iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 Issue: VI Month of publication: June 2025

DOI: https://doi.org/10.22214/ijraset.2025.71986

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 VI June 2025- Available at www.ijraset.com

Hate Speech Classification for Text

Ajay Bhagat, Snehal Bhalerao, Radhika Jagtap, Siddhi More, Dr.Dinesh Bhagwan Hanchate, Prof. Swati Mahadev Atole, Dr.Sachin Sukhadeo Bere
Department of Computer Engineering, Dattakala Group of Institution Facultyo f Engineering, Bhigwan-413130, Savitribai Phule Pune University, Maharashtra, INDIA

Abstract: Hate speechhas turned into anoffense that hasincreased in therecent past, bothonline and offiine. There areseveralfactorsthatexplainwhythisisso.Ononeside,peoplearemoreinclinedtoactviolentlybecause of the anonymity provided by the internet,and social networks in particular.Conversely,people's needto express themselves on the internet has grown, and with this has been the prevalence of hate speech. Given how detrimental this kind of discriminatory speech is to society, detection and prevention by social media companies and governments can both be useful. With this survey, we provide an overall overview of the work that has been accomplished in the field, which addresses this dilemma. The use of many com- plex and non-linear models made this challenge possible,and CAT Boost carried out the others becauseit employed latent semantic analysis (LSA) for dimensionality reduction.Hate speech refers to abusiveor discriminatory language directed at an individual or group based on attributes such as race, religion, ethnicity,gender,orsexualorientation. Itsproliferationcanre sultinconcretedamag eandannihilate the security and inclusivity of online environments. This paper presents a machine learning approach to hate speech categorization in text. Utilizing publicly accessible labeled data, we examine various natural languageprocessing(NLP)techniquesandsupervisedlearningalgorithmslikelogisticregression,support vectormachines, anddeeplearningmodelstocategorizeintohatespeech,offensivelanguage,andinnocu- ouscontent. Keyfeaturessuchasn-grams,TF-IDFscores,andwordembeddingsareexploitedtoenhance model performance.The results validate the effectiveness of utilizing linguistic features in combination with optimal classifiers to achieve high precision and accuracy for hate speech detection.The findings emphasize the importance of using well-balanced datasets and ethical considerations while developing automated content moderation systems.

I. INTRODUCTION

With the digital era, the Internet platforms are now at the center of public discussion, where people can shareviewsandpasson information tolargeaudiences.However,thisopennessalsocreatesconditionsfor the spread of harmful material, such as hate speech, words that attack or demean an individual or group based on characteristics such as race, religion, gender, or sexual orientation. Hatespeechcategorizationistheprocessofautomaticallydetectingandfilteringsuchabusivecontent based on natural language processing (NLP) and machine learning (ML) methods. Creating effective and precise models for hate speech detection is not only important for ensuring platform integrity but also for safeguarding users from harassment and encouraging respectful communication.This work is specifically on the construction and deployment of a hate speech classifier based on supervised learning methodologies. Through feature analysis of annotated datasets and consideration of multiple linguistic and contextual information, the system will be capable of differentiating between hate speech, offensive terms, and innocent content with accuracy and consistency.

Hate speech has become a growing issue over the past decade, both online and offiine.Many things areatplayinthiscase.Theinternetanonymitymakesindividualsmorelikelytobehaveaggressively but also more likely to post their views online, which fuels hate speech.Governments and social media corporations can gain from detection and prevention methods as this type of discriminatory speech can have a catastrophic effect on society.We wish that our survey might enlighten us a little on the vast amount of research that has been conducted in this area.

Hate speech is any speech that has the potential to cause harm to a person or group and that couldlead to violence, insensitivity, or irrational or inhuman conduct. With growth in the usage of social media siteslikeTwitterand Facebookcome increasesintheirusageofhatespeech. Theexistenceofhatespeech has a proven connection with the increase of hate crimes. As hate speech becomes a controversial topic, multiple government-led endeavors are being instituted, such as the Council of Europe's campaign No HateSpeech. EUHateSpeechCodeofConductthathastobesignedandobeyedbyallthesocialnetworks in24hoursisonesuch method bywhichithascomeintoimplementation.Manyproblemsraisedbyithave caused serious questions about dataset quality, which this work tries to remedy. This piece also addresses the second issue, that is, prior to making an effective classifier, the optimal features used to classify hate speech need to be studied and determined.

Hate speech is abusive or threatening language that is directed at an individual or a group of people based on features like race, religion, gender, or sexual orientation.Hate speech can have severe adverse effects on users and online conversations.

This thesis tackles the issue of detecting hate speech by creating a machine learning-based classifiertoidentify hatefulmessagesin text. Theresearch investigatesarange oflinguisticand contextualfeatures and tests various supervised learning models to determine their performance.The aim is to develop a strong model that will be part of the continued attempt to make digital spaces safer and more tolerant.

Theriseofhatespeechonlineposesanurgentchallengetobothplatformoperatorsandsocietyingen- eral. Hate speech detection through automated means has emerged as a vital research problem in natural language processing (NLP), working towards limiting the spread of offensive content while maintaining freedom of expression.

With the age of fast digital communication, the spread of user-generated content on the web has brought sophisticated challenges in sustaining healthy discussion.Of these challenges, hate speech has been a particularly insidious type of abusive content.Hate speech is any form of communication that insults an individual or a group on the basis of characteristics like race, religion, ethnicity, gender, sexual orientation, nationality, or other identity features.It not only discredits social cohesion but also adds to actual-world violence, discrimination, and exclusion.As social networking sites, forums, and comment fields increase in usage, so does the amount of content to be moderated. Manual moderation is inefficient and labor-intensive at large scale.

Classificationofhatespeechisadifficulttaskforanumberofreasons:

- SubjectivityandContext-Dependence:

Whatishatespeechmaydifferbetweencultures,communities,andindividuals.Sarcasm,contextual subtleties, and implicit bias make it hard for machines to accurately determine intent.

- DataImbalance:

Hate speech occurrences are frequently sparse relative to neutral or non-hateful material, resulting in highly skewed datasets that can skew learning algorithms.

- Evolving Language:

Users often adopt coded language, slang, or euphemisms to bypass moderation, necessitating con- tinuous updates to models and lexicons.

A. TerminologiesofHate SpeechClassification forText

Datasetsusuallyfallintooneofthesecategoriesbecauseof:

1) Hate Speech:

Hate speech is any message that insults, discriminates, incites violence or prejudice against, or de- meansordemoralizes individualsorgroupsonthebasisofimmutableorinherentcharacteristicslikerace,religion,ethnicorigin,gender,nationality,sexualorientation,ordisability. Hatespeechcanbeovert(e.g., slurs or insults) or veiled (e.g., coded language or racist stereotypes). It is uniquely dangerous because it may lead to offiine harm such as social exclusion, violence, and institutionalized discrimination.

2) Offensive Language:

Offensivelanguageconsistsofabusive,profane,vulgar,ordisrespectfullanguage,butitdoesnotnec- essarily target protected groups or promote hatred.It usually involves personal insults or rude remarks andcanviolatecommunity standardswithout qualifyingashate speech. Examplesincludegeneralcursing or trolling.Differentiating between offensive language and hate speech is essential for ethical AI imple- mentation and effective moderation.

3) Toxicity:

Toxicity is a more general term encompassing any content that is hateful, inflammatory, or likely to degradethequalityofdiscourse. Itincludeshatespeech,abusivecomments,harassment,threats,bullying,andotherformsofverbalabuse.

ModelssuchasthePerspectiveAPIbyJigsaw/Googleusetoxicityscores to assess comment harmfulness and support real-time moderation.

4) Classification:

Classification is the process of labeling input data with predefined categories.In text, this means analyzingwrittencontentand predictingitsclass—suchas“hatespeech,”“offensivelanguage,”or“neutral.”

It is a supervised learning task trained on annotated datasets. Classification tasks can be binary or multi- class/multi-label based on the application.

5) Machine Learning:

Machine Learning (ML) enables systems to learn from data and make predictions.In hate speech detection, ML models learn linguistic patterns from labeled data. Common ML methods include:

- LogisticRegression:Suitableforbinaryclassificationtasks.
- SupportVectorMachines(SVM):Effectiveforhigh-dimensionaltextdata.
- DecisionTrees: Simpleandinterpretable,butpronetooverfitting;oftenenhancedthroughensemble methods like Random Forest or XGBoost.

6) NaturalLanguageProcessing (NLP):

NLP is a subfield of AI focused on enabling machines to understand and generate human language.In hate speech classification, NLP provides tools for extracting syntactic, semantic, and contextual infor- mation from text.Techniques such as named entity recognition, part-of-speech tagging, and semantic similarity are instrumental in creating accurate classification systems.

7) Text Preprocessing:

Textpreprocessingpreparesrawtextdataforanalysisandmodeltraining.Typicalpreprocessingsteps include:

- Tokenization:Dividing text into words or phrases.
- StopWordRemoval:Eliminatingcommonbutnon-informativewords.
- StemmingandLemmatization:Reducingwordstotheirrootorbaseform.
- Lowercasingandpunctuationremovaltonormalizetext.

Thesestepsreducenoiseandstandardizeinputformachinelearningmodels.

8) Tokenization:

Tokenization is the process of breaking a text string into smaller units called tokens, which can be words, subwords, or characters. This is a foundational step in NLP, as most models operate ontokens.AdvancedmethodslikeByte-PairEncoding (BPE)andWordPiece (usedinBERT)helpmanage rare and compound words more effectively.

9) SentimentAnalysis:

Sentiment analysis detects emotional tone or opinion in text and classifies it as positive, negative, or neutral. While not equivalent to hate speech detection, it can act as a supplementary tool. Negative senti- ment often coexists with toxic or hateful language, although sentiment analysis alone lacks the specificity to detect targeted or ideological hate.

10) Precision and Recall:

Thesearecoremetricsinevaluatingclassificationmodels:

- **Precision**= $\frac{\text{TruePositives}}{\text{TruePositives+FalsePositives}}$:

Measurestheproportionofpredictedhatespeechinstancesthatwereactuallyhatespeech.

- **Recall**= $\frac{\text{TruePositives}}{\text{TruePositives+FalseNegatives}}$:

Measureshowmanyactualhatespeechinstanceswerecorrectlyidentifiedbythemodel.

High precision minimizes false positives, while high recall minimizes false negatives.A good balance is vital for fair and effective hate speech detection.

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 VI June 2025- Available at www.ijraset.com

11) F1Score:

TheF1Scoreistheharmonicmeanofprecisionandrecall,offeringabalancedperformance metric:

$$F1=2\times\frac{\text{Precision}+\text{Recall}}{\text{Precision}\times\text{Recall}}$$

It is especially useful in scenarios with imbalanced datasets, where accuracy may be misleading.A high F1 score indicates the model is neither too lenient nor too strict in classifying hate speech.

II. LITERATURE SURVEY

A literature survey of hate speech classification for text would report the current research, methods, datasets, challenges, and advancements in this area.

A. IntroductiontoHateSpeechClassification:

Hate speech isn't just about harsh words—it's about language that targets people because of whothey are. Whether it's based on race, religion, gender, or other core identities, hate speech aims to insult, exclude, or harm.

It's different from offensive or rude language, which might be inappropriate but not necessarily dis- criminatory.It also overlaps with things like online harassment and abuse, but not every insult is hate speech.

As social media becomes a bigger part of our daily lives, the need to detect and filter hate speech automatically has grown. These platforms allow messages to spread quickly—sometimes with real-world consequences. But building automated tools to identify hate speech is tricky.

Definitionsofwhatcountsas"hatespeech"varydependingonthecultureorcountry,andthere's the added challenge of keeping these systems fair and unbiased.We also have to be careful when deal- ing with sensitive issues related to protected groups, making it vital to approach this work ethically and thoughtfully.

B. Datasets for Hate Speech Classification:

To teach a computer how to recognize hate speech, we first need examples.That's where datasets come in. Some of the most well-known datasets include:

- Hatebase:A multilingual collectionof hate-related terms.
- Twitterdatasets:Collectionsoftweetslabeledashatespeech,offensive,or neutral.
- Facebook and YouTube datasets: Where comments are annotated based on their toxicity or harm- fulness.

Thesedatasetshelptrainandtestdetectionmodels,butthey'renotperfect.Manyare*platform-specific*, meaning a model trained on Twitter might not work well on Reddit.

Some datasets also suffer from *bias* or a *lack of diversity*, especially when they focus only on English or specific regions. That's why expanding and improving our data sources is an ongoing challenge.

C. Preprocessing Techniques:

Before we feed text into a machine learning model, we need to clean it up—just like editing a rough draft before turning it in. Here's what that usually involves:

- TextCleaning: Lowercasingallwords,removingspecialcharacters,punctuation,orfillerwordslike "the" or "is" that don't add much meaning.
- Tokenization,Stemming,andLemmatization:
  - Tokenizationbreaksthe textinto wordsor phrases.
  - Stemming and lemmatization reduce words to their root forms—so "running" and "ran" both become "run," for example.
- Handling Imbalanced Data:Hate speech is often rare compared to regular,non-harmful content.This imbalance can confuse models, so we use techniques like:
  - Oversamplingtheminorityclass.
  - Undersamplingthemajorityclass.
  - Data augmentationtocreatesyntheticexamples.

D. Feature Extraction Techniques:

Machines don't understand language the way we do—they need numbers. So we have to convert text into numerical formats through *feature extraction*.

- BagofWords(BoW)andTF-IDF:Thesearetraditionalmethodswhereeachwordbecomesanumber based on how often it appears. TF-IDF adjusts scores by giving less importance to common words.
- WordEmbeddings:Toolslike*Word2Vec*,*GloVe*,and*fastText* createwordmapsthatshowhowwords relate to each other.For example, "king" and "queen" might be closer together than "king" and "banana."
- ContextualEmbeddings:Modelslike*BERT*and*RoBERTa*don'tjustlookatindividualwords—they understand the meaning based on surrounding words. That means they can tell the difference be- tween "bad dog" (as a scolding) and "bad guy" (as a villain).

E. MachineLearningModelsforHateSpeechClassification:

Once we have features,it's time tochoose a model.Early on, researchersused simpler methodslike:

- NaïveBayes, LogisticRegression, andSVM:Theseclassicalmodelsarefastandeasytouse, espe- cially with BoW or TF-IDF features.
- Thencameneuralnetworks:
- CNNsandRNNs(LSTM,GRU):Thesearebetteratunderstandingsequencesoftextandpickingup patterns that older models miss.
- Now,transformershavetakenover:
- ModelslikeBERT,GPT,andRoBERTahavebeenpre-trainedonmassiveamountsofdataandfine tunedforspecifictaskslikehate speech detection. They'regreatatcapturingcontext, whichmakes them highly accurate.

Finally, some researchers are combining different approaches—hybrid models—that mix rule-based filters, traditional ML, and deep learning to better capture the complexity of human language and hate speech.

III. METHODOLOGY

The hate speech classification methodology is a systematic approach that brings together data-driven nat- ural language processing methods with moral considerations.

Presented below is a step-by-step detailed account of the steps used in developing an efficient text-based hate speech detector.

A. Data Collection and Annotation:

Theinitialstepinthemethodologyistocollectdiverseandpertinenttextdata.Thisencompasses:

- SocialMediaSites: TextpostsorcommentsfromsiteslikeTwitter,Facebook,Reddit,orYouTube.
- OnlineForums: Specializedsitesthatcanhostideologicallymotivateddiscussions(e.g.,extremistorconspiracy forums).
- PublicDatasets: DatasetsfoundonrepositorieslikeKaggle,GitHub,orresearchlibraries(e.g.,Hatebase,Davidson's Twitter dataset).

AnnotationStrategy:

- Textsamplesarecategorizedintotypessuchas:HateSpeech,OffensiveLanguage,andNeutral.
- Multiple annotators are employed to ensure consistency, and inter-annotator agreement (e.g., Co- hen's Kappa) is calculated.
- Clearannotatorguidelinesareestablishedtohelpdistinguishhatespeechfromgeneralprofanity or sarcasm.

B. Data Preprocessing

Preprocessingisnecessarytoremoverawtextcontaminantsandpreparedataformachinelearningmodels.

- CleaningofText:
  - Lowercasingtext
  - Eliminationofpunctuation,HTMLtags,emojis,andURLs
  - Removal ofextraspacesand non-alphanumericcharacters
- Tokenization:

Splitting sentences into words or subwords using tools like SpaCy, NLTK, or Transformers' Tok- enizers.

- StemmingandLemmatization:

Reducing words to their root forms (e.g., "arguing" → "argue", "better" → "good") to consolidate lexical variations.

- HandlingImbalancedData:

- Oversampling:Copyorsynthesizenewexamplesfortheminorityclass.
- Undersampling:Eliminatesamplesfromthemajorityclass.
- DataAugmentation:Utilizeparaphrasing,synonymsubstitution,orback translation.

C. Feature Extraction

Thenextstepistorepresenttextinnumericalformsthatcanbeunderstoodbymachinelearningalgo- rithms.

- ClassicalTechniques:
  - BagofWords(BoW):Simplycountsthefrequencyofwords,irrespectiveofcontext.
  - TF-I*DF:*Assignsweightstowordsdependingonhowsignificanttheyareinthecorpus.
- WordEmbeddings: *Word2Vec*,*GloVe*,*FastText*:Pre-trainedvectorsthatcapturesemanticwordsimilarity.
- ContextualEmbeddings: Modelssuchas*BERT*,*RoBERTa*,and*DistilBERT*produceembeddingsthattakethecontextofword meaning into account, allowing detection of subtle hate speech.

D. ModelChoiceandTraining

Thetransformedfeaturesareinputintodifferentmodelsforclassification.Commonmodelsinclude:

- TraditionalMachineLearningModels:
  - LogisticRegression
  - SupportVectorMachines(SVM)
  - NaiveBayes

Idealforsmalldatasetsandrapidbaselinetesting.

- DeepLearningModels:
  - ConvolutionalNeuralNetworks(CNNs):Extract spatialpatterns(n-grams).
  - RecurrentNeuralNetworks*(RNNs)*,*LSTMs*,*GRUs:*Capturesequenceandcontextinformation.
- .Transformer-BasedModels:
  - *BERT*, *RoBERTa*, *XLNet*,*GPT*: Fine-tunedon hatespeech datasetsfor state-of-the-artresults.
- HybridandEnsembleApproaches:
  - *Combine rule-based filters (e.g., slur detection) with machine learning classifiers.*
  - *UseensemblemethodslikeRandomForest*+BERTforenhancedaccuracyandrobustness.

E. EvaluationMetrics

The models are evaluated based on their ability to accurately detect hate speech while minimizing mis-classification.

- Accuracy: Overallcorrectness(lessusefulinimbalanceddatasets).
- Precision: Howmanypredictedhatespeechcaseswereactuallyhatespeech.
- Recall: Howmanyactualhatespeechcasesweredetected.
- F1-Score: Harmonicaverageofprecisionandrecall.
- ConfusionMatrix: Graphicalrepresentationoftrue/falsepositivesandnegatives.
- ROC-AUC: Evaluatesclassifierperformanceacrossdecisionthresholds.

F. FairnessandBiasAnalysis

Ethicalconsiderationsarecrucial.Detectionsystemsmustavoidperpetuatingexistingsocietalbiases.

- BiasTesting: Evaluatemodelperformanceacrosssubgroupsbasedonrace,gender,religion,etc.
- FairnessMetrics: Measureequityusingstatisticaltestssuchasequalizedodds,demographicparity,anddisparate impact.
- BiasMitigation: Applymethodssuchasadversarialdebiasing,bias-awarelossfunctions,ordatabalancing.

G. Deployment Considerations

For real-worldapplication,modelsmustbescalableandcontinuallymonitored.

- Real-timepipelinescanbedeployedusingAPIsorcloudinfrastructure(e.g.,AWS,GCP,Azure).
- Continuousmonitoringisrequiredtodetectdatadriftandensureperformancedoesnotdegrade.

- Incorporatefeedbackloopsfromuserstoimprovemodelaccuracyandfairnessovertime.

IV. IMPLEMENTATION DETAILS

A. Data Collection and Annotation:

CollecttextdatafromsocialmediaplatformssuchasTwitter,Facebook, onlineforums, orfrompublicly available hate speech datasets.Examples include various Kaggle datasets that provide labeled data in multiplelanguages. Ifconstructingacustomdataset ,annotateeachtextinstanceintopredefinedcategories such as *hate speech*, *offensive language*, and *neutral*.

Annotation should ideally be performed by multiple annotators to ensure consistency and minimize bias.The dataset should be cleaned to remove duplicates, bot-generated texts, and noisy data that could degrade model performance.

B. Data Preprocessing:

- Text Cleaning: Removespecialcharacters,emojis,URLs,andconvertalltexttolowercase.Optionally,removestop words that do not contribute to classification.
- Tokenization: SplitthetextintoindividualwordsorsubwordsusinglibrariessuchasSpaCyorNLTK.
- StemmingandLemmatization: Convertwords to their base or root form(e.g., “running” →“run”) to reduce redundancy.
- HandlingImbalance:
  - *Oversampling:* Increasethenumberofhatespeechexamplesbyduplicatingorsynthesizingnewsamples.
  - *Undersampling:* Reducethenumberofnon-hatespeechinstances.
  - *DataAugmentation:* Use techniques such as synonym replacement,back translation,or paraphrasing to expandthe hate speech class.

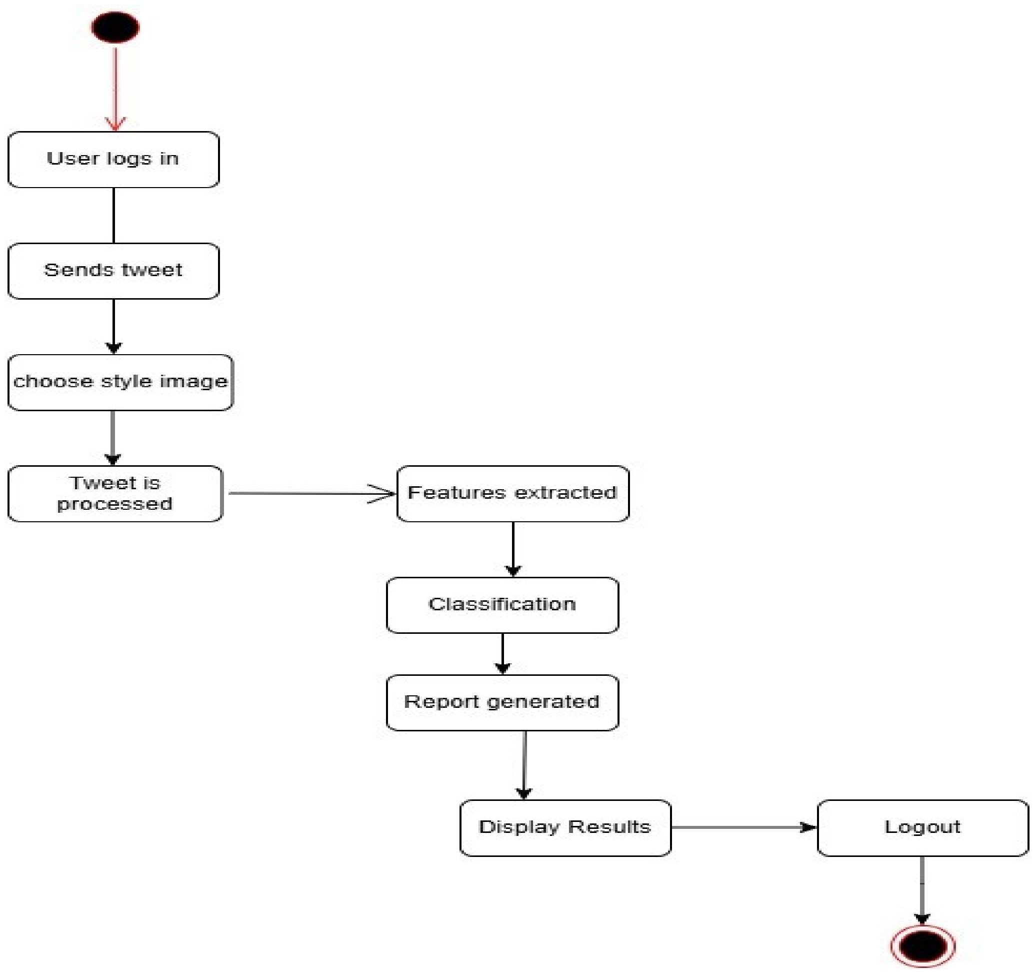


Figure:ActivityDiagram

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 VI June 2025- Available at www.ijraset.com

C. FeatureExtraction:

- BagofWords(BoW):Representstextasamatrixofwordoccurrencecounts.
- TF-IDF (Term Frequency-Inverse Document Frequency):A refinement of BoW that gives higher weight to less frequent but more informative words.
- WordEmbeddings:Usepre-trainedwordvectorslikeWord2Vec, GloVe, orFastText tocapturese- mantic relationships between words.
- Transformer-basedEmbeddings: Leveragemodelssuchas*BERT*,*RoBERTa*,or*DistilBERT*thatgen- erate context-aware word embeddings.

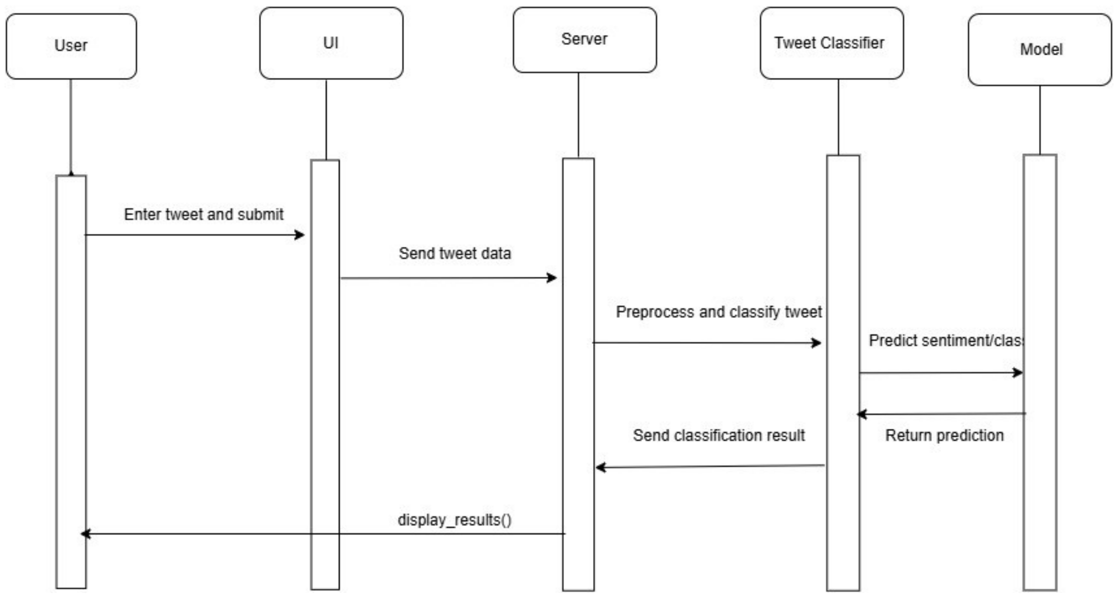


Figure:SequenceDiagram

D. EvaluationMetrics:

- Precision: Proportionofpredictedhatespeechsamplesthatareactuallyhatespeech.
- Recall: Proportionofactualhatespeechsamplescorrectlyidentifiedbythemodel.
- F1Score: Harmonicmeanofprecisionandrecall,especiallyvaluableinclass-imbalanceddatasets.
- ConfusionMatrix: Visualizes true positives, false positives, true negatives, and false negatives, allowing for a betterunderstanding of model errors.

E. HandlingBiasandFairness:

- BiasDetection: Evaluate model performance across diverse demographic subgroups to ensure consistent accuracy and recall across races, genders, and other protected categories.
- FairnessMetrics: Applystatisticalfairnesstestssuchas*equalizedodds*or*disparateimpact*toassessandmitigate potential biases in the model.

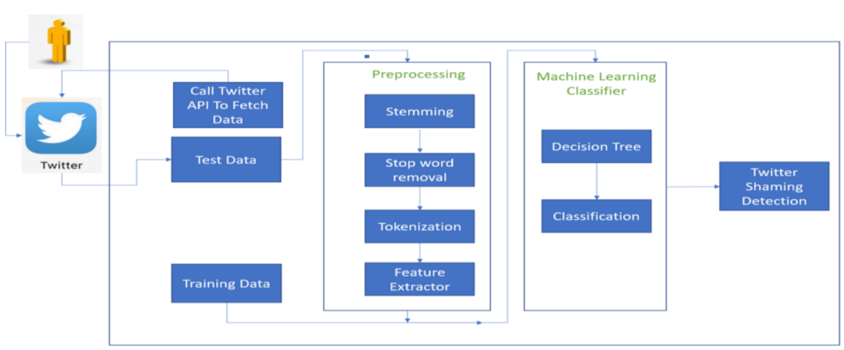


Figure: System Architecture

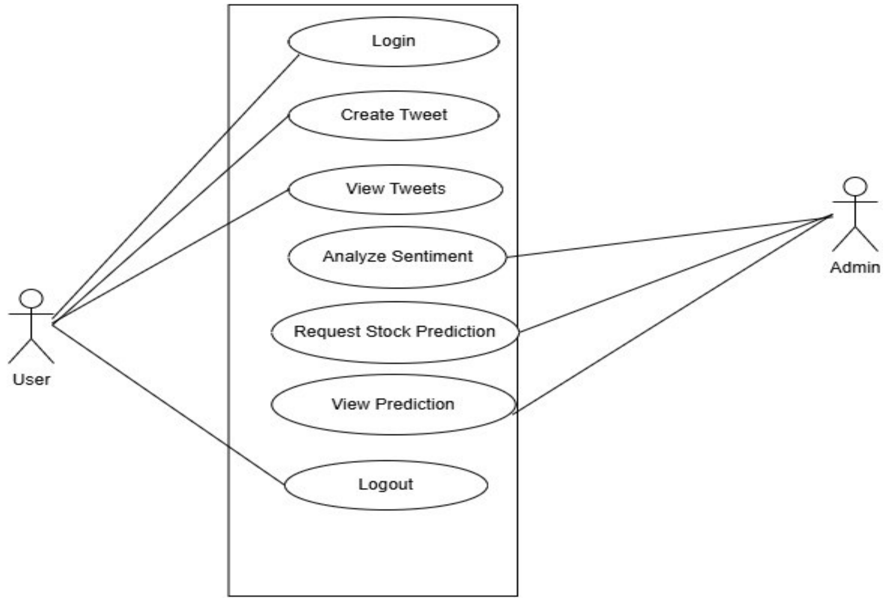


Figure:UseCaseDiagram

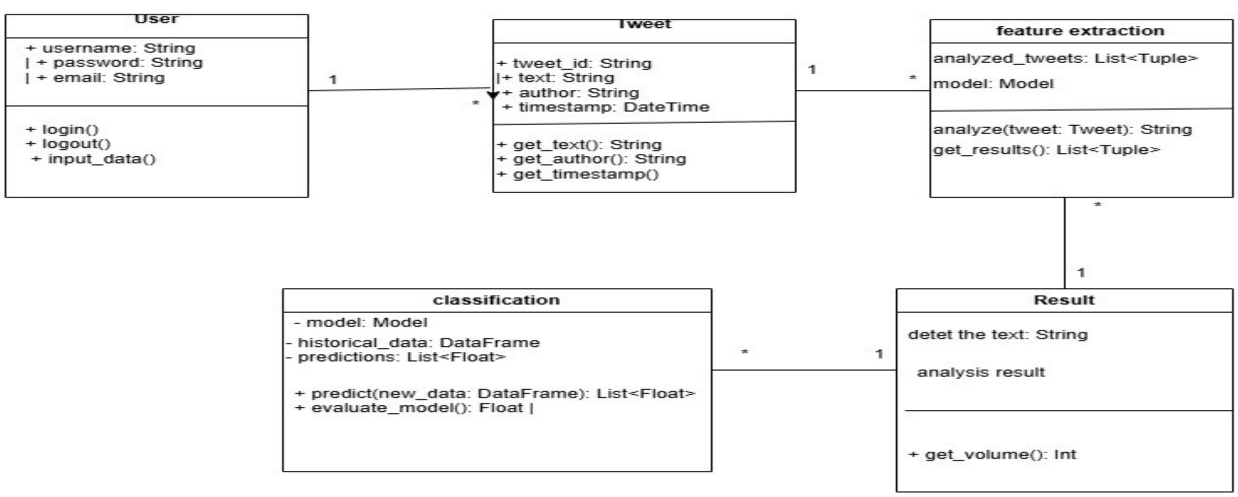


Figure:ClassDiagram

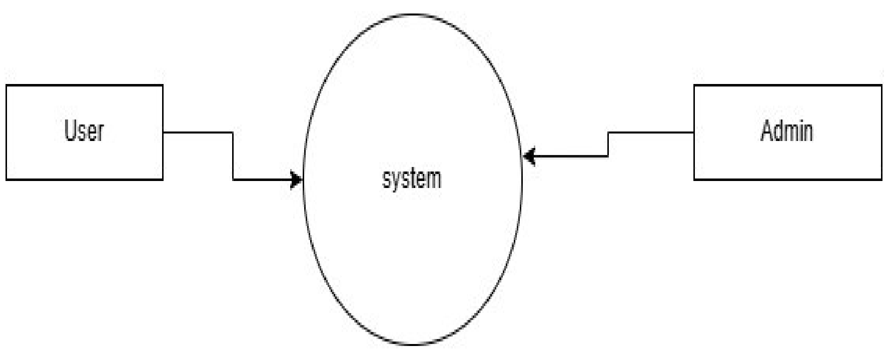


Figure:DFDLevel0

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 VI June 2025- Available at www.ijraset.com

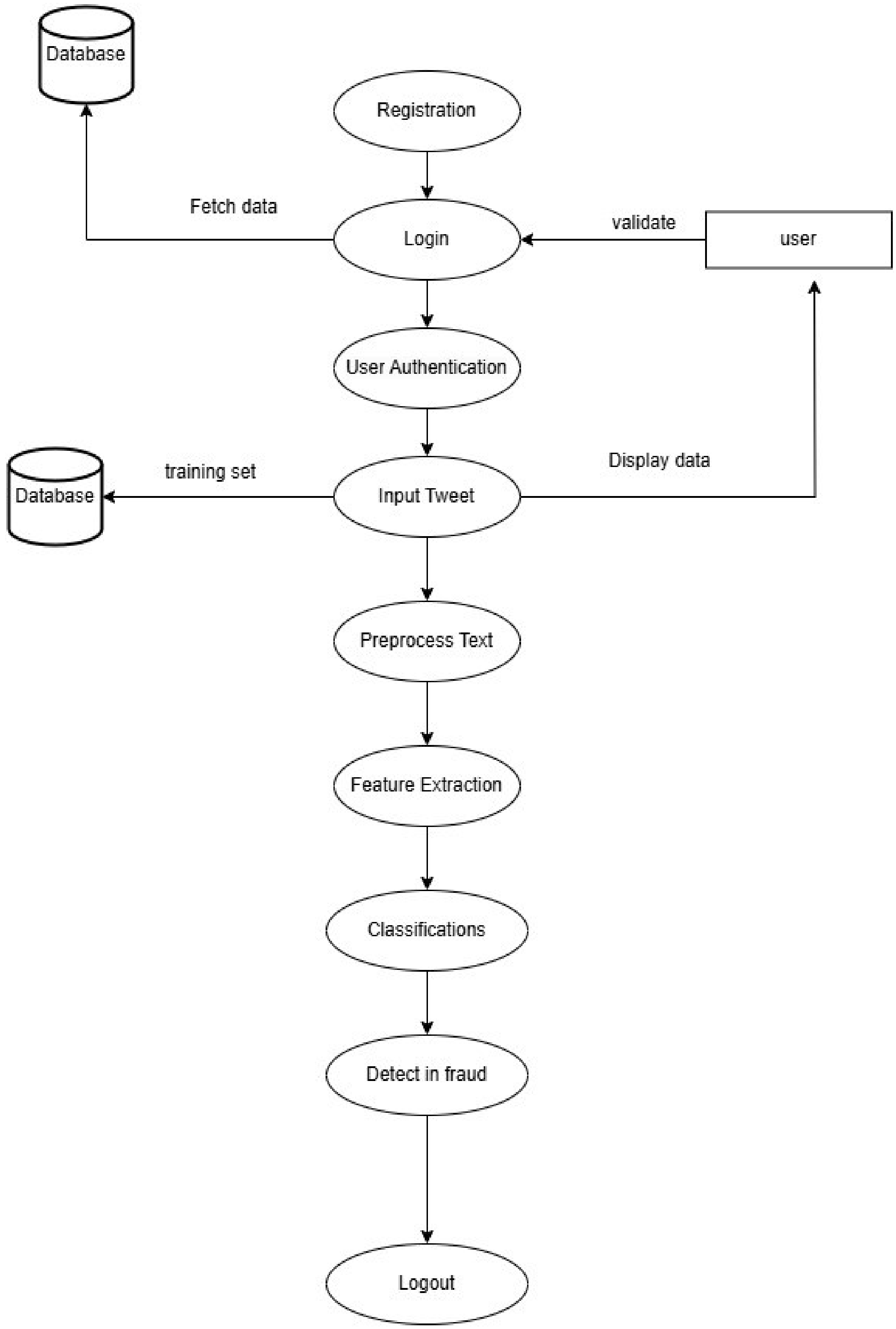


Figure:DFDLevel1

V. RESULTS & DISCUSSION

Detection of hate speech in Tex was overlooked in earlier technology as there was no survey of auto-maticdetection.IntheWhiteSupremacyForum, therearealotmoresentenceswhicharenotusedforhate speech compared to 'hateful' sentences. There is a good chance that the boost in the F1-score in the two datasetswasmoderated bythesinglefeature (countof)'Followers',whichalsoboostedthesubsetimprove- ment.These patterns and unigrams may be applied as pre-compiled dictionaries not part of the proposed hate speech detection dictionaries as pre-existing dictionaries to be used in future research projects.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue VI June 2025- Available at www.ijraset.com

VI. CONCLUSION

The conclusion emphasizes the success of this project in overcoming such challenges with a multi-class classification approach.The success factor was the development and utilization of ten separate binary datasets, eachdealingwithaparticulartypeofhatespeech. Ratherthangroupingeverythingtogether, this fine-grained approach allowed models to concentrate on distinctive features of each hate category. Each dataset was annotated with great care by domain experts strictly adhering to guidelines, so labeling was highlyconsistentandaccurate.Takingsuchcareenhancedtrainingandtestingofthemodels,resulting in improved generalization and practical application. Also, the datasets were balanced, which in machine learning is important to avoid bias towards the majority class. Hate speech has been underrepresented in mostoftheexistingdatasets, andsuchbiascanoccurinmodeloutputs. Equalrepresentationhereallowed classifiers to perform better on all the classes.

VII. ACKNOWLEDGEMENTS

Thispaperwouldnothavebeenwrittenwithoutthesupportandencouragementof Asst. Prof. Dr.D.B.Hanchate, guide of BE Dissertation work. Author's special thanks go to all the professors of Computer Engineering Department of DGOI FOE Swami-Chincholi, for their guidance and for giving her an opportunity to work onHateSpeechClassificationonText.

REFERENCES

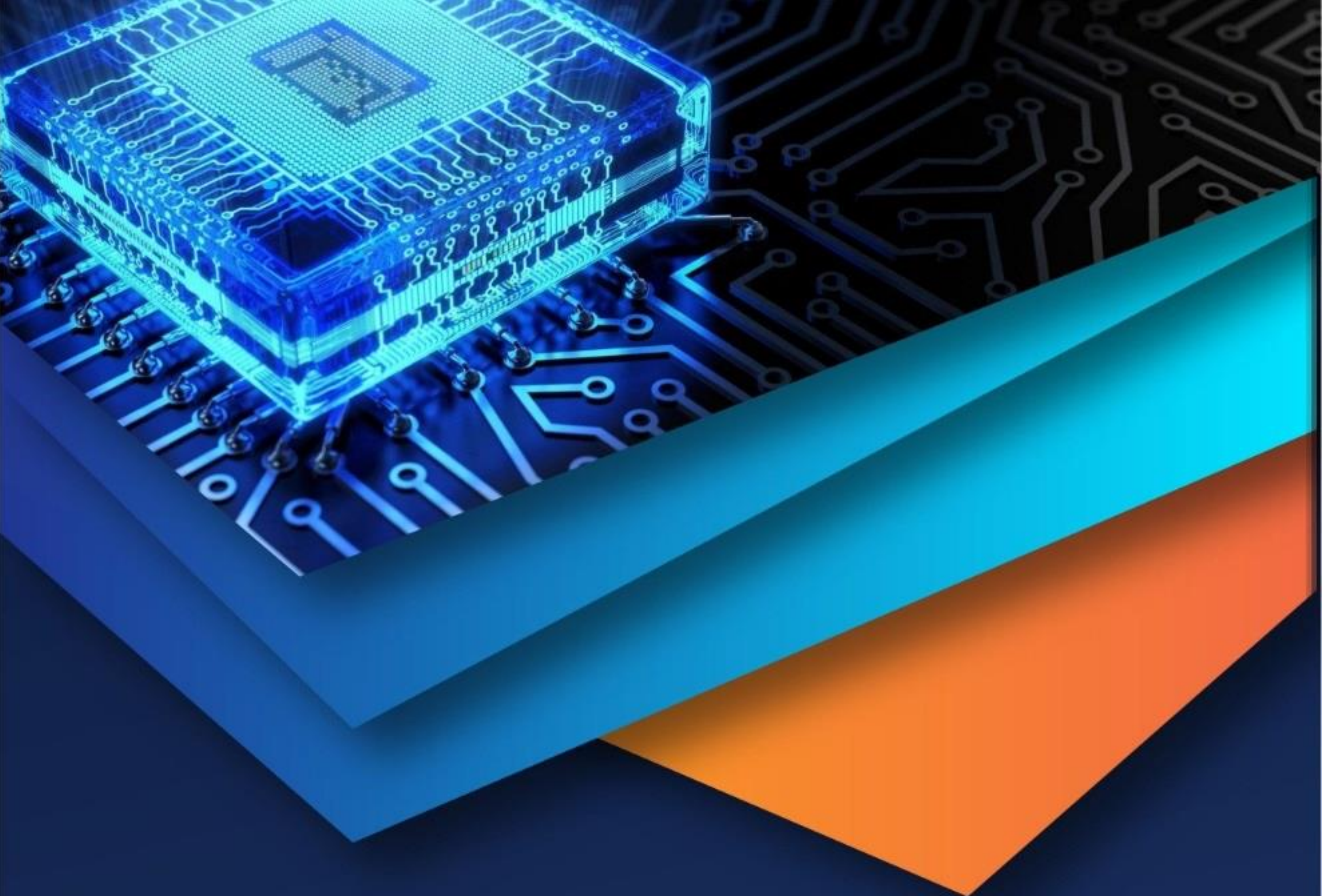
[1] HateSpeechExplained: AToolkit,vol. 19,London,U.K.,2015.
[2] K.Saha,E.Chandrasekharan,andM.DeChoudhury,"Prevalenceandpsychologicaleffectsofhateful speechinonlinecollegecommunities,"inProc.10thACMConf.WebSci.,Jun.2019,pp.255–264.
[3] M.BilewiczandW.Soral,"HatespeechEpidemic. Thedynamiceffectsofderogatorylanguageon intergrouprelationsandpoliticalradicalization,"PoliticalPsychol.,vol. 41,no. S1,pp. ,Aug. 2020.
[4] E. Blout and P. Burkart, "White supremacist terrorism in charlottesville: Reconstructing unite the Right,"Stud. ConflictTerrorism,pp. 1–22,Jan. 2021.
[5] R. McIlroy-Young and A. Anderson, "From 'welcome new gabbers' to the Pittsburgh synagogue shooting: The evolution of gab," in Proc.Int.AAAI Conf.Web Social Media, vol.13, 2019, pp. 651–654.
[6] A. Warofka, "An independent assessment of the human rights impact of Facebook in Myanmar," Facebook Newsroom, vol. 5, Nov. 2018.
[7] T.H.Paing,"ZuckerbergurgedtotakegenuinestepstostopuseofFbtospreadhateinMyanmar," Irrawaddy.
[8] Z. Waseem and D. Hovy, "Hateful symbols or hateful people?Predictive features for hate speech detectiononTwitter,"inProc. NAACLStudentRes. Workshop,2016,pp. 88–93.
[9] T.Davidson,D.Warmsley,M.Macy,andI.Weber,"Automatedhatespeechdetectionandtheprob- lemofoffensivelanguage,"Proc.Int.AAAIConf.WebSocialMedia,vol.11,no.1,pp.512–515,May 2017
[10] A.SchmidtandM.Wiegand,"Asurveyonhatespeechdetectionusingnaturallanguageprocessing," in Proc. 5th Int. Workshop Natural Lang. Process. Social Media, 2017, pp. 1–10.
[11] P. Fortuna and S. Nunes, "A survey on automatic detection of hate speech in text," ACM Comput. Surv., vol. 51, no. 4, pp. 1–30, 2018.
[12] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. R. Pardo, P. Rosso, and M. Sanguinetti, "SemEval-2019 task 5: Multilingual detection.
[13] hatespeechagainstimmigrantsandwomeninTwitter,"inProc. 13thInt. WorkshopSemanticEval. Vancouver, BC, Canada: Association for Computational Linguistics, 2019, pp. 54–63

AuthorBiblography

	AjaySureshBhagat BirthPlace:Bhavaninagar(2002). DiplomainComputerTechnologyfromS.B.PatilPolytechnic,Indapur. Pursuing B.E. in Computer Technology from DGOI FOE. CertifiedinRoboticProcessAutomationfromProAzureTechnology,Pune.
	SnehalJagdishBhalerao BirthPlace:Bhiwandi(2001). Diploma in Computer Technology from Shivajirao S. Jondhale Polytechnic, Asangaon. PursuingB.E.inComputerTechnologyfromDGOIFOE. CertifiedinRoboticProcessAutomationfromProAzureTechnology,Pune.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue VI June 2025- Available at www.ijraset.com

	RadhikaSuryakantJagtap BirthPlace:Barshi(2004). DiplomafromGovernmentPolytechnic,Pune. PursuingB.E.inComputerTechnologyfromDGOIFOE. CertifiedinRoboticProcessAutomationfromProAzureTechnology,Pune.
	SiddhiVijayMore BirthPlace:Bhiwandi,Thane(2002). DiplomainInformationTechnologyfromK.K.WaghPolytechnic,Nashik. Pursuing B.E. in Computer Technology from DGOI FOE. CompletedinternshipinRPAatProAzure Company. CertificationsinJava,AdvancedJava,andCyberSecurity.
	Dr.Dinesh Bhagwan Hanchate Birth- Solapur (MH- India), B.E. Comp. (Walchand College of Engg., Sangli, MH, (India)), M. Tech. Comp.(Dr. Babasaheb Ambedkar Technological University, Lonere, MH (India)). Ph.D. Comp. (SGGSIET, Nanded and SRTMU, Nanded, MH (India)). Former HOD of Comp. and IT. Head PG section, PG teacher. Did STTP, QIP programs sponsored by IIT, Kanpur, AICTE, ISTE, SPPU andUG. Positions: PhD Guide, Dean Industry Institute Interaction Cell.R & D coordinator, Head Media Cell, Editor ArtMagazine, Students Development Officer, System Officer, Pune Division Head, Career Katta, GoM DTE,Author, Poet, Currently Professor in Dattakala Group of Institutions, Daund. Interest in ML, Software Engineering, AI, IR, Math Modelling, Usability Engg., Optimization, Soft Computing.. Email:dineshbhanchate@gmail.com
	SwatiMahadev Atole Birthplace: Korti(Karmala) Education: BE (Computer science and Engineering) fromKarmaveerBhauraoPatilCollege Of Engineering, Satara (2013) Currently pursuing M.E. in Computer Technology from DGOI FOE. Currently working as Assistant Professor(Computer & IT) dept of Dattakala Group of Institutions Faculty of Engineering, Bhigwan. Email:-swati.atole03@gmail.com
	Dr. Sachin Sukhadeo Bere Currently working as Associate Professor inDattakala Group of InstitutionsFacultyofEngineeringBhigwan. CompletedM.Tech (CSE)withFirstclass&Distinction from JNTU Hyderabad affiliated college. CompletedPh.D. in Computer Science and Engineering, from theShri JagdishprasadJhabarmal Tibrewala University, Rajasthan India .14 years of Teaching experience and 7 years of Research experience.Maharashtra. Currently working as Associate Professor(Computer & IT) dept of Dattakala Group of Institutions Faculty of Engineering, Bhigwan. Email:-sachinbere@gmail.com



10.22214/IJRASET

INDEX COPERNICUS 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)