

# GLOBAL ACADEMY OF TECHNOLOGY

An Autonomous Institute, Affiliated to VTU Belagavi, Approved by AICTE, Accredited by NAAC,’A’ Grade,

Ideal Homes Township, Rajarajeshwari Nagar, Bengaluru – 560 098 Tel: +91 80 28603158 (Ext - 300), Email: [hodcse@gat.ac.in](mailto:hodcse@gat.ac.in) Web: [www.gat.ac.in](http://www.gat.ac.in/)

# Department of Computer Science and Engineering

##### (Accredited by NBA 2022 - 2025)

A

Mini Project Report on

**“HATE SPEECH ANALYSIS”**

*Submitted in partial fulfilment of the requirement for the Artificial Intelligence and Machine Learning Laboratory with Mini Project (20CSEL58) of V Semester*

#### Bachelor of Engineering in

**Computer Science and Engineering**

Submitted By

#### GAGAN DEEP B S - 1GA20CS046

#### HITESH KUMAR V N- 1GA20CS056

Under the Guidance of

#### Prof. Shridhar B. Devamane Prof. Snigdha Sen

Associate Professor, Assistant Professor,

Dept. of CSE. Dept of CSE.

**2022-2023**

# GLOBAL ACADEMY OF TECHNOLOGY



#### Department of Computer Science and Engineering

**(Accredited by NBA 2022 - 2025)**

**Rajarajeshwari Nagar, Bengaluru – 560 098**

**CERTIFICATE**

This is to Certify that V Semester Mini Project in Artificial Intelligence and Machine Learning Laboratory Entitled **“HATE SPEECH ANALYSIS”** carried out by **Mr. Gagan Deep B S**, bearing **USN 1GA20CS046, Mr. Hitesh Kumar V N**, bearing **USN 1GA20CS056** is submitted in partial fulfilment for the award of the **Bachelor of Engineering** in Computer Science and Engineering during the year **2022-2023**. The Artificial Intelligence and Machine Learning Mini Project report has been approved as it satisfies the academic requirements in respect of the mini project work prescribed for the said degree.

|  |  |  |
| --- | --- | --- |
| **Prof. Shridhar B. Devamane Associate Professor,** | **Prof. Snigdha Sen**  **Assistant Professor,** | **Dr.Kumaraswamy S**  **Professor & HOD,** |
| **Dept of CSE,** | **Dept of CSE,** | **Dept of CSE,** |
| **GAT,Bengaluru** | **GAT,Bengaluru** | **GAT,Bengaluru** |

##### External Exam

Name of the Examiners Signature with date

1.

2.

# GLOBAL ACADEMY OF TECHNOLOGY



#### Department of Computer Science and Engineering

**(Accredited by NBA 2022 - 2025)**

#### Rajarajeshwari Nagar, Bengaluru – 560 098

**DECLARATION**

We, **Gagan Deep B S** , bearing USN **1GA20CS046**, **Hitesh Kumar V N**, bearing USN **1GA20CS056**, students of Fifth Semester B.E, Department of Computer Science and Engineering, Global Academy of Technology, Rajarajeshwari Nagar, Bengaluru, declare that the Mini Project entitled **“HATE SPEECH ANALYSIS”** has been carried out by us and submitted in partial fulfilment of the course requirements for the award of degree in Bachelor of Engineering in Computer Science and Engineering during the academic year **2022-2023**.

**GAGAN DEEP B S 1GA20CS046 ……………**

**HITESH KUMAR V N 1GA20CS056 ……….……**

Place: Bengaluru

Date:

**ACKNOWLEDGEMENT**

The satisfaction and euphoria that accompany the successful completion of any task would be incomplete without the mention of the people who made it possible and whose constant encouragement and guidance crowned our efforts with success.

I consider myself proud, to be part of **Global Academy of Technology** family, the institution which stood by the way in endeavors. I express my deep and sincere thanks to our Principal **Dr. N. Ranapratap Reddy** for his support.

I am grateful to **Prof. Shridhar B. Devamane, Associate Professor**, Dept of CSE who is source of inspiration and of invaluable help in channelizing my efforts in right direction.

I wish to thank my internal guide **Prof. Snigdha Sen, Assistant Professor,** Dept of CSE for guiding and correcting various documents of mine with attention and care. he has taken lot of pain to go through the document and make necessary corrections as and when needed.

We would like to thank the faculty members and supporting staff of the Department of CSE, GAT for providing all the support for completing the Project work.

Finally, we are grateful to my parents and friends for their unconditional support and help during the course of my Project work.

**GAGAN DEEP B S (1GA20CS046)**

**HITESH KUMAR V N (1GA20CS056)**

## ABSTRACT

Hate speech is a growing problem on social media platforms and online communities. Hate speech is a form of verbal or written communication that aims to degrade, humiliate, or discriminate against a particular group based on their race, ethnicity, gender, sexual orientation, religion, or other characteristics. The prevalence of hate speech in online platforms such as social media has become a growing concern in recent years due to its negative impact on individuals and communities. In recent years, machine learning algorithms have been used to automatically detect hate speech in online content. Hate speech analysis involves developing automated techniques and algorithms to identify and categorize hate speech in large volumes of online content. The process involves several steps, including data collection, preprocessing, feature extraction, and model training. Natural Language Processing (NLP) techniques such as sentiment analysis, topic modeling, and named entity recognition are used to extract meaningful features from the text data. This project explores the use of Support Vector Machines (SVM), Logistic Regression, Random Forest, and Naïve Bayes algorithms in recognizing hate speech in social media posts using TF-IDF vectorization. The performance of these algorithms was evaluated using various metrics such as accuracy, precision, recall, and F1 score. Our results show that the Random Forest algorithm achieved the highest accuracy of 90.35% in identifying hate speech using the TF-IDF vectorization method.

## TABLE OF CONTENTS

|  |  |  |  |
| --- | --- | --- | --- |
| **Chapter No.** | - | ACKNOWLEDGEMENT | **i** |
|  | - | **ABSTRACT** | **ii** |
|  | - | LIST OF CONTENTS | **iii** |
|  |  | LIST OF FIGURES | **iv** |
|  |  | LIST OF TABLES | **iv** |
| **1.** | **-** | **INTRODUCTION** | 1 |
|  | 1.1 | Introduction of Project | 1 |
| **2.** | - | **REQUIREMENT SPECIFICATION** | **2** |
|  | 2.1 | Software Requirements | 2 |
|  | 2.2 | Hardware Requirements | 2 |
| **3.** | **-** | **PROBLEM STATEMENT** | **3** |
|  | 3.1 | Project Scope | 3 |
| **4.** | **-** | **OBJECTIVE OF THE PROJECT** | **4** |
| **5** | **-** | **DATSETS** | **5** |
| **6.** | - | **METHODS AND ALGORITHMS USED** | **6** |
|  | 6.1 | Logistic Regression | 6 |
|  | 6.2 | Naïve Bayes | 6 |
|  | 6.3 | Random Forest | 7 |
|  | 6.4 | Support vector machine | 7 |
| **7.** | - | **EXPERIMENTS** | **8** |
|  | 7.1 | Data Preparation | 8 |
|  | 7.2 | Exploratory Analysis | 9 |
| **8.** | - | **EVALUATION METRICS** | **10** |
|  | 8.1 | Accuracy | 10 |
|  | 8.2 | Recall | 10 |
|  | 8.3 | Precision | 11 |
|  | 8.4 | F1-Score | 11 |
| **9.** | - | **DISCUSSION ON RESULTS** | **12** |
|  |  | **CONCLUSION** | **13** |
|  |  | **REFERENCES** | **14** |

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| **Figure No.** | **Title** | **Page No.** |
| Figure 5.1 | Original Dataset snapshot | 5 |
| Figure 7.1 | Modified Dataset | 8 |
| Figure 7.2 | Basic visualization of data using histograms | 9 |
| Figure 9.1 | Accuracy Score of Different Algorithms | 12 |

## LIST OF TABLES

|  |  |  |
| --- | --- | --- |
| **Table No.** | **Title** | **Page No.** |
| Table 9.1 | Discussion on Results | 12 |

**CHAPTER 1**

## INTRODUCTION

### 1.1 INTRODUCTION OF PROJECT

Hate speech has been defined as any form of speech or expression that seeks to attack, degrade, or marginalize an individual or a group based on their race, ethnicity, gender, sexual orientation, religion, or other characteristics. In recent years, the rise of social media platforms has made it easier for individuals to engage in hate speech, making it a significant societal issue. As a result, there has been a growing interest in developing automated techniques to identify and analyze hate speech in online platforms. Hate speech analysis involves the use of natural language processing (NLP) techniques to identify patterns of hate speech in large volumes of text data. By analyzing these patterns, researchers can develop a better understanding of the nature and extent of hate speech online, as well as identify effective strategies for combating it. Our project aims to develop an automated system for hate speech analysis using four machine learning algorithms: logistic regression, naive Bayes, support vector machine (SVM), and random forest. Additionally, we will be using TF-IDF vectorization to convert the text data into numerical features that can be used as input for the machine learning models.

The goal of this project is to develop a system that can accurately identify and classify hate speech in online platforms. We will be collecting data from various online sources and manually labeling them as either hate speech or non-hate speech to create a training dataset. We will then preprocess the data by removing stop words, stemming, and other text cleaning techniques to ensure that the data is in a suitable format for analysis.Once we have a preprocessed dataset, we will use TF-IDF vectorization to convert the text data into numerical features. We will then apply the four machine learning algorithms: logistic regression, naive Bayes, SVM, and random forest, to develop models that can accurately classify hate speech in online communication. We will evaluate the performance of each model using various metrics such as accuracy, precision, recall, and F1 score.

The outcome of this project will be an automated system that can accurately identify and classify hate speech in online platforms. This system can be used to monitor online communication and flag instances of hate speech, allowing for more effective intervention and prevention strategies. Additionally, the insights gained from this project can contribute to the development of policies and strategies aimed at reducing the prevalence of hate speech in online communication.

**CHAPTER 2**

## SYSTEM REQUIREMENTS SPECIFICATION

#### Hardware Requirements

* System : Intel i3 or above
* Hard Disk : 120 GB or above
* Monitor : 15’’ LED or above
* Input Devices : Keyboard, Mouse
* Ram : 8 GB or above

#### Software Requirements

* Operating system : Windows 10 or above.
* Coding Language : PYTHON 3.6 or above
* Tools : Anaconda

**CHAPTER 3**

## PROBLEM STATEMENT

### The problem we are addressing in this project is the prevalence of hate speech in online communication. Hate speech has a negative impact on individuals and communities, leading to harassment, discrimination, and even violence. With the rise of social media platforms and online communication, hate speech has become more widespread and difficult to monitor.

### 3.1 PROJECT SCOPE

Machine learning techniques have been around us and has been compared and used for analysis for many kinds of data science applications. The major motivation behind this research-based project was to explore the feature selection methods, data preparation and processing behind the training models in the machine learning. With first hand models and libraries, the challenge we face today is data where beside their abundance, and our cooked models, the accuracy we see during training, testing and actual validation has a higher variance. Hence this project is carried out with the motivation to explore behind the models. There are many potential developments that could improve bank note authentication in the future, making it more dependable and effective. Using biometric authentication, which may identify a person's distinctive physical attributes, such as fingerprints, retinal scans, or facial recognition, is one potential development. With the help of this technology, banks would be able to more securely and in real time confirm the identity of the cash bearer. Moreover, machine learning and artificial intelligence (AI) may be used to accelerate and more accurately authenticate bank notes as well as to spot fake currency. To identify patterns in possibly fraudulent notes and warn the user or financial institution, AI systems could be trained. Eventually, blockchain technology might be utilized to produce unchangeable transaction records and guarantee the security and traceability of all transactions. The future of bank note authentication is looking brighter and more effective with the correct security measures and technology.

**CHAPTER 4**

## OBJECTIVE OF THE PROJECT

The main objective of developing this project are:

1. Identifying hate speech: One of the primary objectives of a hate speech analysis project is to identify instances of hate speech. This involves developing algorithms or models that can automatically detect and classify hate speech.
2. Understanding the nature and scope of hate speech: Another objective of a hate speech analysis project is to gain a better understanding of the nature and scope of hate speech. This may involve analyzing the content of hate speech, as well as the demographics of the individuals or groups targeted by hate speech.
3. Developing strategies to counter hate speech: A hate speech analysis project may also aim to develop strategies to counter hate speech. This may involve developing tools or techniques to identify and report hate speech, as well as strategies to promote tolerance, understanding, and respect for diversity.
4. Evaluating the effectiveness of interventions: Another objective of a hate speech analysis project is to evaluate the effectiveness of interventions aimed at countering hate speech. This may involve analyzing the impact of anti-hate speech campaigns, educational programs, or other interventions designed to reduce the incidence of hate speech.

**CHAPTER 5**

## DATSETS

Twitter is a common source for hate speech datasets as it is a platform where users can express their opinions and emotions freely. However, using Twitter data for hate speech analysis can also be challenging because of the noise, sarcasm, and ambiguity present in the tweets. Also, it is important to obtain ethical clearance and consider issues of privacy, consent, and bias when collecting Twitter data.

If the dataset for hate speech analysis is taken from Twitter tweets, some preprocessing steps should be taken to remove irrelevant information and to ensure that the tweets are representative of the type of hate speech being targeted. These steps may include removing retweets, removing irrelevant hashtags or usernames, and filtering out tweets that are not related to the target group or topic of interest. It is also important to perform thorough labeling of the dataset to ensure that the models can accurately classify hate speech.

Overall, while Twitter can be a useful source for hate speech datasets, it is important to carefully consider the challenges and limitations of using Twitter data and to take appropriate measures to preprocess and label the data.

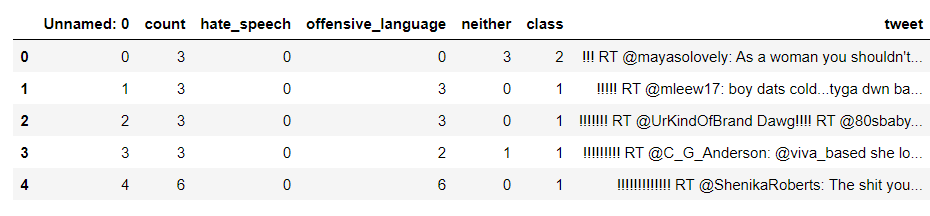


Figure 5.1: Original Dataset snapshot

**CHAPTER 6**

## METHODS AND ALGORITHMS USED

Algorithms are step-by-step procedures used to solve problems or perform tasks. In the context of machine learning and data analysis, algorithms are used to create models that can predict or make decisions based on input data.

Different machine learning algorithms can be used to build hate speech analysis models, like logistic regression, decision trees, random forests, svm, naïve bayes and neural networks.

### Logistic Regression

Logistic Regression is a supervised classification algorithm. It is a predictive analysis algorithm based on the concept of probability. It measures the relationship between the dependent variable (TenyearCHD) and the one or more independent variables (risk factors) by estimating probabilities using underlying logistic function (sigmoid function). Sigmoid function is used as a cost function to limit the hypothesis of logistic regression between 0 and 1 (squashing) i.e. 0 ≤ hθ (x) ≤ 1.

Logistic Regression relies highly on the proper presentation of data. So, to make the model more powerful, important features from the available data set are selected using Backward elimination and recursive elimination techniques.

### Naïve Bayes

Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems.

It is mainly used in text classification that includes a high-dimensional training dataset.

Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.

It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.

Some popular examples of Naïve Bayes Algorithm are spam filtration, Sentimental analysis, and classifying articles

.

### Random Forest

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that

dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

### Support vector machine

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n- dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane. SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine.

### CHAPTER 7

* 1. **Data Preparation**

## EXPERIMENTS

The dataset consists of 6 attributes in total . This dataset contains a collection of 7,000 tweets that were scraped from Twitter . The tweets cover a wide range of topics, including politics, sports, and entertainment. The purpose of this dataset is to provide a resource for sentiment analysis and machine learning tasks. The dataset contains tweets in English language only.

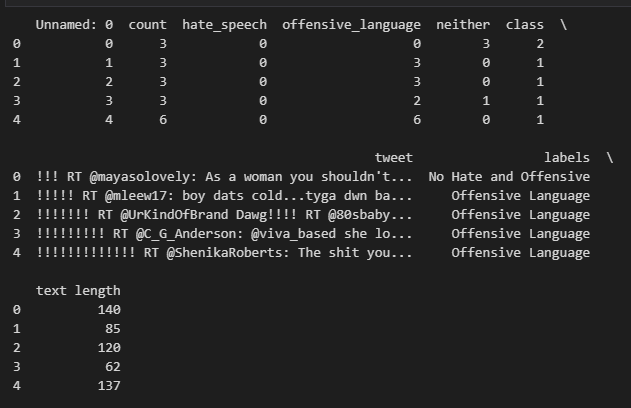


Fig 7.1: Modified Dataset

#### Exploratory Analysis

It shows that there is no single feature that has a very high correlation with our target value. Also, some of the features have a negative correlation with the target value and some have positive. The data was also visualized through graphs

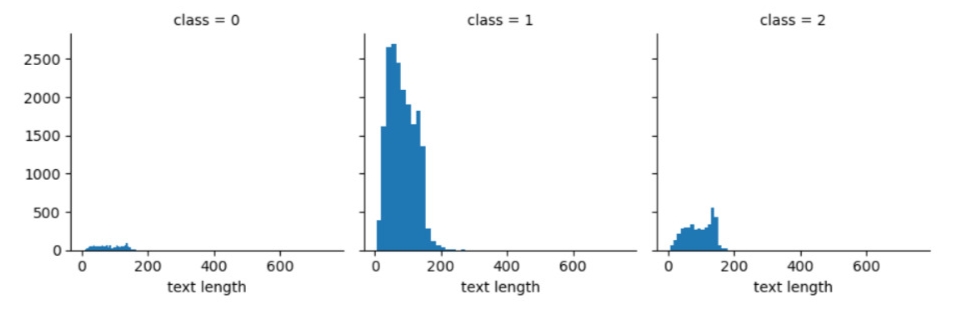


Figure 7.2: Basic visualization of data using histograms

**CHAPTER 8**

**EVALUATION METRICS**

For the evaluation of our output from our training the data, the accuracy was analysed “Confusion matrix”.

### Accuracy

The accuracy is calculated as:

Accuracy = TP+TN TP+TN+FP+FN

Where,

* + True Positive (TP) =Observation is positive, and is predicted to be positive.
  + False Negative (FN) = Observation is positive, but is predicted negative.
  + True Negative (TN) = Observation is negative, and is predicted to be negative.
  + False Positive (FP) =Observation is negative, but is predicted positive

### Recall

Recall can be defined as the ratio of the total number of correctly classified positive examples divide to the total number of positive examples. High Recall indicates the class is correctly recognized (a small number of FN). Recall is calculated as:

Recall = TP TP+FN

The obtained recall during training the data after feature selection using backward elimination was and during testing was 0.99.

### Precision

To get the value of precision we divide the total number of correctly classified positive examples by the total number of predicted positive examples. High Precision indicates an example labelled as positive is indeed positive (a small number of FP). Precision is calculated as:

Precision = TP TP+FP

The obtained precision during training the data after feature selection using backward elimination was 0.86 and during testing was 0.84.

The obtained precision during training the data after feature selection using REFCV method and during testing was 0.86.

* 1. **F1-Score**

F1 score is a commonly used metric for evaluating the performance of a classification model, particularly when the classes are imbalanced. It is the harmonic mean of precision and recall, and it provides a balanced view of the model's performance. The F1 score is calculated as follows:

F1 Score = 2 \* ((Precision \* Recall) / (Precision + Recall))

It ranges from 0 to 1, with 1 indicating perfect precision and recall, and 0 indicating poor performance. A higher F1 score indicates better performance of the model in terms of both precision and recall.

**CHAPTER 9**

## DISCUSSION ON RESULTS

|  |  |  |
| --- | --- | --- |
| **Sr. No.** | **Algorithm** | **Accuracy** |
| 1. | Logistic Regression | 89.75% |
| 2. | Naïve Bayes | 64.91% |
| 3. | Random Forest | 90.35% |
| 4. | Support Vector Machine Classifier | 89.32% |

Table 9.1: Discussion Results

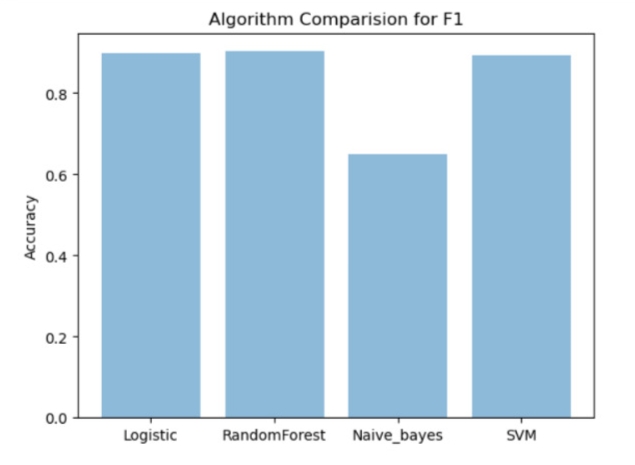


Fig 9.1 : Accuracy Score of Different Algorithms

## CONCLUSION

In our project, we have implemented four different classification algorithms from scratch namely: Logistic Regression, Random Forest Classifier, Naïve Bayes and Support Vector Machine Classifier and calculated the accuracies of each algorithm on the Hate Speech dataset.

After analyzing the accuracies of these algorithms in Hate Speech Analysis, we conclude that that Random Forest Classifier is the best algorithm with an accuracy of 90.35% and Naïve Bayes is the worst algorithm having an accuracy of 6491% for the given dataset.

Thus, TFIDF Vectorizer along with Random Forest Classifier can be used to create a model which can accurately analyse hate speech.

In future, this work can be extended by implementing some of the other classification algorithms apart from the ones already implemented in this project or by varying some of the parameters for the above implemented algorithms so as to achieve higher accuracy.

## REFERENCES

1. Priyanka Meel and Dinesh. Vishwakarma, “Fake News Rumor Information Pollution in Social Media and Web: A Contemporary Survey of State-of-the-arts Challenges and Opportunities”, Expert Systems with Applications, vol. 153, 2019.
2. Priyanka Meel, Harsh Agrawal, Mansi Agrawal and Archit. Goyal, Analysing Tweets for Text and Image Features to Detect Fake News Using Ensemble Learning, vol. 10, 2020.
3. Mansi Dhawan and M. L. Sharma, “HATE SPEECH DETECTION AND SENTIMENTANALYSIS”, International Research Journal of Engineering and Technology (IRJET), 2020.
4. Mohit Dagar, Abhishek Kajal, Pardeep Bhatia, “Twitter Sentiment Analysis using Supervised Machine Learning Techniques”, 2021 5th International Conference on Information Systems and Computer Networks (ISCON), pp.1-7, 2021.
5. Mona Khalifa A. Aljero, Nazife Dimililer, “Genetic Programming Approach to Detect Hate Speech in Social Media”, IEEE Access, vol.9, pp.115115-115125, 2021.