**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:**

**Name: Gauri Gupta**  **University Roll No. : 2018809**

***Under the Mentorship of***

**Mrs. Garima Sharma**

**Asst. 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”** 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, Asst. Professor**, Department of Computer Science and Engineering, Graphic Era (Deemed to be University), Dehradun.

Name: Gauri Gupta University Roll no: 2108809

**Table of Contents**

|  |  |  |
| --- | --- | --- |
| **Chapter No.** | **Description** | **Page No.** |
| Chapter 1 | Introduction | 4-6 |
| Chapter 2 | Literature Survey | **7-**8 |
| Chapter 3 | Methodology | 9-10 |
| Chapter 4 | Result and Discussion | 11 |
| Chapter 5 | Conclusion and Future Work | 12 |
|  | References | 13 |

**Chapter 1**

**Introduction**

**1.1 What is Hate Speech?**

Hate speech can be defined as "public speech that demonstrates hate or encourages violence towards a person or group grounded on commodity such as race, religion, sex, or sexual exposure”. Hate speech is "generally allowed to comprise dispatches of animosity or dimishment of an individual or a group on account of a group characteristic such as race, color, national origin, sex, disability, religion, or sexual orientation. Legal delineations of hate speech vary from country to country.

Susan Brison (1998a) offers a disjunctive description that centers on a kind of abuse to targets. She defines detest speech as “speech that vilifies individualities or groups grounded on attributes such as race, sex, ethnicity, religion, and sexual orientation, which (1) constitutes face-to-face condemnation, (2) creates a hostile or intimidating environment, or (3) is a kind of group calumniation” (313). ‘Harm’ as used by Brison mentions that what Joel Feinberg describes “as a wrongful setback to (or invasion of) someone’s interests” (Brison, 1998b, 42).

**1.2 Harms off Hate Speech**

In recent times, the world has witnessed civil mass atrocities. In numerous of these cases detest speech was identified as a "precursor to atrocity crime including genocide". While the use of social media and digital platforms to spread abomination is fairy recent, the weaponization of public discourse for political gain is unfortunately not new. As history continues to show, detest speech coupled with this disinformation can lead to stigmatization, discrimination, and large-scale violence. The damages that have been attributed to hate speech comprise a long and varied list, ranging from the immediate psychological damages endured at the moment by the person(s) targeted by an instance of hate speech to much more long-term impacts that affect not only those targeted but whole communities, and indeed the strength of an entire nation.

**1.3 Is all ‘Hate Speech’ the same?**

“Hate speech” can range from ignorant discriminatory commentary and obnoxious jokes to unequivocal calls for demarcation against a group, or at worst, calling for mass murder. International human rights law singles out specific categories of severe “hate speech” that governments must enjoin. This includes speech that actively promotes discriminative abomination in such a way that it incites the audience to take action to harm the targeted group, simply because of who they are. That harmful action might be violence, demarcation, or another hostile act. If some “hate speech” is merely insulting or offensive, and other “hate speech” incites people to murder, it is important to accurately identify what kind of “hate speech” we are dealing with in any particular script. To identify appropriate and effective responses to hate speech, we need to understand the root causes of hate first.

**1.4 Hate Speech Laws in India**

The [hate speech](https://en.wikipedia.org/wiki/Hate_speech) laws in [India](https://en.wikipedia.org/wiki/India) aim to turn aside conflict among its many ethnic and religious communities. The laws permit a citizen to seek the punishment of anyone who shows the citizen disrespect "on grounds of religion, race, place of birth, residence, language, caste or any other ground whatsoever". Section 153A of the Indian penal code prohibits citizens from generating disharmony or feelings of enmity, hatred, or ill will between different groups of people.

**1.4.1 Section 153(A)**

Section 153A of the Indian correctional code says:

Whoever (a) by words, either spoken or written, or by signs or by visible representations or otherwise, promotes or attempts to promote, on grounds of religion, race, place of birth, hearthstone, language, caste or community, or any other ground whatsoever, discord or passion of enmity, hatred or ill-will between different religious, racial, language or regional groups or castes or communities, or (b) commits any act which is prejudicial to the maintenance of harmony between different religious, racial, language or regional groups or castes or communities, and which disturbs or is likely to disturb the public tranquillity, shall be penalized with imprisonment which may extend to three times, or with fine, or with both.

**1.4.2 Section 295(A)**

[Section 295(A) of the Indian Penal Code (IPC)](https://en.wikipedia.org/wiki/Section_295A_of_the_Indian_Penal_Code) enacted in 1927 says:

Whoever, with deliberate and malicious intention of outraging the religious feelings of any class of [citizens of India], [by words, either spoken or written, or by signs or by visible representations or otherwise], insults or attempts to insult the religion or the religious beliefs of that class, shall be punished with imprisonment of either description for a term which may extend to [three years], or with fine, or with both.

**Chapter 2**

**Literature Survey**

**2.1 Related Work**

Recent years we have seen an increasing number of research on hate speech detection as well as other related areas. As a result, the term ‘hate speech’ is often seen to co-exist or become mixed with other terms such as ‘offensive’, ‘profane’, ‘abusive languages’, and ‘cyber bullying’. To distinguish them, we identify that hate speech 1) targets individuals or groups based on their characteristics; 2) demonstrates a clear intention to incite harm, or to promote hatred; 3) may or may not use offensive or profane expression. The Internet enables the access and sharing of information at an unprecedented rate. This potential combined with the opportunity to remain anonymous also makes it an effective vehicle for the spread of hateful or offensive content. Because of this, many researchers have examined the task of automatically detecting this phenomenon, and there were also many competitions dedicated to solving problems like that (e.g. detecting insults , aggression , or hateful/offensive content in German, Spanish, or English ). A fundamentals approach is using simple templates (e.g. I [intensity] [userintent] [hate target]) or keywords). An additional popular approach is the combination of feature extraction and classical machine learning algorithms.

**2.2 Challenges of Detecting Hateful and Offensive Speech**

There are many layers to the difficulty of automatically detecting hateful and/or offensive speech, particularly in social media. Some of these difficulties are closely related to the shortcomings of keyword-based approaches. For one, words can be obfuscated in many different ways, both in an intentional attempt to avoid automatic content moderation or as a consequence of the use of social media for communication .

**2.3 Approaches for Detecting Hateful and Offensive Speech**

An introductory approach for relating hate speech is using a keyword-grounded approach. By using an ontology or wordbook, textbook that contain potentially spiteful keywords are identified. For instance, Hate base maintains a database of derogatory terms for numerous groups across 95 languages. Fresh information from social media can help further understand the characteristics of the posts and potentially lead to a better identification approach. Information similar as demographics of the posting user, position, timestamp, or indeed social engagement on the platform can all give further understanding of the post in different granularity.

However, this information is not often frequently available to external experimenters as publishing data with sensitive user information raises privacy issues. External researchers might at most have part or even none of the user information. Therefore, they conceivably break the wrong mystification or learn based on wrong knowledge from the data. For case, a system trained on these data might naturally poison towards flagging content by certain druggers or groups as hate speech based on incidental dataset characteristics.

**Chapter 3**

**Methodology**

**3.1 Procedure for Hate Speech Detection**

This study centered on machine learning classiﬁcation models hate speech datasets,

and the most signiﬁcant quality of hate speech models in light of such a widespread phenomenon.

The dataset I’m using for the hate speech detection mini project task is downloaded from Kaggle. The Dataset using Twitter data 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. This dataset contains the following columns:

1. Index: This column has the index value
2. Count: It has the number of users coded for each tweet.
3. Hate speech: This column has the number of users who judged the tweet to be hate speech.
4. Offensive language: This column has the number of users who judged the tweet to be offensive speech.
5. neither: This column has the number of users who judged the tweet to be neither hate speech nor offensive speech.
6. class: It has a class label for the majority of the users, in which 0 denotes hate speech, 1 means offensive speech and 2 denotes neither of them.
7. tweet: This column has the Text tweets.

**3.2 Importing the required libraries-**

Python libraries made it very easy for us to handle the data and perform typical and complex tasks with a single line of code.

[**Pandas**](https://www.geeksforgeeks.org/python-pandas-dataframe/)– This library helps us to load the data frame in a 2D array format and has multiple functions to perform analysis tasks in one go.

[**Numpy**](https://www.geeksforgeeks.org/python-numpy/)– Numpy arrays are quiet fast and can perform large computations in a very short time.

[**Matplotlib**](https://www.geeksforgeeks.org/matplotlib-tutorial/)/[**Seaborn**](https://www.geeksforgeeks.org/introduction-to-seaborn-python/)/**[Wordcloud](https://www.geeksforgeeks.org/generating-word-cloud-python/)–** This library is used to extract visualizations.

**Nltk-** – Natural Language Tool Kit provides various functions to exercise the raw textual data.This Library is used for symbolic and statistical natural language processing for english written in the Python programming language.

**Chapter 4**

**Result and Discussion**

The model that I created resulted in an accuracy of **86.4%**.

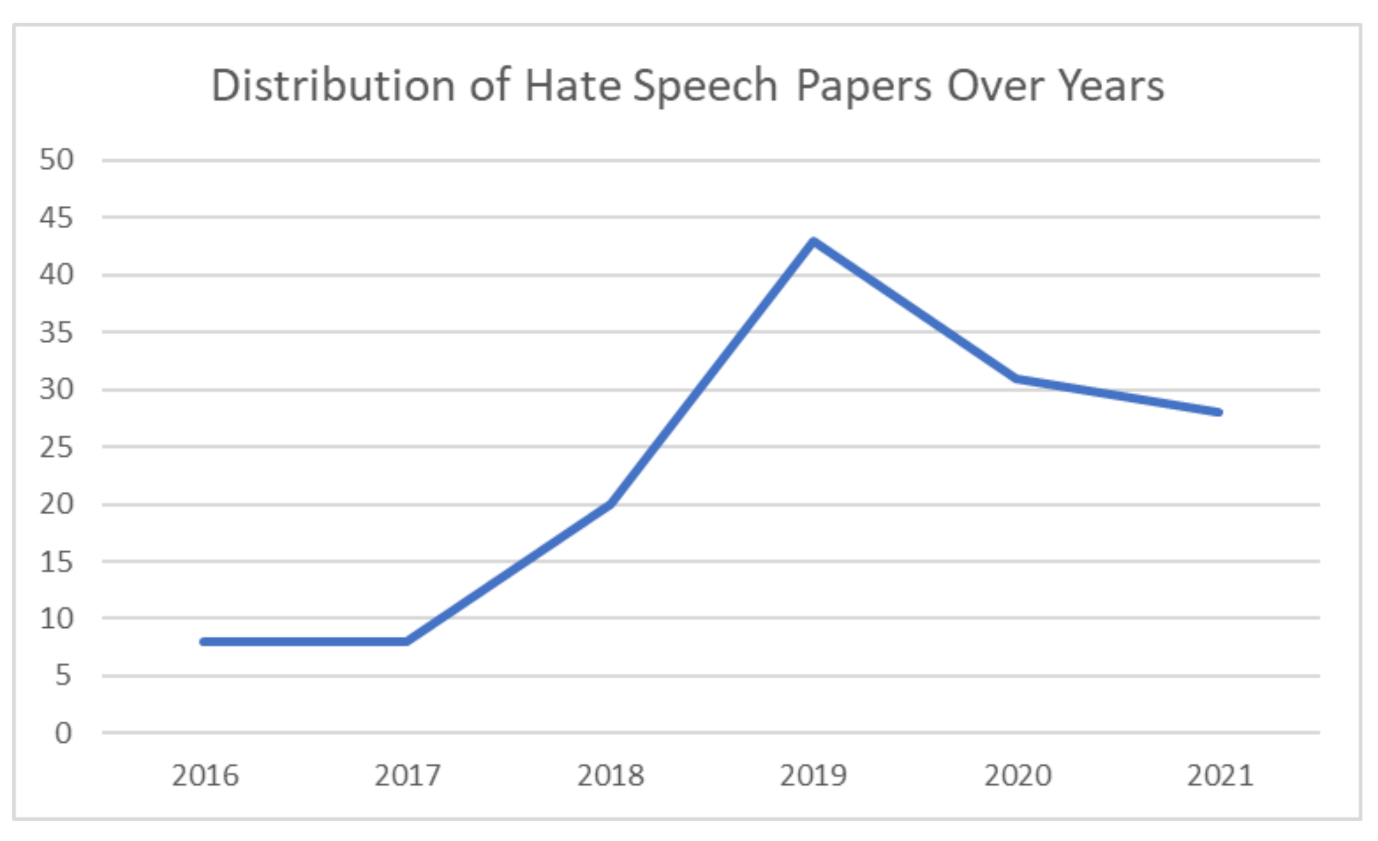


Fig 4.1: Distribution of Hate Speech Papers over Years

Figure 4.1 shows an increase in hate speech detection over the years.

The results show features and classifiers performed well for three classes (i.e. hate speech ,offensive speech, and neither hate speech nor offensive speech). For illustration , bigrams like “ lame nigga, white trash, whine made ” are more constantly appearing in class “ Offensive but not detest Speech ” as compared to class “ detest Speech ”. Hence, it might be possible that the classifier learned weak literacy rules.

**Chapter 5**

**Conclusion and Future Work**

**5.1 Conclusion**

 We have construct a project for **Hate Speech detection using Machine Learning**.  The survey work is mainly on Deep Learning and Hybrid approaches for hate speech identification, including significant activities in those fields Taking limited and public datasets for training the hate speech detection model is one of the limitations found, and the model can be improved by using real-time big data sets.  Hate speech is one of the significant issues we see on social media platforms like Facebook and Twitter. Our techniques and experiments will only address hate speech, due to both dataset availability and the goal of this work.

**5.1 Future Work**

We aim to explore the following directions of research in the future:

First, we will explore other branches of styles that aim at compensating the lack of training data in supervised literacy tasks. styles similar as transfer literacy could be potentially promising, as they study the problem of conforming supervised models trained in a resource-rich environment to a resource- dread environment. We will probe, for illustration, whether features discovered from one hate class can be transferred to another, therefore enhancing the training of each other.

Second, as shown in our data analysis as well as error analysis, the presence of abstract generalities similar as ‘ sexism ’, ‘ racism ’ or indeed ‘ detest ’ in general is veritably delicate to descry if solely grounded on verbal content. thus, we see the need to go beyond pure textbook bracket and explore possibilities to model and integrate features about druggies, social groups, collective communication and indeed background knowledge(e.g., generalities expressed from tweets) decoded in being semantic knowledge bases.

**References**

[1] <https://en.wikipedia.org/wiki/Hate_speech>

[2]ttps://en.wikipedia.org/wiki/Hate\_speech\_laws\_in\_India#:~:text=Section%20153A%20of%20the%20Indian,between%20different%20groups%20of%20people.

[3] <https://link.springer.com/article/10.1007/s42979-021-00457-3#Sec14>

[4] https://arxiv.org/pdf/1803.03662.pdf

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

[6] <https://www.mdpi.com/2078-2489/13/6/273>

[7] <https://arxiv.org/pdf/1803.03662.pdf>

[8]<https://www.researchgate.net/publication/331433247_DETECTION_OF_HATE_SPEECH_IN_SOCIAL_NETWORKS_A_SURVEY_ON_MULTILINGUAL_CORPUS>

[9] <https://vitalflux.com/hate-speech-detection-using-machine-learning/>

[10]https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0221152#:~:text=A%20basic%20approach%20for%20identifying,many%20groups%20across%2095%20languages.