**Mini Project Report on**



**HATE SPEECH DETECTION USING MACHINE LEARNING**



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

**BACHELOR OF TECHNOLOGY**

**IN**

**COMPUTER SCIENCE & ENGINEERING**

**Submitted by:**

**Deepshi Jindal** **2016723**

***Under the Mentorship of***

**Ms. Vishu Tyagi**

**Assistant Professor**



**Department of Computer Science and Engineering**

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

**Dehradun, Uttarakhand**

**January 2023**



**CANDIDATE’S DECLARATION**

I hereby certify that the work which is being presented in the project report entitled **“Hate Speech Detection Using Machine Learning”** 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 **Ms. Vishu Tyagi, Assistant Professor**, Department of Computer Science and Engineering, Graphic Era (Deemed to be University), Dehradun.

**Deepshi Jindal 2016723 signature**

**Table of Contents**

|  |  |  |
| --- | --- | --- |
| **Chapter No.** | **Description** | **Page No.** |
| Chapter 1 | Introduction | **4** |
| Chapter 2 | Literature Survey | **8** |
| Chapter 3 | Methodology | **12** |
| Chapter 4 | Result and Discussion | **15** |
| Chapter 5 | Conclusion and Future Work | **17** |
|  | References | **18** |

**Chapter 1**

**Introduction**

* 1. **Introduction**

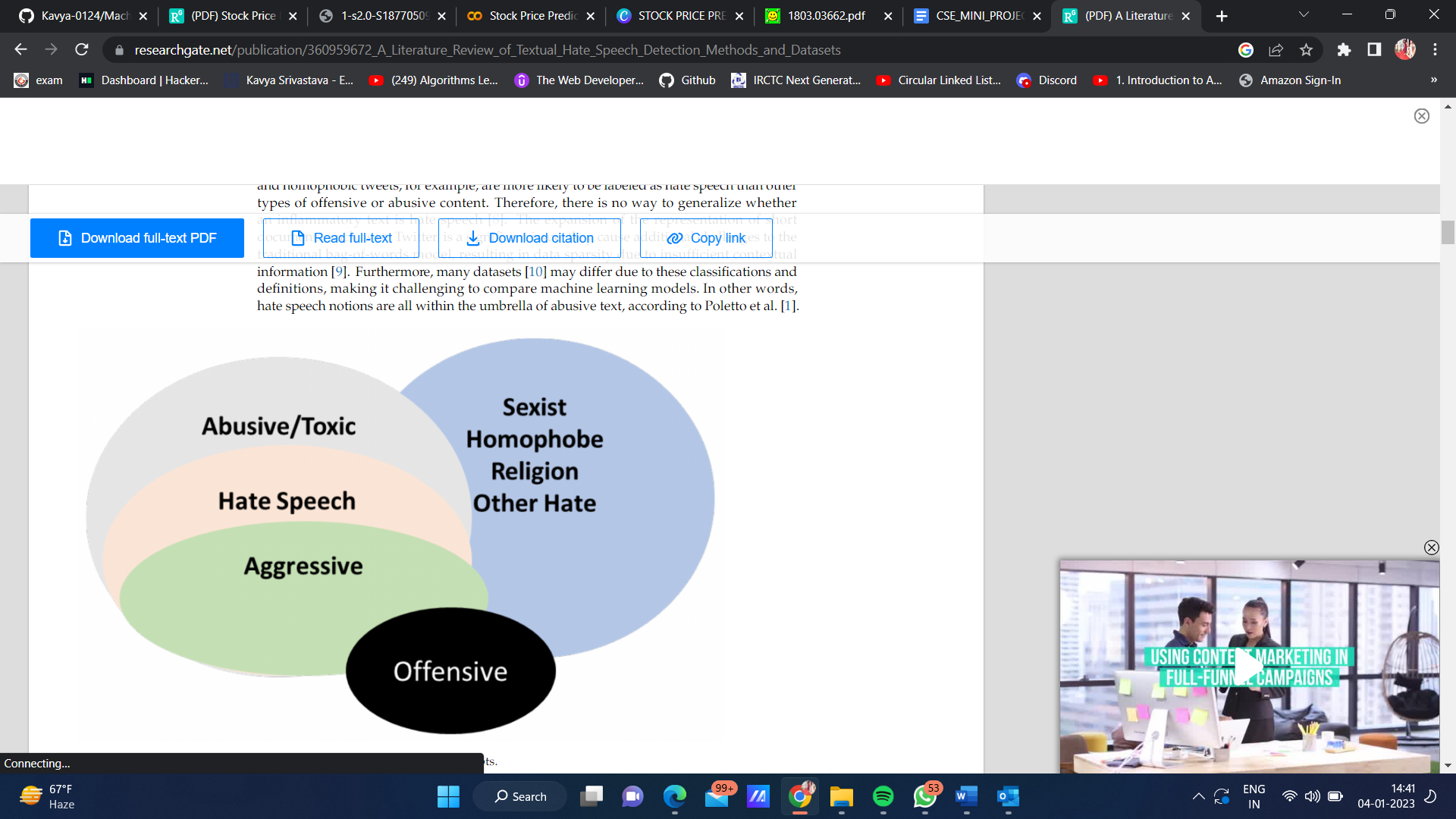
With the widespread use of social networks over the past few decades, people have become more involved. In-person and online hate speech has become more prevalent. The development and dissemination of hateful content, which eventually results in hate crimes, is greatly aided by social media and other internet platforms. The ability to express and exchange opinions widely and in real time was made possible by the development of microblogging programs. These expressions give academics the opportunity to look at the social emotions experienced online throughout various occasions. People may now communicate freely, which has enabled them to share a wide range of ideas, feelings, and information. Cyberspace is not always secure, and it sometimes serves as a venue for the distribution of offensive and dangerous material. Online expressions of prejudice and hostility frequently take the shape of hate speech. This could imply verbal aggression that is racial, xenophobic, or other. The act of disparaging a person or group of people because of one or more traits, such as race, ethnicity, sexual orientation, gender, religion, or nationality, is commonly referred to as hate speech. Social media sites have an uncontrolled volume of comments and posts posted every second, making it hard to monitor or manage the information on those platforms. Social media networks therefore struggle to control these posts while maintaining a healthy level of speech freedom. Additionally, the diversity of people, their histories, cultures, and worldviews can fan the flame of hate speech. On the other hand, every culture has its own distinctive definitions and traits of cyber-hatred. Therefore, it is anticipated that each culture would behave differently and have its own means of intervening in ways that are appropriate for that culture. Due to differences over various definitions of hate speech, detecting hate speech can be difficult.

Consequently, depending on their own definitions, some people may find certain things to be hateful while others may not.

"Content that advocates violence against individuals or groups based on race or ethnic origin, religion, handicap, gender, age, veteran status, and sexual orientation/gender identity”.

The suggested approaches used ML algorithms and various feature engineering techniques to categorize information as hate speech. Despite all this effort, it is still challenging to compare how well any method for classifying hate speech content performs. To the best of our knowledge, no research has been done that compare various feature engineering methods with ML algorithms.

**1.2 Problem Statement**

A lot of money has been invested recently by governments, businesses, and researchers because of the rising prevalence of hate speech on social media and the urgent need for efficient defenses. Automated hate speech detection techniques have been developed in a variety of ways. This tries to classify textual information into non-hate or hate speech, in which case the approach may additionally identify the targeting features (i.e., categories of hatred, such as race, and religion) in the hate speech. But we see a big difference between the two in terms of performance (i.e., non-hate vs. hate). In this work, we make the case for a practical focus on the latter issue. Our research of the language in the common datasets demonstrates that it is a considerably more difficult process since hate speech lacks distinctive, discriminative traits and is thus present in the dataset's "long tail" where it is more difficult to find.

**Chapter 2**

**Literature Survey**

On social media, hate speech refers to writing that denigrates the target and could endanger or cause harm to that person. It is a biased, hostile, and spiteful discourse that singles out a person or a group of people based on either their consciously held or unconsciously held intrinsic qualities. It is a style of speech that expresses a clear desire to injure others, incite conflict, or spread hatred. The social media and collaborative web environments provide a favorable setting for the creation, dissemination, and exchange of hate speech against a purported enemy group. The human process of identifying and removing hate speech articles or comments takes a lot of time and costs a lot of computation. Due to these problems and the availability of offensive information on social media, there is a solid argument in favor of hate speech detection.

Many algorithms, such as LSTMs and GRUs, are included in RNNs. LSTMs were created to solve the problem of vanishing gradients that can occur while training standard RNNs. The GRU is similar to long short-term memory (LSTM) with a forget gate but lacks an output gate. Hence, it has fewer parameters. Variations of LSTM were also used for many tasks of hate speech detection, such as hate ideologies, COVID-19 and the US election, the hate detection of hybrid CNN and RNN, HASOC LSTM models, Bi LSTM models, and generative pretrained transformer models. Some studies have added additional features to Twitter network analysis such as user proﬁles for hate speech detection. For example, Fountas et al. proposed two-layer RNNs. The uniﬁed model was built on tweet characteristics (the Glove technique) and metadata on persons, networks, and content. They used many datasets to test their model, including the cyberbullying dataset, the hateful dataset, the offensive dataset, the sarcasm dataset, and the abusive dataset. The model produced variable results depending on the input characteristics and dataset utilized; nevertheless, the RNN and metadata interleaved model were the best, with an average accuracy of 90.2. The dataset of Waseem and Ho was used by Fount et al.to detect abusive cyberbullying language on Twitter. Their model blends the LSTM architecture with social network analysis to detect hate speech sources. In order to train the LSTM model, they used Fast Text embeddings. An F1-score of 0.823 was reported. Corozal et al. used

the dataset of Waseem and Hoy with the following algorithms: LSTM, GRU, and Bi LSTM. N-grams of words; tweet features, such as emojis and emotion lexica; and social network-speciﬁc features are among the features employed. They reported an F1-score of

0.823; the best method was LSTM.

Bidirectional Encoder Representations from Transformers (BERTs) is a transformer- based machine learning technique for natural language processing which was pre-trained and developed by Google based on the knowledge extracted from text (vectors) using the surrounding text to establish the context. Several complex models that used BERT are discussed in the literature such as hateful meme challenges, the study of Ron Ahu et al.; ensembles of BERT; Yu’s model of knowledge enhanced vision-language representations; Facebook Hateful Memes (FHM); Liu et Aland BERT over Zamperini three tasks; Caselli et al. who re-trained the BERT for hate speech (HATEBERT), the Abu’s Eval, and the Hat Eval models; Nguyen et al. who employed the BERT model of Roberta by training on 80 GB of uncompressed texts of 850 M tweets (16B-word tokens), and BER Tweet which outperforms strong base-lines Roberta-base and XLM-R-base. Caselli et al. used the Reddit Abusive Language English dataset, which contains over one million Reddit messages (43 billion tokens). They compared the following datasets from Seme Val 2019: Task 6 Offense Eval 2019, the Abu’s Eval, and the Hat Eval. They found that the HATEBERT model was more efﬁcient than the datasets they evaluated, with a 5% increase in precision. Nguyen et al. used the Roberta BERT model to train 80 GB of raw texts from 850 M tweets (16 B-word tokens). It was found that BER Tweet surpasses the Roberta-base and XLM-R-base, which are both powerful baselines.

Consequently, RNNs and CNNs have commonly used machine learning methods in

deep learning methods: however, research has also shown that hybrid and complex models

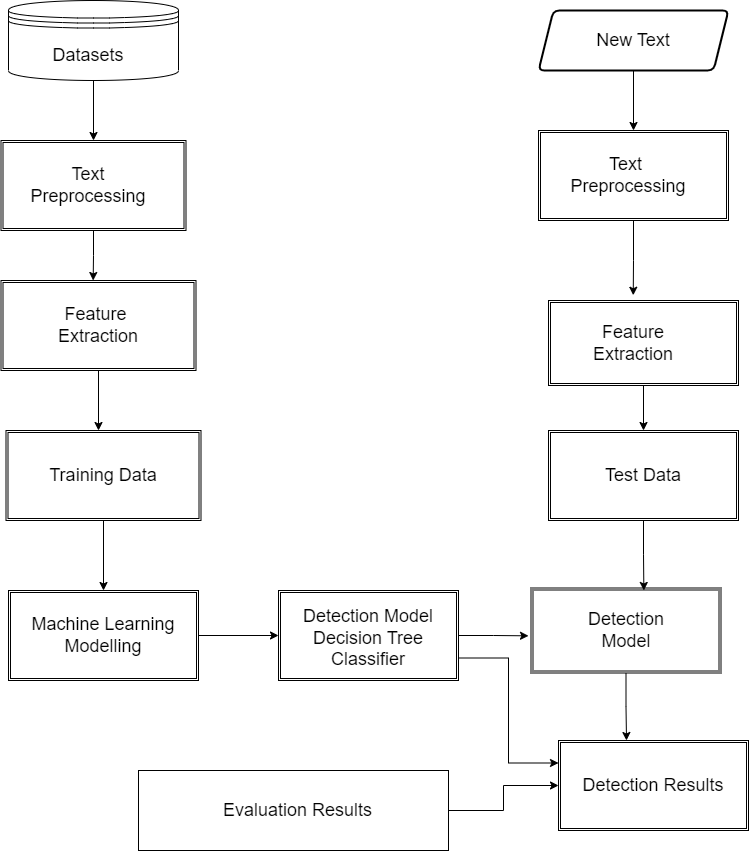
cover approximately 29% of studied models. For example, Jahan and Oussama showed that the BERT method covers approximately 33% of studies, followed by LSTM and CNN with 20%.

Hybrid methods combine more than one machine learning method to improve performance, including ensemble models. In addition, hybrid models are considered robust as they incorporate more than one data source, such as the metadata from Twitter. Pistils et al. tested an ensemble of RNN classiﬁers in the datasets of Waseem and Hoy. They used the word frequency vectors and the user sentiment towards each Information class of hate speech (neutral, racism, or sexism) based on user tweet history. The result showed that the model provided an F1-score of 0.93. Joslin et al. reported the best performance was when using a combination of LSTM + Random Embedding + Boosted Decision Trees (GBDTs) with an F1-score of 0.93 in the dataset of Waseem and Hoy and using Glove embeddings. Paschal ides et al. used the dataset of Davidson et al. to develop a new online Twitter, hate speech detection system called MANDOLA. The system is based on many ensembles of CNNs and RNNs that automatically learn abstract feature representations depending on many features: TF-IDF vectors and word embeddings. MANDOLA allows users to visualize online results. The approach reported a balanced accuracy of 0.770.

**Chapter 3**

**Methodology**

This section describes the suggested system we used to categorize tweets into three groups: "hate speech, offensive speech, and no hate and offensive. The whole technique is shown in Fig.1. Data collecting, data preprocessing, feature engineering, data splitting, classification model development, and classification model evaluation are the six main processes of the study technique as indicated in this picture. The following sections go into detail about each phase.

****

**Fig.1 Hate speech Detection Architecture**

It accepts bilingual data as input and pre-processes it in accordance with the characteristics of the languages, which includes removing punctuation, normalizing, tokenizing, and other fundamental pre-processes. After that, feature extraction uses TF-IDF to extract the feature. A key feature vector (training data) from the dataset used to train the model is the task's output. The models are trained using Decision Tree Classifier machine learning methods following feature extraction. These activities lead to the creation of a detection model for offensive and hateful speech. Based on the findings of the model assessment technique discussion, the detection model is assessed and chosen.

**About the dataset**

Dataset using Twitter data, it was used to research hate-speech detection. The text is classified as: hate-speech, offensive language, and neither. Due to the nature of the study, it’s important to note that this dataset contains text that can be considered racist, sexist, homophobic, or generally offensive.

**Pandas**

Pandas is an open-source library that is made mainly for working with relational or labeled data both easily and intuitively. It provides various data structures and operations for manipulating numerical data and time series. This library is built on top of the NumPy library. Pandas is fast and it has high performance & productivity for users. Pandas is an open-source, BSD-licensed Python library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language. Python with Pandas is used in a wide range of fields including academic and commercial domains including finance, economics, Statistics, analytics, etc. In this tutorial, we will learn the various features of Python Pandas and how to use them in practice.

**NumPy**

NumPy stands for numeric python which is a python package for the computation and processing of the multidimensional and single dimensional array elements. It is an extension module of Python which is mostly written in C. It provides various functions which can perform the numeric computations with a high speed. NumPy provides various powerful data structures, implementing multi-dimensional arrays and matrices. These data structures are used for the optimal computations regarding arrays and matrices.

**NLTK**

The Natural Language Toolkit, or more commonly NLTK, is a suite of [libraries](https://en.wikipedia.org/wiki/Library_(computer_science)) and programs for symbolic and statistical [natural language processing](https://en.wikipedia.org/wiki/Natural_language_processing) (NLP) for English written in the [Python programming language](https://en.wikipedia.org/wiki/Python_(programming_language)). NLTK has been used successfully as a teaching tool, as an individual study tool, and as a platform for prototyping and building research systems. NLTK (Natural Language Toolkit) Library is a suite that contains libraries and programs for statistical language processing. It is one of the most powerful NLP libraries, which contains packages to make machines understand human language and reply to it with an appropriate response.

**Count Vectorizer**

Count Vectorizer is a great tool provided by the scikit-learn library in Python. It is used to transform a given text into a vector based on the frequency (count) of each word that occurs in the entire Count Vectorizer creates a matrix in which each unique word is represented by a column of the matrix, and each text sample from the document is a row in the matrix. The value of each cell is nothing but the count of the word in that particular text sample text. This is helpful when we have multiple such texts, and we wish to convert each word in each text into vectors (for use in further text analysis).

**Decision Tree Classifier**

Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules, and each leaf node represents the outcome. One of the most popular classification approaches is decision tree learning. It is highly efficient and has classification accuracy comparable to other learning methods. A decision tree is a tree that reflects the classification model that has been learned. It's an easy-to-understand decision tree classification paradigm. The method evaluates all feasible data split tests and chooses the one with the highest information gain.

**Chapter 4**

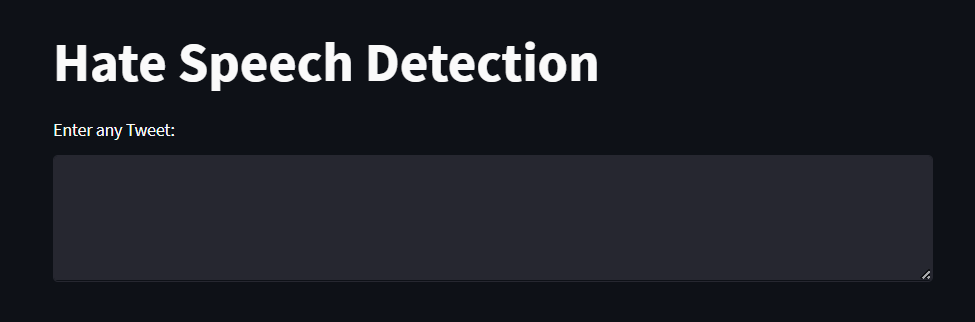
**Result and Discussion**

Hate speech detection is a complex process, partly due to the availability of the datasets

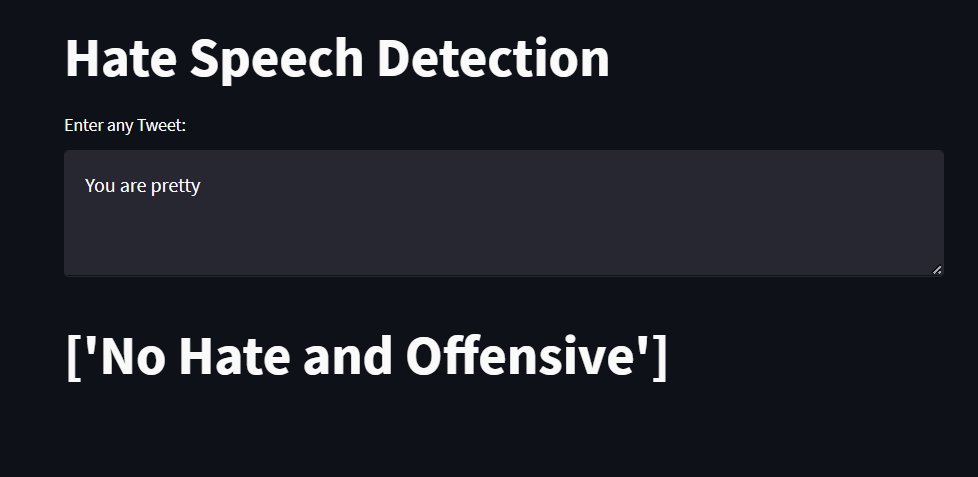
and the task of developing a machine learning model with rigid performance. This section describes the suggested system we used to categorize tweets into three groups: "hate speech, offensive speech, and no hate and offensive.

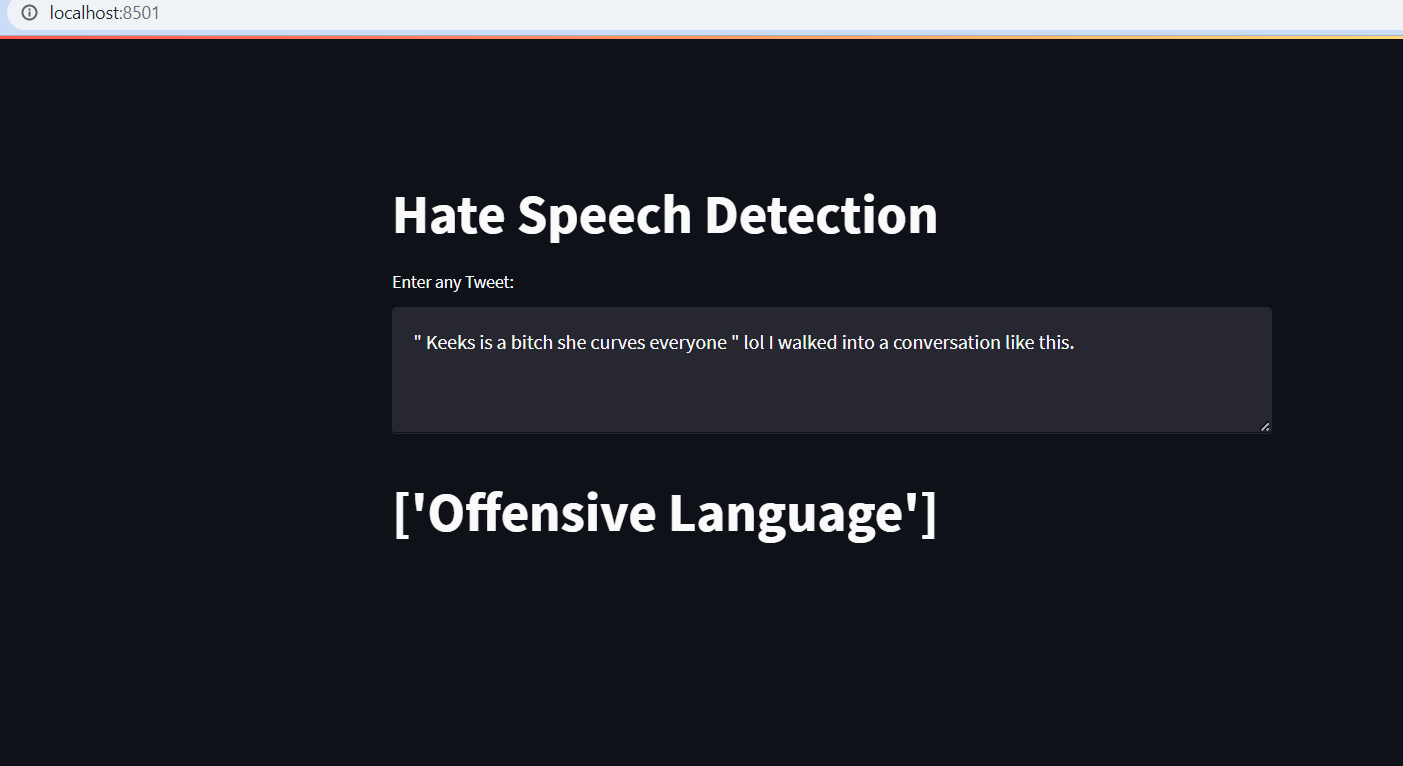
**GUI INTERFACE**

The snapshots below demonstrate the GUI Interface of the hate speech detection, which is divided into a section as mentioned below.



Here are the snapshots of the tweet we detected in our search.





Furthermore, detecting abusive language is problematic for various reasons, including word obfuscation, difﬁculties tracing racist and slurs, and professionally written and abusive language that crosses sentence boundaries. However, small datasets are insufﬁcient for generalizing conclusions or capturing features.

**Chapter 5**

**Conclusion and Future Work**

There is a critical dearth of reporting in the literature on the optimal set of features for hate speech detection that can be applied to both classical and deep learning models. Therefore, extensive research is needed to develop features that work well with diverse datasets with multifaceted hate speech concepts. A successful model should also have features that can be applied to new datasets and previously unseen tweets. A direction could be research, in which more features are added to develop additional features.

Aside from the basic hate/no hate categorization for traditional and deep learning models, the literature lacks a detailed investigation of ﬁne-grained hate speech detection at the label level. According to the studies gathered, there is still a gap in creating a model that successfully performs the multi-classiﬁcation of hate speech, has acceptable performance, and can be generalized across settings. A starting point could be using the models of, where several classes were adopted.

There are no recommendations in the literature to ensure that hate speech detection methods are adequately compared across different datasets. Therefore, a new method- ology for dataset comparison is needed so that datasets can be rigorously compared.

**Conclusions:**

New datasets of hate speech from different regions with various topics of hate speech and offensive content is constantly being developed. However, many datasets are small in size whilst others lack reliability due to how the data were collected and annotated. Additionally, many datasets are small or sparse, lacking linguistic variety. Above all, the language of the content and region where the data were collected from social media make the comparison between various hate speech detection models difﬁcult. One of the signiﬁcant challenges in hate speech detection is the architecture of the machine learning model and the lack of consensus on hate speech deﬁnition. It was reported that creating large and varied hate or abusive datasets that minimize potential bias is laborious and requires specialized experts. Although many machine learning models are developed, feature set selection should be ﬁne-grained at the label level, as well as in the generic hate/no hate classiﬁcation by developing the best set of features for hate speech detection, which is applicable to traditional and deep learning models.

**References**

[1] Zizi Zhang, Hate Speech Detection: A Solved Problem? The Challenging Case of Long Tail on Twitter **(https://arxiv.org/pdf/1803.03662.pdf)**

[2] Fatimah Ankomah 1, 2, \* and Xiaoguang Ma “A Literature Review of Textual Hate Speech Detection Methods and Datasets”

**(https://www.researchgate.net/publication/360959672\_A\_Literature\_Review\_of\_Textual\_Hate\_Speech\_Detection\_Methods\_and\_Datasets)**

[3] (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 11, No. 8, 2020

[4]Ankomah, F.; Ma, X. A Literature Review of Textual Hate Speech Detection Methods and

Datasets. Information 2022,13, 273.

[5] Chen, Y., Detecting offensive language in social media for protection of adolescent online safety. 2011.

[6] Singh, P.K. and M.S. Husain, Methodological study of opinion mining and sentiment analysis techniques. International Journal on Soft Computing, 2014.

[7] Sharma, S., S. Agrawal, and M. Shrivastava, Degree based classification of harmful speech using twitter data. arrive preprint arXiv:1806.04197, 2018.

[8] Natasha Lomas. Facebook, google, twitter commit hate speech action in germany, Last accessed: July 2017.