**Mini Project Report on**



**HATE SPEECH DETECTION**



**Submitted in partial fulfillment of the requirement for the award of the degree of**

**BACHELOR OF TECHNOLOGY**

**IN**

**COMPUTER SCIENCE & ENGINEERING**

**Submitted by:**

**Mansi Mehara 2021305**

***Under the Mentorship of***

**Mr. Kireet Joshi**

**Assistant Professor**



**Department of Computer Science and Engineering**

**Graphic Era (Deemed to be University)**

**Dehradun, Uttarakhand**

**January 2024**



**CANDIDATE’S DECLARATION**

I hereby certify that the work which is being presented in the project report entitled **“Hate Speech Detection”** in partial fulfillment of the requirements for the award of the Degree of Bachelor of Technology in Computer Science and Engineeringof the Graphic Era (Deemed to be University), Dehradun shall be carried out by the under the mentorship of **Mr. Kireet Joshi, Assistant Professor**, Department of Computer Science and Engineering, Graphic Era (Deemed to be University), Dehradun.

Mansi Mehara 2021305

**Table of Contents**

|  |  |  |
| --- | --- | --- |
| **Chapter No.** | **Description** | **Page No.** |
| Chapter I | Introduction | 1 |
| Chapter II | Literature Survey | 4 |
| Chapter III | Methodology | 6 |
| Chapter IV | Result and Discussion | 9 |
| Chapter V | Conclusion and Future Work | 11 |
|  | References | 12 |

**Chapter I**

**Introduction**

* 1. **Introduction**

In our rapidly advancing world, where nearly everything undergoes digitalization or operates online, the societal interactions of individuals are increasingly transitioning from the physical realm to the virtual sphere. This shift has paved the way for the integration of social media platforms into people's daily lives. Despite the numerous benefits or advantages associated with this trend, it also presents a substantial drawback.

* 1. **The Problem**

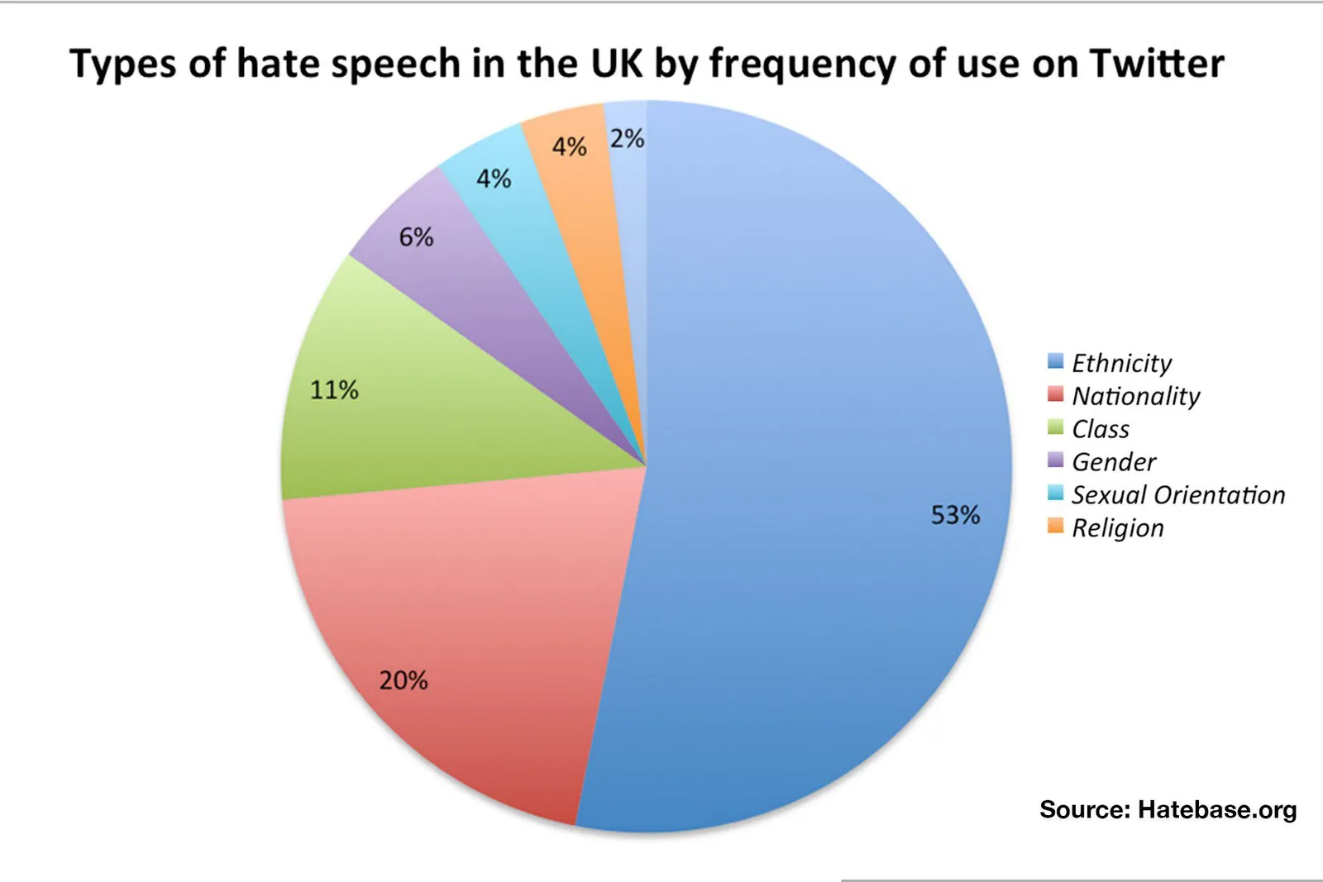
People tend to possess a lot of hate or offence for others and portraying such thoughts over the internet is a much easier for such people as they can do the evils from anywhere in the globe to anyone on the social media platform.

Any kind of hate or offence can be hurtful or saddening for the person victimized with such speech, may it be on the physical or virtual world. The problem with it being on the social platform online is that it cannot be erased by the one to whom it is targeted to. There will always be a hateful or offensive comment for that one person forever. As much as it is non-ethical or immoral, it can be very depressing for the victim. Along with that, the viewers of such comments, may it be anyone related or non-related, may not like seeing such content on their social handles. The social platforms will also be greatly affected by this as people will no longer like to use the platform and will try to find alternatives or quit this online social shift entirely.

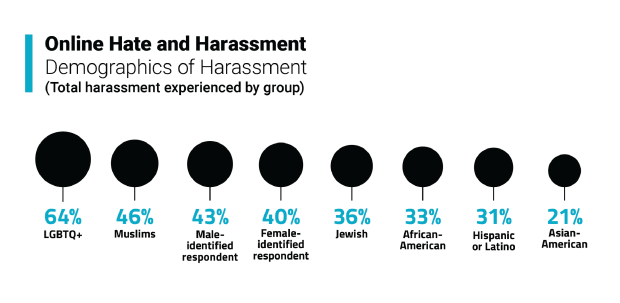
* + 1. **The Extent**

All the information looks true and concerning upon reading, but one needs to know some related stats to really get a proper understanding of the problem in hand to get the idea that how much of a trouble it is for the society.

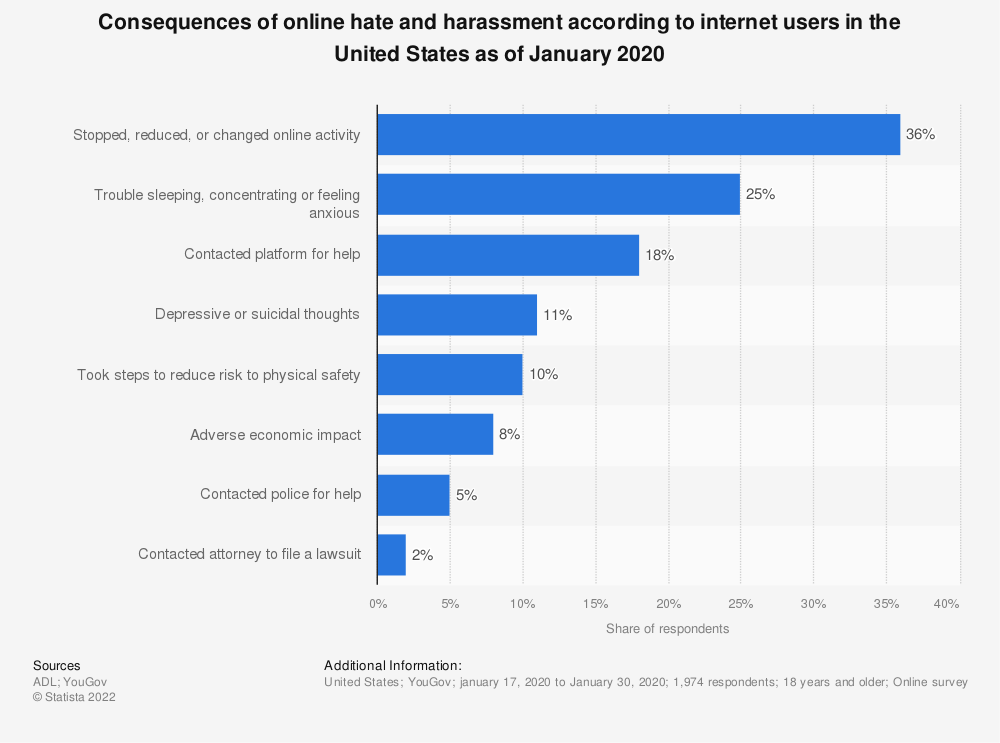
To help with the same, here are some statistics for judging the extent and impact of such hateful/offensive deeds over the internet :-



**Fig. 1.1 Pie-chart representing the types & percentages of hate speech on twitter**



**Fig. 1.2 Percentages per groups of people experiencing hate online**

****

**Fig. 1.3 A bar-graph to show the consequences of hate online**

* 1. **The Need of a Solution**

With this problem there arises a need to resolve or at least to fight the issue of hateful/offensive speech or comments on social media platforms.

But to do so, one first needs to detect the hateful or offensive comments from the non-hateful or non-offensive ones. This is where the need of this project comes into existence.

**Chapter II**

**Literature Survey**

**2.1 First Survey**

Hate speech is becoming more prevalent as internet material expands. The difficulties that online automated methods for text-based hate speech identification encounter are identified and examined. Language barrier, divergent interpretations of what constitutes hate speech and a lack of data availability for these system’s training and testing are some of these challenges. Additionally, a lot of current techniques have a problem with interpretability, which makes it challenging to comprehend why the systems choose the actions they perform. A multi-view SVM technique is suggested that, while being simpler and providing judgements that are simpler to understand than those produced by neural methods, achieves performance that is close to the state-of-the-art. The task's remaining technical and practical difficulties are also explored.[1]

The demand for automatic hate speech detection systems grows as long as hate speech is a social issue. The existing methods for doing this work are discussed but talk of a brand-new technique that achieves respectable accuracy is must. Additionally, a fresh strategy must be put forth that can do better than current systems at this task while also being more comprehensible. There is a need for greater study on this issue, taking into account both technical and practical issues, given all the obstacles still present.[1]

**2.2 Second Survey**

Communication between people with various cultural and psychological backgrounds has grown more direct as a result of the fast rise of social networks and microblogging websites, leading to an increase in "cyber" conflicts between these individuals. As a result, hate speech is employed more frequently, to the point that it has seriously damaged these public areas. A specific group of individuals sharing a characteristic — such as their gender (sexism), their ethnic group or race (racial discrimination), or their beliefs and religion — are the targets of hate speech, which is the use of aggressive, violent, or abusive rhetoric. Even while the majority of online social networks and microblogging platforms restrict the use of hate speech, their magnitude makes it nearly impossible to monitor.[2]

Although the majority of microblogging and online social networks prohibit the use of hate speech, the magnitude of these networks and websites makes it nearly difficult to regulate all of their material. Therefore, it becomes necessary to automatically identify such speech and censor any information that contains unacceptable words.[2]

In the research, they presented a fresh approach to identifying hate speech on Twitter. Their suggested method uses emotive and semantic elements, together with automated detection of hate speech patterns and the most prevalent unigrams, to categorize tweets as hateful, offensive, or clean.[2]

**Chapter III**

**Methodology**

**3.1 Introduction**

The language used for implementing the project is Python due to its wide range of available libraries for training machine learning models and tools for the same. The project is made possible by training a machine learning model with the help of a dataset downloaded from Kaggle which contains a wide range of comments collected from twitter which are classified into ‘Hate Speech’, ‘Offensive Speech’ and ‘No Hate or Offensive Speech’.

This trained model is then used to predict the input text into ‘Hate Speech’, ‘Offensive Speech’ or ‘No Hate or Offensive Speech’. The model so made shows 86.99%+ accuracy in predicting the text accordingly.

**3.2 The Necessities**

There are various things that I required for the completion of this project are :-

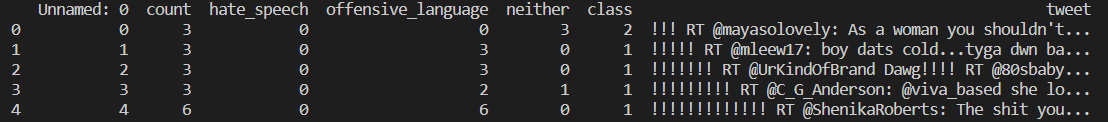
* Visual Studio Code – It is a code editor by Microsoft. I chose it due to its easy to use and helpful interface and the wide variety of extensions it provides us.[3]
* Python – It is a high level and easy to understand programming language. I chose it due to the wide variety of first-party and third-party libraries it provides the support for.[4]
* Dataset - A dataset is must to train any model such that it predicts future inputs. The one I used is downloaded from Kaggle which has a large amount of twitter comments that are classified according to hateful/offensive or not.[5]
* Essential Libraries – Some essential libraries that I installed using the python package-manager ‘pip’ are ‘pandas’, ‘numpy’, ‘sklearn’, ‘nltk’ and ‘streamlit’. Amongst them, ‘pandas’ is used to work on data, ‘numpy’ provides us with various mathematical tools to work on arrays, ‘sklearn’ (scikit-learn) provides us with machine learning tools, ‘nltk’ (natural language toolkit) helps us to apply statistical Natural Language Processing (NLP) and ‘streamlit’ is an open-source app framework used to create interactive web apps for machine learning python codes.[4]

**3.3 Working on the Dataset**

Firstly, the dataset is made appropriate for model training.

**3.3.1 Reading the Data**

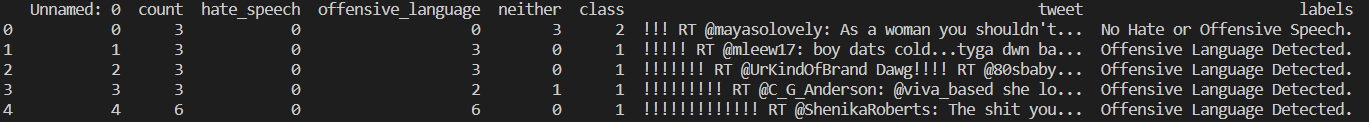
The data is read in the program. The data we initially see is as such :-



**Fig. 3.1 Raw Dataset**

**3.3.2 Labeling the Data**

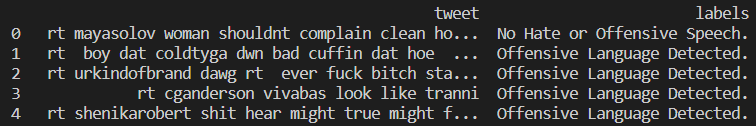
The above data is now labeled as ‘Hate Speech’, ‘Offensive Speech’ and ‘No Hate or Offensive Speech’ accordingly. The data now will be as such :-



**Fi. 3.2 Labeled Dataset**

**3.3.3 Cleaning the Data**

As shown above, the dataset used contains a large number of twitter comments. The comments naturally contains many punctuation marks or symbols which are of no use in order to train the model. So, the dataset is cleaned off such parts like ‘@’, punctuation marks, ‘https’, ‘www’, etc. The cleaned data will be as such :-



**Fig. 3.3 Cleaned Dataset**

The dataset is now finally ready for model training.

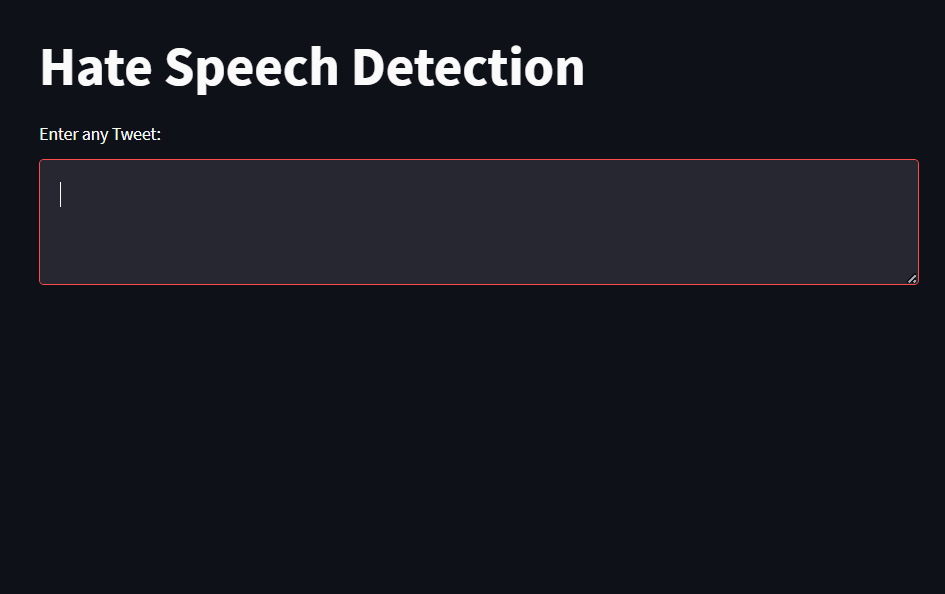
**3.4 Model Training**

The model is now trained using the dataset we just prepared right now. The model looks for the words or phrases that are unique in the hateful/offensive labeled comments and is now trained to detect those particular words.

The model is trained using the ‘train\_test\_split’ function from the ‘sklearn’ library. The dataset is split into two models ‘train’ and ‘test’. 'train’ is used to train the model and ‘test’ is used to test it for accuracy.

**3.5 Resulting Prediction Process**

The model is now put to use. The approach implemented here is using the ‘streamlit’ library. It is an open source app framework used to create easy and interactive web apps for machine learning python projects.



**Fig. 3.4 Interface of Web-app**

**Chapter IV**

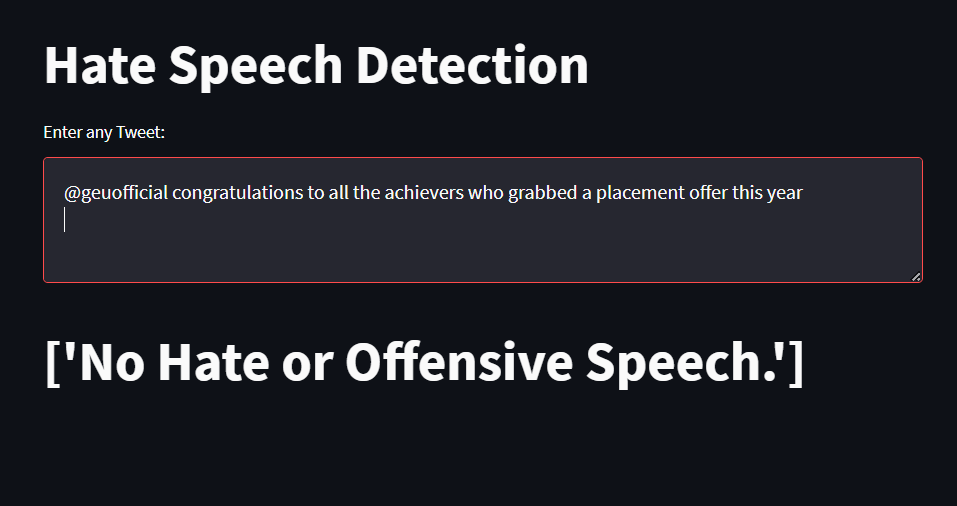
**Result and Discussion**

**4.1 Result**

The result of all this hard work is a model that is ready to predict any input tweet or text into 3 categories :-

* Hate Speech
* Offensive Speech
* No Hate or Offensive Speech

Here is the working of our model in the ‘streamlit’ web app :-



**Fig. 4.1 First output demonstration**

**Graphical user interface, text, application, website

Description automatically generated**

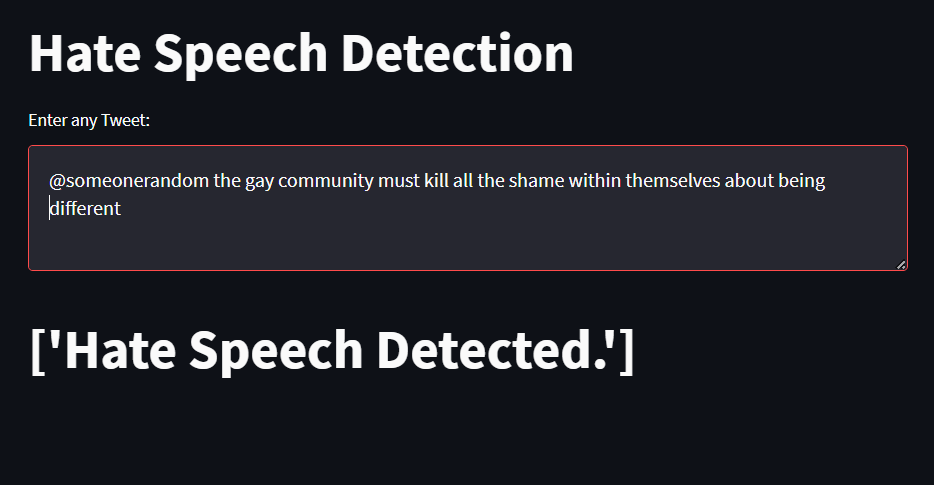
**Fig. 4.2 Second output demonstration**

****

**Fig. 4.3 Third output demonstration**

**4.2 Discussion**

This final model has shown an accuracy of over 86.99% and on the off chance it predicts the wrong output are all close call cases such as :-



**Fig. 4.4 Wrong prediction case example**

**Chapter V**

**Conclusion and Future Work**

**5.1 Conclusion**

This project provides us with a model for detecting hate or offensive speech which is trained on a dataset and is able to make predictions about the nature of future input texts, or in this case tweets, provided to it.

The model has an accuracy of 86.99% on the provided dataset of twitter tweets.

**5.2 Future Work**

This project has the basic concepts of how hate speech is detected. It can detect hate or non-hate speech in simple sentences and do so with high accuracy but it is unable to understand logical sentences for hate or non-hate as shown in **Fig. 4.4**.

There can be many better codes which will be able to work on comments with emojis and check them for vulgar or non-vulgar. There can also be codes which can work and understand such logical hate or non-hate comments as the one in **Fig. 4.4**.

The future works may also include projects which can detect the sentences or comments as the user is typing them and restrict the user to post it so the problem is eradicated even before it becomes a problem.

**References**

[1] Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian and Ophir

Frieder, “Hate speech detection: Challenges and solutions.”

Published: August 20, 2019

[2] Hajime Watanabe, Mondher Bouazizi and Tomoaki Ohtsuki, “Hate Speech on Twitter: A

Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech

Detection” Journal & Magzines IEEE Access (Volume 6).

Published: February 15, 2018

[3] Visual Studio Code website was used to download the setup of VS Code.

<https://code.visualstudio.com/download>

[4] Python website was used to download Python 3.10.9 setup.

<https://www.python.org/downloads/release/python-3109/>

[5] Kaggle website was used to download the dataset.

<https://www.kaggle.com/datasets/mrmorj/hate-speech-and-offensive-language-dataset>