Hate Speech Detection and Analysis in Indian Social Media (Hinglish)

In this project, I've developed a prototype system to detect hate speech in Indian social media, especially focusing on Hinglish and other regional languages. The workflow includes steps like text preprocessing, language detection, basic sentiment analysis, and classification using TF-IDF with SVM. I also explored how hate speech trends change over time through time-series plots. Although the main focus was on Hinglish, the system is built in a way that it can be extended to support other languages like Tamil, Bengali, and Telugu.

Discovery of Dataset

The dataset I used in this project was taken from an online source: https://data.mendeley.com/datasets/snc7mxpj6t/1.
Initially, I wanted to collect real-time data from Twitter (now X), but due to recent changes in their policies, accessing tweet data has become a paid service. So, I had to drop that idea. After that, I tried web scraping comments from a YouTube video made by Dhruv Rathee, a well-known Indian YouTuber who often posts political content. Even though I successfully scraped the data, the comments were mostly in Hinglish, and when I tried using a language detector, most of them were labeled as "unknown" because Hinglish isn't properly supported. This issue is also visible in the notebook, where around 96.4% of the data is tagged as neutral, and only 3.6% is classified as either positive or negative. Eventually, I came across the Mendeley dataset mentioned above, which already had hate speech labels assigned as HS0, HS1, and HSN, and decided to use that for my final analysis.

Key Insights

- The main language used in this project was Hinglish, which is a mix of Hindi and English either Hindi written in English script or sentences blending both languages.
- During the analysis, I noticed that sentiment and hate labels didn't align well because the language detector couldn't accurately identify Hinglish. As a result, most of the comments were marked as "unknown" sentiment, which shows the limitation of standard tools when dealing with code-mixed data.
- For classification, I used an SVM model trained on TF-IDF features. The model was applied to predict the labels HS0, HS1, and HSN based on the comments. This helped in categorizing the comments into non-hate, hate, and extreme hate.
- I also created time-series plots and other visualizations, which gave insights into when hate speech tends to spike. This shows that certain dates or times can have more hateful activity, possibly linked to events.

- The dataset mainly came from an Instagram post and contained a variety of comments, including vulgar and extreme hate speech. This highlights the importance of moderation systems that also factor in how viral or liked such content becomes.

Analysis

In this analysis, we can see that the accuracy after applying the classification model was around **83.7%**, which is quite decent. I used an SVM model with TF-IDF on cleaned Hinglish comments to classify the data. The test set had 227 samples, and the model worked well for HS0 but struggled a bit with HS1 and HSN due to fewer examples.

From the time-based plots, we also get an idea about when people usually post hate comments. Extreme hate was mostly seen during weekends and early hours, while regular hate was more common around mid-week, especially on Wednesday and Thursday. Non-hateful comments (HS0) also appeared more during mid-week. This kind of pattern shows that people might be more aggressive or emotional during weekends, maybe because they have more free time or are more active online. So, based on this, if someone wants to post political content and avoid hate, they should prefer weekdays over weekends.

Another thing I noticed was that the number of comments increased a lot during July. This matched with a period when there was political instability in India, especially in Maharashtra. The wordcloud also shows that words like "Shiv" and "Sena" came up a lot, which points to the influence of the Shiv Sena party (a political party in Maharashtra) and its mention in the comments. Along with that, many curse words in Hindi were visible in the wordclouds, showing that the topic sparked strong reactions from people.

Literature Review with Links

HASOC - Hate Speech and Offensive Content Identification (FIRE 2019)
One of the earliest datasets focused on Hinglish hate content.
https://ceur-ws.org/Vol-2517/T2-1.pdf

Indian Politics Tweets EDA and Sentiment Analysis
https://www.kaggle.com/code/adritpal08/indian-politics-tweets-eda-and-sentiment-analysis

Hate Speech and Offensive Language Dataset
https://www.kaggle.com/datasets/mrmorj/hate-speech-and-offensive-language-dataset/data

Impact & Applications

1. From the time-based plots, we can figure out which days or time periods have more hate comments. This can help link it to real-life events.

2. Platforms like Twitter or YouTube should take steps to detect and reduce hate comments that are getting a lot of likes or going viral.

Limitations of My Project

1. One major limitation is that the dataset is **imbalanced**. There are too many HS0 (non-hate) comments compared to HS1 and HSN. Because of that, the model struggles to detect hate and extreme hate correctly.
2. I only used **TF-IDF and SVM**, which are traditional ML models. They are fast but not very powerful for understanding context or sarcasm like transformer-based models (e.g. MuRIL).
3.The **comments don't have location info**, so I couldn't generate geo heatmaps which were part of the original plan as mentioned in my Project Proposal..