



Post Graduation in Data Science & Analytics

Capstone Project on
Hate Speech Detection

Made by:

Mr. Shivam Sarvajeet Mishra

2023-24

Supervised by:

Mr. Rithik Raj Vaishya

Index:

Abstract	01
1. Introduction	02
1.1 Objective	04
1.2 Motivation	05
1.3 Background	06
2. Work Description	08
3. Technical Specification	10
4. Literature Survey	12
5. Methodology	13
6. Project & Output	16
7. Future Enhancement	23
8. Conclusion	24
9. Reference	25

Abstract

In social media platforms, hate speech can be a reason of “cyber conflict” which can affect social life in both of individual-level and country-level. Hateful and antagonistic content propagated via social networks has the potential to cause harm and suffering on an individual basis and lead to social tension and disorder beyond cyber space. However, social networks cannot control all the content that users post. For this reason, there is a demand for automatic detection of hate speech.

This paper will present a background on hate speech and its related detection approaches. In addition, the recent contributions on hate speech and its related anti-social behaviour topics will be reviewed. Finally, challenges and recommendations for the hate speech detection problem will be presented.



1. Introduction

What constitutes hate speech and when does it differ from offensive language? No formal definition exists but there is a consensus that it is speech that targets disadvantaged social groups in a manner that is potentially harmful to them.

Drawing upon these definitions, we define hate speech as language that is used to express hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group. In extreme cases this may also be language that threatens or incites violence, but limiting our definition only to such cases would exclude a large proportion of hate speech.

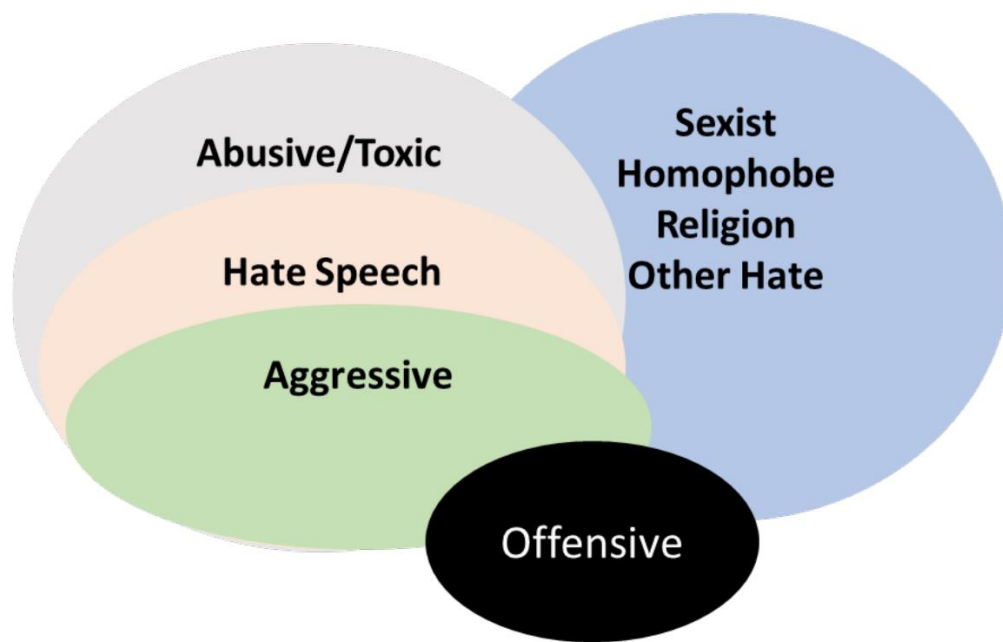
Over the last decades, people are getting more engaged with the wide spread of social networks. Microblogging applications opened up the chance for people around the globe to express and share their thoughts extensively and in a real-time manner. Such expressions afford researchers with the ability to investigate the online social emotions in different events. People now have the potential to speak freely, this allowed them to exchange all sorts of thoughts, emotions and knowledge. However, cyberspace is not always safe, it can be a reason for the dissemination of aggressive and harmful content. Hate speech is an online common form for expressing prejudice and aggression. This may convey racist, xenophobic and many forms of verbal aggression. Hate speech is typically defined as the act that disparages a person or people on the basis of a number of characteristics that may include and not limited to: race, ethnicity, sexual orientation, gender, religion and nationality

[1]. In social media platforms, there are uncontrollable number of comments and posts issued every second which make it impossible to trace or control the content of such platform. Therefore, social platforms are facing a problem in limiting these posts while balancing the freedom of speech

[2]. In addition, the diversity of people and their backgrounds, cultures and beliefs can ignite the flame of hate speech

[3]. In the other hand, each culture has its own different interpretations and characteristics of cyber-hate. So, every culture is assumed to act differently and have their own way of intervention in a manner which best suits the culture.

In many countries, including the United Kingdom, Canada, and France, there are laws prohibiting hate speech, which tends to be defined as speech that targets minority groups in a way that could promote violence or social disorder. People convicted of using hate speech can often face large fines and even imprisonment. These laws extend to the internet and social media, leading many sites to create their own provisions against hate speech. Both Facebook and Twitter have responded to criticism for not doing enough to prevent hate speech on their sites by instituting policies to prohibit the use of their platforms for attacks on people based on characteristics like race, ethnicity, gender, and sexual orientation, or threats of violence towards others.



Objective

Primary objective:

Develop a machine learning model that can accurately detect hate speech in text data. This includes defining what constitutes hate speech for your project and setting specific metrics for accuracy (e.g., precision, recall, F1 score).

Secondary objectives:

- Explore different machine learning algorithms and techniques for hate speech detection. This could involve comparing traditional classifiers like Decision tree, Logistic Regression, Random Forest and Support Vector Machines to deep learning models like Recurrent Neural Networks.
- Analyse the impact of data pre-processing techniques on model performance. This could involve experimenting with different methods for cleaning and feature engineering text data.
- Investigate the challenges of dealing with bias and mis categorization in hate speech detection. This could involve analysing how your model performs on different types of hate speech and different identity groups.
- Develop a user-friendly interface or application for your model. This could make your project more impactful and accessible to a wider audience.
- Contribute to the broader field of hate speech detection research. This could involve publishing your findings in a conference or journal paper, or making your code and data publicly available.



1.2 Motivation

Personal Motivations:

Combating online negativity and fostering open communication: You may be personally affected by the prevalence of hate speech online and feel compelled to contribute to creating a more inclusive and respectful digital environment.

Exploring a challenging technical problem: This project presents a chance to apply your machine learning skills to a complex and nuanced task with immense real-world implications.

Developing critical thinking and ethical awareness: Working with hate speech requires careful consideration of bias, cultural nuances, and free speech, fostering valuable critical thinking and ethical awareness.

Contributing to a broader social movement: You can be part of the collective effort to combat hate speech and online extremism through your research and model development.

Societal and Technical Motivations:

Rising prevalence of hate speech online: The anonymity and reach of online platforms facilitate the spread of hate speech, causing real-world harm and inciting violence.

Limited capacity for manual content moderation: The sheer volume of online content makes manual detection of hate speech inefficient and unsustainable.

Potential for automation and scalability: Machine learning models can provide tools for automatic detection and flagging of hate speech, offering scalability and speed.

Improvement of existing models and datasets: Research in this area can contribute to the development of more accurate and nuanced hate speech detection models and datasets.

1.3 Background

What is Hate Speech

The case of hate speech and violent communication conducted over the internet can be referred as cyber-hate. It is a narrow and specific form of cyber-bullying and it can be defined as “any use of electronic communications technology to spread racist, religious, extremist or terrorist messages” it is different from cyber-bullying in that hate speech can target not only individuals but it also has implications on whole communities. Brown has also defined hate speech as any textual or verbal practice that implicates issues of discrimination or violence against people in regard to their race, ethnicity, nationality, religion, sexual orientation and gender identity.

According to Anis hate speech can occur in different linguistic styles and several acts like insulting, provocation, abusing and aggression. However, according to Chetty and Alathur, hate speech can be categorized into the following categories:

Gendered hate speech:

This category includes Any form of hostility towards particular gender or any devaluation based on person’s gender. This includes any post that offense particular gender. Also, it includes any form of misogyny. Moreover, Jha and Mamidi clarify that sexism may come in two forms: Hostile (which is an explicit negative attitude) and Benevolent (which is more subtle).

Religious hate speech:

This will include any kind of religious discrimination, such as: Islamic sects, calling for atheism, Anti-Christian and their respective denominations or anti-Hinduism and other religions. However, Albadi et al. mentioned that religious hate speech is considered as a motive of crimes in countries with highest social crimes.

Racist hate speech:

Lastly, this category includes is Any sort of racial offense or tribalism, regionalism, xenophobia (especially for migrant workers) and nativism (hostility against immigrants and refugees) and any prejudice against particular tribe or region. For instance, offending an individual because he belongs to a particular tribe or region or country or favouritism of a particular tribe. Add to that, offending the appearance and colour of individual.

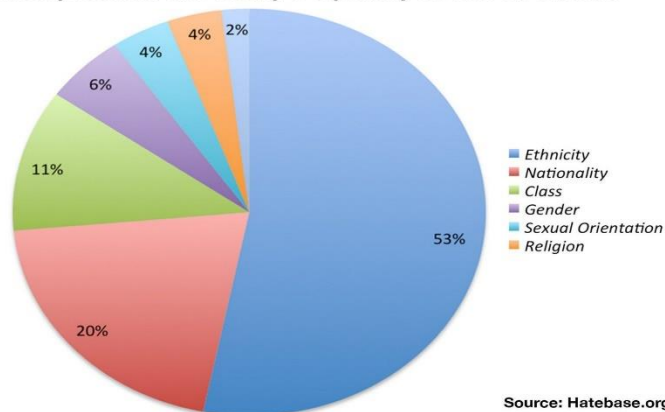
What Constitutes Hate Speech?

Hate speech is hard to comprehend. However, it can be recognized based on specific characteristics that can be distinguished from one culture to another. These characteristics are debatable, some may interpret them as a pure hate and some don’t. This problem is considered as a controversial problem that no one can agree upon. Gelashvili and Nowak argued that it is an obstacle for social media platforms owners to regulate hate speech as many questions will raise to their heads such as what constitute hate speech? And what kind of hate speech need to be countered? Only legitimate people who are actively engaged in the same culture and who can be competent enough can give the answers to these questions.

Some studies have given some necessary terminologies for studying hate speech, for example Fortuna and Nunes have listed some of the main rules for hate speech identification. In brief, hate speech is identified when disparaging stereotype about group. Together with using racial and sexist slurs with intent to harm. Add to that when indecently speak about religion or specific country.

Correspondingly, when identifying hate speech, we need to exclude some conditions. For instance, when trying to explain the meaning of some abusive words or when we use some of racial terms in another context which has no hate undertone. Add to that when writing a news article and referring to a sect which is associated with hate crime “e.g. ISIS” this referral itself won’t be considered as hate speech. In like manner, Waseem and Hovy have proposed 11 parameters to distinguish hate speech specifically in twitter platform, some of which are: usage of sexiest and racial terms, attacking and criticizing minority, promoting violence, distorting the truth with lies and supporting suspicious hashtags.

Types of hate speech in the UK by frequency of use on Twitter



2. Work Description

Project Goal: Develop a machine learning model that can effectively and accurately detect hate speech in online text data or by manually entered data.

Project Scope:

Data Acquisition and Preprocessing:

The Hate speech detection dataset is taken from the Kaggle platform. Identified and acquired a relevant dataset of labelled hate speech and non-hate speech text data (e.g., tweets, hate speech, offensive languages). Data preparation part is covered. In this the data cleaning, encoding, created a chart, etc. Exploratory Data Analysis comes in this phase.

Model Training and Evaluation:

Implement and train various machine learning models which are suitable for text classification (e.g., Decision Tree, Logistic Regression, Random Forest, Support Vector Machines and naïve bayes). Such as –

Decision Tree:

Decision trees are used when we want to make decisions or predictions based on input features. They are particularly useful in scenarios where you need to understand and interpret the decision-making process, as the tree structure is easy to visualize and comprehend.

Logistic Regression:

This model is used for analyse the relationship between independent and dependent variables. It works on the sigmoid which helps to separate the different classes.

Random Forest:

This model is work on the bagging. It's a powerful ensemble machine learning algorithm that combines multiple decision trees to achieve more accurate and robust predictions. It works on the averaging of data which helps to avoid overfitting issues.

SVM (Support Vector Machine):

The primary goal of SVM is to find a hyperplane in an N-dimensional space (where N is the number of features) that distinctly separates the data points into different classes. It works on 2D as well as 3D models also. It uses the Kernels according to data, For Linear data it has linear kernel and for non linear data it has RBF, Polynomial and sigmoid.

Naïve Bayes:

Based on the premise of conditional independence of characteristics given the class label, the probabilistic and generative Naive Bayes model functions. This 'naive' assumption frequently works well, especially for text classification and spam filtering applications.

For Evaluation Metric:

It defines the different classification methods such accuracy, precision, Recall and F1 score. Its totally depend on the values training and testing data.

For Tuning: Tuning is a step which helps to understand the model's behaviour. There are two tuning methods which is used in this model-

GridSearchCV:

GridSearchCV is a function in the scikit-learn library, a popular machine learning library in Python, which performs hyperparameter tuning for a machine learning model using cross-validation. Hyperparameter tuning involves selecting the best set of hyperparameters for a model to optimize its performance. I used this in all the models for tuning.

RandomisedSearchCV:

RandomizedSearchCV is another hyperparameter tuning approach in scikit-learn that, like GridSearchCV, is used for optimizing machine learning models through hyperparameter selection. However, instead of exhaustively searching through all possible combinations of hyperparameter values, RandomizedSearchCV randomly samples a fixed number of hyperparameter combinations from specified distributions. This is also is used in this project.

Analysis and Interpretation:

Analysed the model's strengths and weaknesses, identifying potential biases and limitations. Which helps to understand the model stability.

After making models I got the different accuracy at each algorithm which helps to identify which model is best fir or under fit or overfit. I got best fit in Support Vector Machine Algorithm which is 89%.

3. Technical Specification



Introduction:

Scope: To develop a machine learning model that can effectively and accurately detect hate speech in online text data or by manually entered data.

In this I used all the above-mentioned machine learning algorithms. Which are the main things for this project. It will detect only the English sentences and not works on the other languages.

Hardware Requirements:

For this project minimum requirements are 64 – bit system, 8 GB Ram, i5 core, 2GB Graphics Card, 512 GB SSD or 1 TB HDD.

Software Requirements:

Jupyter Notebook Software or any latest version platform where the python extension is to be installed and Windows 10 OS is required and all the used libraries like pandas, NumPy, sklearn, etc should be installed.

Performance Requirements:

The system speed at least 3 GHz or higher preferred.

Security Requirements:

In future the data will be securely stored in the database (e.g., Email, name, contact details). Right now this won't ask you about any personal info.

Reliability and Availability:

In future it stores all the sentence what you will search in the database but right now it won't save any info whatever you search.

Appendix:

This project is made by step-by-step manner like I followed the Machine learning Life Cycle. I used the different techniques or methods to get the best accuracy and accurate output. For this project all the basic requirements of software and hardware should be full fill.

4. Literature Survey

1. Automated Hate Speech Detection and the Problem of Offensive Language: Author - Thomas Davidson, Dana Warmley, Michael Macy, Ingmar Weber

2. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter: Zeerak Waseem, University of Copenhagen, Denmark; Dirk Hovy, University of Copenhagen, Denmark.

3. DETECTION OF HATE SPEECH IN SOCIAL NETWORKS: A Survey on a Multilingual Corpus by Areej Al-Hassan and Hmood Al-Dossari, Department of Information Systems, King Saud University, Riyadh, Saudi Arabia.

4. A Literature Review of Textual Hate Speech Detection Methods and Datasets by Fatimah Alkomah (Department of Computer Science, University of Idaho, Moscow, USA) and Xiaogang Ma (Department of Information Systems, Princess Nourah bint Abdulrahman University, Riyadh 11671, Saudi Arabia).

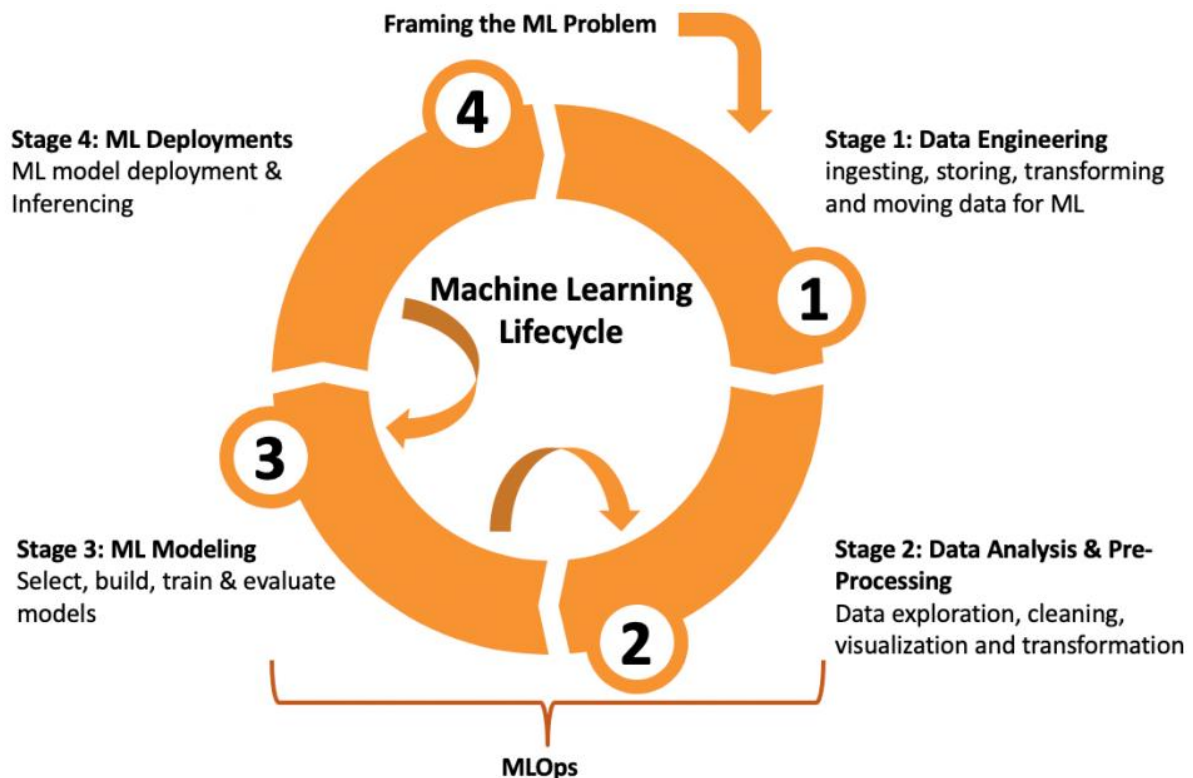
5. Hate Base: A comprehensive database of hate speech and extremist symbols

6. Project Hate-O-meter: A platform for developing and evaluating hate speech detection tools

7. ACL Anthology: A repository of research papers on natural language processing, including hate speech detection

8. Jigsaw Machine Learning for Good: Resources and tools for applying machine learning to social good.

5. Methodology



Data Collection and Preparation:

- Gather a suitable dataset: Source of my dataset is Kaggle. The dataset includes a lot of information about hate speech, offensive language, tweets and labels.
- After getting suitable Dataset we need to prepare the data and for this we put data in suitable place so that we can use it in machine learning training.
- This step includes the below tasks:
 - Identify various data sources
 - Collect data
 - Integrate the data obtained from different sources

Data Processing:

- Data preprocessing is the process of transforming raw data into a form suitable for analysis and model development. It is one of the most critical steps in determining the success of the final model.
- There are several ways to preprocess your data. It may include one or more of the following steps:
 - Removing irrelevant features from your dataset
 - Filling in missing values
 - Reducing the size of the dataset and feature set
 - Transforming categorical variables into numerical variables (or vice versa)
 - Normalizing the data points

Data Wrangling

- Data wrangling is the process of cleaning and converting raw data into a useable format. It is the process of cleaning the data, selecting the variable to use, and transforming the data in a proper format to make it more suitable for analysis in the next step. It is one of the most important steps of the complete process.
- Cleaning of data is required to address the quality issues.

Data Analysis

- Now the cleaned and prepared data is passed on to the analysis step. This step involves:
 - Selection of analytical techniques
 - Building models
 - Review the result
- The aim of this step is to build a machine learning model to analyse the data using various analytical techniques and review the outcome. It starts with the determination of the type of the problems, where we select the machine learning techniques such as Classification, Regression, Cluster analysis, Association, etc. then build the model using prepared data, and evaluate the model.

Train Model

- In machine learning, a train model is a crucial step in the development of a successful program. It refers to the process of feeding a machine learning algorithm with training data so it can learn and improve its ability to make predictions on new data.
- After training the model is also tested to check the accuracy of the model.

Deployment

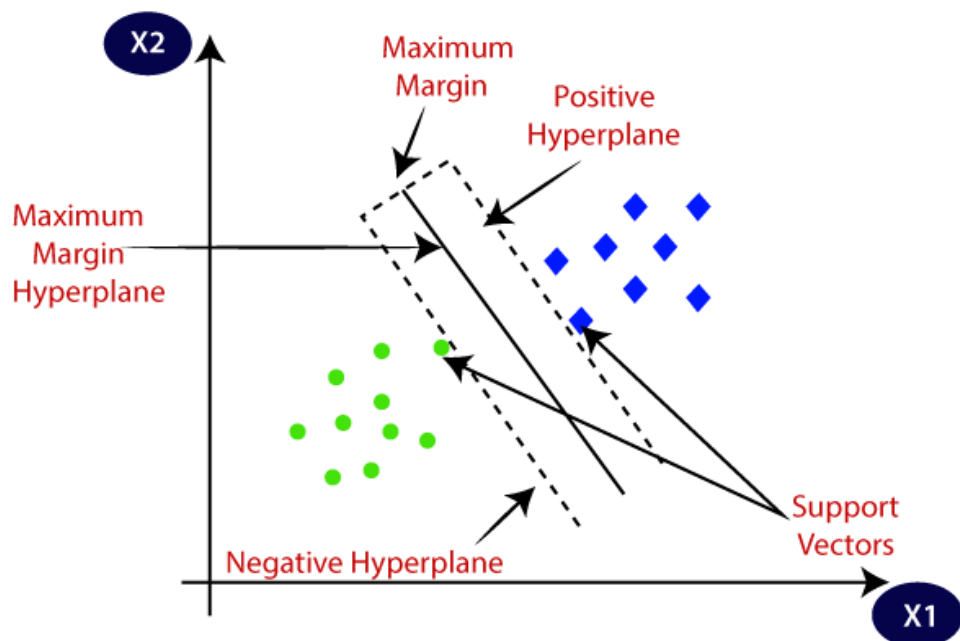
- Deployment in Machine Learning is the method by which you integrate a machine learning model into an existing production environment to make practical business decisions based on data.
- It is the last stage in the machine learning lifecycle.

Based on above-mentioned process, I choose to proceed with SVM (Support Vector Machine) and Logistic Regression for model building process:

SVM (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. 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. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane:



Formula of SVM is:

$$y = w^T x + b$$

Here,

$f(x)$ is the decision function.

W is the weight vector.

X is the input feature vector.

b is the bias term.

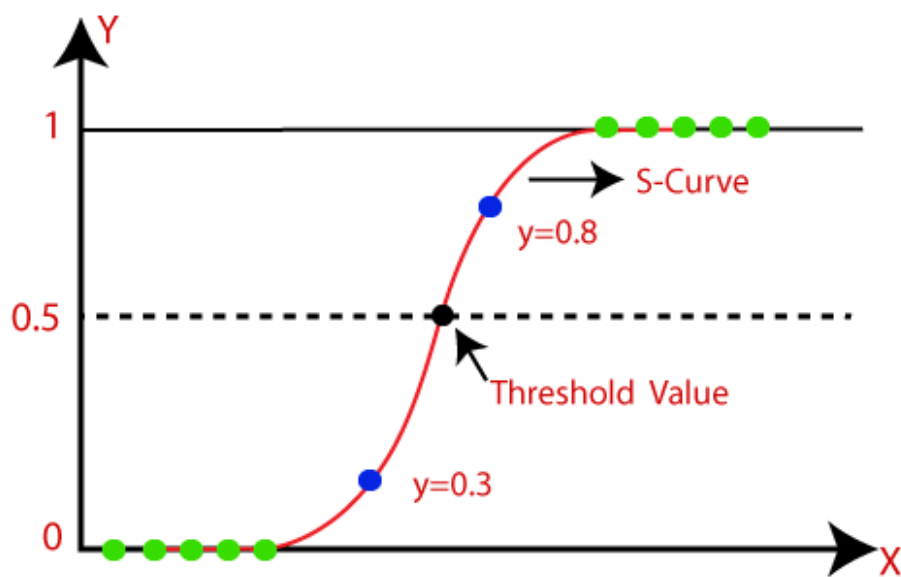
Logistic Regression:

Logistic regression is a statistical method that uses math to analyse the relationship between data factors and predict the value of one factor based on the other. It's a supervised machine learning algorithm that predicts the probability of an outcome, event, or observation. Logistic regression is used to build machine learning models where the

dependent variable is binary. It predicts a dependent data variable by analysing the relationship between one or more existing independent variables.

Logistic regression predicts a binary outcome, such as yes or no, based on prior observations of a data set. The model delivers a binary or dichotomous outcome limited to two possible outcomes: yes/no, 0/1, or true/false. Formula of Logistic Regression is :

$$P(Y = 1) = \frac{1}{1+e^{-(b_0+B*X)}}$$



The Formula of Sigmoid is:

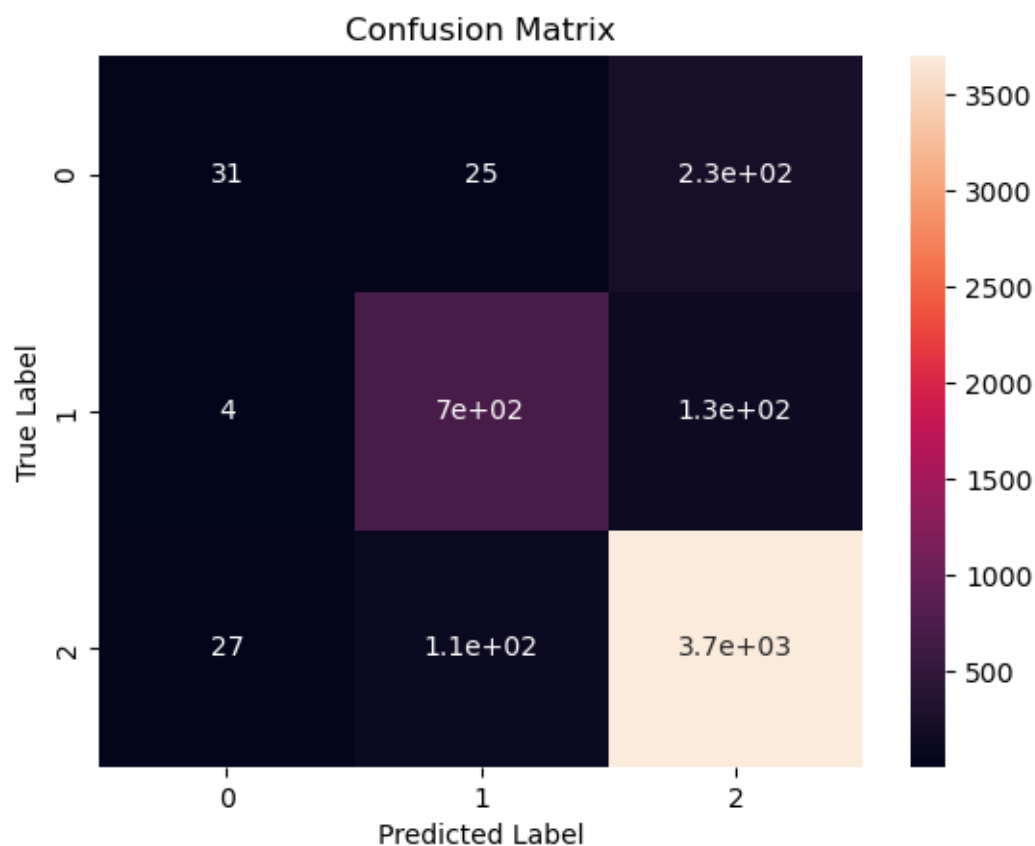
$$p = \frac{1}{1 + e^{-z}}$$

6. Project & Output

There are 24783 numbers of rows and 8 numbers of columns present. The dataset is taken from the Kaggle platform. It contains the tweeter's data which helps me to detect the hate speech and offensive data. There are several models I have used but I move forward with SVM only because it gives the best fit accuracy. Following table shows the accuracy of all the models –

Models	Accuracy before tuning (%)	Accuracy after tuning (%) (GridSearchCV and RandomisedSearchCV)
Decision Tree	88	-
KNN	85	85,85
SVM	89	89,90
Logistic Regression	89	90,90
Random Forest	89	77
Naïve Bayes	84	84,87

As per the table the SVM shows the best accuracy after tuning as well as before tuning. In this project the plotted Heatmap is –



It shows the relationship between the test and predicted values. In SVM I tried to find the best accuracy to I used all the kernels. The following table show the accuracy of the kernels in used SVM model –

Kernels	Accuracy (%)
Linear	88
RBF	89
Sigmoid	89
Poly	80

It means the RBF and Sigmoid gives the best accuracy and stable accuracy and it shows the data are present in the non-linear form.

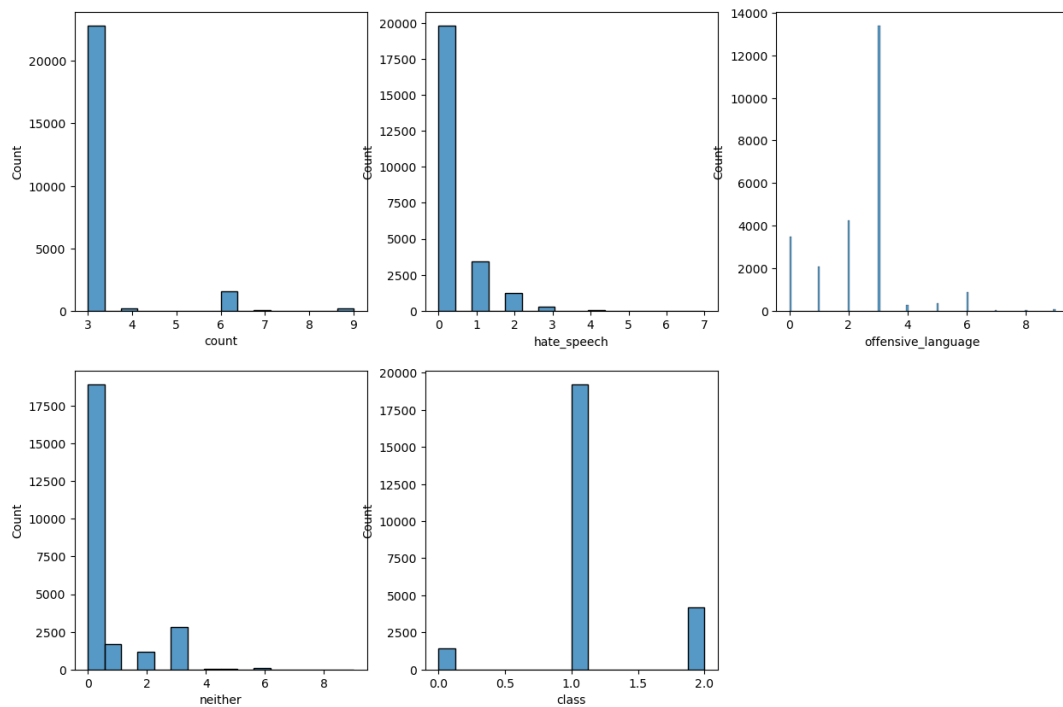
The following tables show all the classification methods values of Hate Speech Detected, N O hate and offensive speech, Offensive language detected respectively–

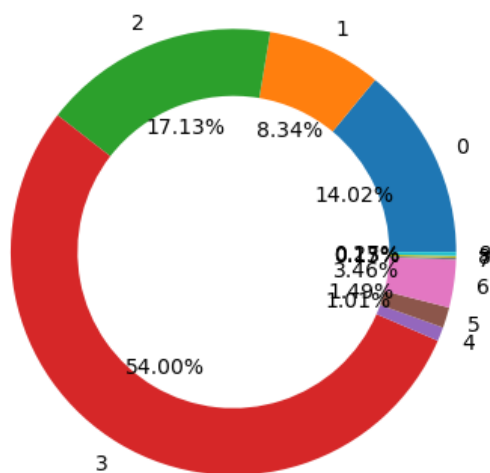
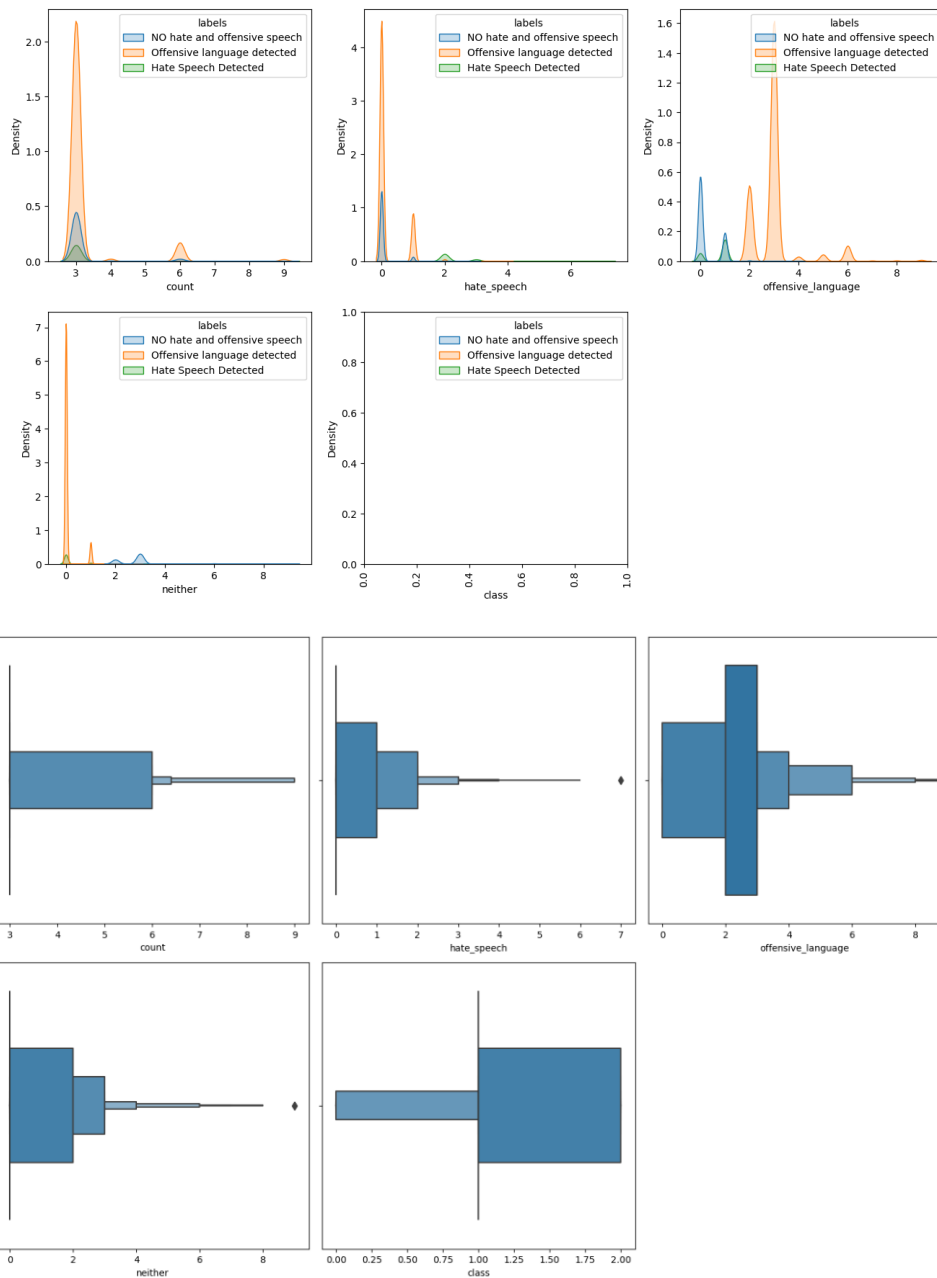
Models	Precision before tuning (%)	Precision after tuning (%) (GridSearchCV and RandomisedSearchCV)
Decision Tree	31, 82, 93	-
KNN	47, 79, 87	40,75, 90 and -
SVM	43, 80, 92	43, 80, 92 and 48, 84, 92
Logistic Regression	97, 97, 97	48, 84, 92 and 49, 84, 92
Random Forest	54, 82, 92	- and 0, 100, 77
Naïve Bayes	100, 84, 83	25, 80, 92 and 30, 77, 90

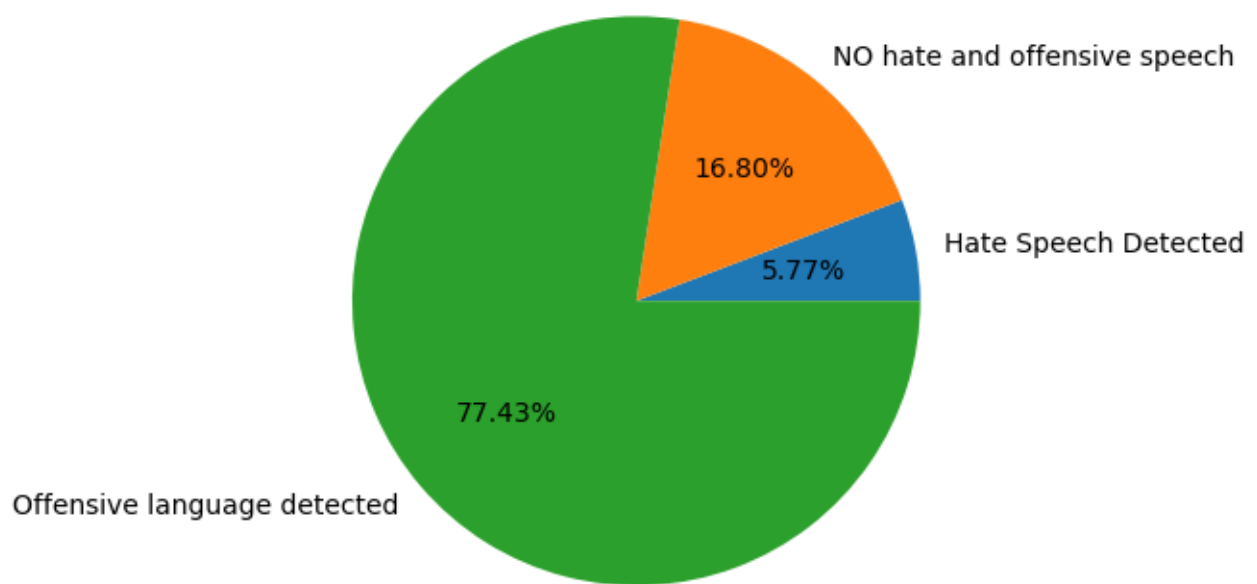
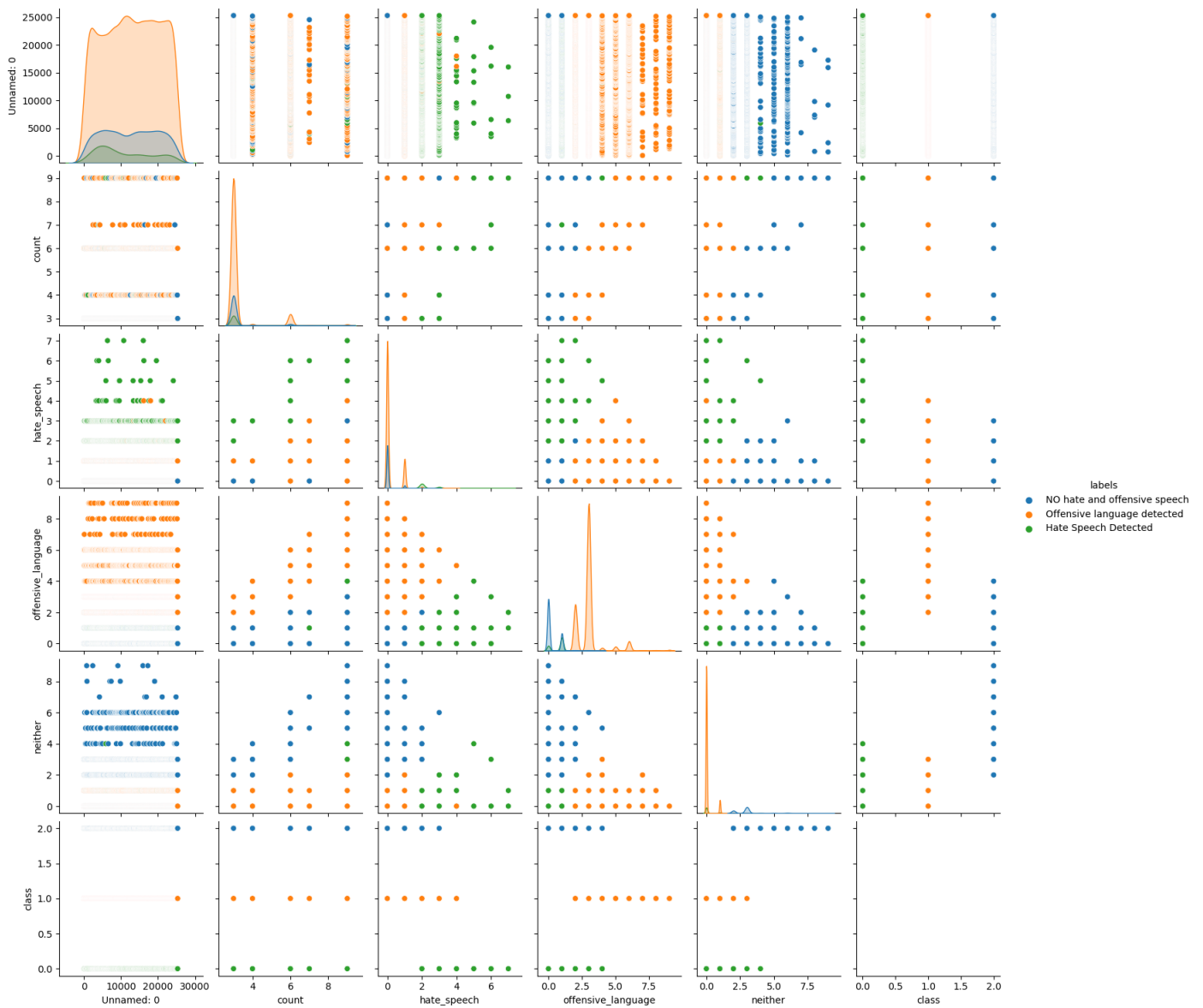
Models	Recall before tuning (%)	Recall after tuning (%) (GridSearchCV and RandomisedSearchCV)
Decision Tree	37, 82, 92	-
KNN	31, 57, 95	33,69, 93 and -
SVM	07, 90, 95	07, 90, 95 and 22, 86, 95
Logistic Regression	66, 98, 100	20, 86, 96 and 20, 86, 96
Random Forest	27, 82, 95	- and 0, 1, 100
Naïve Bayes	0, 42, 99	40, 71, 90 and 11, 75, 95

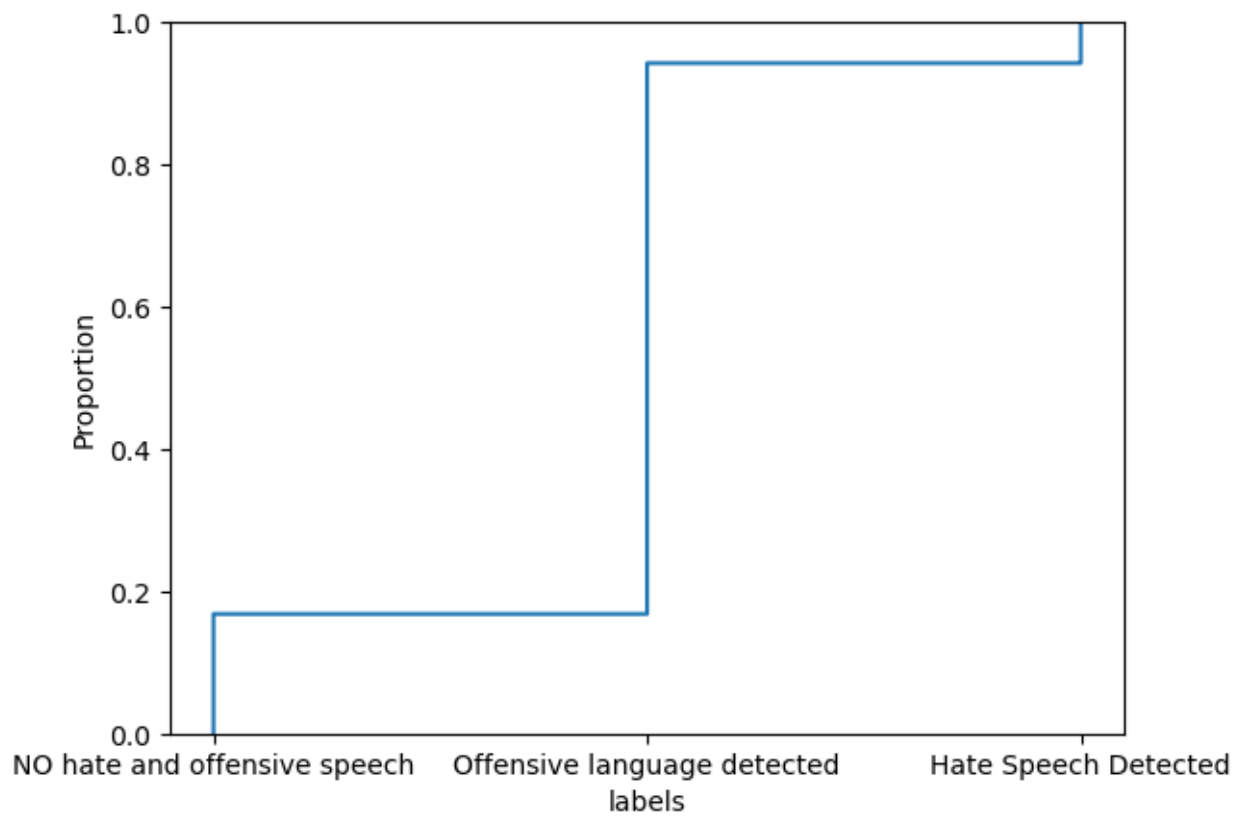
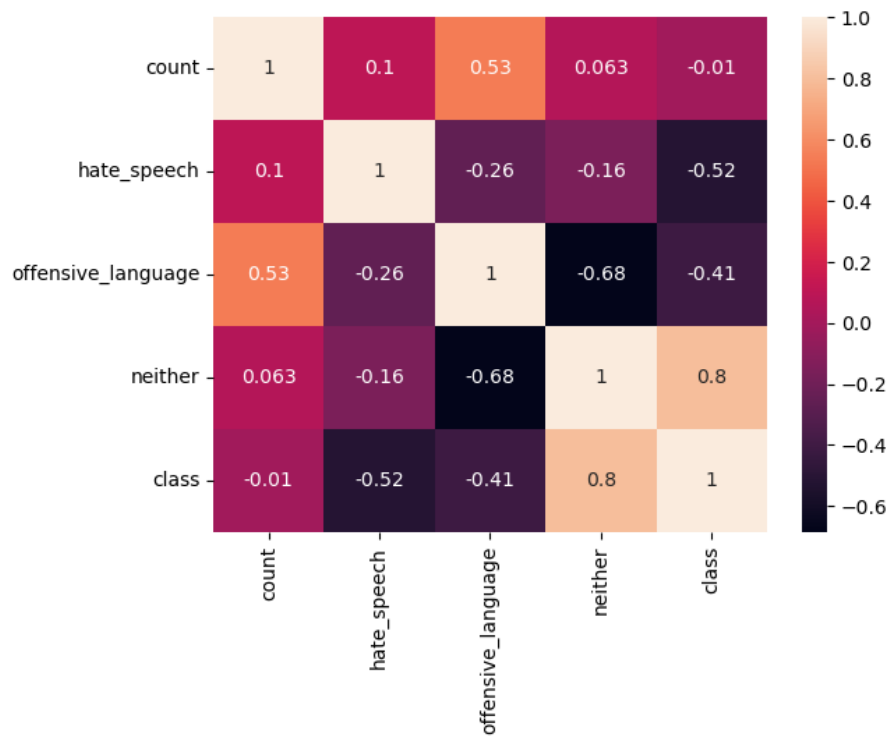
Models	F1 Score before tuning (%)	F1 Score after tuning (%) (GridSearchCV and RandomisedSearchCV)
Decision Tree	34, 82, 93	-
KNN	38, 67, 91	37, 72, 91 and -
SVM	12, 85, 94	12, 85, 94 and 30, 85, 94
Logistic Regression	79, 98, 98	28, 85, 94 and 28, 85, 94
Random Forest	36, 82, 94	- and 0, 1, 87
Naïve Bayes	1, 56, 91	31, 75, 91 and 17, 76, 92

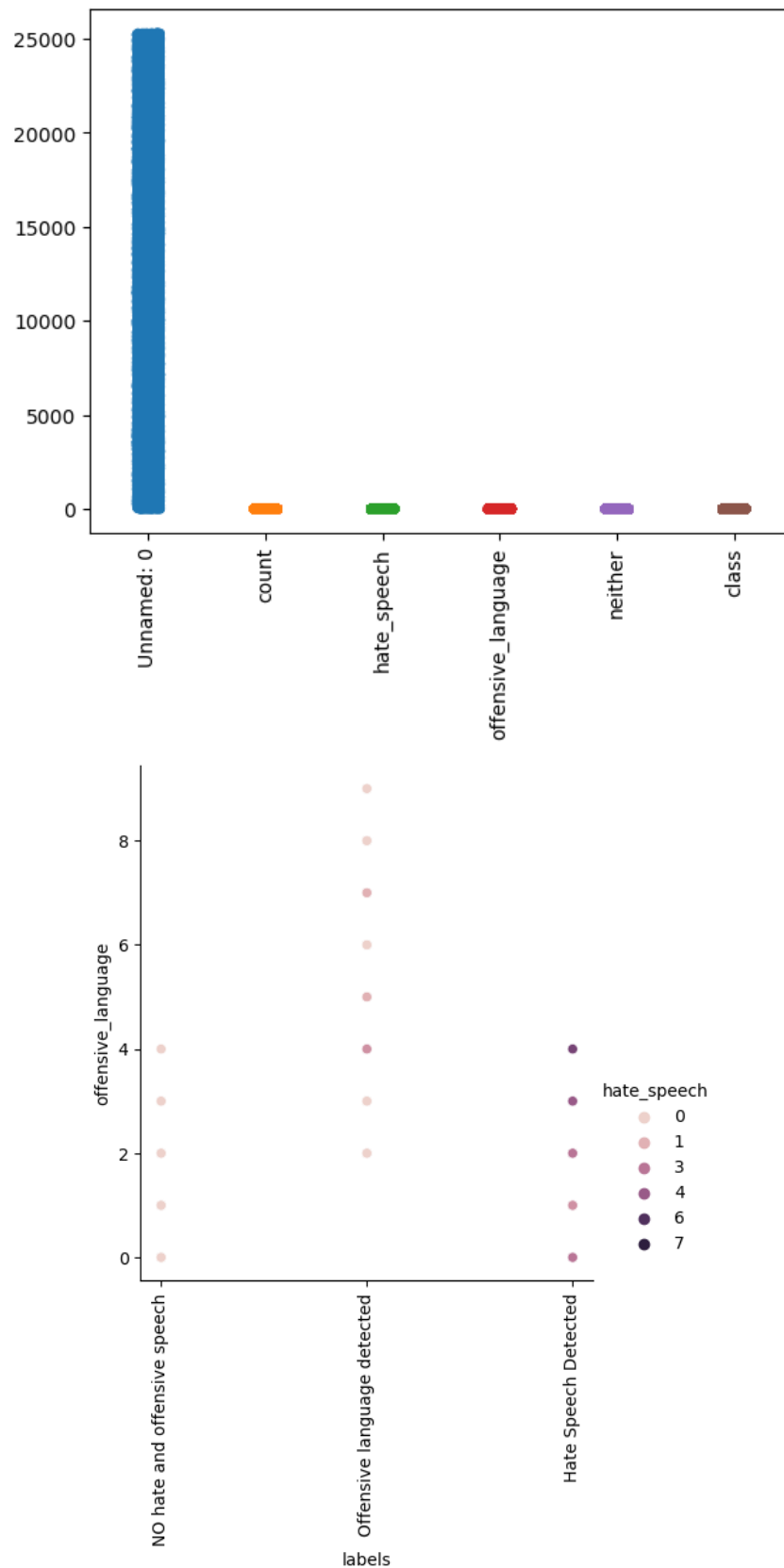
Plotted the various types of graphs which shows the relationship between the values. The graphs are –











These all charts show the relations which helps us to identify the which model is to be use. Pair Plot is most import charts which always shows the relations of each column in the same plot.

7. Future Enhancement:

In future the project will have its own login and signup page which asked the following data –

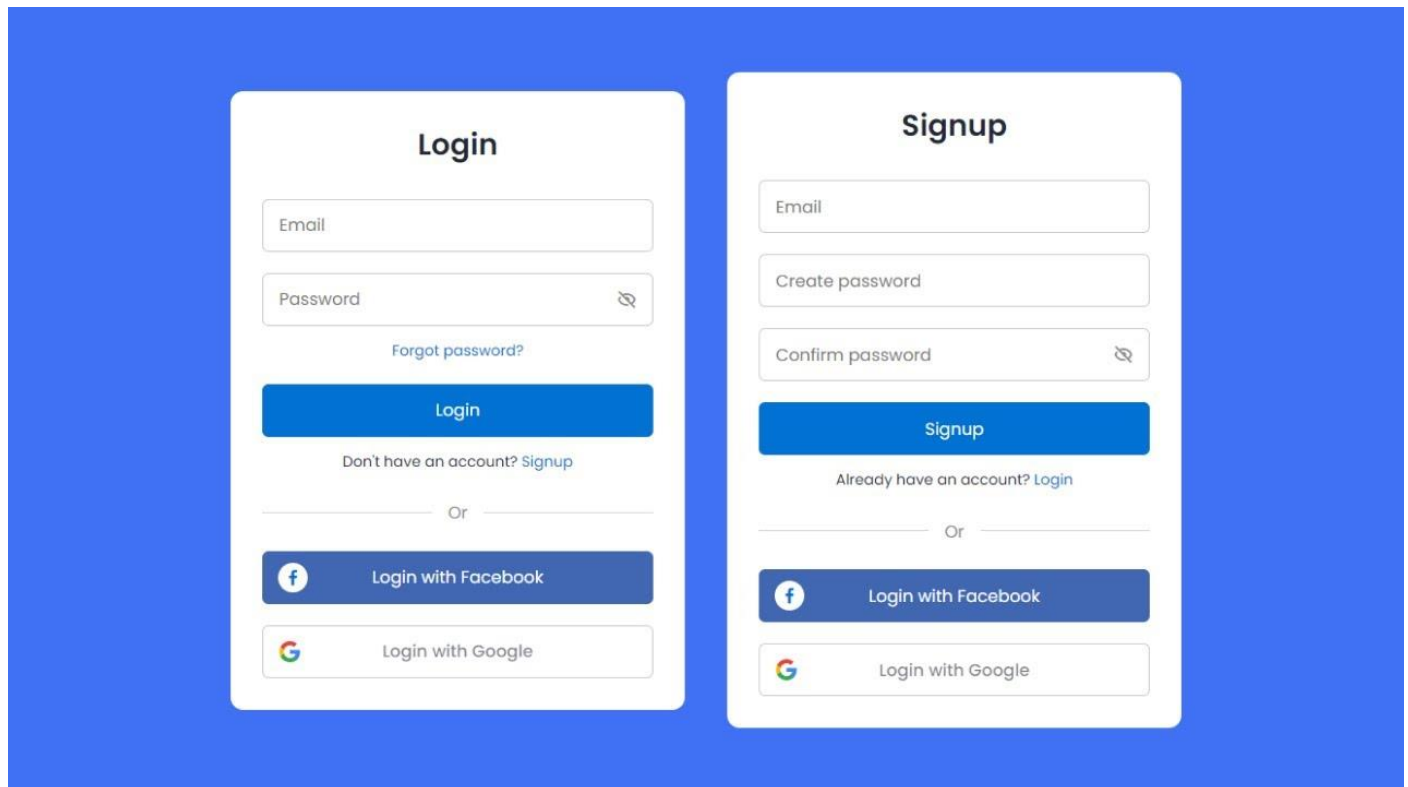
Full Name - It shows on the deployment site.

Mail Id – For login its needed as unique id.

Password – For login purpose.

Forgot password – It can help to reset the password.

There will be a web platform where the user can check the sentence is Offensive or hate speech or not. It will be too interactive.



The image displays two side-by-side web forms, 'Login' and 'Signup', set against a solid blue background. Both forms are white with rounded corners and a subtle drop shadow.

Login Form:

- Header: 'Login' in bold black text.
- Fields: 'Email' and 'Password' (with a toggle icon).
- Link: 'Forgot password?' in blue text.
- Button: A solid blue 'Login' button.
- Text: 'Don't have an account? Signup' in blue text.
- Separator: A horizontal line with 'Or' in the center.
- Buttons: 'Login with Facebook' (blue with Facebook icon) and 'Login with Google' (white with Google icon).

Signup Form:

- Header: 'Signup' in bold black text.
- Fields: 'Email', 'Create password', and 'Confirm password' (with a toggle icon).
- Button: A solid blue 'Signup' button.
- Text: 'Already have an account? Login' in blue text.
- Separator: A horizontal line with 'Or' in the center.
- Buttons: 'Login with Facebook' (blue with Facebook icon) and 'Login with Google' (white with Google icon).

8. Conclusion

In conclusion, after watching all the results the SVM and the Logistic Regression both are best fit with this dataset, but if we compare the F1 Score the Logistic Regression is more better than the SVM.

The Hate Speech Detection Capstone Project represents a significant step towards leveraging machine learning to address the pressing issue of online hate speech. Through meticulous data collection, preprocessing, and the application of advanced models, we have developed a robust system capable of identifying and classifying hate speech with a high degree of accuracy.

Our project has not only contributed to the technical aspects of hate speech detection but has also delved into the ethical considerations surrounding biased predictions and fairness. By implementing strategies for bias detection and mitigation, we aim to ensure that our model is not inadvertently perpetuating discrimination.

The deployment of the hate speech detection model into real-world scenarios is a crucial milestone, offering a practical solution for platforms and communities seeking to foster a safer online environment. The integration of interpretability and explain ability features enhances transparency, providing users with insights into the model's decision-making process.

As we move forward, continuous monitoring and updates will be pivotal in adapting to the evolving landscape of online language use. Our commitment to ethical AI principles remains unwavering, and we recognize the importance of ongoing collaboration with diverse communities to refine and improve our hate speech detection system.

In essence, this capstone project is not just a technological achievement but a conscientious effort to contribute to the well-being of online spaces. By addressing hate speech, we strive to foster a more inclusive, tolerant, and respectful digital environment for users around the globe. As technology evolves, so too must our commitment to harnessing it responsibly for the betterment of society.

If we conflate hate speech and offensive language then we erroneously consider many people to be hate speakers and fail differentiate between commonplace offensive language and

serious hate speech. Given the legal and moral implications of hate speech it is important that we are able to accurately distinguish between the two.

9. Reference

- 1 C. Blaya, “Cyberhate: A review and content analysis of intervention strategies,” *Aggress. Violent Behav.*, no. May, pp. 0–1, 2018.
- 1 F. Miro-Llinares and J. J. Rodriguez-Sala, “Cyber hate speech on twitter: Analyzing disruptive events from social media to build a violent communication and hate speech taxonomy,” *Int. J. Des. Nat. Ecodynamics*, vol. 11, no. 3, pp. 406–415, 2016.
- 2 Brown, “What is hate speech? Part 1: The Myth of Hate,” *Law Philos.*, vol. 36, no. 4, pp. 419–468, 2017.
- 3 M. Y. Anis and U. S. Maret, “Hatespeech in Arabic Language,” in *International Conference on Media Studies*, 2017, no. September.
- 4 N. Chetty and S. Alathur, “Hate speech review in the context of online social networks,” *Aggress. Violent Behav.*, vol. 40, no. May, pp. 108–118, 2018.
- 5 Jha and R. Mamidi, “When does a compliment become sexist? Analysis and classification of ambivalent sexism using twitter data,” *Proc. Second Work. NLP Comput. Soc. Sci.*, pp. 7–16, 2017.
- 6 N. Albadi, M. Kurdi, and S. Mishra, “Are they Our Brothers? Analysis and Detection of Religious Hate Speech in the Arabic Twittersphere,” *2018 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Min.*, pp. 69–76, 2018.
- 7 T. Gelashvili and K. A. Nowak, “Hate Speech on Social Media,” *Lund University*, 2018.
- 8 P. Fortuna and S. Nunes, “A Survey on Automatic Detection of Hate Speech in Text,” *ACM Comput. Surv.*, vol. 51, no. 4, pp. 1–30, 2018.
- 9 BBC. 2015. Facebook, google and twitter agree german hate speech deal.
<http://www.bbc.com/news/world-europe-35105003>. Accessed on 26/11/2016.

- 10 Warner, W., and Hirschberg, J. 2012. Detecting hate speech on the world wide web. In LSM, 19–26.
- 11 Hate Speech and Offensive Content datasets: Kaggle, Jigsaw Toxic Comment Classification Challenge, Twitter Hate Speech and Offensive Language dataset.
- 12 Machine learning libraries: scikit-learn, TensorFlow, PyTorch.
- 13 Research papers on hate speech detection with machine learning.
- 14 <https://towardsdatascience.com/>
- 15 <https://www.datacamp.com/>
- 16 <https://www.javatpoint.com/>
- 17 <https://www.kaggle.com/datasets/mrmorj/hate-speech-and-offensive-language-dataset>