Analyzing Proliferation of Hate Speech since the Coronavirus Lockdown

Team 4: Aayush Shah, Aniruddh Nathani, Ankita Naikdalal and Vaidehi Patil

What is our project about?

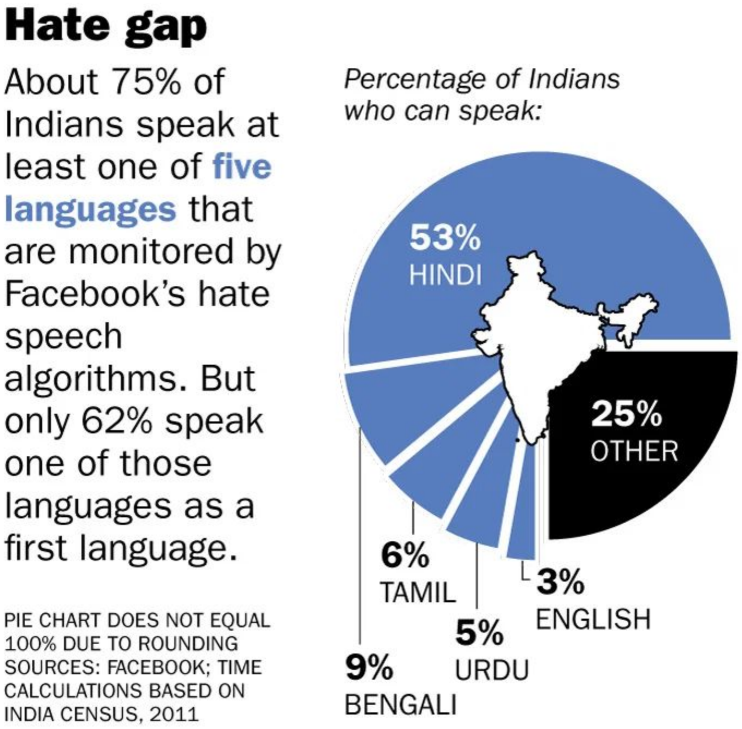
The spread of COVID-19 has created waves of fear, uncertainty, and anxiety across the globe. With social media being the prime platform for people to spread hateful speech amongst the community, the issue of online hate speech seems to be exacerbated by the ongoing COVID-19 pandemic.

The main objectives of our project are:

- To effectively detect hate speech content on social media platforms such as Twitter, Reddit and Youtube
- To analyze whether COVID-19 has led to a proliferation of hate speech since the lockdown
- To identify which social media platforms are more prone to hate speech content

Why is it hard?

- Definition of hate speech varies largely
- API limitations to extract large datasets
- Limited public datasets that identify hateful, offensive speech to train the models
- Subtleties in language and differing definitions on what constitutes hate speech



Hate gap

About 75% of Indians speak at least one of **five languages** that are monitored by Facebook's hate speech algorithms. But only 62% speak one of those languages as a first language.

PIE CHART DOES NOT EQUAL 100% DUE TO ROUNDING
SOURCES: FACEBOOK; TIME CALCULATIONS BASED ON INDIA CENSUS, 2011

Source: https://time.com/5739688/facebook-hate-speech-languages/

How is it done today?

Hate speech detection is a popular albeit challenging problem to solve that has increasingly propagated in online communities due to the exponential growth of social media. Currently, the following methods are used to detect hate speech.

- Classification and deep learning models [1]
- Crowdsourced labeling for hate speech [2]
- Lexical detection methods [3]

Limitations of current practices:

- Lack of trust for the crowdsourced labelled hate speech dataset
- Disregarding the presence of emojis in text
- Absence of research done on hate speech text from different parts of the world

What’s new in our approach?

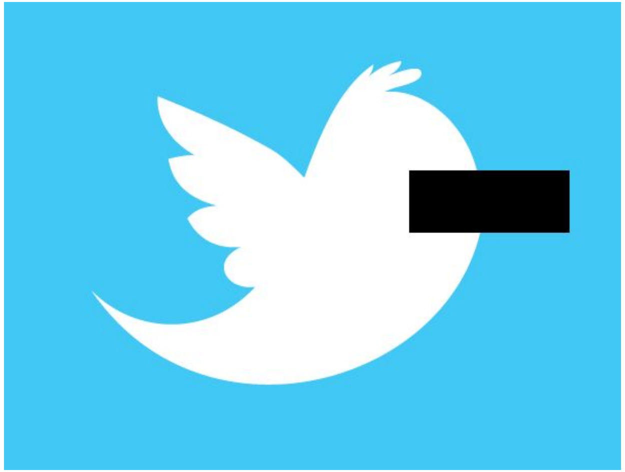
- Combination of multiple data sources (Twitter, Reddit, Youtube)
- After hate speech is aggregated from the above sources, we will focus on stemming, token splitting, character removal and inflection elimination on the data before performing the hate speech recognition process
- We will apply Modern Machine Learning (ML) and Natural Language Processing (NLP) algorithms. Further, we will also plan to examine it using NLP optimization ensemble deep learning approach
- Using publicly labelled datasets for identifying hate speech, we will also be using VADER that interprets other languages and emojis to find more negative speech, as a part of sentiment analysis

Why do we think it will be successful?

- Modern and refined algorithms will pave the way to improve the accuracy and categorization of hate speech as compared to previous research papers who have focused on hate speech detection
- Combination of multiple data sources might affect the accuracy, but will give us more believable results
- We feel that our project statement is impactful for the following reasons:
    - To understand whether COVID-19 has had a negative effect on dissemination of hate speech
    - To provide suggestion for new and innovative features on social media platforms to mitigate the spread of hate speech
    - To identify which social media platforms are more prone to hate speech content and suggest relevant recommendations to mitigate the same

Who cares?

- Social media platforms who want to prevent spread of hate speech on their platform
- Social media users that may be affected by hate speech
- Data scientists and researchers who want to get a good focal point for COVID-19 and hate speech related research
- General public who want a safe experience on social media platforms during COVID-19



Where will you get the data from?

For the purpose of out project, we will be analyzing data from popular social media platforms such as Twitter, Youtube and Reddit to detect hate speech. We will be using data from the below sources:

- Twitter API
- Reddit API
- Youtube API
- Datasets annotated for hate speech, online abuse, and offensive language

We aim to extract a large amount of tweets/comments by using API's during the pre and post coronavirus lockdown period from the above social media platforms. Additionally, we plan to use labelled hate speech datasets from hatespeechdata.com that contain the data used in various research papers such as [3] for their analysis.

Project Planning

Task Name	Due Date	Responsibility
Literature Review	12th February	Ankita, Vaidehi
Data Collection	19th February	Aayush, Aniruddh
Data Cleaning and Preprocessing	19th February	Aayush, Aniruddh
Exploratory Data Analysis	26th February	Ankita, Vaidehi
Data Modelling	26th February	Ankita, Vaidehi
Data Visualization	5th March	Aayush, Aniruddh
Report and Presentation Generation	5th March	Aayush, Aniruddh, Ankita, Vaidehi

References

[1] Hate Speech Detection: A Solved Problem? The Challenging Case of Long Tail on Twitter (n.d.). Retrieved February 4, 2021, from (PDF) Hate Speech Detection: A Solved Problem? The Challenging Case of Long Tail on Twitter

[2] MacAvaney, S., Yao, H., Yang, E., Russell, K., Goharian, N., & Frieder, O. (n.d.). Hate speech detection: Challenges and solutions. Retrieved February 04, 2021, from https://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0221152#sec019

[3] Automated Hate Speech Detection and the Problem of Offensive Language (n.d.). Retrieved February 4, 2021, from http://sdl.soc.cornell.edu/img/publication_pdf/hatespeechdetection.pdf