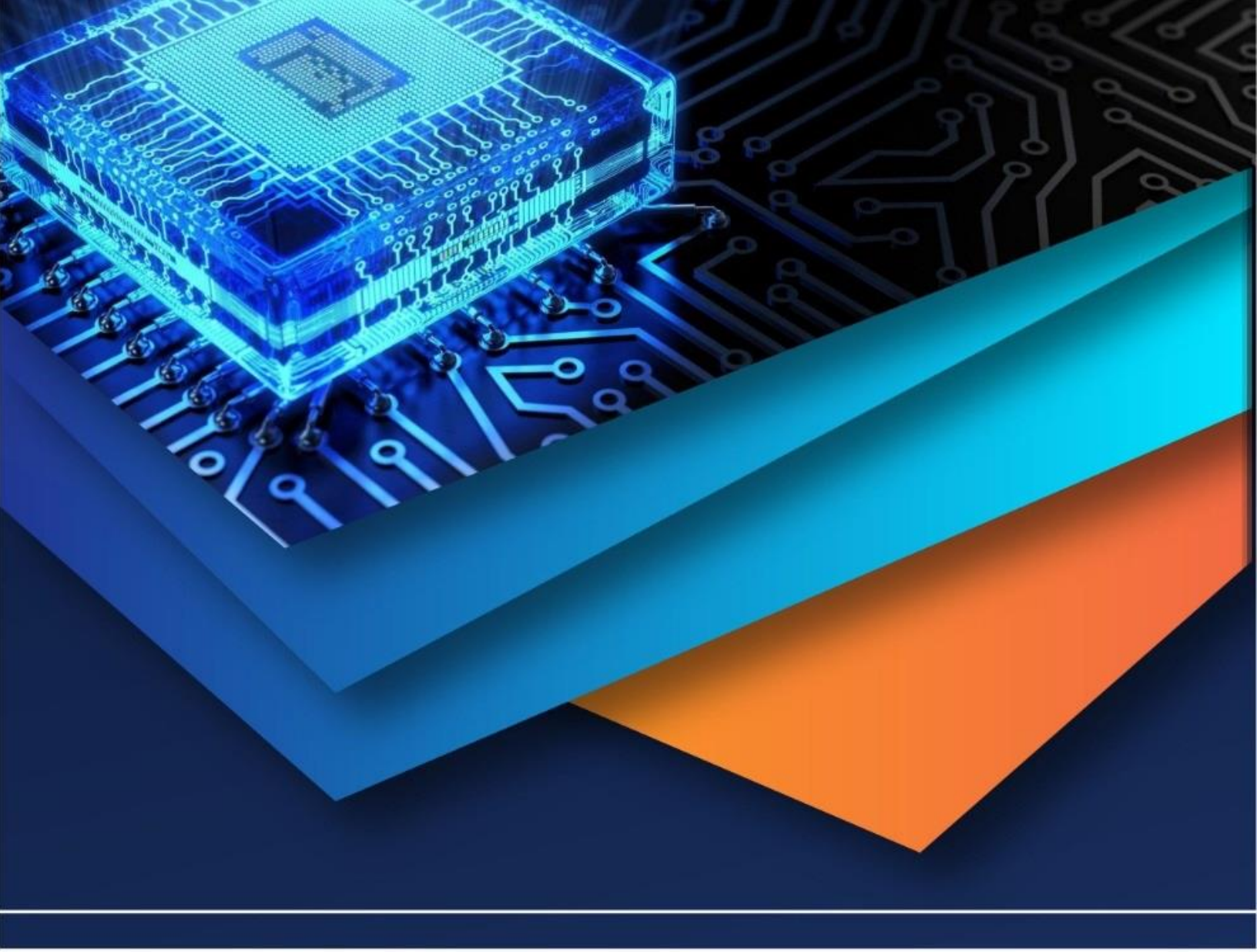
[6] M. Dadvar, F. de Jong, R. Ordelman, and D. Trieschnigg, "Improved Cyberbullying Detection through User Context," in Proc. 35th European Conf. Information Retrieval (ECIR), 2013.

[7] J.M. Xu,K. S.Jun,X. Zhu,and A. Bellmore, "Learning from Bullying Tracesin SocialMedia,"in Proc. North American Chapter of the ACL: HLT, 2012.

[8] C.VanHee,E.Lefever,andV.Hoste,"DetectingCyberbullyinginSocialMedia,"inProc.COLING,2018.

[9] R.Kumar,A. K. Ojha, S.Malmasi,and M. Zampieri, "Benchmarking Aggression Identification in SocialMedia,"in Proc. TRAC Workshop, 2018

[10] V. S. Chavan and S. S. Shylaja, "Machine Learning Approach for Detection of Cyber-Aggressive Comments by Peers on Social Media Network," in Proc. Int. Conf. Adv. Computing, Communications and Informatics (ICACCI), 2015.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY