Social Media Analysis

Name - Sanket S Houde
Institute - IISER Bhopal

Objective

The objective of this project was to analyze the social media content surrounding a recent political event in India. I decided to analyze tweets on Twitter regarding the recently concluded Karnataka state assembly elections, the results for which were declared on 13th May 2023. The tweets were extracted, preprocessed and annotated with appropriate sentiment using the VADER model. Using this sentiment annotation,I analyzed the most frequent words, popular hashtags and influential accounts grouped by sentiment tag.

Extracting data

The data was scraped using a module in Python named '*snscrape*'. The search term was 'karnataka elections' and the time frame was restricted from 10th April 2023 to 9th June 2023. The extracted data was converted to a Pandas dataframe for further analysis.

Sentiment analysis

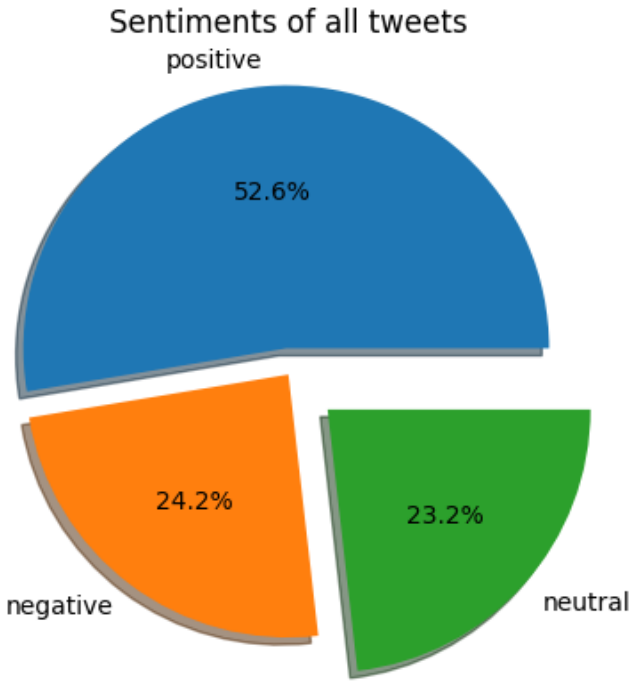
The tweet content was stripped of twitter handles, twitter return handles, punctuation marks, special characters and url links. This was done to preprocess the data for sentiment analysis. Valence Aware Dictionary and sEntiment Reasoner (VADER) which is fine tuned towards annotating social media text was used for determining the sentiment of the tweet. VADER calculates probabilities for each sentiment and then gives a compound score. A compound score of more than 0.05 was considered 'positive', less than -0.05 as 'negative' and between -0.05 and 0.05 as neutral. These cutoff values are routinely used.

Preprocessing for further use

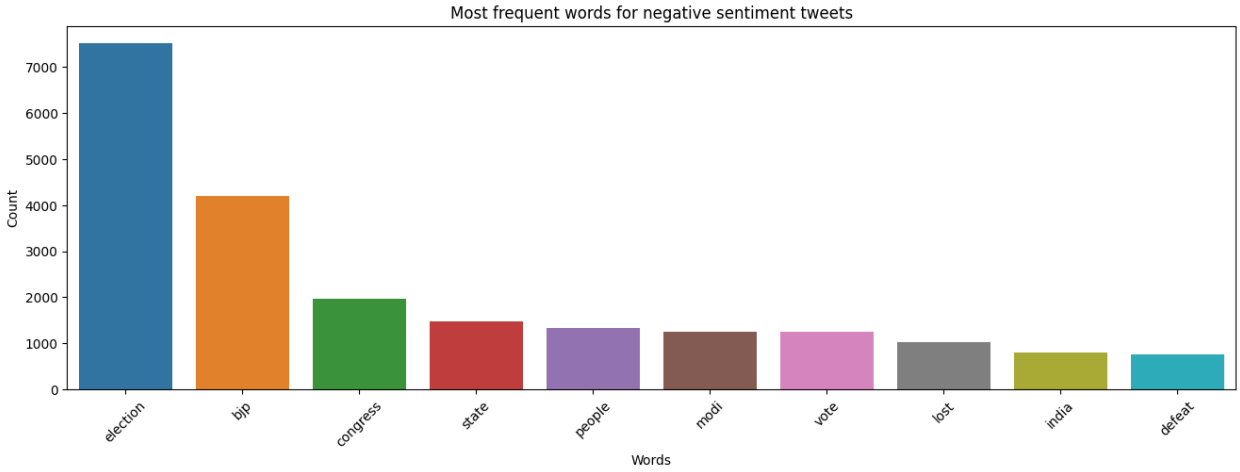
There was further preprocessing done to analyze the tweet contents. In addition to the above preprocessing, the words were lowercased, stop words were removed (using the NLTK package), emojis were removed and the words were lemmatized. Lemmatization groups inflected or variant forms of the same words and replaces it by this base word. Eg - the words 'built', 'building' and 'builds' are all different forms of the word 'build'.

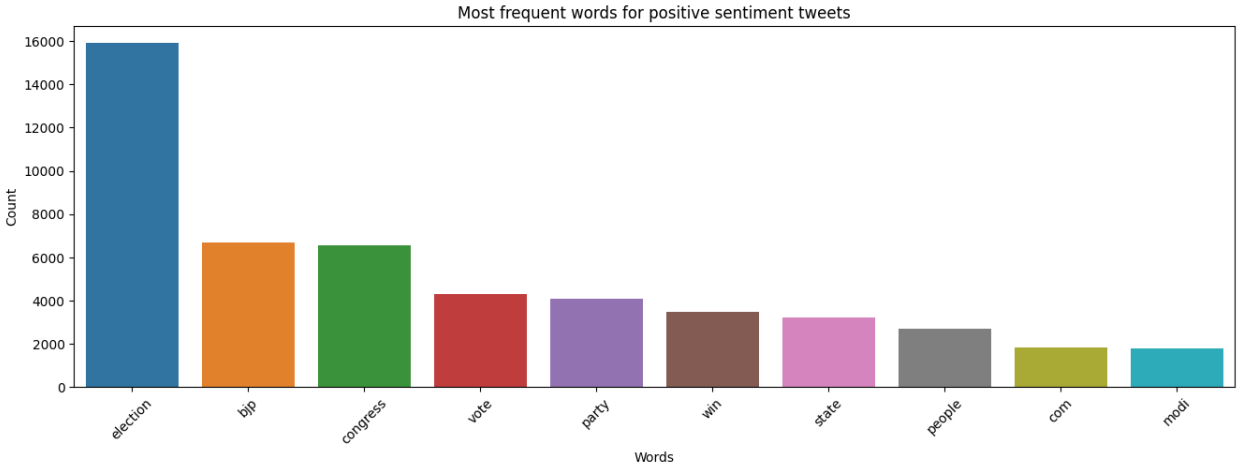
Analysis of the entire dataset

The tweets were preprocessed as described above and I observed that more than half of the tweets were positive (**52.6%**). Negative (**24.2%**) and neutral (**23.2%**) tweets were divided equally among the other half of the tweets.



The frequency of words in these tweets itself were analyzed grouped by sentiment.





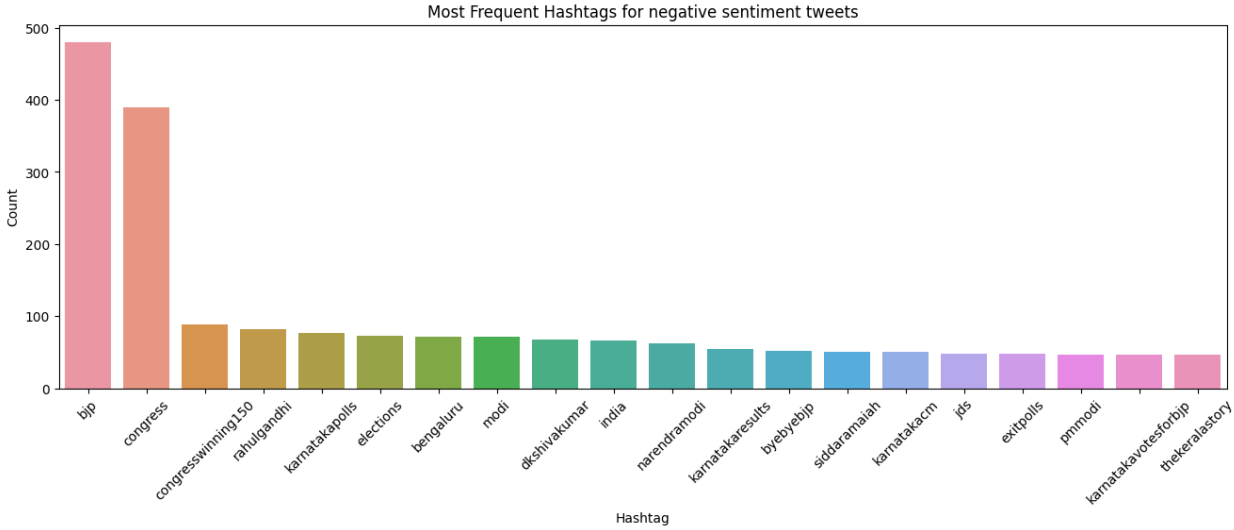
The relevant words to keep an eye on here are *'bjp'*, *'congress'* and *'modi'*. The first are the two main parties in contest even though other regional parties. The third term is an important figure from the BJP political party. Although *Rahul Gandhi* is an important figure from the Congress political party, his name doesn't show up frequently.

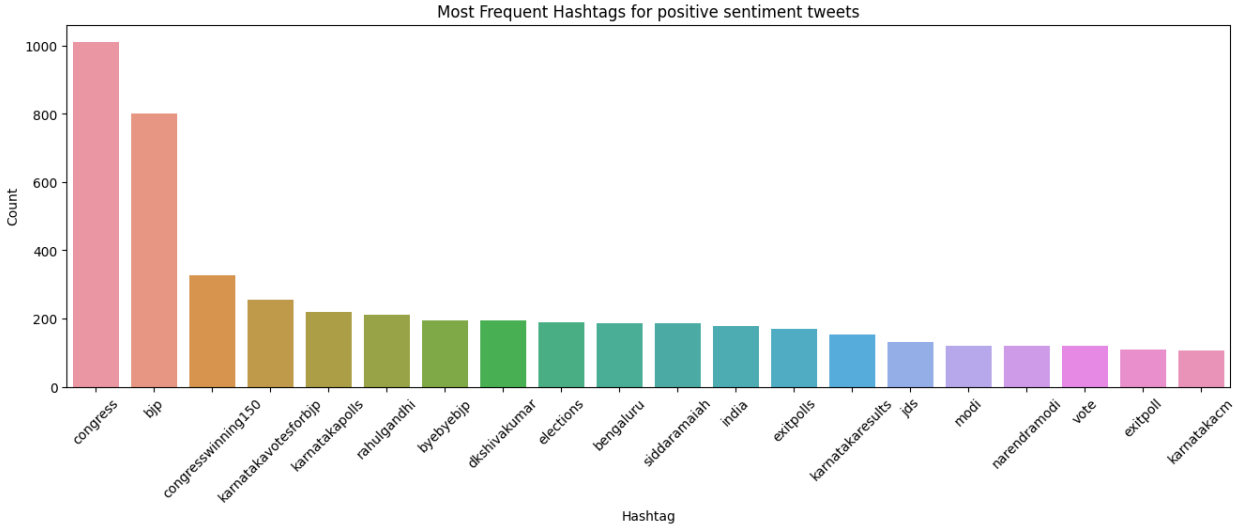
The positive sentiment tweets have almost an equal number frequency for *'congress'* and *'bjp'* and *'modi'* ranks last. However, there's a significant difference in the frequency of '*congress'* and *'bjp'* and the counts are higher for *'bjp'* which indicate a negative sentiment associated with that particular party as compared to the other one. Even *'modi'* jumps in at 6th position which further indicates negative sentiment associated with that party.

Here's a nice Word Cloud visualization in case bar plots bore you :

Analysis of Hashtags

Hashtags can be considered as a proxy for the topic of discussion of the tweet content. But there's one problem, not every user uses hashtags. There were ~6000 tweets with hashtags out of the total ~38k tweets. Although not representative of the entire data, we can get an idea of the topics being discussed.

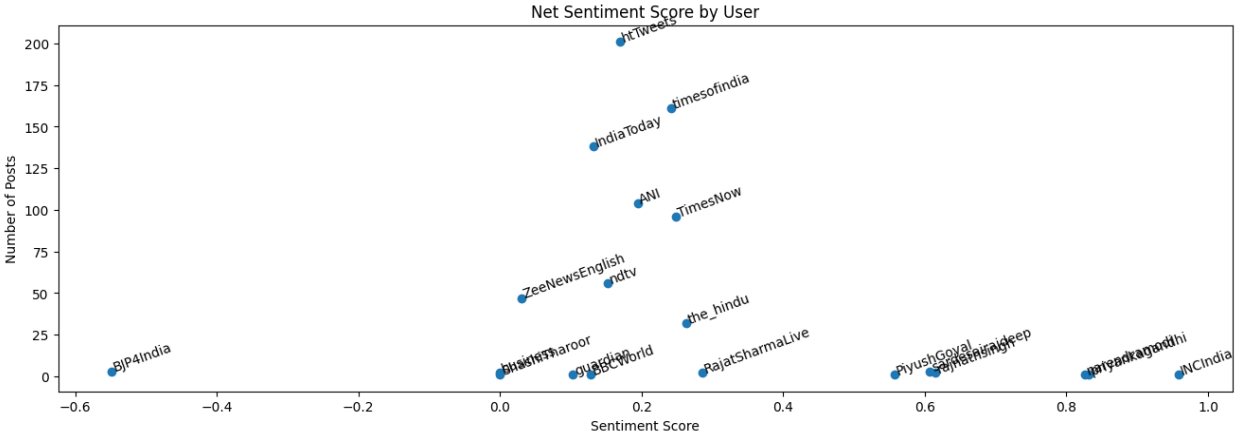




A quick glance shows us that the term *'congress'* has higher mentions in tweets with positive sentiments and *'bjp'* has higher mentions in tweets with negative sentiments. Other than that, the other hashtags are almost similar across the two sentiments with some variations in the frequencies.

Analysis of popular influencers

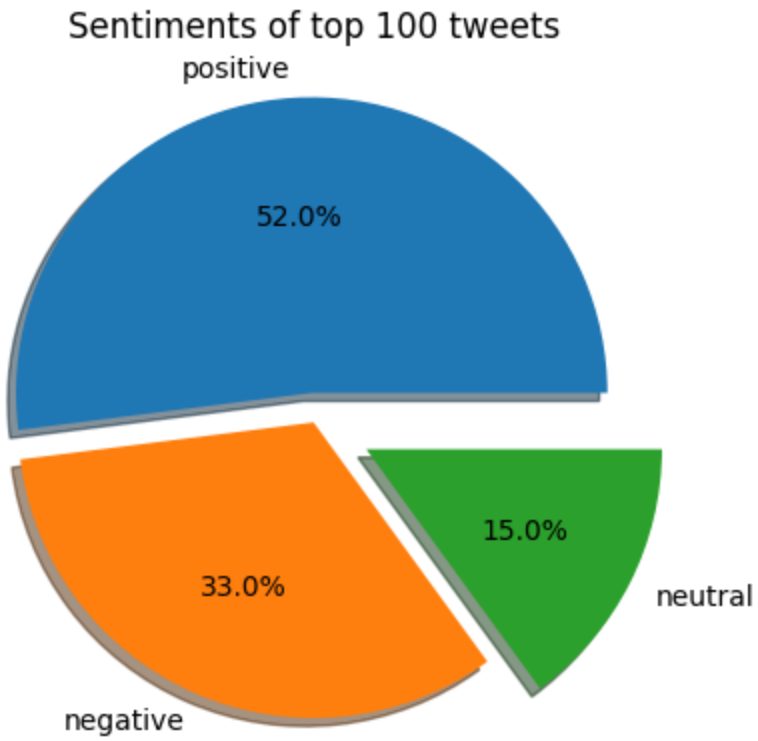
An influencer *score* was calculated by subtracting the number of followers from the number of friends to assess the reach of that account. The rationale here being that a more influential person will have more accounts following them than the number of accounts they follow. The accounts with top 10 scores were considered. The compound sentiment scores were averaged across all the tweets of a particular account.



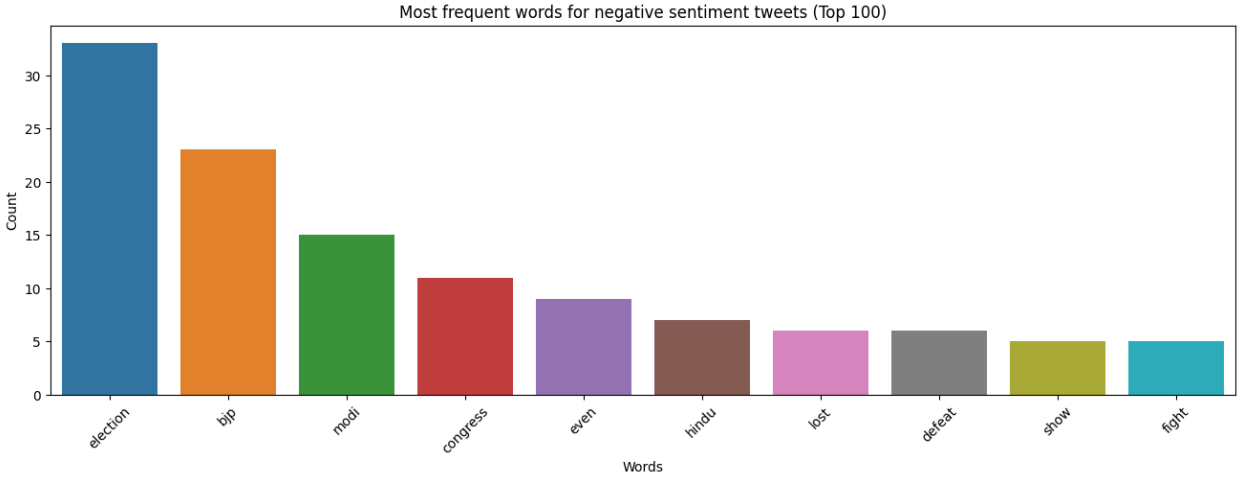
We can observe that accounts associated with news outlets either have very low (< 0.3) positive compound scores or neutral scores. Shashi Tharoor, a MP belonging to the Congress party has a neutral score too unlike other accounts of politicians who have high positive compound scores. Among the accounts associated with the two major national parties, the congress party has a very high positive score whereas the account associated with BJP has a high negative compound score of around -0.6.

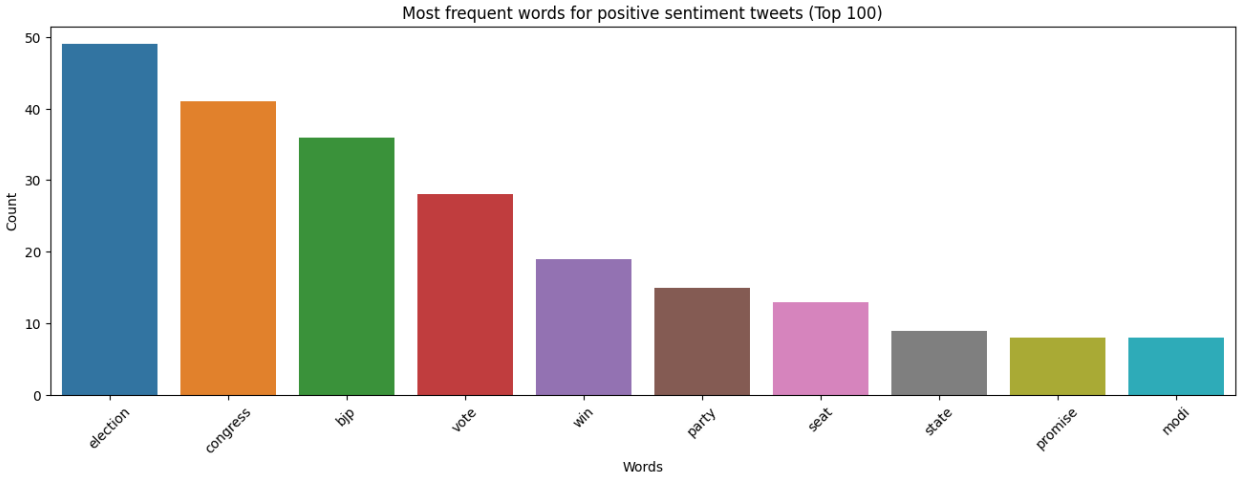
Analysis of the top 100 tweets

A score was calculated by taking the sum of the number of retweets, replies, quotes and likes to assess the reach of the tweet. Higher the score, the better was the reach of the tweet. Using this metric, the top 100 tweets were considered. The top tweets had a similar distribution of the sentiment with around half of the tweets being tweets and the other half distributed among the other two sentiments. Although, the percentage of negative tweets had slightly increased.



The top 10 most frequent words were extracted from these tweets grouped by their sentiment.





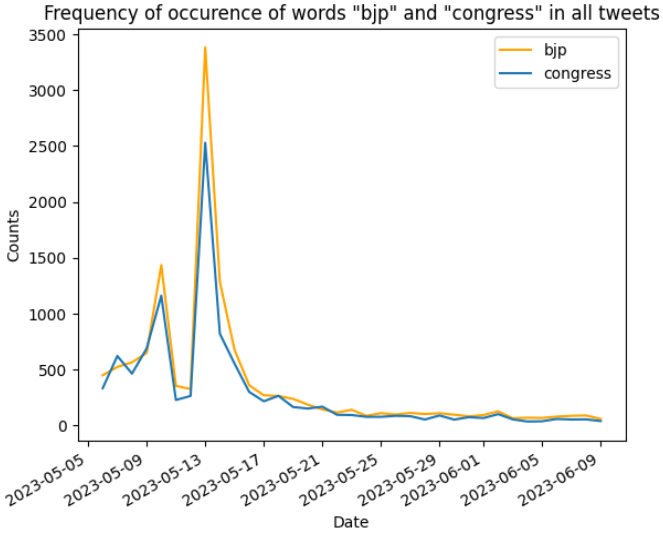
As we can observe here, the negative sentiments have a prevalence of *'modi'* and *'bjp'* as compared to *'congress'*. The term *'hindu'* is observed among the negative tweets. Although, the term *'congress'* has a slightly higher frequency than '*bjp'.* The term *'promise'* is observed among the positive tweets. Other than that , the remaining terms are uninformative about the political event.

Analysis of particular terms over time

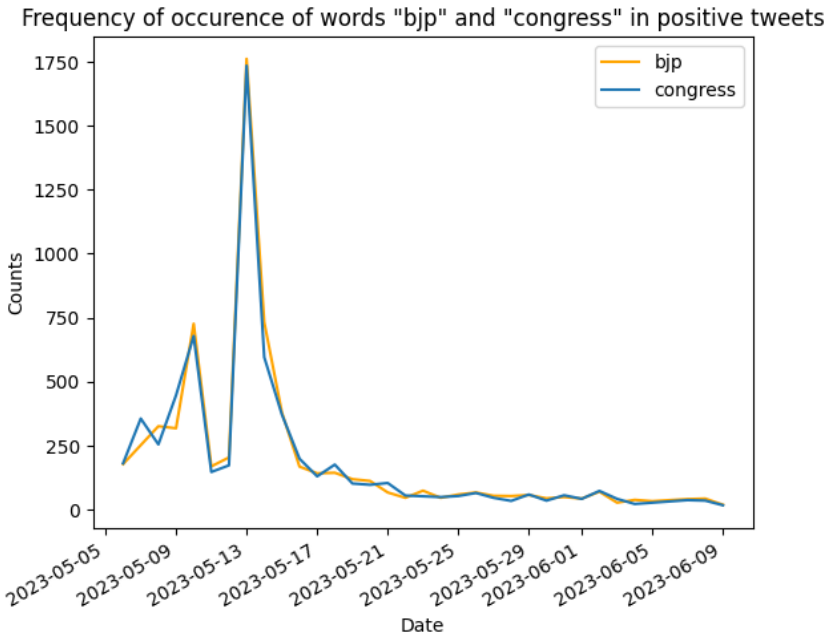
Frequencies of two sets of terms were observed as the function of time. The words ('*bjp', 'congress'*) and (*'modi','rahul'*) were considered since these were the two important parties involved in the elections and the two important figures of those parties. The term *'modi'* is related to *'bjp'* and the same for the other two terms.

BJP vs Congress

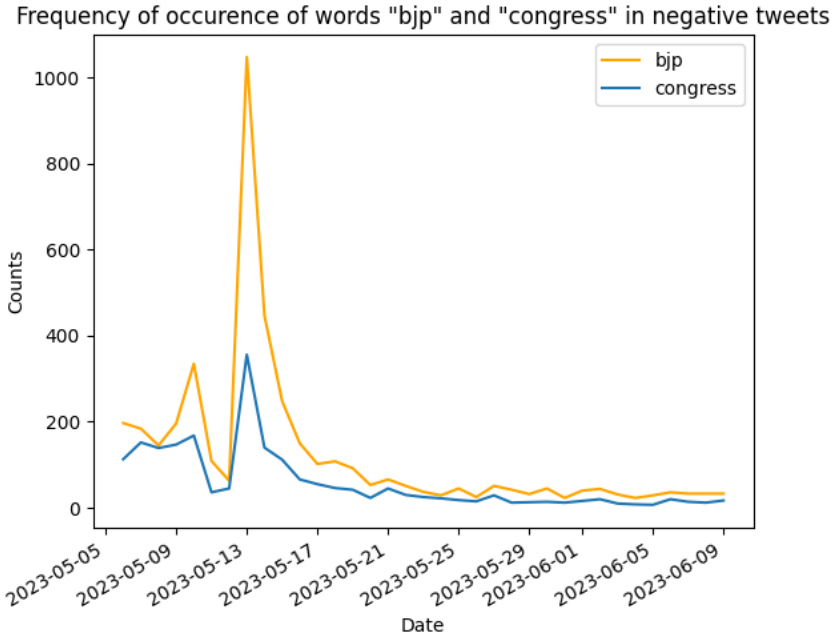
Both the terms had similar frequencies across time when all the tweets were considered regardless of the sentiments



It was the same case when only the positive tweets were considered.

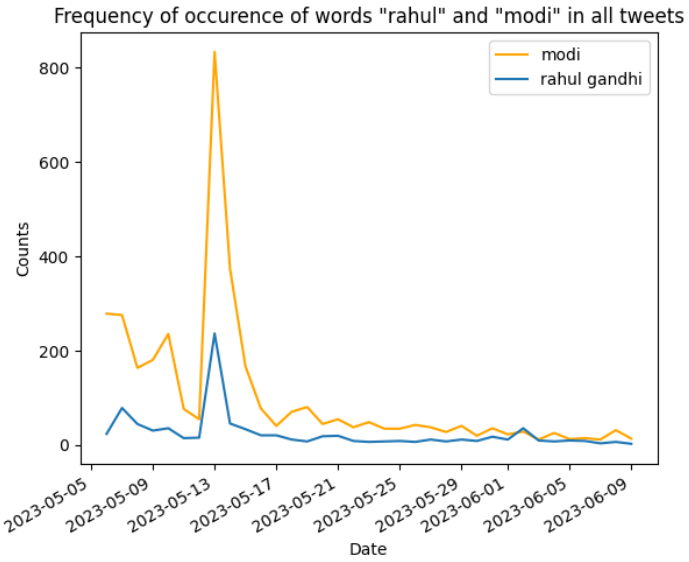


However, the term '*bjp'* had a higher overall frequency across the whole time period but especially on 13th May 2023, there's a huge spike for the term *'bjp'* as compared to the term *'congress'*. This is because of the declaration of the election results. This trend can be seen for all the plots due to this phenomenon but the term *'bjp'* has a very huge prevalence in the negative tweets compared to *'congress'.*

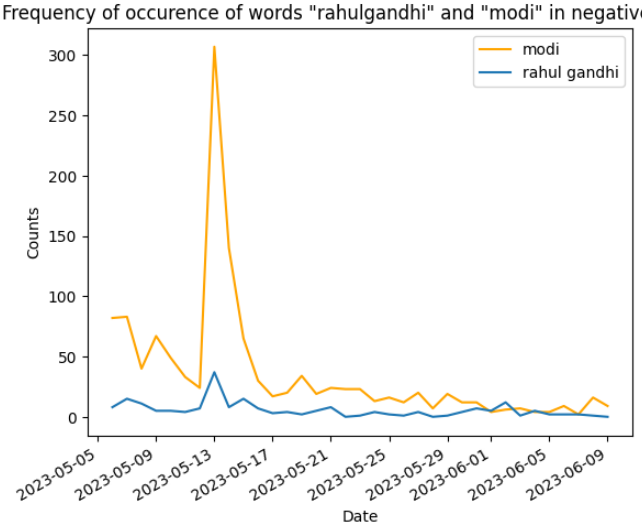


Rahul vs Modi

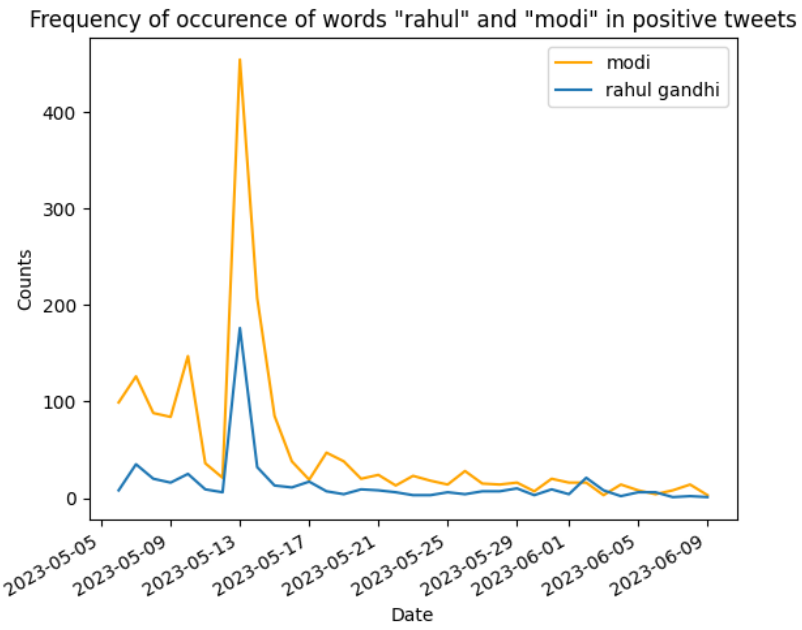
We can immediately observe that the frequency of the term ‘modi’ is higher than that of ‘rahul’ in the first half of the graph. In fact, ‘modi’ has a very high spike when the results are declared as compared to ‘rahul’. However both the terms start having similar frequencies towards the end.



The term ‘modi’ has higher frequency and an even higher spike during the results’ declaration among the negative tweets which tapered towards the end. However ‘rahul’ has a lower frequency and doesn’t experience the peak. The frequency is consistently low.



The term ‘modi’ has higher frequency before the election results and also experiences a higher peak when results were declared among the positive tweets. The frequency tapers and reaches similar levels of the term ‘rahul’. Interestingly, ‘rahul’ gets a nice spike among positive tweets on the day the results were declared.



Conclusions

After analyzing all the data, we can say that although both Congress and BJP parties had similar positive sentiment levels among the audience, BJP had higher negative sentiment associated with them.

Future Directions

1. A bigger dataset with a history of more than 1 month before the elections could have made the analysis a bit more reliable by breaking down the analysis into pre and post election epochs.
2. A better state of the art ML model such as roBERTa model trained on nearly 58 million tweets. This can definitely increase the reliability of sentiment annotation which was an important factor throughout our analysis. Due to time constraints, I had to use the VADER model which uses a user maintained lexicon of social media terms and usually falls behind the latest slangs and sarcasm used in this ever evolving social media space.