**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:**

**Harshvardhan Singh Bisht**  **University Roll No. 2018830**



**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 **Mrs. Garima Sharma, Assistant Professor**, Department of Computer Science and Engineering, Graphic Era (Deemed to be University), Dehradun.



**Harshvardhan Singh Bisht** **2018830** **signature**

**Table of Contents**

|  |  |  |
| --- | --- | --- |
| **Chapter No.** | **Description** | **Page No.** |
| Chapter 1 | Introduction | **1-2** |
| Chapter 2 | Literature Survey | **3-4** |
| Chapter 3 | Methodology | **5-8** |
| Chapter 4 | Result and Discussion | **9-10** |
| Chapter 5 | Conclusion and Future Work | **11** |
|  | References | **12** |

**Chapter 1**

**Introduction**

* 1. **Introduction**

Over the last decade, people have become more engaged with 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. People now have the potential and ability to speak freely at the comfort of their fingertips, and this allowed them to exchange all sorts of thoughts, opinions and knowledge. However, cyberspace is something that’s not always safe, therefore it can be a reason for the circulation of aggressive, harmful, and hateful content.

**1.2 Definition**

Hate speech is an online common form for expressing prejudice, hate, and aggression. It may convey racist, sexist, homophobic, and many other forms of verbal aggression.

According to Law commission report*, “Hate speech is an incitement to hatred primarily against a group of persons in terms of race, ethnicity, gender, sexual orientation, religious belief and the like.”*

In social media platforms, there are numerous uncontrollable number of comments and posts issued every second, which make it almost impossible to trace or control the content over the platform. Therefore, social media platforms are facing a problem in limiting these posts & comments while balancing the freedom of speech. In addition, the diversity of people and their backgrounds, cultures and beliefs can ignite the flame of hate speech. On the other hand, each culture has its own different interpretations of cyber-hate. So, every culture is assumed to act differently on Hate Speech, and have their own way of intervention in a manner which best suits their culture. [2]

A picture containing text, sign, book, red

Description automatically generated

**Figure 1.1** Poster against hate speech

The growth of hateful content online has been tremendously increased, alongside the rise of easily shareable disinformation, supported by digital tools. This raises many challenges for our societies, as governments struggle to enforce national laws in the virtual world's growth scale and speed.

Unlike in traditional media, online hate speech can be produced as well as shared easily, at low cost and anonymously. It has the potential to reach a global and diverse audience in no time. The permanence of hateful & aggressive online content is also problematic, as it can resurface and (re)gain popularity over time in the upcoming years.

Understanding, filtering, and monitoring hate speech across diverse online communities and platforms is key to making social places safe for everyone. But efforts are often stunted by the technological advancements, lack of automated monitoring systems, and the lack of consideration in people. [5]

**1.3 Problem Statement**

We must design a Machine Learning based software/application that helps us detect and filter Hate Speech over online social platforms.

**Chapter 2**

**Literature Survey**

**2.1 Books -**

**2.1.1** ***“Viral Hate: Containing Its Spread on the Internet”***

***book by* Abraham H. Foxman**

- provides a great and clear view of what hate speech really is and what can be its consequences. It depicts the ways how hate speech can affect lives of individuals and is spread uncontrollably across the internet and numerous social platforms.

**2.1.2 *“Python GUI Programming with Tkinter”***

***book by*** **Alan D. Moore**

- provides a deep and simple understanding of the Tkinter library in python used to create attractive Graphical User Interfaces, which can be used in a variety of python-based applications to make applications attractive and user-friendly.

**2.1.3 “*INTRODUCTION TO MACHINE LEARNING*”**

***book by Nils J. Nilsson, Department of Computer Science, Stanford University***

***-*** provides the details on latest developmentsand advancements in the field of Machine Learning & Deep Learning, the technological advancements, and about the fore coming future works in this field.

**2.2 Research Papers-**

**2.2.1 *“Hate speech detection: Challenges and solutions”***

***research article by Sean* MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, Ophir Frieder**

- provides an overview of development of softwares and techniques to detect hate speech on social media platforms, the challenges faced while detecting and monitoring hate content, and the solutions to the issue.

**2.2.2 *“A Research on Machine Learning Methods and Its Applications”***

**research article by Özer Çelik**

- provides a good understanding of how Machine Learning works and how it can be implemented to perform tasks impossible for human to perform efficiently, and many other applications of Machine Learning models.

**2.2.3 *“Deep Learning for Hate Speech Detection in Tweets”***

**research article by P. Badjatiya, Shashank Gupta, Manish Gupta, & Vasudeva Varma**

- provides an insight of the use of machine learning and deep learning methodologies and toolkits to detect Hate Speech in various online platforms such as Twitter, Facebook, etc.

**Chapter 3**

**Methodology**

**3.1 Python Libraries (for Machine Learning) -**

**3.1.1 pandas**

Python library used for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series.

**3.1.2 numpy**

Python library used for working with arrays, domains of linear algebra and metrices. It stands for “Numerical Python”.

**3.1.3 sklearn**

Scikit Learn is a library used for implementing Machine Learning and predictive data analysis.

*CountVectorizer* is used to transform a given text into a vector based on the frequency (count) of each word that occurs in the entire text. [6]

*train\_test\_split* separates data into train and test types. Model is trained using X-train & Y-train and fitted using X-test & Y-test fields.

*DecisionTreeClassifier* is a supervised machine learning algorithm. It uses a set of principles to make judgements, just like humans do, by learning simple decision rules. [6]

**3.1.4 re**

Python library that allows us to check if a particular string matches a given regular expression.

**3.1.5. nltk**

NLTK toolkit is used to deal with NLP. It offers us numerous test data, datasets, and different text processing libraries.

*Snowball Stemmer* is a stemming algorithm. Stemming is the process of reducing inflection in words to their root forms such as mapping a group of words to the same stem even if the stem itself is not a valid word in the Language. [3]

*Stopwords* is a part of NLTK toolkit. Stop words are common words like ‘the’, ‘and’ ‘I’, etc. that are very frequent in text, and so don’t convey insights into the specific topic of a document. We can remove these stop words from the text in a given corpus to clean up the data, and identify words that are more rare and potentially more relevant to what we’re interested in. [3]

**3.2 Python Libraries (for GUI) -**

**3.2.1 Tkinter**

Tkinter is a vast Python library used to create graphic based user interfaces which makes basic python applications interactive and appealing to the user.

*messagebox* is a sub-library of Tkinter used to display a message box/alert box on the interface screen.

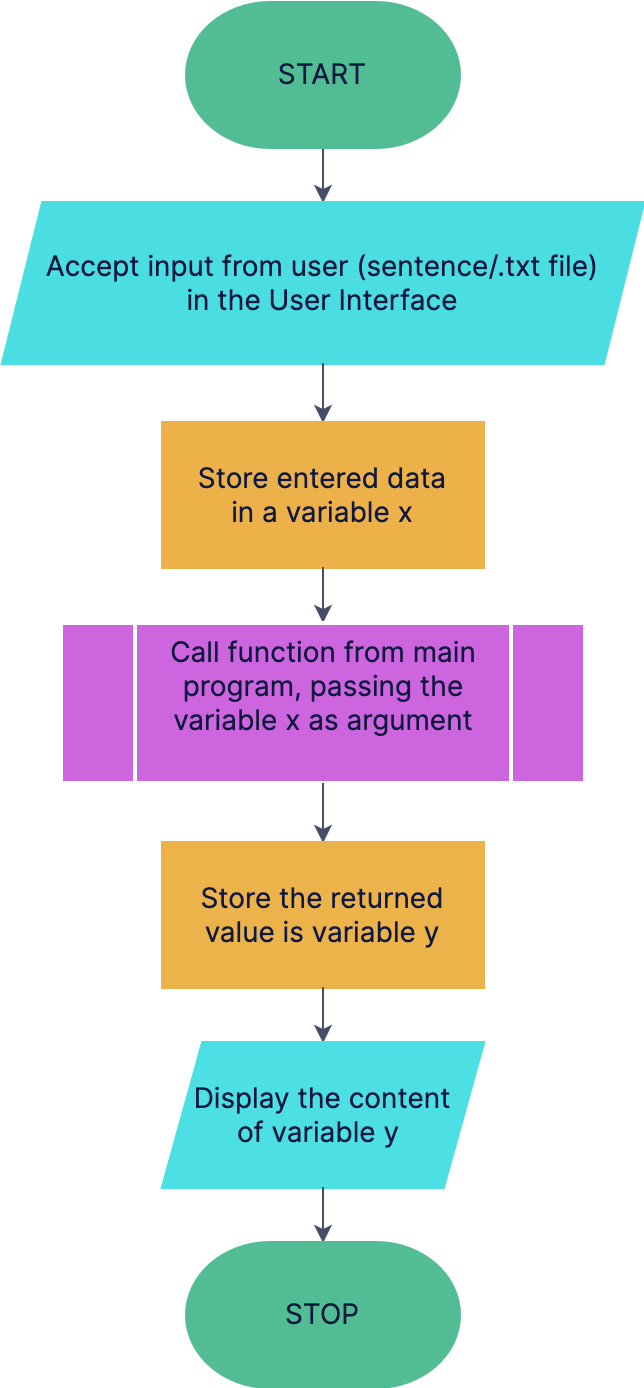
*filedialog* is a sub-library of Tkinter used to browse files in the user system and work on them.[1]

**3.2.2 PIL**

PIL or Pillow is a Python imaging library that adds image processing capabilities to your Python interpreter. [1]

*Image / ImageTk* is a sub-library of Pillow used to open images of various extension types into your interface created using Tkinter.

**3.3 Flowchart depicting the working of program -**

****

**3.3 Flowchart depicting the working of the called function -**

**Shape, polygon

Description automatically generated**

Set A is a set containing a collection of hate words.

Set B is a set containing a collection of offensive words.

**Chapter 4**

**Result and Discussion**

**4.1 Result**

The program is created using Python, implementing the concept of Machine Learning, with various Libraries being used to detect Hate Speech in a message. This software can be used anywhere across multiple social media platforms. With the tremendous rise of social media and the concept of Freedom of Speech being used, the uses of Hate Speech Detection are endless.

The program is accompanied by as graphical user interface which makes it appealing to the eye of the user. It accepts input in form of input of text. The user can also browse a .txt file in their system to be read and scanned for any present Hate Speech.

The program distributes the scanned data into three categories:

1. Hate Speech
2. Offensive Speech
3. No Hate or Offensive Speech

The input is scanned by comparing the data with a dataset stored as a .csv file, which has an enormous collection of tweets posted by various users, which are classified as Hate, Offense, or None. Machine Learning model helps the program compare the input with the dataset for any present Hate Speech.

**4.2 Trial & Testing**

This program was tried as tested with many user inputs, giving an output with working accuracy of over 91%, with the exceptions being the newer hate words/slangs which came into use recently and are not present in older datasets.

It was also tested with 3 other datasets with slight changes in the program, to which it gave an accuracy of over 88%, showing that it can work with a newer and updated dataset as well.

**Chart, bar chart

Description automatically generated**

**Figure 4.1 - Graph depicting accuracy over various datasets**

**Chapter 5**

**Conclusion and Future Work**

**5.1 Conclusion**

The project was completed in a duration of 2 weeks. It provided a deep insight on Hate Speech and its effect on people. The project also helped in getting efficient in making programs using machine learning tools and creating attractive user interfaces using various python libraries.

This project can be implemented in numerous applications, across social media. It has great working efficiency and accuracy and was tested multiple times with hundreds of different test cases and other datasets as well. It can come to great use in these times of rapid development and growth of social media and the internet.

**5.1.1 Applications**

- can be used on Twitter to filter hate tweets & comments.

- can be used on other Social Media networks to detect and stop hate comments.

- can filter hate comments on YouTube videos.

- can be used to filter unnecessary hate remarks over educational websites.

- can be used to filter hate comments on college websites.

**5.2 Future Work**

In the fore coming future, many advancements will be made in the field of Machine Learning and Artificial Intelligence, which will make it quite easy to create applications to detect Hate Speech and solve many other problems & issues as well.

There might be new hate words that come into practice, so the datasets must be frequently updated to keep the software working efficiently.

Hate speech is something that cannot be put an end upon, as social media is only going to rise in the upcoming future, but the least we can do is stop it from reaching other people by filtering everything beforehand. Future applications and toolkits will be powerful enough to stay updated with new hate words and filter everything out easily on their own using AI.

Research can be directed towards creating environments online that have zero tolerance policies against hate speech and offensive language on any basis.

**References**

[1] Alan D. Moore, “*Python GUI Programming with Tkinter*”, 1st ED, Packt Publishing, May 2018, Book on Python GUI applications.

[2] Wikipedia, “*Hate Speech*”, Article on Google published on June 2021.

[3] Wes McKinney, “*Python for Data Analysis*”, Book on Python data analysis.

[4] **“***Deep Learning for Hate Speech Detection: A Large-scale Empirical Evaluation*”. Accessed on 22nd January 2023: (<https://towardsdatascience.com/deep-learning-for-hate-speech-detection-a-large-scale-empirical-evaluation-92831ded6bb6>)

[5] Singh Malik, Jitendra, Guansong Pang, and Anton van den Hengel, “*Deep Learning for Hate Speech Detection: A Comparative Study*”, Research Paper (e-print), May 2022.

[6] Aurélien Géron, “*Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*”, Published Nov 2020, Book on Machine Learning toolkits.