**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:**

**Student Name**  **University Roll No.**

**Navpreet Singh 2017167**

***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”** 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.

Name University Roll no **signature**

Navpreet singh 2017167

**Table of Contents**

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

**Chapter 1**

**Introduction**

In the following sections, a brief introduction and the problem statement for the work has been included.

* 1. **Introduction**

When it comes to online interaction, a person can communicate with freedom of speech. But it also led to harassment or bullying to a certain individual.

In simple terms, hate speech refers to any speech, conduct, writing, or expression that may incite violence or prejudicial action against or by a particular individual or group, or because it disparages or intimidates a particular individual or group. Hate speech can take many forms, including name-calling, slurs, or other derogatory language or behavior. It can be directed at individuals or groups based on their race, ethnicity, national origin, religion, sexual orientation, gender identity, or other characteristics. Hate speech can have serious consequences for those targeted by it, as it can lead to social ostracism, discrimination, and even violence. It is important to recognize and actively oppose hate speech, as it goes against the values of a respectful and inclusive society.

**1.2 Causes of hate speech**

There are many potential causes of hate speech, and the reasons why individuals engage in hate speech can vary. Some people may engage in hate speech as a way to express their own feelings of anger, frustration, or resentment towards a particular group of people. Others may engage in hate speech as a way to gain power or attention, or to align themselves with a certain ideology or belief system. In some cases, hate speech may be motivated by a desire to spread fear or to intimidate certain individuals or groups. It is important to recognize that hate speech can be a symptom of deeper social, cultural, or psychological issues, and addressing the root causes of hate speech is often necessary in order to effectively combat it.

**1.3 potential consequences of hate speech**

Hate speech can have serious consequences for individuals and communities. It can contribute to a culture of intolerance and discrimination, and can lead to violence and other harmful actions. Hate speech can also have a negative impact on the mental health and well-being of those who are targeted by it, as it can cause feelings of fear, anxiety, and isolation. In some cases, hate speech can even lead to physical harm or death. It is important to recognize the potential consequences of hate speech, and to take steps to address and combat it in order to create a more inclusive and respectful society.

**1.4 importance of recognizing and addressing hate speech**

Recognizing and addressing hate speech is important for several reasons. First and foremost, hate speech can have serious consequences for individuals and communities, as it can contribute to a culture of intolerance and discrimination, and can lead to violence and other harmful actions. Addressing hate speech is therefore necessary in order to protect the safety and well-being of those who may be targeted by it.

In addition, hate speech can undermine the principles of respect, equality, and inclusion that are vital to a healthy and functioning society. By recognizing and addressing hate speech, we can work towards creating a more inclusive and respectful society, where all individuals are treated with dignity and respect, regardless of their race, ethnicity, religion, gender, sexual orientation, or any other characteristic.

Finally, addressing hate speech can also help to promote dialogue and understanding between different groups, and can foster a sense of community and belonging. By working together to combat hate speech, we can help to create a more cohesive and harmonious society.

**Chapter 2**

**Literature Survey**

In the previous chapter, we saw that, according to the definitions of hate speech, its targets are groups or individuals based on their specific attributes, such as origin, religion, disability, gender identity, age, veteran status, sexual orientation or other. Studies have been conducted with the goal of describing online hate speech and which groups are more threatened. This section presents the main conclusions found on the articles that we labeled as having a more descriptive approach to the problem of hate speech detection. We found descriptive articles about Racism, Sexism, Prejudice toward refugees, Homophobia, and General hate speech.

**2.1 categories**

1. **Racism**

In one study [1], the authors tried to understand when hate speech occurs and why

messages in social networks are cataloged as racist. They concluded that in the majority of the cases (86%) this is because of the “presence of offensive words.” Other motives are “references to painful historical contexts” and “presence of stereotypes or threatening.” The authors of another study [2] describe racism across the United States and tried to understand the geographic distribution of racist tweets. They used information gathered in Twitter to describe the frequencies of tweets in the several states, using the geographic reference of the messages.

1. **Sexism**

In a study [3] about sexism, a very simplistic approach was conducted. Tweets using offensive words toward woman were collected using the Twitter search API. Approximately 5,500 tweets were gathered and coded by one researcher, using a simple binary model. Despite the limitations of the study (e.g., many of the tweets were repeating the title or lyrics from popular songs that included the searched offensive words), it was still relevant for understanding that offensive communication toward woman is a reality in Twitter. A second study also describes misogynistic language on Twitter [4]. The main conclusions were that 100,000 instances of the word rape used in UK-based Twitter accounts were found, from which around 12% appeared to be threatening. Moreover, approximately 29% of the rape tweets appeared to use the term in a casual or metaphorical way. However, this study also points out that women are as almost as likely as men to use offensive terms against women on Twitter.

1. **Prejudice toward refugees**

Another study [5] was focused on the annotation of a dataset in German for hate speech against refugees. The main goal of this study was to point out the difficulties and challenges when annotating a dataset.

1. **Homophobia**

Some other study [6] , using an ethnographic methodology, was conducted in Africa. Data was collected from several sources (e.g., newspapers, sites) to conclude that homophobic discourses were using arguments related with Abnormality, Xenophobia, Racism, Barbarism, Immorality, Unpatriotism, Heterosexism, AntiChristianity, UnAfrican, Animalistic behaviour, Inhumane, Criminality, Pathology, and Satanism.

1. **General hate speech**

Finally, other studies take into consideration several types of hate speech at the same time. In one particular case [7], two social networks (Twitter and Whisper) were crawled with expressions that follow a rigid pattern: I < intensity >< userintent >< hatetarget >. One message following this pattern would be “I really hate people.” After collecting the messages, the researchers tried to infer the target of hate in the tweets. With this method, they concluded that “race,” “behavior,” and “physical” were the most hated categories. Finally, an analysis of data recorded by the FBI in 2019 [8] for victims in the USA of single-bias hate crime incidents, showed that the offender’s bias was toward different targets in different proportions (Figure 1).

**Chart

Description automatically generated**

**Fig.1**

**Chapter 3**

**Methodology**

1. **Development environment**

When it comes to development environment, we can use jupyter notebook and google colab. But for this particular project, we are using jupyter notebook within anaconda navigator (it provides tools for data science and analysis which include jupyter).

1. **Collecting data**

As the data should be in textual form, so we will choose the data represented by twitter( as it is among those platform where hate speech is on peek.) The dataset for building our hate speech detection model is available on [www.kaggle.com](http://www.kaggle.com). The dataset consists of Twitter hate speech detection data, used to research hate-speech detection. The text in the data is classified as hate speech, offensive language, and neither.

1. **Creating project**

Create project is mainly divided into many steps-

1. **Libraries**

For this we are including pandas(it is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool, built on top of the Python programming language), numpy (Numerical Python and it is a core scientific computing library in Python), scikit-learn(Simple and efficient tools for predictive data analysis),NLTK ( The Natural Language Toolkit) library, used for symbolic and statistical natural language processing for many languages written in the Python programming language.

1. **Preprocessing of data**

In Data preprocessing, we prepare the raw data and make it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model. When creating a machine learning project, it is not always a case that we come across clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put it in a formatted way.

1. **Splitting of data**

The next important step is to explore the dataset and divide the dataset into training and testing data.

1. **Building the model**

After segregating the data, we must work apply the algorithm which is best suited, and we have chosen for it.

1. **Evaluating the results**

In this step, we measure the performance of the model we built.

**Chapter 4**

**Result and Discussion**

1. **Accuracy**

The accuracy of the model will be determined by scikit-learn library we included.

The library we are using is “from sklearn. metrics import accuracy\_score”

In my model it is 0.8911847414109304

1. **Detections**

****

****

****

Fig-2

in the fig-2, we can see it detects hate speech.

**Chapter 5**

**Conclusion and Future Work**

1. **Conclusion**

In this project, we developed a hate speech detection model using Python and several machine learning algorithms. Our model was able to achieve a high level of accuracy in detecting hate speech in online text. We also identified several key factors that contributed to the effectiveness of our model, including the use of large annotated datasets, the incorporation of relevant features such as word embeddings, and the selection of appropriate machine learning algorithms.

1. **Future Improvement**

There are several areas where our hate speech detection model could be improved in the future. One potential improvement would be to incorporate more advanced natural language processing techniques, such as fine-tuning pre-trained transformer models, to better capture the context and semantics of the text. Additionally, we could experiment with different types of neural network architectures and optimization techniques to further improve the performance of our model. Finally, we could also consider expanding our dataset to include a greater diversity of languages and cultural contexts, in order to improve the generalizability of our model.

**References**

[1] Irene Kwok and Yuzhou Wang. 2013. Locate the hate: Detecting tweets against blacks. In Proceedings of the Association for the Advancement of Artificial Intelligence

[2] Matthew Zook. 2012. Mapping racist tweets in response to President Obama’s re-election. Retrieved from <https://www.theguardian.com/news/datablog/2012/nov/09/mapping-racist-tweets-president-obama-reelection.>

[3] Sarah Hewitt, Thanassis Tiropanis, and Christian Bokhove. 2016. The problem of identifying misogynist language on Twitter (and other online social spaces). In Proceedings of the 8th ACM Conference on Web Science. ACM, 333–335.

[4]Jamie Bartlett, Richard Norrie, Sofia Patel, Rebekka Rumpel, and Simon Wibberley. 2014. Misogyny on Twitter. Technical Report. Demos

[5] Bjorn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2017. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. arXiv Preprint arXiv:1701.08118 (2017)

[6] Vasu Reddy. 2002. Perverts and sodomites: Homophobia as hate speech in Africa. South. African Linguist. Appl. Lang. Studies 20, 3 (2002), 163–175

[7] Leandro Silva, Mainack Mondal, Denzil Correa, Fabrício Benevenuto, and Ingmar Weber. 2016. Analyzing the targets of hate in online social media. arXiv Preprint arXiv:1603.07709 (2016)

[8] FBI. 2019. 2019 hate crime statistics. Retrieved from <https://ucr.fbi.gov/hate-crime/>