Language And Society-Project Report

- Eshika Khandelwal 2020114018
- Adith John Rajeev 2020114010

Link to the dashboard: https://twitter-data-analysis-ls.herokuapp.com/
Github Repository: https://github.com/esh04/Twitter-Data-Analysis-LS

Introduction

We aim to study what the diverse linguistic population of India thinks their Official Language or Lingua Franca should be. How the attitudes, views and opinions of the public led to the backlash against Hindi as the Official Language of India and in favour of other languages.

Data Collection

We first attempted to use 'tweepy' to collect the data for our survey but were limited to tweets only from the past week, which was insufficient for our study. Therefore, we relied on **'snscrape'** to collect tweets and obtained various tweets from a span of almost two years. We scraped tweets across English as well as various Indian languages based on specific keywords and hashtags.

To generate the map displaying the users' locations, we separated all the tweets with the appropriate coordinate tags. All the Non-English tweets in this category were then translated for better understanding and to generate and interpret word clouds.

Literature Survey

- *"Should South Indians learn Hindi? Why or why not?", Anna Rácová.*
  In this paper, the author surveys an online discussion on whether South Indians should learn Hindi or not. The author organises the points made by the respondents into political, pragmatic, and cultural fronts. Finally, the author analyses the controversial view that English must be the only official language of India. This paper was relevant to our Project as we performed a similar study, the only difference being that we observed tweets rather than an online discussion.

- *A Note on Rural Attitudes Towards Hindi: A Mysore Village Study, Helen E. Ullrich*
  This paper, published in 1969, looks at the differences in the attitudes of the people residing in rural areas versus those in urban. While the people in rural areas preferred using Hindi over English, urbanisation led people to believe that English was necessary for obtaining jobs. This occupational threat led them to favour English.
  The paper emphasises the difference in views across regions; we too made similar observations regarding how the people's opinions differ with time and location by plotting a map and identifying patterns within specific areas.

- *Language as an Identity: Hindi-Non-Hindi Debates in India, Amit Ranjan*

  Language is considered an inherent component of a person's identity, and any attempt to subjugate this identity would be resisted. The mentioned paper looks at such resistance from a historical point of view. It examines the politics of language-based identity, the related tensions and debates between Hindi and non-Hindi speakers since the years of the anti-colonial movement in India.

  We were able to observe such backlashes in our data as well. Every time a policy that favoured Hindi was instituted, we located spikes in the number of tweets. We tried to deduce what policy or news caused the spike and have mentioned some of those in the observations below.

Observations

- From the data visualisation produced, we observed a significantly large number of tweets from southern India. On studying these tweets, we found that most of them were a backlash against Hindi imposition and the disparity between their respective mother tongues and Hindi. The popularity of the tweets like **#StopHindiImposition** confirms this inference.

- We also noticed spikes in the number of tweets on specific dates. Below are our reasonings for some of those spikes:
  - **9th August 2020**: This was related to the tweet by the Dravida Munnetra Kazhagam party MP Kanimozhi who shared her experience at Chennai International Airport. Her Indian Nationality was questioned by a CISF officer posted in the airport because she conversed in Tamil and English, but not Hindi. *[link]*
  - **13-14 September 2020:** This was traced back to the celebration of Hindi Divas. People expressed many contradicting views, which sparked a lot of discussion and debate regarding the same.
  - **20th of November 2020:** This was caused when the current CM of Tamil Nadu, MK Stalin, tweeted that a letter he sent to the Central Ministries in English was replied to in Hindi. He demanded that they respond to it in the language it was initially sent in. [link]

- We also found hashtags related to specific events that created a buzz on Twitter.
  - For example, **#RejectZomato** was trending during late October when a Zomato executive refused to cooperate with a customer because he did not know Hindi. *[link]*

- We also made a word cloud from all the English and the translated Indian Language tweets. The word cloud was made to include occurrences of bigrams which led to some interesting observations.
  - We observed that the term **'national language'** occurs a lot more than the term **'official language'**. On studying these occurrences in our tweet contents, we noticed that there is a misconception among people that Hindi is our National Language. For example,
    - *"If u r Indian u have to know and speak Hindi and accept it as national language…"*

        - *"...let the entire south Indians know that their national language is Hindi"*

- The case of Tamil Nadu
    - We observe that the tweets concentrated in the Southern Part of India belong majorly to Tamil Nadu.
    - After English, we observe that the most significant number of tweets are from the Tamil script.
    - It is also interesting to note that Tamil Nadu follows a **2-language policy –** of Tamil and English, which is perhaps why they have such strong opinions against Hindi being used for official purposes.
    - Also, some events occurred after we extracted the data, like the postponement of 2021 KVPY because of regional language issues in Tamil Nadu. The Madras High Court demanded that the exam be held in regional languages along with Hindi and English. This event could also have been due to their strong opinions regarding Hindi.

Problems Faced

- **Language Identification:**

  Users tend to use the Roman alphabet instead of the language's script; for example, people rarely use the Devanagari for Hindi while texting. Due to this, the tool used for identifying the language failed at various instances and assigned a foreign or undetermined language to that tweet leading to a lot of discrepancies in the data.

- **Sentiment Analysis:**

  We attempted to tag all the tweets with the bound sentiments so that they could be used for further analysis. The VADER Sentiment Analyser was used for this process. However, its outputs were limited to positive, negative and neutral. These sentiments alone could not be used to determine anything due to the varying subjects across the tweets. For example, some tweets had a negative sentiment while referring to Hindi, while some had a negative sentiment with respect to English. So the data could not be classified on the basis of their sentiments alone, and so we decided to not tackle this issue in our project.

- **Translation of Tweets:**

  Due to the presence of more than 100000 tweets, the task of translating the tweets became difficult. We tried using many translators for this task like the Python-Translator, GoSlate and GoogleTranslate modules. Python-Translator and GoogleTranslate took too long to run with our extensive dataset and GoSlate gave a 429 (too-many-requests) error. We, therefore, had to limit the number of tweets for which we performed the translation. These limited tweets were used for the generation of the word clouds.