iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** V **Month of publication:** May 2023

DOI: https://doi.org/10.22214/ijraset.2023.52206

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 11 Issue V May 2023- Available at www.ijraset.com

Social Media based Hate Speech Detection using Machine Learning

Dr. Nisha Auti[1], Sumit Ranaware[2] Shreeraj Ghadge[3], Rajdatta Jadhav[4], Prajwal Jagtap[5]
Department of Computer Engineering, JSPM Narhe Technical Campus, Pune, Maharashtra, India

Abstract: Hate speech is a crime that has been increasing in recent years, not only in person but also online. There are several causes for this.
There is tremendous growth in social media that promotes full freedom of expression through anonymity features. Freedom of expression is a human right, but hate speech directed at individuals or groups on the basis of race, caste, religion, ethnicity or nationality, gender, disability, gender identity, etc. is a violation of that sovereignty.
Freedom of expression is a human right, but hate speech directed at individuals or groups on the basis of race, caste, religion, ethnicity or nationality, gender, disability, gender identity, etc. is a violation of that sovereignty. It promotes violence and hate crimes, creates social imbalances, and undermines peace, trust and human rights. Revealing hate speech in social media discourse is a very important but complex task.
On the one hand, the anonymity provided by the Internet, especially social networks, makes people more likely to engage in hostile behaviour. On the other hand, the desire to express one's thoughts on the Internet has increased, leading to the spread of hate speech.
Governments and social media platforms can benefit from detection and prevention technologies, as this kind of bigoted language can wreak havoc on society. We help resolve this dilemma by providing a systematic overview of research on this topic in this survey.
This project aims to accurately predict various forms by addressing different categories of hate individually and examining a set of text mining functions. Hate speech detection
Keywords: Hate Speech, Machine Learning, Social Media, Social Network, Multi-Class Hate Speech, Natural Language Processing, Hate Speech Classification, Social Media Microblogs, Multi-Class Hate Speech Dataset, Twitter Hate Speech, Text Mining, Features Exploration

I. INTRODUCTION

Hate speech is a crime that has been on the rise in recent years, not just in face-to-face contacts but also online. Social media is exploding in popularity, and its anonymity aspect fully fosters freedom of expression.
Hate speech directed at an individual or group based on race, caste, religion, ethnic or national origin, sex, handicap, gender identity, or other factors is an abuse of this sovereignty.
It actively promotes violence or hate crimes and disrupts society by jeopardizing peace, credibility, and human rights, among other things. Detecting hate speech in social media discourse is crucial, but it's a difficult undertaking.
This study aims to address the quality of datasets, which is a major concern raised by many of the problems that have been brought to light.
This paper also addresses the second issue, which is that the best characteristics for hate speech identification must be investigated and determined before developing a suitable classifier. For this reason, datasets tend to fall into one of these categories.

The work is divided into two parts: Hate speech tweets are categorized into five types.

1) Toxic
2) Obscene,
3) Threat,
4) Insult,
5) Identity hate.
6) Hate Tweet is classified into one of these types or as Non-hate tweets.

II. LITRATURE SURVEY

1) Muhammad sabih Et.al. presented “UnCompromised Credibility: social media based MultiClass Hate Speech Classification for Text” In this paper their work was that to identify the problem which is had speech towards a person or a group because of that it promotes violence or hate crimes and create and imbalance in society. Addressing different categories of head separately this paper aims to correctly predict their forms. • Data set used: - a. Hate Based Twitte b. Hat Eval c. Waseem A d. Waseem B In this study, major challenges are identified first and the complex problem of multi-class automated hate speech classification for text is accomplished with much better results. Ten separate binary classified datasets consisting of different hate speech categories are constructed. Each dataset was annotated by experts with the strong agreement of annotators under comprehensive, clear definition and well-defined rules. Datasets were well balanced and broad. They were also supplemented with language subtleties. Compilation of such dataset was achieved as necessary requirement for filling the gap of the field. After the development of high-quality datasets, a list of effective, commonly used and recommended features extracted from related studies under the field of text mining were identified. In addition to these features our own potential features were also proposed. These features were then explored and identified with respect to their problem objective. It is found that character 2 to 4- grams, word 1 to 5- grams, dependency tuples, sentiment scores, and count of 1st, and 2nd person pronouns were very effective. There are a total of ten separate datasets compiled with binary labels each. Different features together with a different set of models are explored over each dataset. Best features are identified and ten independent models are trained. Each tweet will be passed to all ten models and therefore it may have multiple hate classes identified by each model.

2) Idris et.al published "Detecting Hate Speech on Social Media Using Deep Learning Techniques". Her work shows that hate speech is a recurring problem on social media platforms, attacking specific groups of people based on certain common characteristics. Online data is created by users so quickly that it has become a daunting task to manually moderate a user's comments, including hate speech, in order to reduce the negative impact on the platform. In our previous research, we were able to create a model that can detect hate speech with high accuracy when detecting hate speech in user comments and posts (called tweets) on her Twitter, a social media platform. rice field. two base classifiers Long-short-term memory labels were used for naive-based SVM b. classification. H. “Hate Speech” and “No Hate Speech”. David Sonnet. al. has shown in his work that this has led to the "hate speech" label, which includes forms of offensive language other than hate. B. Offensive language. The ensemble model was developed using a soft-voting combination technique with his two base classifiers, NBSVM (Naive Bayes Support Vector Machine) and his LSTM (Long Short-Term Memory). We created a data representation using Facebook's Fast Text and TF-IDF (Term Frequency Inverse Document Frequency) using character n-grams known for rare detection.

3) Raquel Fernandez et al. Published "Hate Speech Corpus Research for Detecting Hate Speech and Predicting Popularity". Her research showed that, as a result, her discussion environment online can become abusive, hateful, and toxic, especially when user anonymity is added. In order to identify, study, and ultimately contain this problem, such negative environments and the language used within them are studied under the name of hate speech. In this post, the last Focus on her two points. Consider a specific hate speech corpus, the Twitter corpus collected by Waseem and Hovy (2016). This corpus is gaining momentum as a resource for training models to detect hate speech. Manually annotated to distinguish between two types of hate speech (sexist and racist), allowing for more nuanced insight and analysis. In addition, as a Twitter corpus, we offer all sorts of analysis and research possibilities for typical characteristics of Twitter corpora, such as: B. User and Tweet metadata, user interactions, etc.

4) Eileen Kwok et al published "Localizing Hate: Detecting Tweets Against Black People". Their study found that Twitter has a sizeable Black following, but anti-Black people We've found that racist tweets are particularly damaging to the Twitter community.1 In doing so, they can provide data on the sources of hate speech against black people. Nov 2012 On January 1st, a Twitter user wrote, "So my tweet caused an 11-year old black girl to commit suicide? It has been retweeted 77 times, has 17 favourites, and the user currently has 14,959 followers. They processed this balanced training data set of 24582 tweets by removing URLs, mentions, stop words, and punctuation. lowercase; and equate alternate spellings of insults with their properly spelled equivalents. They showed that our bag-of words model was insufficient to classify anti-Black tweets accurately. Their research is becoming increasingly relevant, such as how often tweets are woven into various conversations.

5) Mohyaddin et al. published Automatic Hate Speech Detection: A Literature Review. Their study provided a comprehensive review of the different approaches to detect hate speech on social media platforms that have been deployed in recent years, along with a brief description of their analysis. Language on Twitter. Data was collected and preprocessed using the Twitter API. We then applied a nearby classification algorithm and got an accuracy of 93%. Thus, a dangerous statement can be observed as follows: Dangerous speech is offensive speech that encourages the audience to participate in acts of violence against a particular group of people. Therefore, the most common hate speech online is related to religion, race, sexual

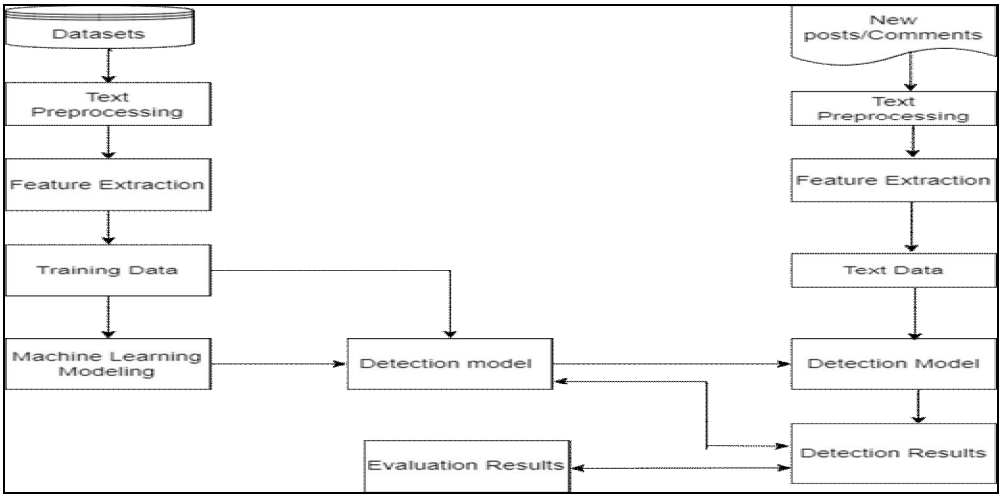
orientation, nationality, class, and gender. b) Hate speech may contain one of the pillars of dangerous speech. c) Dangerous speech often incites listeners to support or commit acts of violence against a particular group. The six most common calls to action in dangerous language are kill, riot, beat, loot, kick out, and discriminate. The Internet is inherently open and dynamic, but communities have their own rules that define language boundaries. there is. These boundaries vary by culture and are shaped by historical events and cultural norms.

6) Resmi-Regnathan et al. With the announcement of Hate Speech Detection in Conventional Languages on Social Media Using Machine Learning, the need to automate the process of classifying hate speech data arises. They also use Malayalam for hate speech. For Malayalam, we basically developed Malayalam data for the system, the system detects hate speech based on the dataset applied to the English system, and uses SVM, logistic regression, and random forest machine learning algorithms. did. This methodology describes a proposed device set up to classify speech into two specific classes, specifically "hate speech, clean speech". Suggest a perfect learning method. Specifically, as this figure shows, the research methodology consists of six major steps: dataset acquisition, pre-processing, feature extraction, model training, evaluation run, and model checking.

7) Resmi-Regnathan et al. "Detection of customary language hate speech in social media using machine learning" was published. Therefore, it becomes necessary to automate the process of classifying hate speech data. They also use Malayalam for hate speech. For Malayalam, we basically developed Malayalam data for the system, the system detects hate speech based on the dataset applied to the English system, and uses SVM, logistic regression, and random forest machine learning algorithms. did. Methodology describes a proposed deprecated device for classifying speech into two specific classes, specifically "hate speech clean speech". Suggest a perfect learning method. Specifically, as this figure shows, the research methodology consists of six major steps: dataset acquisition, pre-processing, feature extraction, model training, evaluation run, and model checking. In everyday life, with the increasing use of social media, it seems that everyone thinks they can speak and write as they please. Because of this thinking, hate speech is on the rise. Hate speech can hurt individuals and communities. However, manually identifying hate speech is very difficult. Therefore, it becomes necessary to automate the process of classifying hate speech data. To simplify the process of classifying hate speech, we used a machine learning approach to detect hate speech from speech Most machine learning algorithms require data to be formatted in a very specific way, so usually some preparation is required to get useful insights from datasets. Stemming is the process of creating morphological variants of root/base words. Remove word prefixes or suffixes. Stemming programs are commonly called stemming algorithms or stemming algorithms. The stemming algorithm reduces the words "chocolates", "chocolatey" and "choco" to the stem "chocolate" and "retrieval", "retrieved" and "retrieves" to the stem "retrieve". Stemming is an important part of pipeline processing in natural language processing. A stemmer's input is a tokenized word. Lemmatization The process of grouping different forms of a word so that they can be analysed as a single element. It usually refers to doing things right using vocabulary and morphological analysis of words. This is usually intended to remove only inflections and return the base or lexical form of the word, known as the lemma, which is the root word rather than the stem, which is the output of word stemming.

8) Aliji al-Hassan et al. "Detecting Hate Speech in Social Networks: A Multilingual Corpus Research" was published. They also distinguished various antisocial behaviors (cyberbullying, abuse, abusive language, hate speech). After Differentiation, they also published a comprehensive study on the use of text mining, examining several challenges for detecting Arabic hate speech. In this paper, they also present a table containing the different algorithms used to detect different hate categories and the accuracy of this model. The table summarizes all the work discussed, arranged according to their chronological order. English anti-social behavior, English hate speech and finally Arabic anti-social behavior. They serve as a quick reference for the important work done in auto-discovering social media. All approaches and their respective test results are clearly listed. Consolidate all terms related to hate speech and related posts. The Challenge of Hate Speech Detection in Arabic Hate speech detection is more than a simple keyword detection, it is a complex task with many challenges. Based on the review done in the previous section, they can identify some research challenges in automated detection of Arab hatred in social media. The first obstacle is that there is some research into hate speech detection. This can lead to high precision and high recall. There is growing awareness of the problem of hate spreading through social networks in the Arab region and around the world.

III. SYSTEM ARCHITECTURE



IV. MODULE EXPLANATION

A. *Dataset Preprocessing*

Dataset preprocessing refers to the steps taken to prepare raw data for analysis or modeling. It involves transforming the data into a format that can be easily understood and used by machine learning algorithms. The goal of dataset preprocessing is to clean and transform raw data so that it is more accurate, consistent, and usable.

The preprocessing steps can include:

1) Data Cleaning: Removing or fixing any errors, inconsistencies, or missing values in the dataset.
2) Data Transformation: Converting data into a standard format, normalizing or scaling data, and handling outliers or anomalies.
3) Data Integration: Merging or combining data from multiple sources.
4) Data Reduction: Reducing the size of the dataset, removing irrelevant features, or extracting important features.
5) Data Discretization: Converting continuous data into discrete data.
6) Data Sampling: Selecting a subset of data to be used for analysis or modeling.

These preprocessing steps are essential for ensuring that the data used for analysis or modeling is accurate, consistent, and reliable. By preparing the data properly, we can improve the performance of machine learning algorithms, reduce errors, and generate more accurate results.

B. *Feature Engineering*

During the exploratory data analysis, it is found that many attributes of comments outside of the words themselves may be useful in predicting whether they are toxic. The features added to the dataset are:

1) Comment length in characters
2) Percent of letters in a comment that are capitalized
3) Average length of words in a comment
4) Number of exclamation marks in a comment
5) Number of question marks in a comment

C. *Feature Extraction*

Feature extraction for hate speech classification involves identifying and extracting relevant features from text data that can be used to distinguish between hate speech and non-hate speech. Here is a general approach to feature extraction for hate speech classification:

Tokenization: The next step is to break down the text into Feature Extraction aims to reduce the number of features in a dataset by creating new features from the existing ones (and then discarding the original features). These new reduced set of features should then be able to summarize most of the information contained in the original set of features. In this way, as summarized version of the original features can be created from a combination of the original set.

D. Classification

The purpose of classification in machine learning (ML) is to develop algorithms that can automatically assign predefined categories or labels to new, unseen data based on patterns and relationships learned from labeled training data. The goal is to build a predictive model that can accurately classify or categorize data instances into distinct classes.

Classification is defined as the process of recognition, understanding, and grouping of objects and ideas into pre-set categories also known as "sub-populations." With the help of these pre-categorized training datasets, classification in machine learning programs leverage a wide range of algorithms to classify future datasets into respective and relevant categories.

E. Vectorization

In this using a term frequency – inverse document frequency (tf-idf) statistic to vectorize text. The number of features and presence of character n-grams is a parameter to tune for model optimization.

Vectorization in the context of hate speech detection refers to the process of representing text data in a numerical format that machine learning algorithms can understand and process. It involves converting textual information into numerical vectors that capture the semantic and syntactic properties of the text. Here's an explanation of vectorization for hate speech detection:

1) *Bag-of-Words (BoW) Vectorization:* The bag-of-words approach represents text by creating a vocabulary of unique words present in the dataset. Each word in the vocabulary is assigned a unique index. For each document or text sample, a vector is created where each element represents the frequency or occurrence of a specific word from the vocabulary. This representation ignores the order of words but captures their frequency.
2) *Term Frequency-Inverse Document Frequency (TF-IDF) Vectorization:* TF-IDF takes into account the importance of words in a document relative to the entire corpus. It assigns a weight to each word based on its frequency in a document (term frequency) and its inverse frequency across all documents (inverse document frequency). TF-IDF vectorization captures the significance of words in a document while reducing the influence of commonly occurring words.
3) *Word Embeddings:* Word embeddings are dense, low-dimensional vector representations that capture semantic and contextual information of words. Popular word embedding models such as Word2Vec, GloVe, or FastText learn to map words into continuous vector spaces based on their co-occurrence patterns in large text corpora. These pre-trained word embeddings can be used to convert individual words or entire text sequences into fixed-length vectors.

F. Feature Scaling

The engineered features are normalized from 0.0 to 1.0. The tf-idf features are not scaled. Feature scaling is the process of standardizing or normalizing the numerical features in a dataset to ensure that they have the same scale and range. It is a common pre-processing step in machine learning, including for hate speech detection. Here's an explanation of feature scaling for hate speech detection:

1) *Standardization:* In standardization, each feature is scaled to have zero mean and unit variance. This is achieved by subtracting the mean of the feature from each data point and then dividing by the standard deviation of the feature. Standardization works well when the data is normally distributed and has outliers.
2) *Normalization:* In normalization, each feature is scaled to have values between 0 and 1. This is achieved by subtracting the minimum value of the feature from each data point and then dividing by the range of the feature (i.e., the difference between the maximum and minimum values). Normalization is effective when the data is not normally distributed or has a nonlinear distribution.

V. MOTIVATION

Today, social networking sites involve billions of users around the world.

1) User interactions with these social sites, like Twitter, have a huge and sometimes unwanted impact on everyday life.
2) Vandals disrupt meaningful discussions in online communities by posting irrelevant comments.
3) Victims receive punishment disproportionate to the extent of the crime they clearly committed

VI. OBJECTIVE OF THE SYSTEM

1) Automatically reduce and categorize Hate Speech tweets.
2) Helps block haters from attacking victims on social media.
3) Provides insight into embarrassing events and shameful people.
4) Attempts to improve classification accuracy using machine learning and real-time Twitter data.

VII. METHODOLOGY

In the proposed systems approach, we formulate a problem classifying task to identify and mitigate the side effects of public shame on networks.

Two major contributions:

1) Classification and automatic classification of embarrassing tweets.

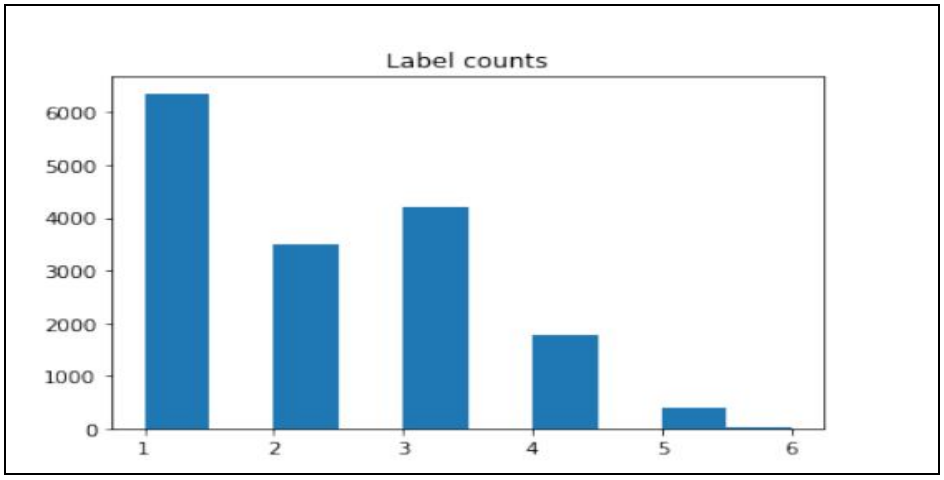
2) Develop a web application that allows Twitter users to identify Shamers.

The goal is to automatically classify tweets into 8 categories. For each category, the labelled training and test sets undergo pre-processing and feature extraction. The training set is used to train the random forest (RM). Tweets marked as negative by all classifiers are not considered shameful.

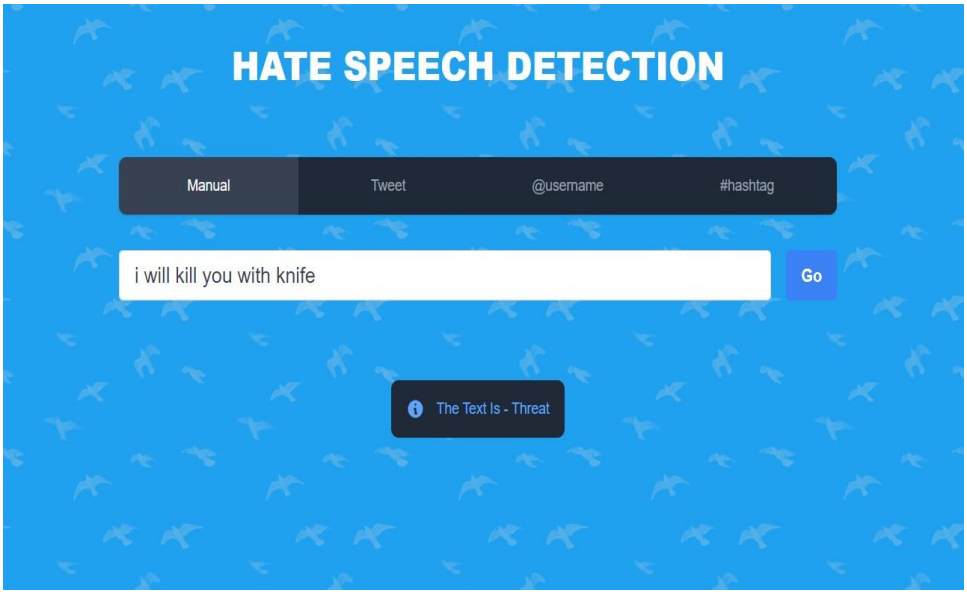
VIII.DATASET

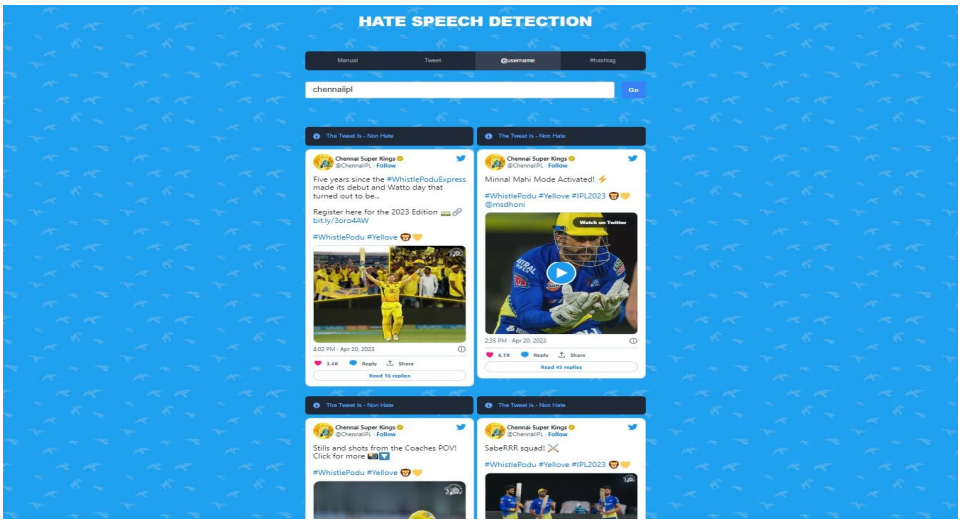
The dataset contains 159,571 comments from Wikipedia. The data consists of one input feature, the string data for the comments, and five labels for different categories of toxic comments: toxic, obscene, threat, insult, and identity hate.

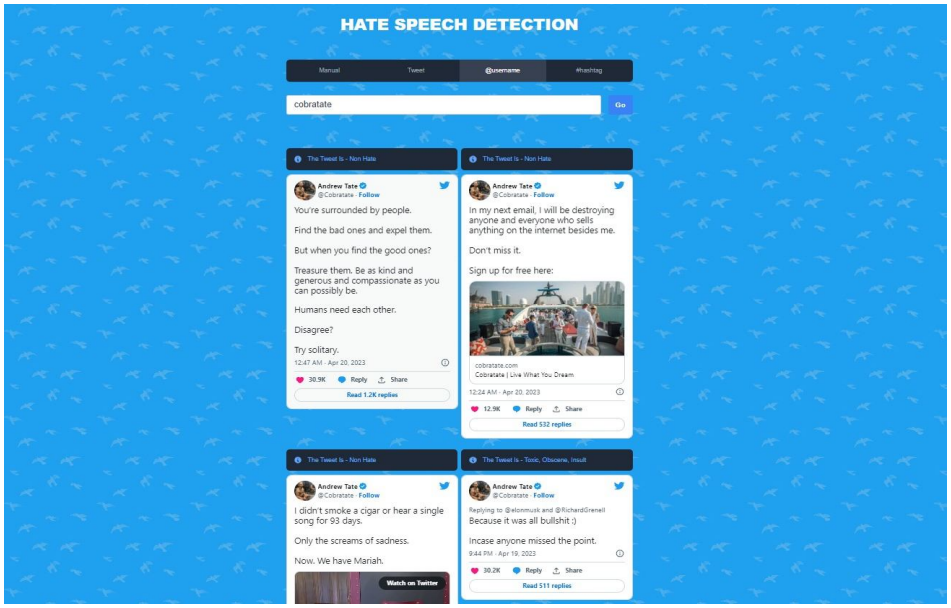
The figure on the following page contains a breakdown of how the labels are distributed throughout the dataset



IX.RESULTS







X. ADVANTAGES

1) Real time twitter classification.
2) Accurate text classification in various categories.
3) Real time fetching particular twitter account or using hashtag.

XI. LIMITATIONS

1) Only classified English language tweets.
2) Prediction is not possible for paragraphs.

XII. APPLICATIONS

1) Different Social media platforms like,
    - Facebook,
    - Instagram,
    - Share chat, etc.
2) A.I Chatbots

XIII.CONCLUSION

After identifying the primary challenges, the multi-class automated hate speech categorization for text problem is solved with significantly better results. Potential solution for countering the menace of online public shaming in Twitter by categorizing shaming comments in eight types, choosing appropriate features, and designing a set of classifiers to detect it. The propagation of hate speech on social media has been increasing significantly in recent years and it is recognized that effective counter-measures rely on automated data mining techniques. Our work made several contributions to this problem. First, we introduced a method for automatically classifying hate speech on Twitter using a machine learning that empirically improve classification accuracy.

XIV. FUTURE SCOPE

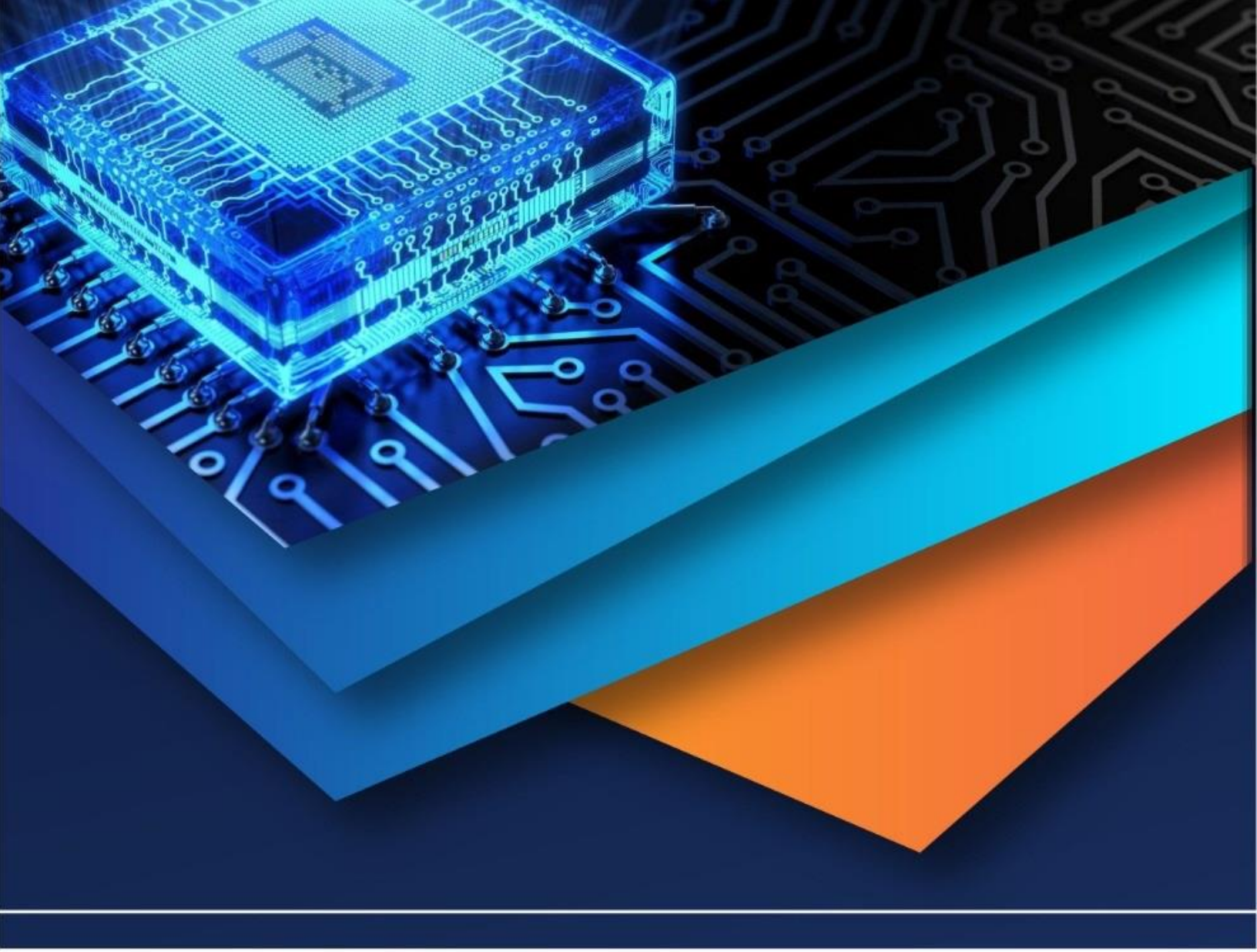
1) Furthermore, we intend to continue to explore new problems from the point of view of a social network service provider, such as Facebook or Instagram, to improve the well-being online social network of users without compromising user participation.
2) In future we can also use audio and video datasets to detect hate speech on various social media platforms.
3) Classification of newspaper text also be done in future.

XV. ACKNOWLEDGMENT

We take this occasion to thank God, almighty for blessing me with his grace and taking our Endeavor to a successful Culmination. We extend my Sincere and heartfelt thanks to my esteemed guide, Dr. Nisha Auti and the industry people, for providing me with the right guidance and advice at the crucial junctures and for showing me the right way.
Above all, I thank the Almighty, the source of all knowledge, understanding and wisdom.

REFERENCES

[1] Muhammad Sabih “Un-Compromised Credibility: Social Media Based Multi-Class Hate Speech Classification for Text” January 2021IEEE Access 9:109465-109477 DOI:10.1109/ACCESS.2021.3101977 License CC BY-NC-ND 4.0
[2] Dris, David, Ogunseye, Elizabeth Oluyemisi and Akinola, Solomon Olalekan. (2020). “Detecting Hate Speech on social media Using Deep Learning Techniques”, University of Ibadan Journal of Science and Logics in ICT Research (UIJSLICTR), Vol. 5 No. 1, pp. 22 - 38. ©U IJSLICTR Vol. 5, No. 1, June 2020.
[3] Filip Klubicka, Raquel Fernandez “Examining a hate speech corpus for hate speech detection and popularity prediction” arXiv:1805.04661v1 [cs.CL] 12 May 2018.
[4] Irene Kwok and Yuzhou Wang “Locate the Hate: Detecting Tweets against Blacks”
[5] Thomas Davidson, Dana Warmsley, Michael Macy, Ingmar Weber “Automated Hate Speech Detection and the Problem of Offensive Language” arXiv:1805.04661v1 [cs.CL] 12 May 2017
[6] Mohiyaddeen, Dr. Shifaulla Siddiqui “Automatic Hate Speech Detection: A Literature Review” e-ISSN: 2250-0758 | p-ISSN: 2394-6962 Volume-11, Issue-2 (April 2021)
[7] Resmi Reghunathan, Asha A S “Hate Speech Detection in Conventional Language on Social Media by using Machine Learning” International Journal of Engineering Research & Technology) http://www.ijert.org ISSN: 2278-0181 Vol. 11 Issue 06, June-2022
[8] Areej Al-Hass an ,Hmood Al-Dossari “Detection of hate speech in social networks: a survey on multilingual corpus” Conference Paper · February 2019 DOI: 10.5121/csit.2019.90208
[9] Sindhu Abro , Sarang Shaikh , Zafar Ali Sajid Khan , Ghulam Mujtaba “Automatic Hate Speech Detection Using Machine Learning: A Comparative Study” (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 11, No. 8, 2020.
[10] Pete Burnap and Matthew L. Williams “Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modelling for Policy and Decision Making .” 1944-2866 # 2015 The Authors. Policy & Internet published by Wiley Periodicals, Inc. on behalf of Policy Studies Organization.
[11] Sreelakshmi ka , Premjith Ba , Soman K.Pa “Detection of Hate Speech Text in Hindi English Code mixed Data” Procedia Computer Science 171 (2020) 737–744 .
[12] Mathew, Binny, et al. "Analyzing the hate and counter speech accounts on twitter." arXiv preprint arXiv:1812.02712 (2018).
[13] Gaydhani, Aditya, et al. "Detecting hate speech and offensive language on twitter using machine learning: An n-gram and tfidf based approach." arXiv preprint arXiv: 1809.08651 (2018).
[14] Watanabe, Hajime, Mondher Bouazizi, and Tomoaki Ohtsuki. "Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection." IEEE access 6 (2018): 13825-13835.
[15] Wich, Maximilian, Jan Bauer, and Georg Groh. "Impact of politically biased data on hate speech classification." Proceedings of the Fourth Workshop on Online Abuse and Harms. 2020.J. Padhye, V. Firoiu, and D. Towsley, “A stochastic model of TCP Reno congestion avoidance and control,” Univ. of Massachusetts, Amherst, MA, CMPSCI Tech. Rep. 99-02, 1999.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY