iJRASET
International Journal For Research in Applied Science and Engineering Technology



# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

**Volume:** 10 **Issue:** V **Month of publication:** May 2022

**DOI:** https://doi.org/10.22214/ijraset.2022.41196

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 10 Issue V May 2022- Available at www.ijraset.com*

# Negative Sentiment Analysis: Hate Speech Detection and Cyber Bullying

Prof. Jaya Jeswani[1], Vedant Bhardwaj[2], Bhavin Jain[3], Baljot Singh Kohli[4]
[1, 2, 3, 4]*Department of Information Technology, Xavier Institute of Engineering, Mumbai, India*

***Abstract: Sentiment Analysis or opinion mining in general, is the use of natural language processing, computer linguistics and related technology used to determine whether the text data is positive, neutral or negative. Hate Speech, on the other hand is abusive or threatening speech or writing that expresses prejudice, hate and / or encourages violence towards a person or a group of persons based on their race, religion, sex, or sexual orientation. With the exponential rise in the number of people making use of social media, where tons of content is posted daily, which is visibly harmless in nature, there has been a sharp rise in hate speech as well. The need and interest for identifying and detecting hate speech on such social media platforms, especially Twitter has risen significantly, as now major firms move towards development of systems which help tackle hate speech online. This project aims to cater to the need of identifying negative tweets which promote hate speech.***

## I. INTRODUCTION

When we look back in history, true social media began on May 24, 1844 as a series of electronic dots and dashes tapped out by hand on a telegraph machine. Fast forward to 1997, the first true social media platform came into existence and it was invented by Andrew Weinreich with his launching of ”SixDegrees”, and moving forward, came giants like MySpace, Facebook and Twitter.
Ever since the introduction of social media to the world, the number of users joining such sites increased rapidly, and within days there were millions of people on Facebook, Twitter, Instagram etc. With this rapid growth came a huge problem, the lack of capability to monitor what is being tweeted / posted on such platforms, Twitter in this case, which led to a massive increase in hate speech, cyber-bulling, cyber-harassment, and the use of hate on such platforms to incite violence against individuals belonging to a certain racial or linguistic group. There are variations in definition of hate speech across the world, but in the simplest way put, ”A communication that encourages violence, prejudice or discrimination against antarget group of people based on their race, ethnicity, sexuality and religious affiliation.”

Before even we begin, one has to understand the following three:

1) *Hate Speech:* Hate speech is defined by the Cambridge Dictionary as ”public speech that expresses hate or encourages violence towards a person or group based on something such as race, religion, sex, or sexual orientation”. Hate speech is ”usually thought to include communications of animosity or disparagement of an individual or a group on account of a group characteristic such as race, colour, national origin, sex, disability, religion, or sexual orientation”. A legal definition of hate speech varies from country to country.
2) *Cyber Bullying:* Cyber-bullying or Cyber-harassment is a form of bullying or harassment using electronic means. Cyber-bullying and Cyber-harassment are also known as online bullying. It has become increasingly common, especially among teenagers, as the digital sphere has expanded and technology has advanced.
3) *Offensive Language:* Language that is unambiguous in its potential to be abusive, for example language that contains racial or homophobic slurs. The use of this kind of language doesn’t imply hate speech, although there is a clear correlation.

The datasets used are labeled as:

- 0: Hate Speech
- 1: Abusive Language
- 2: Neither

The paper proposes a system which takes in a custom input from a user in the jupyter notebook, and run it to check whether the input string is hateful or not, using the models trained using the annotated data.

©IJRASET: All Rights are Reserved 

*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538*
*Volume 10 Issue V May 2022- Available at www.ijraset.com*

## II. ANALYSIS OF RELATED WORK

In Data We Trust: A Critical Analysis of Hate Speech Detection Datasets

In this particular paper, the author explains briefly about how text can be classified into different apsect based on their innate forms and also gives the idea whether the datasets is balanced or not and if not then how can we manage imbalanced datasets. The author presents paper which sets an perfect example for selecting benchmark dataset which has consistent train-test-validation split,accessible data format and has less bias data.

### *A. Characterizing and Detecting Hateful Users on Twitter*

The author of this technical paper tries to classify and detect hateful Twitter users in this research, as defined by Twitter's hateful behaviour guidelines. With a random-walk-based crawler on Twitter's retweet graph, we generate a dataset of 100, 386 individuals, each with up to 200 tweets. This paper finds users who used words from a hate speech-related lexicon and create a sub-sample of users who are at varied distances from these users. Through crowd-sourcing, these are manually tagged as hateful or not.This research used Crowdflower, a crowd-sourcing platform, to manually annotate 4, 988 users, 544 (11 percent) of whom were deemed to be nasty. We argue that this methodology addresses two flaws in previous research: it allows the researcher to strike a balance between a generic sample and a sample prejudiced toward a set of words in a vocabulary, and it provides annotators with realistic context, which is sometimes required to identify hate speech.

### *B. Hate Speech Dataset from a White Supremacy Forum*

In this paper, the author presents the first public dataset of hate speech annotated at the level of the sentence on Internet forum posts in English. Storm-front, the largest online group of white nationalists, is the source forum, which is renowned for pseudo-rational talks on race featuring different degrees of offensiveness. (Schafer, 2002) Storm-front is credited as being the first hate website. The generated dataset contains 10,000 statements that have been classified as hate speech or not. Several features of the generated dataset have also been investigated, such as the annotators' need for extra context in order to make a decision, or the distribution of the vocabulary used in the dataset.

### *C. Detecting Online Hate Speech Using Context Aware Models*

Through this paper, the author briefs us the detailed information about hate-speech detection models and explains how hatespeech is used in different forms.The author showed how important it is to use context information when detecting hate speech online. Initially this paper started by presenting a corpus of hate speech made up of entire threads of internet discussion topics. The author tried introduced two types of models, feature-based logistic regression models and neutral network models, for incorporating context information into hate speech detection performance. Furthermore, it ensemble models that combine the capabilities of both types of models get the greatest results for detecting hate speech online automatically.

### *D. Application of Sentiment Analysis Using Machine Learning Techniques*

Through this paper, the author explains that this paper deals more of applications of Sentiment analysis and also gives detailed analysis about algorithmic approaches.The author makes sure that this paper anticipates that sentiment analysis applications will continue to expand in the future, and that sentiment analytical approaches will be standardised across diverse systems and services. Future study will concentrate on three distinct characteristics that will be used to analyse diverse data sets using a combination of logistic regression and SVM methods.

### *E. A Study on Sentiment Analysis Techniques of Twitter Data*

Through this paper, the author wants to present the current methods for sentiment analysis of twitter data and provide in-depth study on these methods through thorough comparisons. Different kinds of approaches are used by the author for this detailed study. At the start of the paper, the author define about what is sentiment analysis and different classification methods used in machine learning. The author further concentrates on extensive study on document level and 4 types of sentence level sentiment analysis approaches of twitter data: supervised machine learning approaches, ensemble approaches, lexicon based approaches (unsupervised methods) and hybrid approaches. Lastly, comparisons have been done of all these approaches to provide a detailed outlook on sentiment analysis techniques.

©IJRASET: All Rights are Reserved 

*F. Twitter Sentimental Analysis*

Through this paper, the author uses sentiment analysis as a method of analysing a human's opinions and polarity of thoughts. The data gives different types of polarity indications such as positive, negative, or unbiased values. It mainly focuses on the person's tweets and hash tags to have an idea about the situations in every aspect of the existing criteria. As per the author, the goal of this paper is to see and analyse renowned people's twitter id's or hashtags and get an idea of the thinking of the people in a situation when the respective person has tweeted on it. In this paper, the system analyses the sentiments of people using Python, Twitter API and Text Blob. Lastly, in the paper various types of visualisation techniques are implemented and used for further analysis and to also get it done more accurately.

*G. HATECHECK: Functional Tests for Hate Speech De-tection Models*

Through this paper, the author detects online hate using HATECHECK, a set of functional tests for hate speech detection models instead of typical use of metrics like accuracy and F1 score due to their inability to detect weak points in the data. HATECHECK consists of 29 model functionalities wherein test cases are made to check the quality of the models through an extensive and structured annotation process. The functional tests were selected on the basis previous research data of hate speech and also through civil society stakeholders. Since usually models are examined using held-out test data, it becomes quite difficult to assess them properly and hence through this paper, the author shows HATECHECK's targeted insights make the understanding of the model limits better which in turn allows developments of stronger models in the future.

*H. Are You a Racist or Am I Seeing Things?*

Through this paper, the author investigates the impact of annotator knowledge of hate speech on classification models by comparison of results of classification obtained through extensive training on expert and amateur annotations. The author gives evaluation through his own data set using the Waseem and Hovy dataset (2016) on which the models are run. By this paper the author also reveals that amateur annotators are more likely to categorise items as hate speech than expert annotators, and also the systems trained on amateur annotations are most likely to lose out to the systems trained on expert annotations. Lastly in the paper, tables of different metrics are shown for accurate realisation of the purpose of the paper.

*I. How Will Your Tweet Be Received?*

Through this paper, the author predicts the dominating sentiment among tweet replies (first-order) to an English source tweet. The author uses a large dataset called RETWEET which contains tweets and responses with sentiment labels manually added. The author proposes a Deep Learning approach as a starting point for solving this problem. The author first predicted the overall polarity of the tweets, i.e., whether they are received positively, negatively, or neutrally.The author then creates automatic labels for replies, trained a network which predicts the reaction of the twitter audience. Through this method, the author shows that it makes an upper-bound baseline for the polarity of the overall first-reaction of the respective tweet.

## III. EVALUATION

This project makes use of three different classification algorithms, one neural net architecture and four evaluation metrics.

*A. Understanding The Data*

For this project, two data sets were used, namely:

*1)* Hate Speech and Offensive Language Dataset
*2)* Twitter Sentiment Analysis Dataset

Both of the datasets are two publicly available datasets, however both of them have different labels. For the "Hate Speech and Offensive Language" dataset, there were three labels, namely:

- 0: Hate Speech
- 1: Abusive Language
- 2: Neither

While the other dataset, "Twitter Sentiment Analysis" had only the "0" label. Both datasets had different columns and needed to be merged into one common dataset.

*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538*
*Volume 10 Issue V May 2022- Available at www.ijraset.com*

From the ”Hate Speech and Offensive Language Dataset”, we take labels 1 and 0 and copy the labels of ’0’ on to ’1’ and then rename the label ’1’ to label ’0’ and the label ’2’ to label ’1’.

The second dataset, had only one label, ’0’, and thus it didn’t need any such changes. In the end both the datasets were concatenated into one dataset, and data cleansing / cleaning was performed, which involved removal of stop words, symbols, misplaced punctuations etc.

### *B. Analysis of Algorithms*

Classification Algorithms

A Classification Algorithm is a supervised learning technique that is used to predict the category of new observations based off of the training data. In simple terms, this technique can be used to predict which category does an observation belong to, example, whether the answer will be Yes or No. In this case, simply put, whether the user input tweet will be hateful in nature or not. The classification algorithms utilised for this project are:

1) *XGBoost:* XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. In prediction problems involving unstructured data (images, text, etc.) artificial neural networks tend to outperform all other algorithms or frameworks. However, when it comes to small-to-medium structured/tabular data, decision tree-based algorithms are considered best-in-class right now. It was first presented in 2016. This algorithm generates decision trees in a sequential fashion. Weights are very important in XGBoost. All of the independent variables are given weights, which are then fed into the decision tree, which predicts results. The weight of variables incorrectly predicted by the tree is increased, and these variables are then fed into the second decision tree.
2) *Logistic Regression:* Logistic regression is a machine learning algorithm for classification. In this algorithm, the probabilities describing the possible outcomes of a single trial are modelled using a logistic function. In Statistics, the use of Logistic regression is directed toward determining the probability of a certain class or event existing, example pass or fail, win or lose, alive or dead etc. This can also be further extended to model classes of several events such as predicting if the organism is a cat or a dog. In industry, it is a widely used classification algorithm. The logistic regression model, like the Adaline and perceptron, is a statistical method for binary classification that can be generalised to multiclass classification. Scikit-learn has a highly optimised logistic regression implementation that supports multiclass classification tasks.
3) *Decision Tree Classifier:* Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. Given a data of attributes together with its classes, a decision tree produces a sequence of rules that can be used to classify the data. It is simple to understand and visualise, requires little data preparation, and can handle both numerical and categorical data. The set of data on which the model must be trained is used to generate decision trees (decision trees are a part of supervised machine learning). This training dataset will be continuously spliced into smaller data subsets. The divide principle is at the heart of tree classification. It wins where any new example fed into the tree is organised and given a class label after passing through a series of tests.

For the following project, after a trial and error while processing the original 70,000 tweets dataset, the final dataset size was set at 15,000 tweets, after revaluating the model accuracy for different data set sizes. The difference was clearly evident from the results themselves, as there was a significant difference in outputs, with the 15,000 tweets delivering the best possible scores. 7500 tweets belonged to each of the two datasets.

TABLE I
Initial Results

| Algorithm | LABEL | PRECISION | RECALL | ACCURACY | F1-SCORE |
|---|---|---|---|---|---|
| XGBoost | 0 | 0.84 | 0.90 | 0.84 | 0.88 |
| | 1 | 0.9 | 0.80 | | 0.76 |
| Logistic Regression | 0 | 0.80 | 0.85 | 0.85 | 0.90 |
| | 1 | 0.78 | 0.81 | | 0.81 |
| Decision Tree Classifier | 0 | 0.83 | 0.79 | 0.88 | 0.80 |
| | 1 | 0.81 | 0.75 | | 0.85 |

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

For the following project, Recurrent Neural Networks architecture.

Recurrent Neural Network(RNN)

Neural networks, also known as artificial neural networks (ANNs) or simulated neural networks (SNNs), are a subset of machine learning and are at the heart of deep learning algorithms. Their name and structure are inspired by the human brain, mimicking the way that biological neurons signal to one another.

Recurrent Neural Network (RNN) is a type artificial neural network which uses sequential data or time series data. These deep learning algorithms are commonly used for language translation, natural language processing (NLP), speech recognition, image captioning and it has been incorporated into popular applications such as Siri, Google Translate etc.

They take information from prior inputs to influence the current input and output. While traditional deep neural networks assume that inputs and outputs are independent of each other, the output of recurrent neural networks depend on the prior elements within the sequence.

Evaluation Metrics

*a) Precision*

In a dataset, when the classes are imbalanced, accuracy is not a reliable metric for measuring our performance. Precision is the fraction of relevant instances among the retrieved instances, while recall (also known as sensitivity) is the fraction of relevant instances that were retrieved.

P recision = T rueP ositive/P redictedY es (1)

*b) Recall*

In information retrieval, recall is the fraction of the relevant documents that are successfully retrieved.

T rueP ositivityRate = T rueP ositive/ActualY es (2)

*c) Accuracy*

Accuracy is the proportion of true results among the total number of cases examined.

Accuracy = (T P +T N)/(T P +F P +T N +F N) (3)

Where

- TP is True Positive
- TN is True Negative
- FN is False Negative
- FP is False Positive

*d) F1 Score*

F1 Score is a number between 0 and 1 and it is the harmonic mean of precision and recall

F 1 = 2∗((precision∗recall)/precision+recall) (4)

F1 score sort of maintains a balance between the preci-sion and recall for your classifier.

*C. Table of Results*

*1) XGBoost Label '0' Scores*

- Precision = 0.88
- Recall = 0.98
- F1 Score = 0.93

Label '1' scores:

- Precision 0.97
- Recall 0.83
- F1 Score 0.90

Accuracy in both the cases, is 0.91

*2) Logistic Regression Label '0' Scores:*

- Precision = 0.90
- Recall = 0.97
- F1 Score = 0.93

Label '1' scores:

- Precision 0.96
- Recall 0.86
- F1 Score 0.91

Accuracy in both the cases, was 0.92

*3) Decision Tree Classifier Label '0' Scores:*

- Precision = 0.92
- Recall = 0.94
- F1 Score = 0.93

Label '1' scores:

- Precision 0.92
- Recall 0.89
- F1 Score 0.90

Accuracy in both the cases, was 0.92

TABLE II
FINAL RESULTS

| Algorithm | LABEL | PRECISION | RECALL | ACCURACY | F1-SCORE |
|---|---|---|---|---|---|
| XGBoost | 0 | 0.88 | 0.98 | 0.91 | 0.93 |
| | 1 | 0.97 | 0.83 | | 0.90 |
| Logistic Regression | 0 | 0.90 | 0.97 | 0.92 | 0.93 |
| | 1 | 0.96 | 0.86 | | 0.91 |
| Decision Tree Classifier | 0 | 0.82 | 0.94 | 0.92 | 0.93 |
| | 1 | 0.92 | 0.89 | | 0.90 |

*D. Sample Test Cases*

Sample test cases are tested with the algorithm to check their respective nature on the context of the output produced which is a prediction score i.e. if its more than 0.5 then its of a hateful nature and and if its less then 0.5 then otherwise.

*1) Test Case 1: Input of a Negative Tweet*

A sample tweet of the text - "I hate you and want you dead you filth" is tested with the algorithm. The statement is of a negative nature and is then tested with the algorithm to check the produced classified output. Upon testing a prediction score of 0.6719118 is produced which conveys that this tweet if of a hateful and negative nature.

*2) Test Case 2: Input of a Positive Tweet*

Another sample tweet of the text - "I like middle-eastern food and appreciate the culture as well" is tested with the algorithm. The statement is of a fairly positive nature and is then tested with the algorithm to check the produced classified output. Upon testing a prediction score of 0.1969788 is produced which conveys that this tweet if of a positive and fairly non-hateful nature.

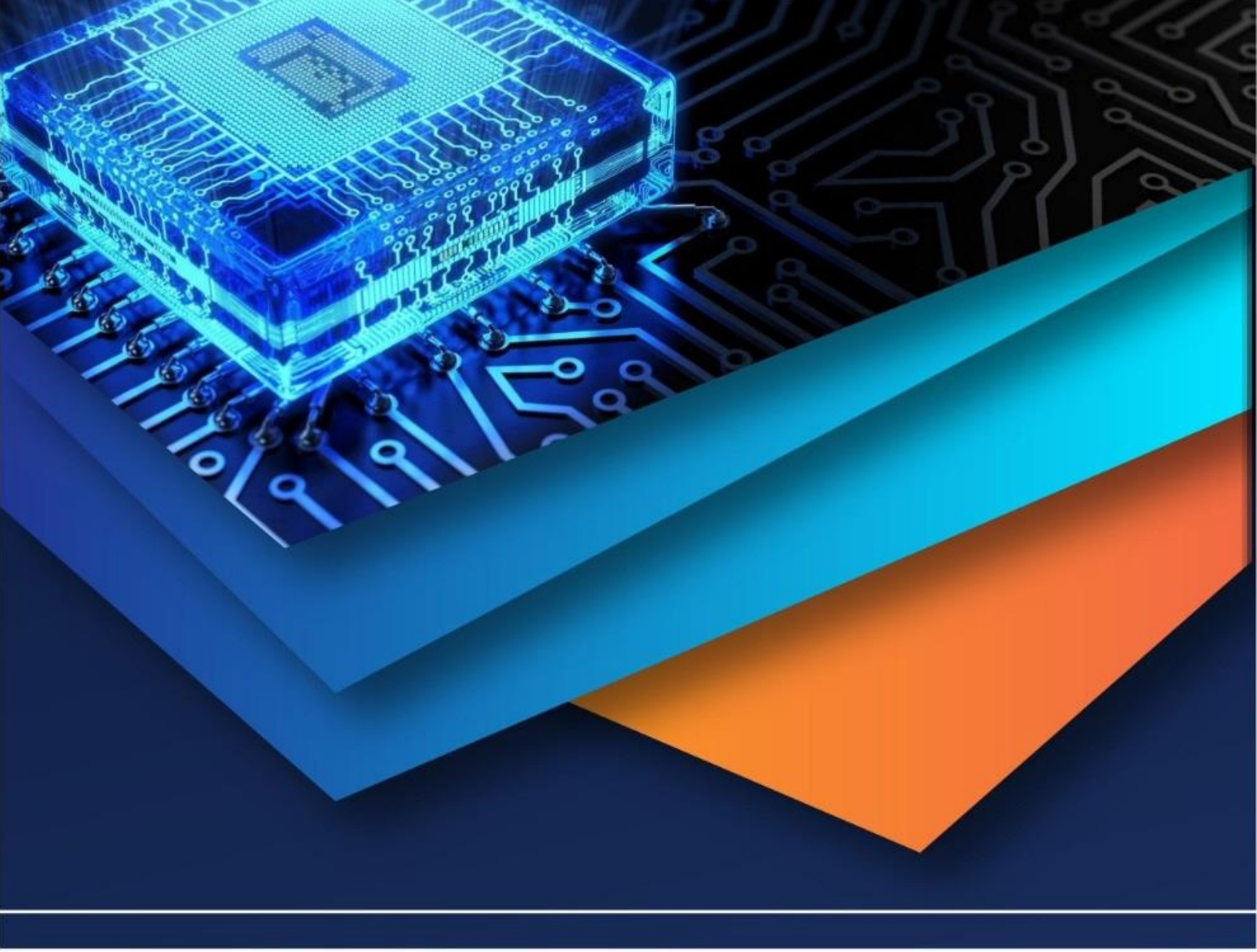
## IV. CONCLUSION

Overall out of all the social media platforms, Twitter is the world's most employed platform for discussion of issues in every context possible. It has become a center for all the digital beings to come and participate together in discussions and hence propose their ideas and ideologies towards a certain topic. So, Twitter is basically a very large collection of data which can be used for sentiment analysis which in turn becomes an inspiration for this paper. It predicts whether a tweet is of a negative sentiment or not for research as well as easy of use purposes. The algorithm only tests tweets in the English language, but future developments and research could expand it to multiple languages as well. Also, the label imbalance between hate and no hate is unbalanced so a need comes for a proper sophisticated manual labelling system which takes learning into consideration which in turn would benefit all the efforts put into labelling. Sarcasm handling can also improved upon further developments. This paper will in turn prove to be a benchmark for understanding the sentiment and real nature behind a certain tweet which would certainly benefit the community by improving their reasoning towards the respective topic or issue. Hence one can find data related to different topics on Twitter such as white supremacy, Hinduphobia, racism, blasphemy, hate speech, etc. and hence use this data to check whether the respective tweet is of a hateful and negative nature or not using the algorithm.

## REFERENCES

[1] Paul Rottger, Bertram Vidgen, Dong Nguyen, Zeerak Waseem,Helen Margetts, Janet B. Pierrehumbert, Uni-versity of Oxford, The Alan Turing Institute, Utrecht University, University of Sheffield: HATECHECK: Functional Tests for Hate Speech Detection Models, May 2021.
[2] Soroosh Tayebi Arasteh, Mehrpad Monajem, Vin-cent Christlein, Philipp Heinrich, Anguelos Nicolaou, Hamidreza Naderi Boldaji, Mahshad Lotfinia and Ste-fan Evert, Friedrich-Alexander-Universitat Erlangen-Nurnberg, Germany, Harvard Medical School, United States, Sharif University of Technology, Iran: How will your Tweets be Recieved? April 2021.
[3] Abdullah Alsaeedi, Mohammad Zubair Khan, A Study on Sentiment Analysis Techniques of Twitter Data - International Journal of Advanced Computer Science and Applications (IJACSA) Volume No. 2, November 2019.
[4] Ona de Gibert, Naiara Perez, Aitor Garc´ıa-Pablos, Montse Cuadros, Hate Speech Dataset from a White Supremacy Forum HSLT Group at Vicomtech, Donos-tia/San Sebasti´an, Spain, September 2018
[5] Pedro H. Calais, Yuri A. Santos, Virg´ılio A. F. Almeida, Wagner Meira Jr., Universidade Federal de Minas Gerais Belo Horizonte, Minas Gerais, Brazil, Characterizing and Detecting Hateful Users on Twitter - Manoel Horta Ribeiro, March 2018
[6] Lei Gao, Ruihong Huang, Texas AM Univer- sity, Detecting Online Hate Speech Using Context Aware Models, May 2018.
[7] Kosisochukwu Judith Madukwe, Xiaoying Gao, Bing Xue, School of Engineering and Computer Science, Victoria University of Wellington, In Data We Trust: A critical analysis of hatespeech detection datasets, November 2020.
[8] Zeerak Waseem, University of Copenhagen, Copen-hagen, Denmark, Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter, 2016.
[9] Shobana G, Vigneshwara B, Maniraj Sai A, Twitter Sentiment Analysis, International Journal of Recent Technology and Engineering (IJRTE), November, 2018.
[10] Anvar Shathik J and Krishna Prasad K,International Journal of Applied Engineering and Management Letters (IJAEML), A Literature Review on Application of Sen-timent Analysis Using Machine Learning Techniques, 2020.

©IJRASET: All Rights are Reserved 



10.22214/IJRASET

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)