IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 12, Issue 7, July 2023

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 7, July 2023 ||

| DOI:10.15680/IJIRSET.2023.1207171 |

Toxic Speech Classification using Machine Learning Algorithm

Dattatri, Prof C S Swetha

Master of Computer Applications, Dept. of MCA, Bangalore Institute of Technology, Bengaluru, India

Assistant Professor, Department of MCA, Bangalore Institute of Technology, Bengaluru, India

ABSTRACT: There has been a tremendous increase in the spread of harmful contentpeech in today's era of internet social media platforms. They provide several enhancements. However, people with significant differences in their points of view have led to a rise in the lethality of people in internet posts and discussions.Since the pandemic's spread, companies, educational institutions, students, and the general public have all expanded their use of websites. For a long time, the increasing popularity of online platforms such as Twitter and Facebook has been a source of concern. They additionally improve interactions but additionally give users the opportunity for communicating their opinions with everyone else of the world right away.Furthermore, given the variety of these platforms'users' backgrounds, beliefs, ethnicity, and cultures, many of them opt to utilise derogatory, abusive, and threatening language. The innovations afforded by these social media platforms in this rising world beneath the shadow of anonymity have increased online toxicity immensely. This problem, unlike others, can be solved utilising Machine Learning. Phrases such as "Obscene," "Toxic," "Severe Toxic," "Threat," "Insult," and "Identity Hate" Frequently utilised and have therefore been categorised as "Toxic" speechcontent. As a consequence, it is critical to distinguish and organically eradicate poisonous speech from internet-based social media networks. This study investigates a diverse array of Machine Learning methodologies, including classic Machine Learning and ensemble approaches. We employ a corpus acquired from the internet platform Twitter to perform binary and multi-class classification and study two techniques: (a) a method that involves collecting word embeddings and then building the model; and (b) improving current models such as RF, DT, VC, LR, and KNN.

KEYWORDS: Machine Learning, Toxic Speech, KNN, Linear Regression, Decision Tree

I. INTRODUCTION

The internet likewise on social networking sites are now fundamental components of how people disseminate and receive information. The use of social media has undergone significant transformation. over time, and now roughly half of people use it to communicate their ideas and thoughts. The way that individuals communicate has changed substantially over the past ten years, partly as a outcome of social media's pervasive rise. It has made it possible for a world that is more connected and informed, Poisonous expression, however, has also made way for a phenomenon that is entirely novel. Since the availability of an open platform for the creation, discussion, and sharing of content, some quite opportunistic people have taken part in poisonous speech and generally unfavourable comments.

This design was the inspiration for our project.We want to build a high accuracy classifier on hazardous speech using the Random Forest Classifier method in order to efficiently identify the existence of toxic speech in comments and texts. Because the terms "Obscene," "Toxic," "Severe Toxic," "Threat," "Insult," and "Identity Hate" are commonly used together, they have been classified as "Toxic" speech content. Consequently, that's the case critical to recognise and naturally eliminate hazardous speech from internet-based social media networks. As a consequence, we developed a model with higher precision, recall, and accuracy scores for categorising given remarks into various categories of toxicity.

II. LITERATURE SURVEY

The separation of hate speech[1] from other instances of objectionable language is a significant difficulty for automatic hate-speech detection on social media. Lexical detection approaches have a poor accuracy since they identify all communications containing certain phrases as hate speech, and earlier work employing supervised learning failed to

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 7, July 2023 ||

| DOI:10.15680/IJIRSET.2023.1207171 |

differentiate between the two groups. We gathered tweets containing hate speech terms using a crowd-sourced hate speech lexicon. We utilise crowdsourcing to categorise a sample of these tweets into three groups: those including hate speech, those containing merely foul language, and those having neither. To discriminate between these several categories, we train a multi-class classifier. A close examination of the predictions and mistakes reveals when we can consistently separating slanderous words from different kinds objectionable words and when we cannot.

when the distinction is more complex. We discover that racist and homophobic tweets sexist messages, however, have a higher probability to be classified as racist remarks than other types of messages. labelled as offensive. Tweets that lack clear hate phrases are likewise more difficult to categorise.

This study aims[2] to investigate the many forms of hate speech that arise on social media by detecting profane phrases used in hate speech. This study also examines profane terms used throughout generations to help determine the user's profile. Five hundred (500) YouTube comments on hostile themes were collected. Hate speech is categorised into eight groups based on profanity. According to the findings, 35% of the profane terms detected in our sample are connected to sexual orientation. A comparison of keywords from 1970 to 2017 reveals that sexual orientation is associated with a significant percentage of profane words. Though the present study's results are based on just 500 comments collected via a YouTube link, they are valuable in building the list.

This study aims to investigate the many forms by identifying the use of profanity in the language of slanderous language that emerges on the web. This study also examines profane terms used throughout generations to help determine the user's profile. Five hundred (500) YouTube comments on hostile themes were collected. Hate speech is categorised into eight groups based on profanity. According to the findings, 35% of the profane terms detected in our sample are connected to sexual orientation. A comparison of keywords from 1970 to 2017 reveals that sexual orientation is associated with a significant percentage of profane words. Though the present study's results are based on just 500 comments collected via a YouTube link, they are valuable in building the list.

S Nuens Propsed that [3] From the standpoint of computer science, the scientific study of hate speech is relatively new. This survey organises and summarises the present status of the area by offering a comprehensive summary of prior attempts, covering basic algorithms, methodologies, and key characteristics. This paper also analyses The difficulty in question idea of hate speech, as defined across several platforms and settings, and offers a unified definition. This field has undeniable societal effect potential, particularly in online communities and digital media platforms. The creation and systematisation of shared resources, such as guidelines, annotated datasets in different languages, and algorithms, is a critical step in the encouragement of hate speech identification.

Toxic online material [4]has become a serious concern in today's world as the usage of the internet by individuals of all ethnicities and educational backgrounds has increased exponentially. Separating hate speech from offensive language is a significant difficulty in computerised identification of hazardous text content. In this research, we present a method for automatically categorising tweets on Twitter into three categories: hateful, offensive, and clean. We do tests on the Twitter dataset, using n-grams as features and sending their term frequency-inverse document frequency (TFIDF) values to several machine learning models. We conduct a comparative study of the models using n-grams and TFIDF normalisation techniques with various values of n. After fine-tuning the model to get the best outcomes, we reach 95.6% accuracy while testing it.

While Social Network Sites facilitate communication and information exchange, they are sometimes[5] used to start damaging campaigns against certain organisations and people. Cyberbullying, encouragement to self-harm, andassaults on women are only a few of the serious consequences of enormous internet offensives. Furthermore, assaults against groups of victims might develop into physical violence. The goal of this endeavour is to restrict and prevent the worrisome spread of such hate campaigns. Using Facebook as a model, we examine the linguistic content of comments posted on a collection of public Italian sites. To differentiate the type of hatred, we first suggest a number of hate types.Crawled suggestions are then annotations by no fewer than five different humans in accordance with the specified classification.

We create and put into operation two sorters for the Italian language that utilise different training methods, including Support Vector Machines (SVM) and the Long Short Term Memory, or LSTM, Recurrent Neural Network. These

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 7, July 2023 ||

| DOI:10.15680/IJIRSET.2023.1207171 |

sorters have their foundations on morpho-syntactical features, sentiment polarity, and word integration vocabulary. We put both of those machine learning techniques through a test to see how well they're doing in classifying sexist remarks. The outcomes demonstrate the efficacy of all two categories examined throughout the initial corpus of social media text that was by hand classified for Italian expressions of hatred.

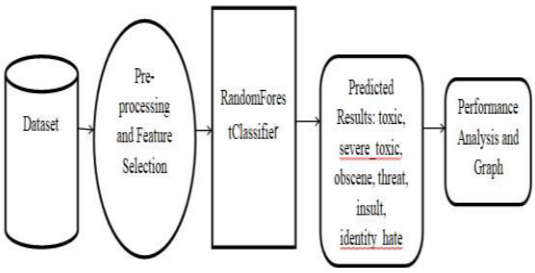


Fig. 1. Proposed Architecture

III. EXISTING MODEL

Waseem used a 16K tweet dataset to classify them as gender stereotypes, racial prejudice, or none of the above. When compared to other approaches like as character and word n-grams, he performed the best utilising the LR algorithm.

Gaydhani et al. performed logistic regression using a combination of three datasets. She observed that utilising logistic regression, a term frequency and inverse document frequency vectorizer, and a term frequency and inverse document frequency vectorizer resulted in a 95.6 percent accuracy.

Davidson examined a 24k tweet corpus that he classified into three categories: hatred, offensive, and neither. He then used NLP methods to the tweets, such as extracting the base form of the phrase, text clustering, term frequency vectorization, and Darth Vader emotive procedures, before running several supervised learning algorithms, the best of which was represented LR with L-2 regularisation. They discovered, however, that using linguistic techniques, it was difficult to distinguish between Hate and Offensive content.

The accuracy of the current system is determined on the quality of the data.

The prediction step may be sluggish in situations where there is a lot of information.

The current approach is sensitive to data size and irrelevant aspects.

To keep track of all of the training data, this approach requires a lot of recollection.

Because it saves training, it might be computationally costly, much like the previous system.

IV. PROPOSED METHODOLOGY

∟ The project's goal is to categorise a fixed of comments into one of six categories, which are:

Toxicsuggestions consist of those nasty, insulting, or unreasonable, and are likely to drive other users away from a discourse. The categorization of toxic comments is a subtask of sentiment analysis.

Severe Toxic: Symptoms that develop after ingesting a test examples of repeatedly or continuously for a large amount of a person's life are referred to as Severe Toxic.

Obscene: When you say something is obscene, you're expressing that it offends you because it contains sex or violence in an inappropriate or unsettling way.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 7, July 2023 ||

| DOI:10.15680/IJIRSET.2023.1207171 |

Insult: An insult is a deliberately impolite deed or way of speaking, as well as a lack of regard, esteem or courteous behaviour.

Threat: It refers to a threatening message delivered in a terrible manner or a threatening message delivered in a horrifying manner.Following preprocessing, We'll use the Random Forest Classifier for learning the algorithm. This idea is additionally referred to as the anticipatory approach. This method, which operates on a collection of data the fact that was manually categorised with or without instructions, may be employed for preparing the mathematical model. The Random Forest Classifier model, and these serves to categorise the feedback into multiple categories, is trained using this assigned information set.

∟ Our suggested system model is built with Random Forest Classifier, a An approach for controlled learning that gains knowledge from practise information predicts the output for test data. This method produces a clustering.

Each feature has An increasingly precise algorithm with less mistakes overfitting situations. It is the most widely used algorithm and is utilised in all ML-related real-time problems. It consists of a number of decision trees drawn from subsets of the datasets, with the average used to increase model accuracy. It also works with bigger datasets with high dimensionality.

∟ We created the system with the Flask web framework, and the user is prompted to submit a remark. The mentioned values are sent into theExample for random forest classification in AI. Considering the aforementioned input variables, our algorithm predicts the harmful class and displays the toxic level to the user.

Our suggested system model's accuracy is typically extremely good.

Its efficiency is especially noticeable with large data sets.

Provides an estimate of crucial categorization variables.

The generated forests can be stored and reused.

Unlike other models, it does not overburden itself with features.

It gives an efficient method of dealing with missing data.

Noise has less of an influence on Random Forest.

Random Forest is typically resistant to outliers and can deal with them automatically.

V. IMPLEMENTATION

Data Collection:

Data Collection procedure isinitially created in the part of the module. The initially real incrementally in actually building a framework for machine learning is gathering data. It is a vital phase that will have a knock-on effect on how successful the model is; theOur algorithm will become more accurate depending on how much and higher quality information we have.

It has a variety of techniques for gathering data, including online scraping and manual interventions. Our dataset may be found in the model folder.The dataset is from the well-known dataset repository kaggle. The dataset's URL is provided below.

Data Preparation:

Gather and arrange data for training. Clean up everything that needs it (eliminate redundant information, correct errors, address absent figures, normalise, and choose a data kind conversions, and so on).

Randomise data to remove the impacts of the sequence in which we acquired and/or otherwise prepared our data.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 7, July 2023 ||

| DOI:10.15680/IJIRSET.2023.1207171 |

Visualise data to aid in the detection of meaningful correlations between variables or class imbalances (bias alert!), or do other exploratory analysis.

Model Selection:

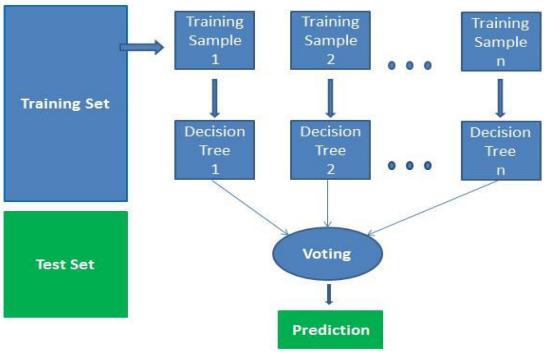
We used RandomForestClassifier machine learning algorithm.

The Random Forests Algorithm

begin to explain the procedure using everyday language. Let's say you want to take a trip and you require to go somewhere you're going to enjoy.

So how do you locate a location that you are going to enjoy? You can perform a web search, read reviews on travel blogs and portals, or ask others you know for recommendations.

Let's say you decided to ask your friends regarding their prior travel experiences, but you did so by having a conversation alongside it. Each of your acquaintances will give you a few ideas. You now need to compile a list of those suggested locations. Then you ask them to choose one location from your list of suggested locations as the most suitable established to visit.



VI.CONCLUSIONS

In this examination, we focused primarily on the remarks that are popular among teenage conversations on social media sites. The Random Forest Classifier was utilised to generate the model. We were able to categorise the comments into six distinct groups, and the Random Forest Classifier produced more accurate findings. The model not only classifies a statement as toxic or non-toxic, but also gives percentages of Obscene, Toxic, Severe Toxic, Hate, Threat, and Identity Hate.Score(toxic) 0.838055, Score(severe_toxic) 0.934874, Score(obscene) 0.909091, Score(insult) 0.883993, Score(threat) 0.795539, Score(identity_hate) 0.768448 were detected on the trained model.We demonstrated that our suggested approach is quite effective at detecting hazardous comments.

VII. FUTURE ENHANCEMENT

In the future, Convolutional neural networks that are alongside recurrent networks, which are methods of extracting features utilised for potentially dangerous expression verification, and other algorithms and frameworks will be tested and looked into. We believe that thecascading model is a good concept. Despite the fact that we justbuilt a simple version of the model, a more complex versionmight increase the model's efficiency and accuracy.

REFERENCES

[1] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hatespeech detection and the problem of offensive language," arXiv, no.
[2] P. L. Teh, C. Bin Cheng, and W. M. Chee, "Identifying and categorising profane words in hate speech," ACM International Conference Proceeding Series, pp. 65-69, 2018.
[3] P. Fortuna and S. Nunes, "A survey on automatic detection of hate speechin text," ACM Comput.Surv., vol. 51, no. 4, 2018.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 7, July 2023 ||

| DOI:10.15680/IJIRSET.2023.1207171 |

[4] A. Gaydhani, V. Doma, S. Kendre, and L. Bhagwat, "Detecting hate speech and offensive language on Twitter with machine learning: An N-gram and TFIDF-based approach," arXiv, 2018.
[5] F. Del Vigna, A. Cimino, F. Dell'Orletta, M. Petrocchi, and M. Tesconi, F. Del Vigna, A. Cimino, F. Dell'Orletta, M. Petrocchi, and M."Hate me, hate me not: Hate speech detection on Facebook," CEUR WorkshopProc., vol. 1816, 2012.
[6] V. Balakrishnan, S. Khan, T. Fernandez, and H. R. Arabnia, "*Cyberbullying detection on Twitter using big five and dark triad features," Personality and Individual Differences, vol. 141, no. 4, pp. 252-257, April 2019.
"A deep learning strategy for automatic detection of racist remarks in the Saudi Twittersphere," R. Alshalan and H. Al-Khalifa, 2020, Appl. Sci., vol. 10, no. 23, pp. 1-16.
[8] M. A. Al-Garadi, M. R. Hussain, N. Khan, G. Murtaza, H. F. Nweke, I. Ali,G. Mujtaba, H. Chiroma, H. A. Khattak, and A. Gani, ‘‘Predictingcyberbullying on social media in the big data era using machine learningalgorithms: Review of literature and open challenges,’’ IEEE Access, vol. 7,pp. 70701-70718, 2019.
[9] B. Ross, M. Rist, G. Carbonell, B. Cabrera, N. Kurowsky, and M. Wojatzki, "'Measuring the Reliability of Hate Speech Annotations: The Case of B. Cabrera,"'

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.423

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details