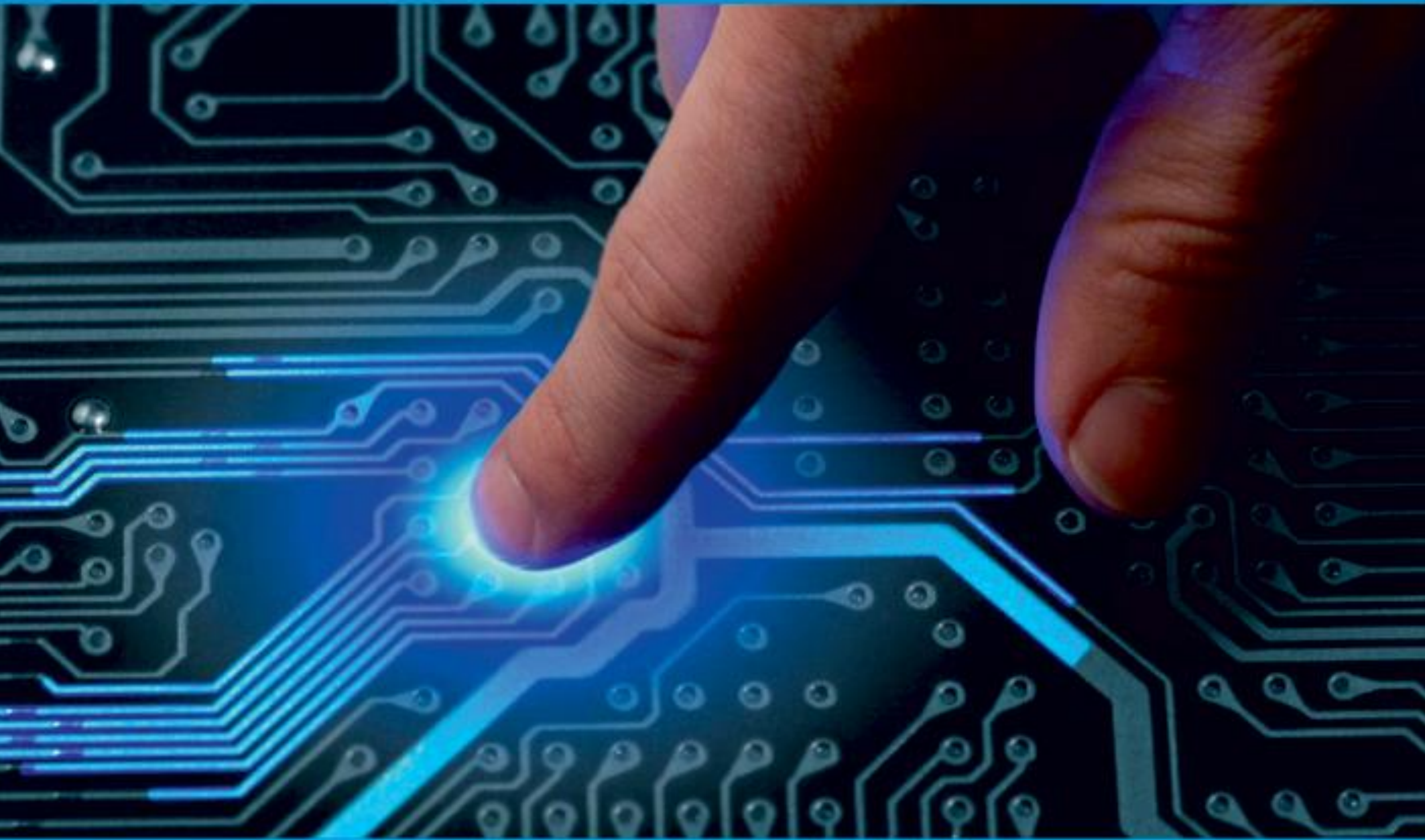
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 4, April 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

Hate Speech Detection for Twitter Data Using Decision Trees

Mr.N. Srinivas [1], P.Sathwika[2], R.Tharun Goud[3], U.Rishika[4]

Associate Professor, Department of Computer Science and Technology, Vignana Bharathi Institute of Technology, Hyderabad, India [1]

Student, Department of Computer Science and Technology, Vignana Bharathi Institute of Technology, Hyderabad, India [2 - 4]

ABSTRACT: This abstract provides a comprehensive overview of the work on Hate speech detection for twitter data using decision tree. In today’s context, People now have access to a variety of forums to freely express their ideas and opinions on a wide range of subjects thanks to the development of the internet and social media. But when this right to free speech is misused to inspire hatred towards specific people or groups of people because of their gender, colour, or other characteristics, it raises issues. Therefore, recent studies have automatically detected hate speech posts on diverse datasets using a variety of feature engineering techniques and machine learning algorithms in an effort to address this developing problem on social networking platforms. Researchers looking for and applying solutions to the hate speech problem have been captivated by the developments in machine learning. Right now, we use text data and the decision tree algorithm technique to identify hate speech.

KEYWORDS: Decision tree, twitter data, machine learning algorithms, networking platforms.

I. INTRODUCTION

Hate speech is defined as any speech that disparages or attacks a group of individuals because of their sexual orientation, gender identity, race, religion, ethnicity, or nationality. Hate speech is widely used to spread prejudice and hatred. Additionally, it can be used to threaten and intimidate others. Classifying emotions is frequently necessary in order to identify hate speech. Therefore, training on data that is generally used to classify sentiment can yield a model that can identify hate speech from a given text sample. Therefore, in this study, we used data from Twitter to help with the challenge of creating a hate speech recognition model. The amount of hate content on social media keeps growing. Facebook, Twitter, and Google have attempted a number of strategies to counteract this nasty content, but the majority of them run the danger of impairing free speech. Conversely, hate speech provides a powerful means of countering hate speech online without compromising one's right to free speech.

Therefore, in order to combat hateful content, these platforms may also choose to encourage hate speech. However, effective propagation of this kind of hate speech necessitates a thorough comprehension of its online dynamics. This knowledge is severely limited by the dearth of well-preserved data. A methodology called "Hate Speech Detection" is used to find and follow harmful and hostile content on the Internet. A large number of people post hurtful and abusive remarks about other people on social media.

Hate speech identification has therefore grown in importance as a problem-solving technique in the modern online environment. Our perspective of the world has been profoundly altered by widespread Internet access.

A result of the World Wide Web, social media (SM) can take many different forms, including social networks, online news services, forums, dating apps, and online gaming platforms.

Different social networks are used for different things: sharing photographs on Instagram, sharing videos on YouTube, holding meetings on Tinder, exchanging ideas on Twitter or Facebook, making professional connections on LinkedIn, etc.

One thing unites them all: they are all interested in fostering connections. Because of social media's immense power, 3.02 billion people will be using it every month by 2021. This amounts to almost one-third of the global populace. Among the many social networks available today, Twitter is one of the most widely used and is a vital source of information for researchers. Twitter is a popular public network of real-time microblogs where 500 million tweets are sent every day and news frequently appears before official media.

Any communication that disparages one or more people for belonging to a group—which is typically defined by traits like race, ethnicity, sexual orientation, gender identity, disability, religion, political affiliation, or opinion—is considered hate speech. Guidelines for distinguishing hate speech from free speech are outlined in the UN's Rabat Plan of Action. Three categories of expression are proposed to be distinguished: "Speech which constitutes a crime; or administrative sanction; civil or administrative sanction, but which nevertheless draws attention to tolerance, civility, and respect for the rights of others."

This happens when offenders choose their victims according to whether or not they belong to a group that is essentially characterised by the traits listed above. It turns out that single, highly publicised incidents—such as terrorist attacks, unchecked immigration, protests, riots, etc.—have an impact on hate crimes.

II. PROPOSED SYSTEM

The problem statement for decision tree-based hate speech detection is to create a model that can reliably determine whether a given text contains hate speech given a set of tweets with various features and points of view. Based on a set of criteria, a prediction model that can identify text as hate speech or non-hate speech is created using a decision tree algorithm. taken out of the text. These features could include any number of lexical, syntactic, or semantic cues that point to hate speech, like the target group's identity, the use of profanity, or the presence of specific terms. The model's architecture places a strong emphasis on precision, recall, and F1 score in order to ensure that hate speech on social media and other online platforms may be accurately identified. It's critical to identify hate speech for a number of reasons. Hate speech, in the first place, exacerbates prejudice and discrimination, which is detrimental to both social cohesiveness and personal wellbeing. Second, hate speech seriously jeopardises public safety by encouraging violence and other negative behaviours. Third, hate speech restricts people's freedom of expression by inciting fear and intimidation. This undercuts the values of democracy and freedom of speech.

Machine learning algorithms can be trained to identify patterns and traits associated with hate speech, including the use of racial slurs, abusive language, and threats of violence. Machine learning can be used to monitor a vast amount of content in order to automatically detect hate speech content and identify possible instances of hate speech more quickly and effectively than manual methods. This lessens the spread of damaging content and shields people and communities from hate speech, which can have negative effects on both people and society at large. The primary goal of hate speech detection is to locate and report potentially objectionable or harmful content in order to stop it from spreading and causing harm to the people or groups that the speech is intended to target.

In this project, we begin by outlining the current problems raised by the debate topic, which is the freedom to express oneself online and the improper use of social media sites like Twitter. These inquiries are included into the driving force. We perform extensive and in-depth research by consulting previous studies in this area, and we suggest a decision tree method to address the issue. We also locate holes in the current body of work and devise solutions for them.

The suggested system aims to achieve the following main goals, in summary:

- Improving the accuracy of identifying hate speech is the main objective.
- Distilling normal speech from hate speech in the provided text.

III. METHODOLOGY

When creating tree structures, a decision tree subject is a representation technique in which every node denotes a test on an attribute value and every branch denotes the test result. The leaves are speaking to the class. The selection tree assessed against the training dataset we utilised for the project is depicted in the figure. The relationships discovered in the training dataset are displayed. This process is speedy unless the planning is extremely beneficial.

Regarding the probability distribution for these specific data, he makes no guesses. We refer to the process of creating a shaft as enticement. Constructing decision trees: The decision tree algorithm creates a tree with leaves that are as uniform as is reasonable to expect because it is a top-down greedy algorithm.

The project mainly comprises three consecutive tasks to be performed. These four tasks are as follows:

a Set up the development environment.
b Understand the data.
c Import the required libraries.
d Process the data.
e Splitting the data.

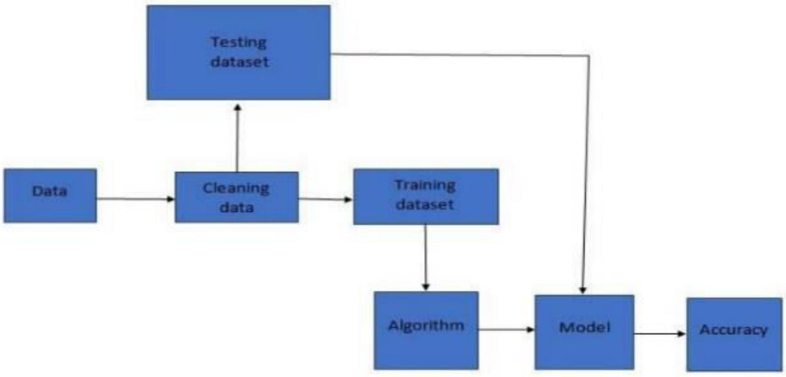


Fig – 1: Hate Speech Detection

System Architecture

- A structured collection of data is called a **dataset**. It might be as basic as a spreadsheet with rows and columns, or it can be more intricate and contain a variety of data kinds.
- We are running the algorithm on the Twitter data collection.
- Cleaning Data: Finding and fixing mistakes or inconsistencies in a dataset is known as data cleaning.
- We employ data preprocessing to identify data that is clear and prepared even though we might not find it that way.
- **Testing dataset:** This is a distinct subset of a dataset that isn't utilised by a machine learning model during training. Thirty percent of the original dataset is used in the testing.

3.1 Set up the development environment

Setting up a development environment is the initial step in developing a Python hate speech detection software. It is necessary to have Jupyter notebook software installed on your computer in order to create a hate speech detection project. If not, you can utilise Google Collab for project creation at https://colab.research.google.com/.

3.2 Understand the data

It is crucial to highlight that, given the nature of the work, this dataset contains texts that may be construed as offensive in general or as racist, sexist, or homophobic. The Hate Speech Detection dataset consists of seven columns: index, count, offensive language, hate speech, none, class, and tweet.
Index: There is an index number in this column.
Count: indicates how many people each coded a tweet.
HateSpeech: This column shows the total number of users who have reported a tweet as hateful.
Offensive Language: This indicates how many people felt the tweet was offensive.
neither: This represents the proportion of users who gave the tweet a neither rating.
Offensive nor non-offensive class: Its class designation is 0. For most people, this means hate speech; 1 means offensive language; and 2 means neither.
tweet: The tweet's text is shown in the column's.

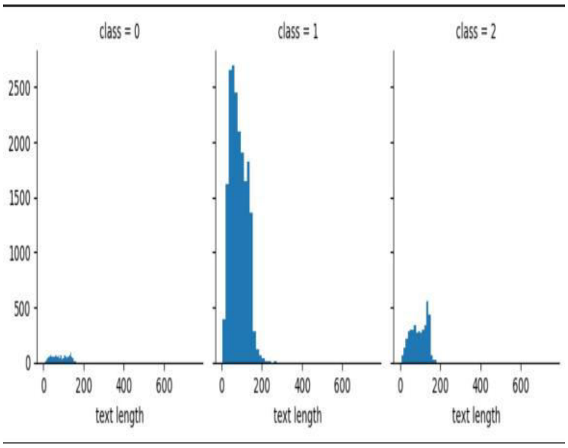


Fig – 2: data classification

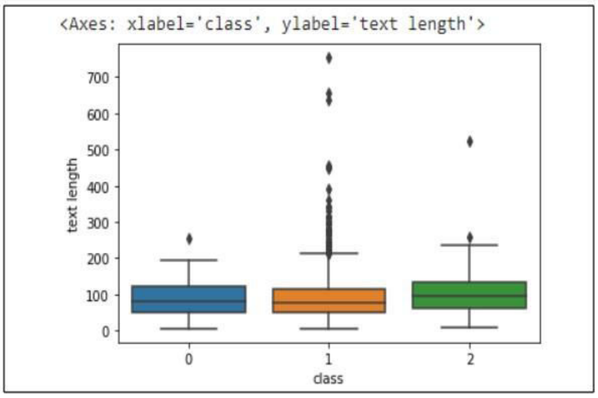


Fig – 3: Box-plot Visualization

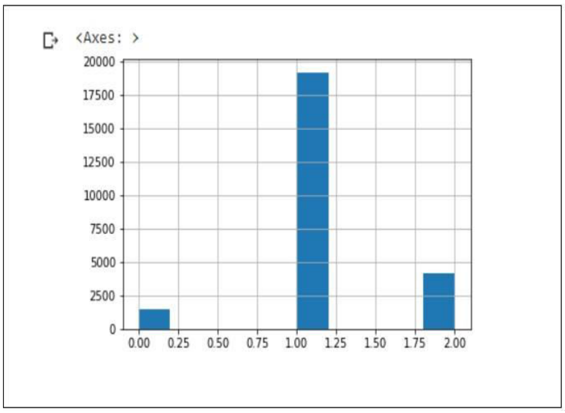


Fig – 4: Histogram Visualization

3.3 Importing the libraries .

These are the few libraries that we employed in Hate Speech Detection.
Namely,
import pandas as pd: This command imports the pandas library, which is frequently needed for analysis and data manipulation.

from sklearn.feature_extraction.text import CountVectorizer: A metric called TF-IDF (Term Frequency-Inverse Document Frequency) quantifies a term's relative relevance in a document. imports the TF-IDF Vectorizer from scikit-learn using the TfidfVectorizer.
sklearn.model_selection train_test_split: This command imports the function train_test_split, which divides the dataset into training and testing sets.
From sklearn.treeimportDecisionTreeClassifier: This imports the scikit-learn version of the Decision Tree Classifier. One kind of machine learning method that can be applied to classification is the decision tree.
Sklearn.metrics: The metrics for model evaluation, such as the accuracy score and the classification report, are imported from sklearn.metrics.
import nltk: This command imports a library for natural language processing called the Natural Language Toolkit (NLTK).

3.4 Processing the data

We prepare the raw data and compare it with an automatic learning model as part of the initial data processing. This is an important initial step in the creation of an ML model. when an ML project is created. We might not locate well-formatted, comprehensible data. Cleansing and formatting the data is necessary before doing anything with it. We employ data preparation procedures to do that. Stop words and stems are two key keywords in natural language processing. Stop words are terms (data) that are not applicable to natural language processing. We don't have to type these words. The process of morphologically varying stem words is known as stemming. For every text, we must locate more reliable and consistent roots.

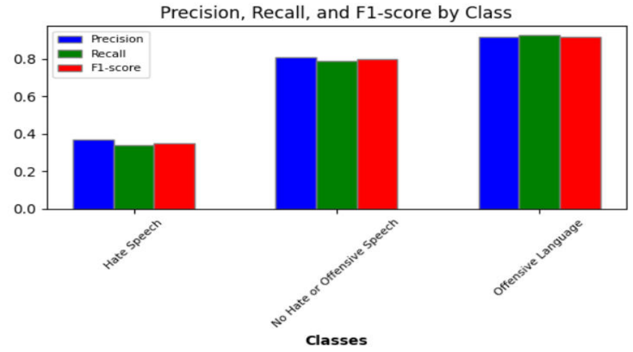
3.5 Splitting the data

To evaluate how successfully your model generalises to new, unknown data, it is standard procedure in machine learning to divide your dataset into training and testing sets. This procedure assists you in assessing the model's performance on untrained data. To detect hate speech, divide your data as follows:
Prepare and load data:
Load your dataset first, which should comprise text samples along with labels indicating whether or not they constitute hate speech.
Tokenize, vectorize, and clean the text data as part of the preprocessing step.
Divide the Data Into Labels and Features:
Sort your data according to labels (y) and features (X). The text examples are in X, and the labels that go with them are in Y.
Divide into Training and Testing Sets: Scoop your data using the scikit-learn train test split function to create training and testing sets.

This procedure offers you an idea of how well your model works on fresh, untested data and guarantees that it is not overfitting to the training set. Depending on the features and size of your dataset, change the test size and other settings. For a more thorough review, think about utilising methods like cross-validation.

IV. RESULTS

An efficient and accurate predictive model that detects the hate speech with accuracy of 87.0% .



```
Decision tree, Accuracy Score: 0.8708888617190366
```

Fig – 5: Output

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204170 |

V. CONCLUSION

Scholars have a long tradition of disregarding hate speech and derogatory words. In this work, we are using social media platforms to detect hate speech by using a decision tree algorithm. When compared to earlier models, we see that our model performs exceptionally well. Using decision trees, we developed a project to detect hate speech. Hate speech is a significant problem on social media sites like Facebook and Twitter, and the most specialised models can identify it.

REFERENCES

[1] Sumaira Ghazal, Umar S. Khan "Human Posture Classification Using Skeleton Information" 2018 International Conference on Computing, Mathematics and Engineering Technologies-ICOMET 2018.

[2] Wenchao Xu, Yuxin Pang, Yangin Yang and Yanbo Liu "Human Activity Recognition Based On Convolutional Neural Network sensors." IEEE International Conference on Consumer Electronics-AsiaIEEE, 2018.

[3] Zhenguo Shi, J. Andrew Zhang, Richard Xu, and Gengfa Fang "Human Activity Recognition Using Deep Learning Networks with Enhanced Channel State information" IEEE 2018.

[4] Akbar Dehghani. Tristan Glatard and Emad Shihab "Subject Cross Validation in Human Activity Recognition Conference 17, July 2017, Washington, DC, USA[5] Erhan BÜLBÜL Aydin ÇETIN and Ibrahim Alper DOĞRU "Human Activity Recognition Using Smartphones Gazi University Ankara, TURKEY IEEE 2018

[5] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang, "Abusive language detection in online user content," in Proceedings of the 25th international conference on world wide web. International World Wide Web Conferences Steering Committee, 2016, pp. 145– 153.

[6] W. Warner and J. Hirschberg, "Detecting hate speech on the world wide web," in Proceedings of the second workshop on language in social media. Association for computational Linguistics, 2012, pp.19-26.

[7] A. Heydarzadegan et al, "Evaluation of Machine Learning Algorithms in Artificial Intelligence", International Journal of Computer Science and Mobile Computing (IJCSMC), Vol.4 Issue.5, pg. 278-286, May 2015.

[8] A. Navlani, "Understanding Random Forests Classifiers in Python", 2019. [Online].Available at: https://www.datacamp.com/community/tutorials/randomforestsclassifi er-python. [Accessed: 10 April 2019]. Amal M. Almana, Mehmet Sabih Aksoy, Rasheed Alzahran, "A Survey on Data Mining Techniques In Customer Churn Analysis for Telecom Industry", International Journal of Engineering Research and Applications, Volume-4, Issue-5, May 2014 .

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details