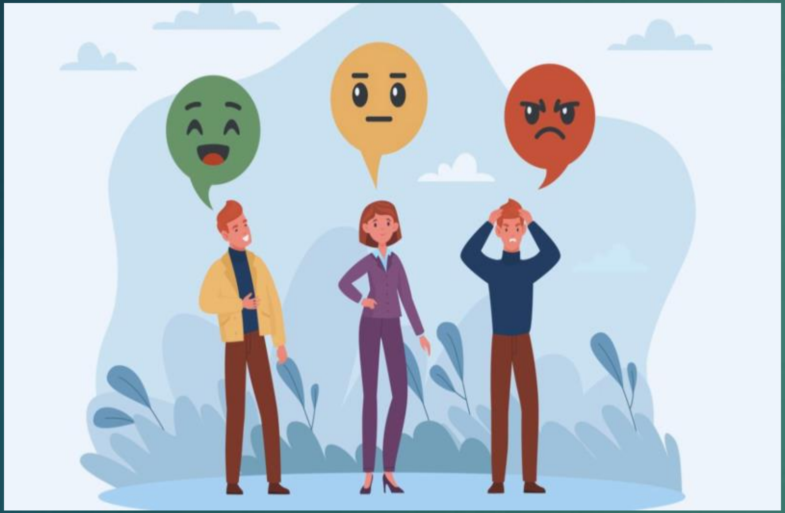
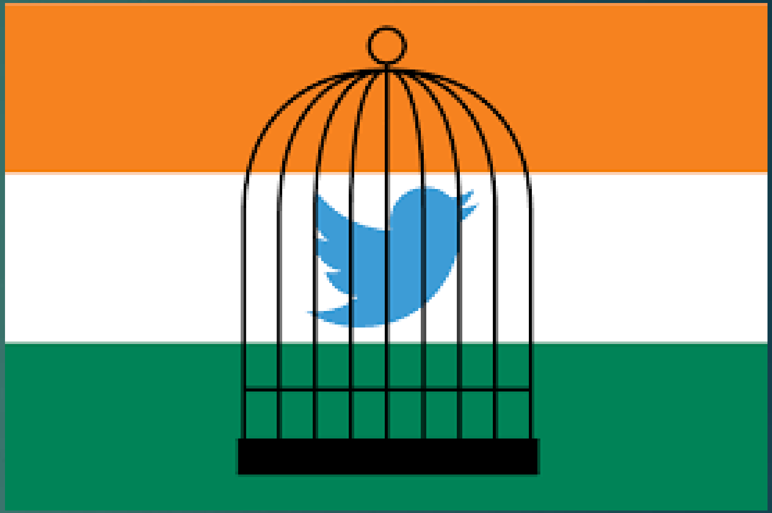
Sentiment Analysis of Indian Political Tweets





GROUP 7

NIKHIL ANNE

SRI VATSAV BUSI

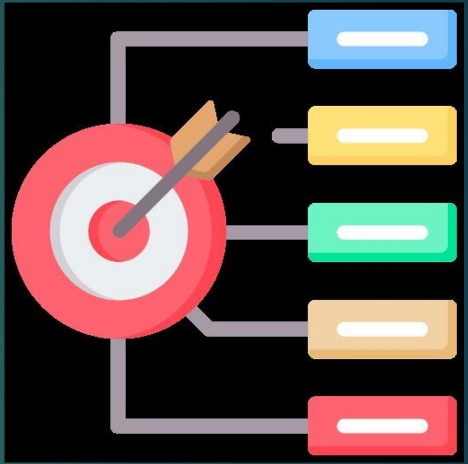
VENKATA SAHITHI NALLANI

SAI CHAITANYA NAGANDLA

HARIPRIYA JANARDHAN RAO

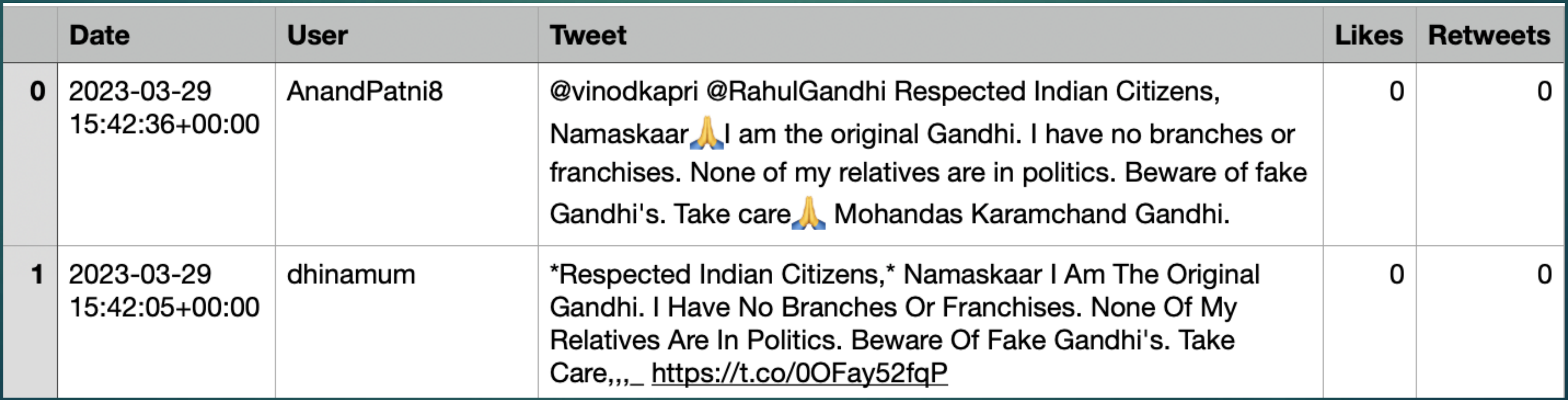
Objective

- Identify the predominant topics of discussion among the public on Twitter.
- Evaluate the overall sentiment of these political tweets to gauge public opinion.
- Understand the sentiment of the public's response to these tweets by assessing likes, retweets, and comments.
- Extract the core subjects and issues being discussed in the chosen timeframe.
- Translate the results into actionable insights for political campaigns, public policy recommendations, and election predictions based on real-time data and sentiment.



Data Source & Overview

- Our dataset is from **www.kaggle.com** a trusted platform for datasets, ensuring the reliability of our data.
- The dataset offers a deep dive into the pulse of Indian politics on Twitter, one of the most vibrant and dynamic platforms for political discourse.
- Beyond tweets, the dataset captures reactions (likes, retweets), giving a fuller picture of public
- Contains a substantial volume of 49,570 unique tweets, allowing for diverse analyses.



	Date	User	Tweet	Likes	Retweets
0	2023-03-29 15:42:36+00:00	AnandPatni8	@vinodkapri @RahulGandhi Respected Indian Citizens, Namaskaar🙏I am the original Gandhi. I have no branches or franchises. None of my relatives are in politics. Beware of fake Gandhi's. Take care🙏 Mohandas Karamchand Gandhi.	0	0
1	2023-03-29 15:42:05+00:00	dhinamum	*Respected Indian Citizens,* Namaskaar I Am The Original Gandhi. I Have No Branches Or Franchises. None Of My Relatives Are In Politics. Beware Of Fake Gandhi's. Take Care,,,_ https://t.co/0OFay52fqP	0	0

Methodology

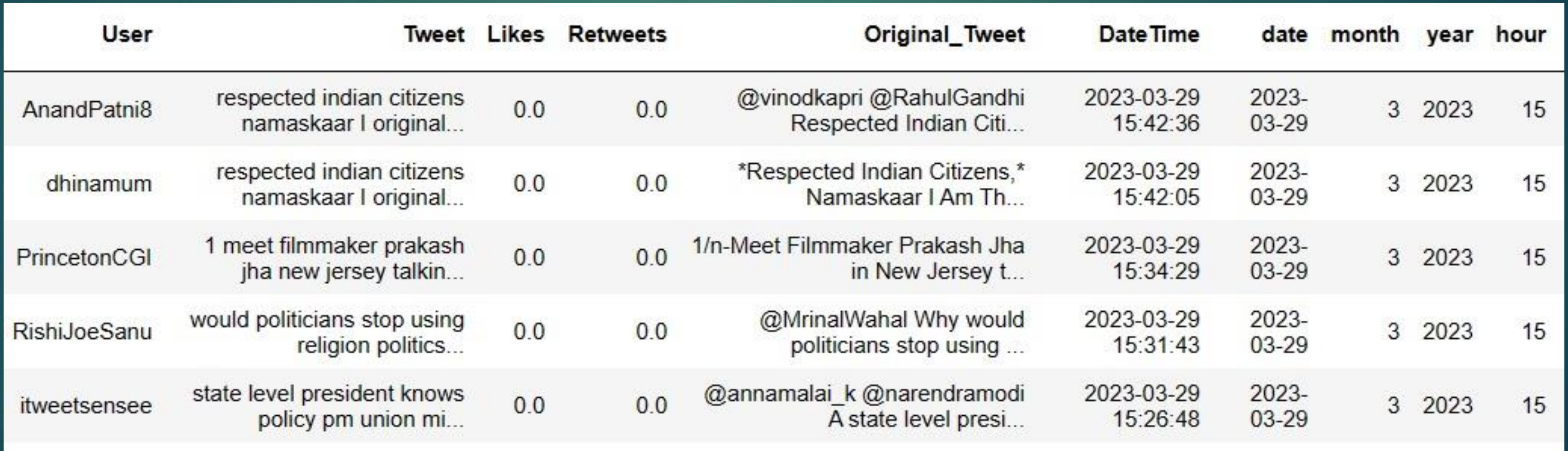
- Clean and preprocess tweet content, stemming and lemmatization.
- Generate word clouds for visual frequency representation.
- Implement TF-IDF for text vectorization and determine word significance.
- Conduct LDA analysis for topic discovery and sentiment modeling.
- Evaluate and validate model performance

Data Cleaning

- Quickly pinpoint and evaluate missing data using df.isna().sum().
- Maintain dataset integrity by removing rows containing null values with df.dropna(inplace=True).
- Ensure data uniqueness by identifying duplicates using df.duplicated().sum().
- Split 'DateTime' into specific segments like day, month, year, and hour, while rectifying any inconsistencies.
- Enhance tweet readability by effectively excluding Twitter handles
- Strip away any URLs to keep the content focused on text alone.

Text Preprocessing & Normalization

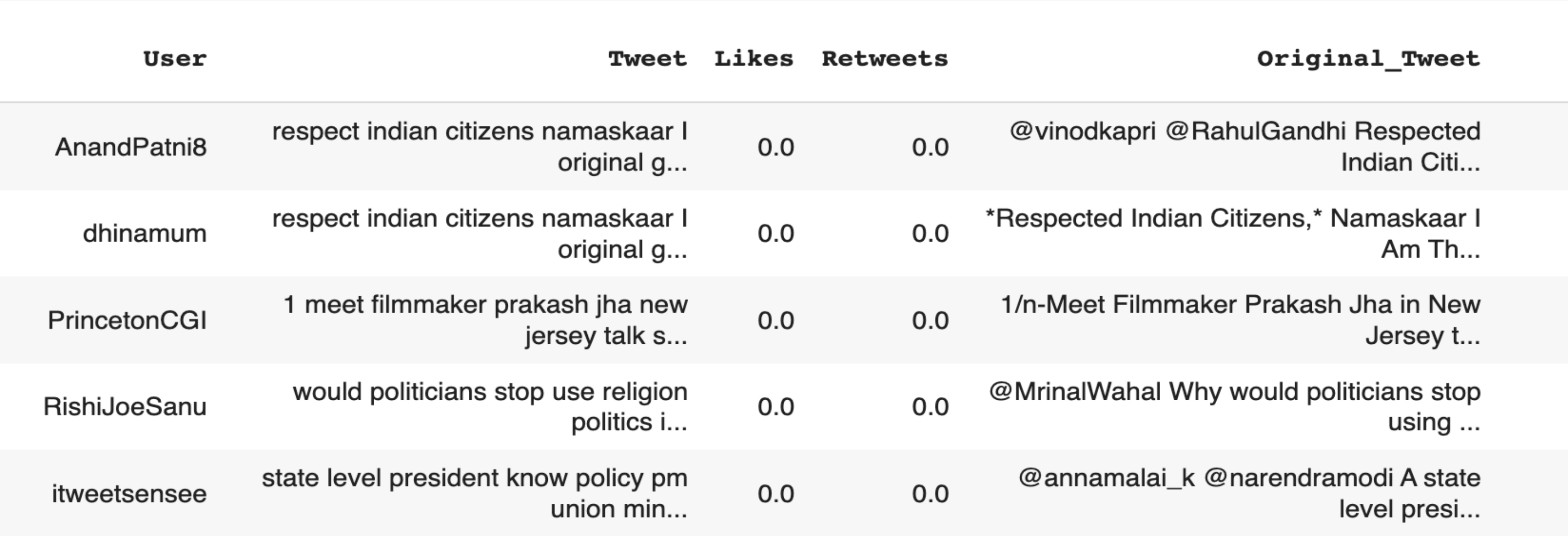
- Refine tweet structure by discarding unwanted characters and punctuations and Streamline white spaces
- Undertake **stopwords elimination** to filter out frequently occurring words, such as "and" or "the", that might detract from the tweet's core message.



User	Tweet	Likes	Retweets	Original_Tweet	DateTime	date	month	year	hour
AnandPatni8	respected indian citizens namaskaar I original...	0.0	0.0	@vinodkapri @RahulGandhi Respected Indian Citi...	2023-03-29 15:42:36	2023-03-29	3	2023	15
dhinamum	respected indian citizens namaskaar I original...	0.0	0.0	*Respected Indian Citizens,* Namaskaar I Am Th...	2023-03-29 15:42:05	2023-03-29	3	2023	15
PrincetonCGI	1 meet filmmaker prakash jha new jersey talkin...	0.0	0.0	1/n-Meet Filmmaker Prakash Jha in New Jersey t...	2023-03-29 15:34:29	2023-03-29	3	2023	15
RishiJoeSanu	would politicians stop using religion politics...	0.0	0.0	@MrinalWahal Why would politicians stop using ...	2023-03-29 15:31:43	2023-03-29	3	2023	15
itweetsensee	state level president knows policy pm union mi...	0.0	0.0	@annamalai_k @narendramodi A state level presi...	2023-03-29 15:26:48	2023-03-29	3	2023	15

Text Preprocessing

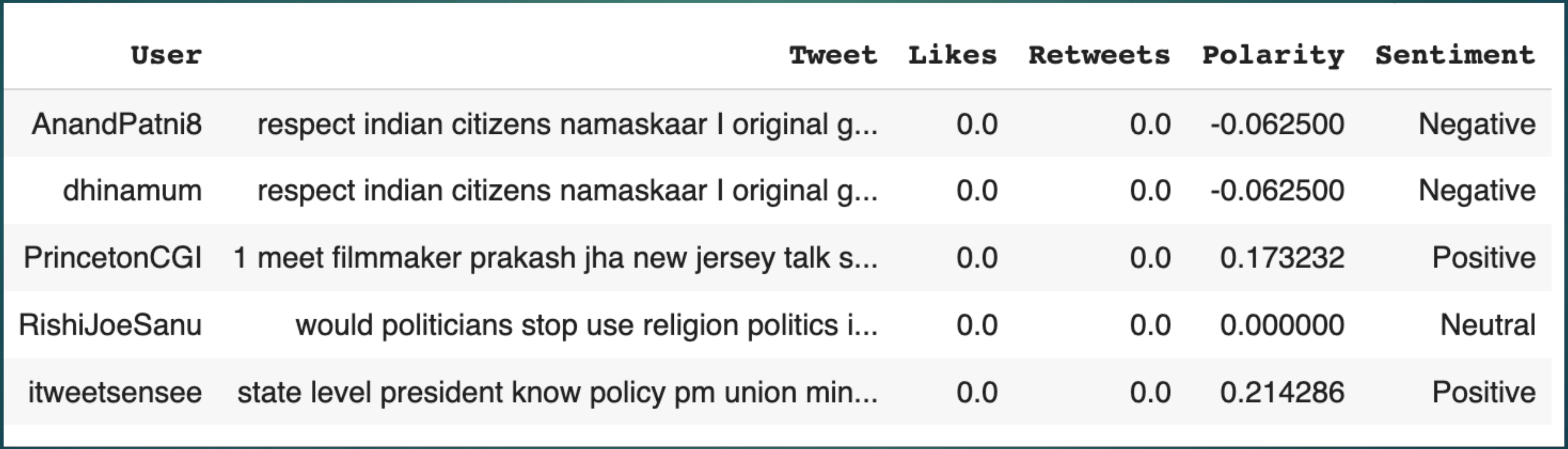
- Implement the WordNetLemmatizer tool to standardize word variations for consistency.
- Apply the efficient Snowball Stemmer for precise word truncation.



User	Tweet	Likes	Retweets	Original_Tweet
AnandPatni8	respect indian citizens namaskaar I original g...	0.0	0.0	@vinodkapri @RahulGandhi Respected Indian Citi...
dhinamum	respect indian citizens namaskaar I original g...	0.0	0.0	*Respected Indian Citizens,* Namaskaar I Am Th...
PrincetonCGI	1 meet filmmaker prakash jha new jersey talk s...	0.0	0.0	1/n-Meet Filmmaker Prakash Jha in New Jersey t...
RishiJoeSanu	would politicians stop use religion politics i...	0.0	0.0	@MrinalWahal Why would politicians stop using ...
itweetsensee	state level president know policy pm union min...	0.0	0.0	@annamalai_k @narendramodi A state level presi...

Sentiment Analysis

- Determining the sentiment polarity using TextBlob for each tweet in the 'Tweet' column .
- Based on the calculated polarity, the tweets are classified as 'Positive', 'Negative', or 'Neutral' and stored in a new 'Sentiment' column

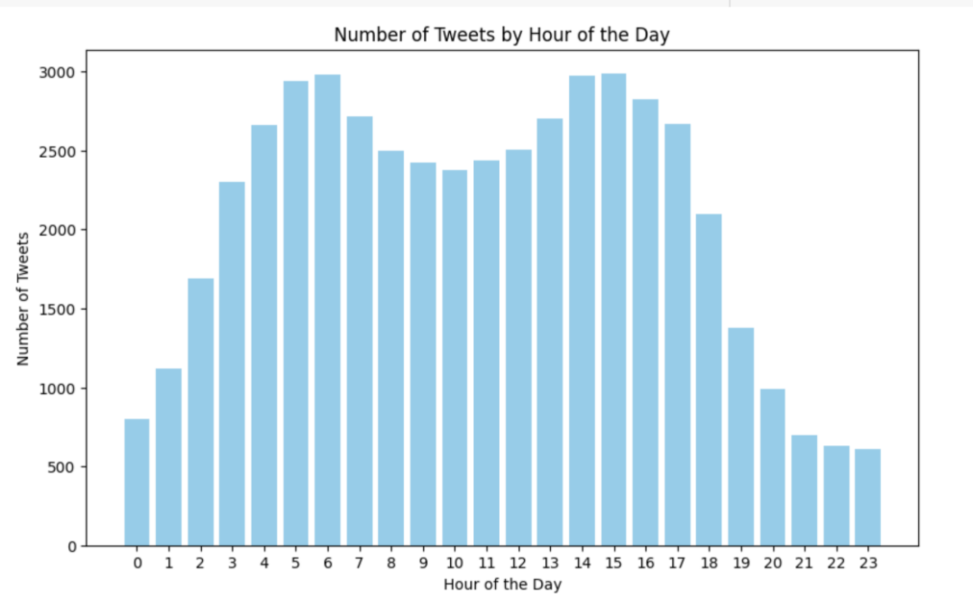


User	Tweet	Likes	Retweets	Polarity	Sentiment
AnandPatni8	respect indian citizens namaskaar I original g...	0.0	0.0	-0.062500	Negative
dhinamum	respect indian citizens namaskaar I original g...	0.0	0.0	-0.062500	Negative
PrincetonCGI	1 meet filmmaker prakash jha new jersey talk s...	0.0	0.0	0.173232	Positive
RishiJoeSanu	would politicians stop use religion politics i...	0.0	0.0	0.000000	Neutral
itweetsensee	state level president know policy pm union min...	0.0	0.0	0.214286	Positive

Exploratory Data Analysis

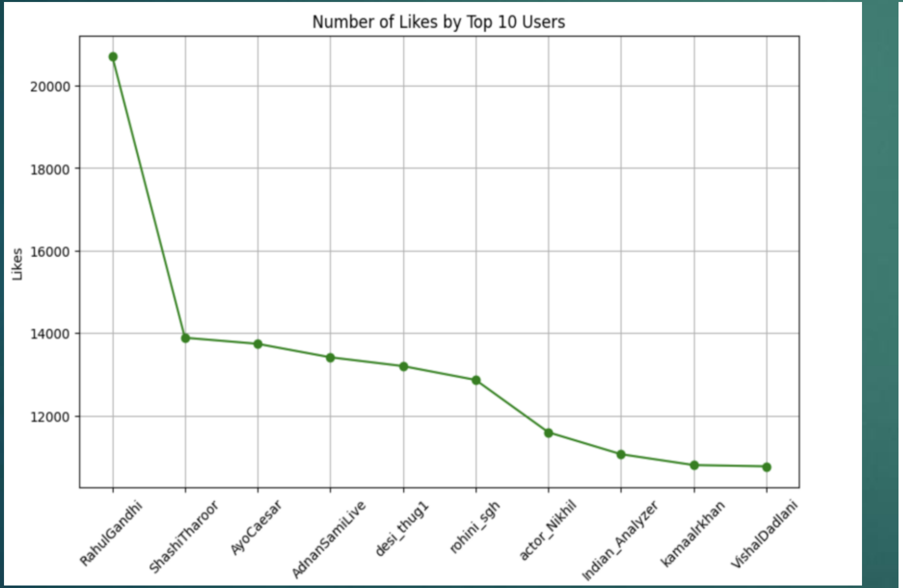
1. Number of tweets posted by hour of the day

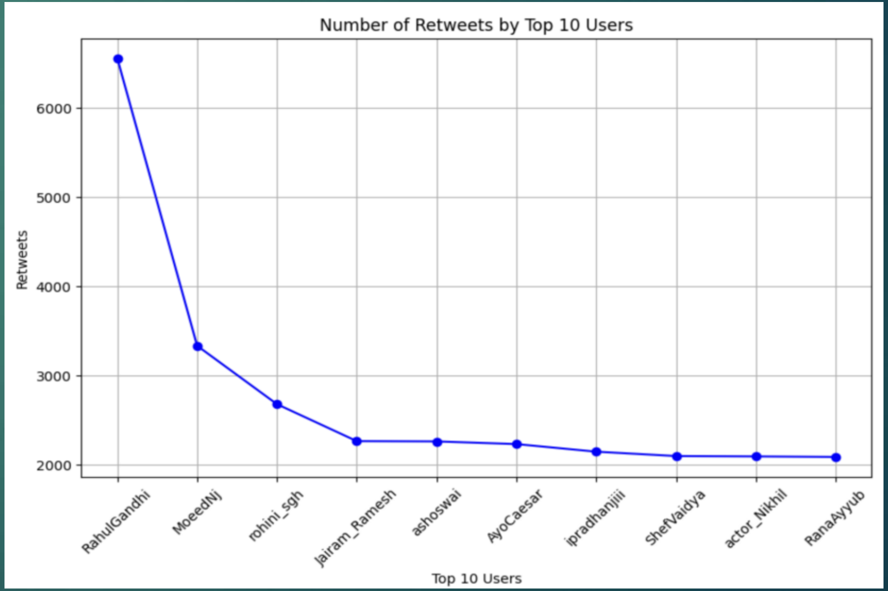
- There's a prominent spike in tweet activity between 2 PM and 3 PM.
- The early hours, notably around midnight, experience the lowest number of tweets.
- Following the mid-day surge, there's a gradual decline in tweets into the evening.
- Starting from the early morning, tweet counts consistently increase, leading up to the mid-day high point.



2. Analysis of Top User Interactions

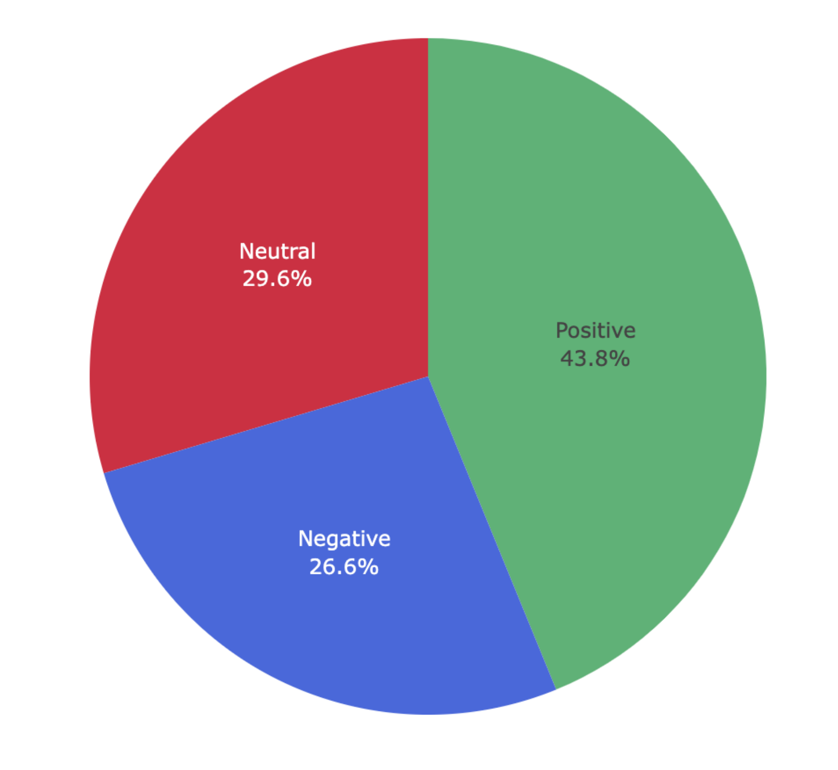
1. The first line graph visualizes the sum of likes received by the top 10 users, highlighting their popularity.
2. The subsequent graph showcases the aggregate retweets of the top 10 users, offering insight into the users' influential content.
3. We can clearly observe ‘Rahul Gandhi’ who is the opposition leader has the most number of likes & retweets signifying his positive presence on Twitter





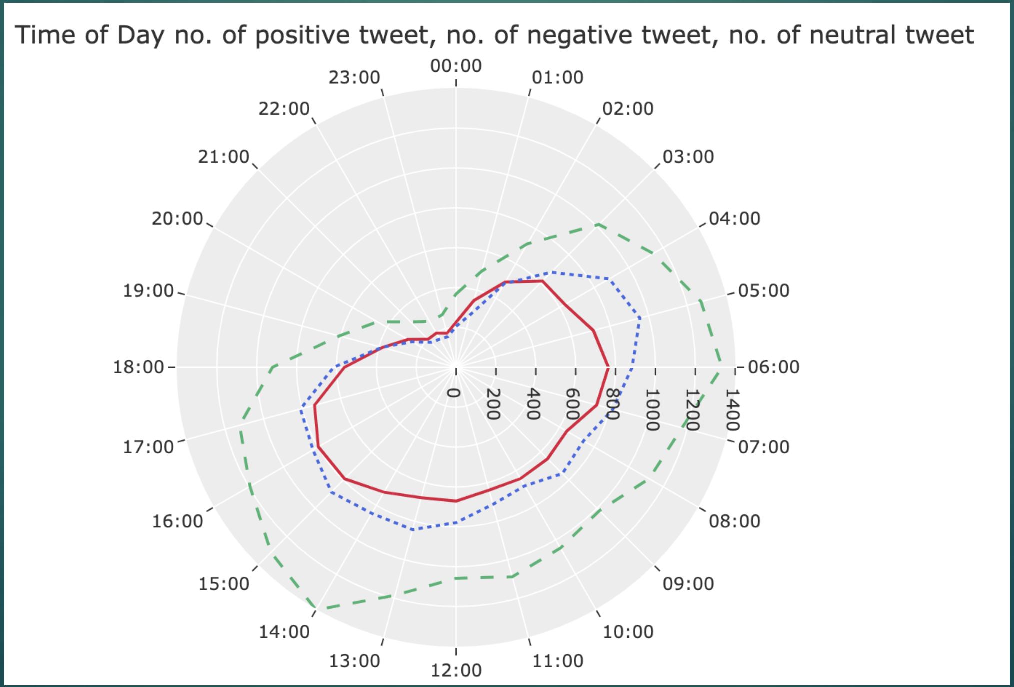
3. Distribution of tweet sentiments

- Analyzing the distribution of tweet sentiments in the dataset, calculates their respective percentages
- Almost 44% of the tweets carry a positive sentiment.
- Neutral sentiments, at nearly 30%, slightly surpass negative sentiments which stand at about 26%.
- The distribution indicates a diverse range of expressions and opinions on the platform.



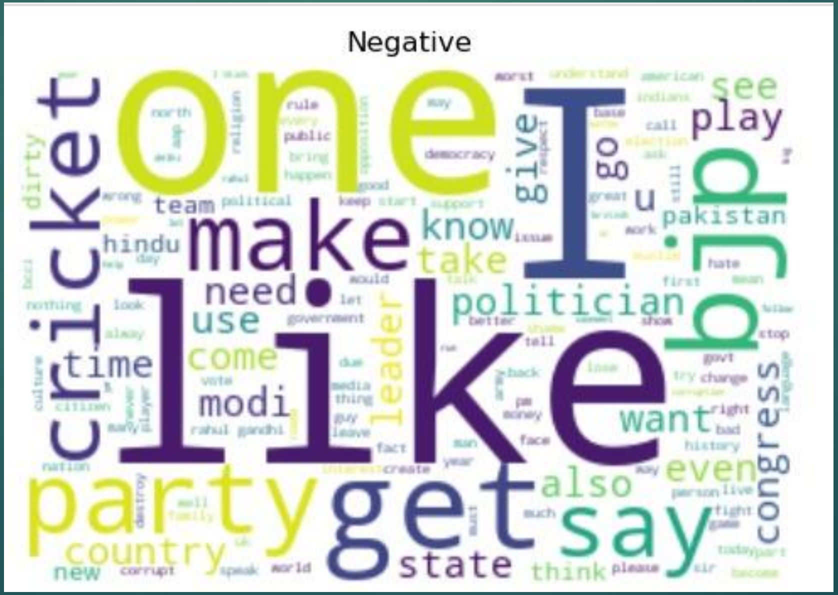
4. Daily Sentiment Trends in Tweets

- Tweets are grouped by sentiment and hour, counting the number of tweets for each combination.
- Most positive tweets are observed around 2 PM.
- Tweet numbers across all sentiments decline during the late-night hours.
- There's a surge in activity from midnight to midday
- Positive tweets generally outnumber negative and neutral ones for most of the day.



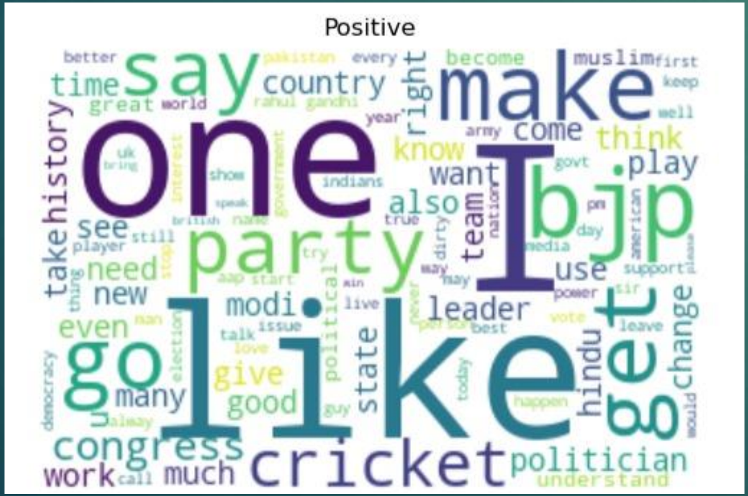
Word Cloud Generation for Tweets

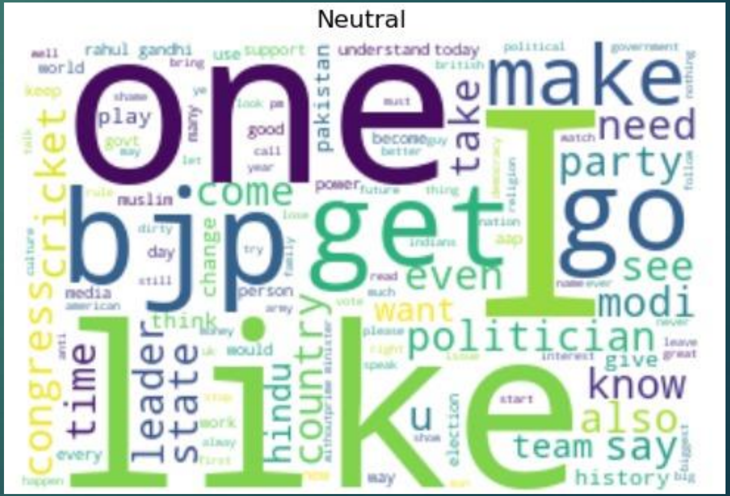
- Visualizes the most frequent words for each sentiment in tweets using word clouds.
- Words such as "BJP," "Congress," "cricket," and "Modi" are recurrent across all sentiments, indicating widespread discussions on these topics.
- In the negative sentiment, words like "politician" and "need" appear, possibly highlighting dissatisfaction or concerns with present politicians and perceived shortcomings.


Negative

Word Cloud Generation for Tweets

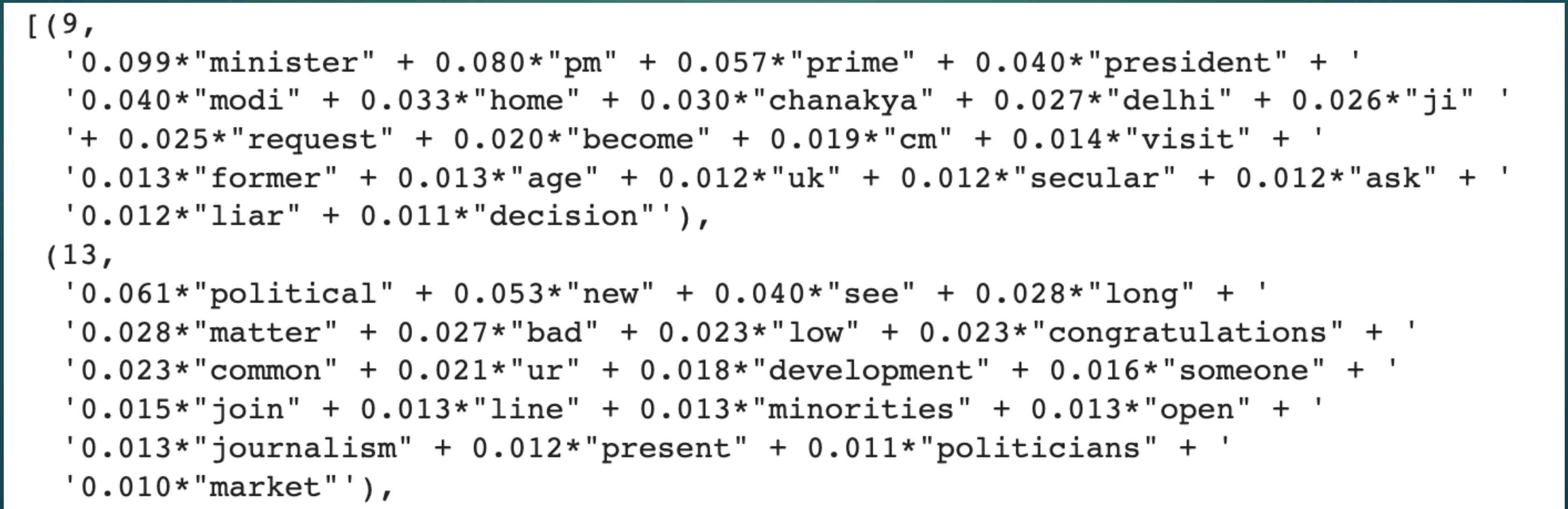
- The positive sentiment prominently features "good" and "team," hinting at positive acknowledgments or praises, possibly in contexts like teamwork or sports achievements.
- Common verbs like "go," "get," "make," and "say" are used across different sentiments, making them neutral in conveying emotion on their own.


Positive


Neutral

Topic Modeling of Tweets using LDA

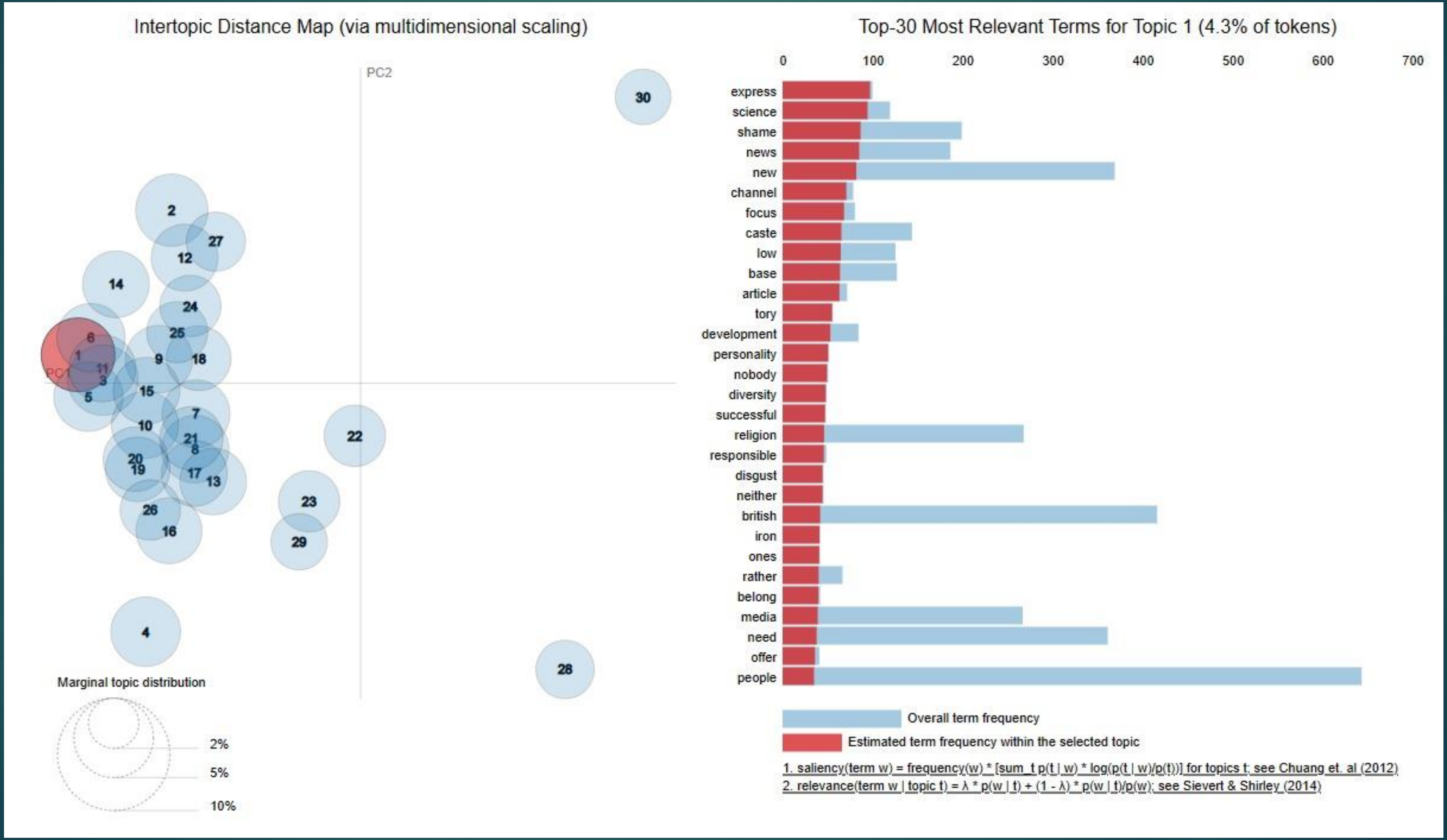
- Preprocessing the tweets from a dataframe, converting them into a list of words
- Creating a dictionary and corpus for the words using the Gensim library
- An LDA model is built to identify 30 topics from the corpus, with each topic showcasing a set of 20 most significant words.



```
[(9,
  '0.099*"minister" + 0.080*"pm" + 0.057*"prime" + 0.040*"president" + '
  '0.040*"modi" + 0.033*"home" + 0.030*"chanakya" + 0.027*"delhi" + 0.026*"ji" '
  '+ 0.025*"request" + 0.020*"become" + 0.019*"cm" + 0.014*"visit" + '
  '0.013*"former" + 0.013*"age" + 0.012*"uk" + 0.012*"secular" + 0.012*"ask" + '
  '0.012*"liar" + 0.011*"decision"'),
 (13,
  '0.061*"political" + 0.053*"new" + 0.040*"see" + 0.028*"long" + '
  '0.028*"matter" + 0.027*"bad" + 0.023*"low" + 0.023*"congratulations" + '
  '0.023*"common" + 0.021*"ur" + 0.018*"development" + 0.016*"someone" + '
  '0.015*"join" + 0.013*"line" + 0.013*"minorities" + 0.013*"open" + '
  '0.013*"journalism" + 0.012*"present" + 0.011*"politicians" + '
  '0.010*"market"'),
```

Inferences from the LDA

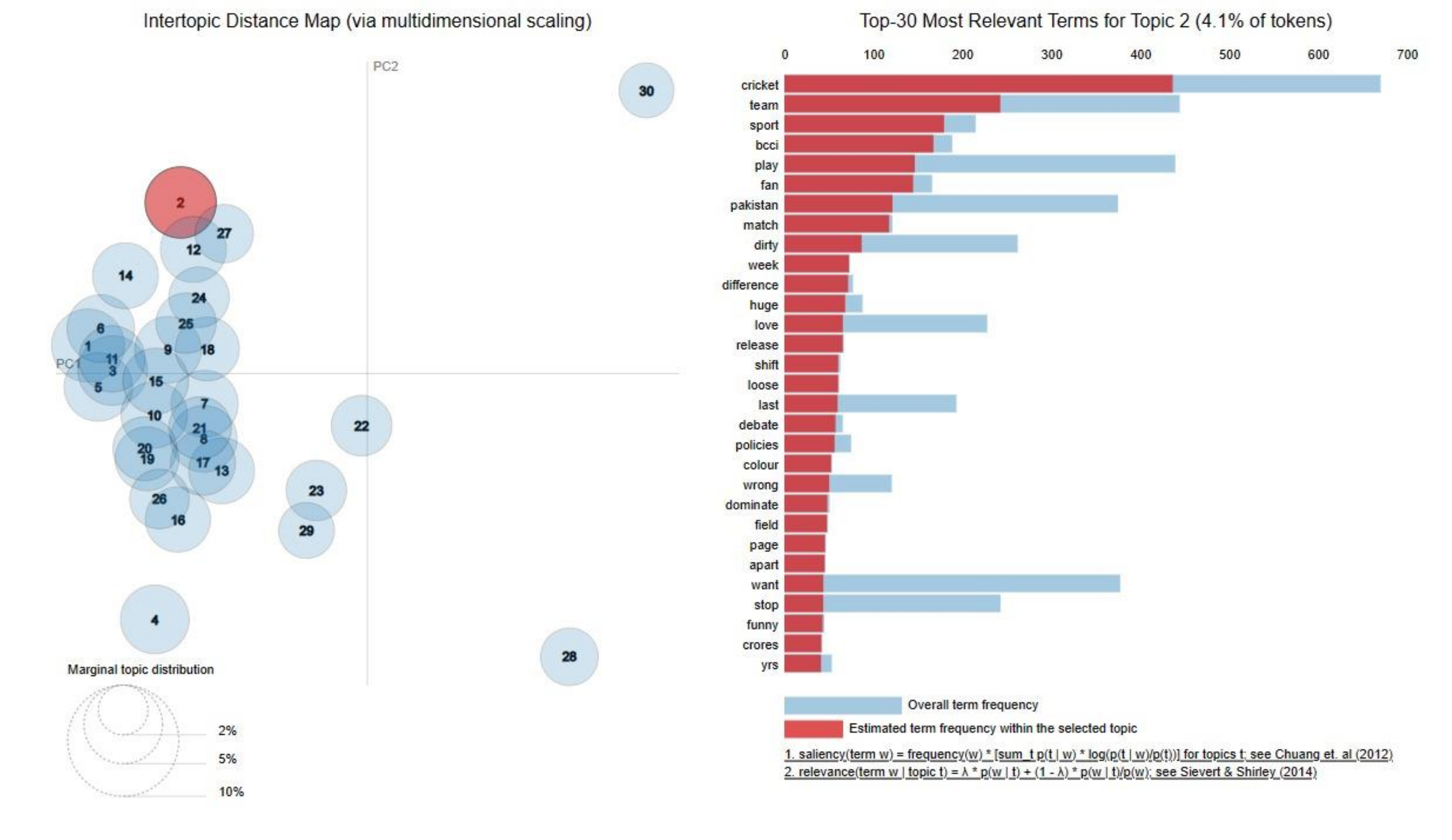
Topic-1: Social Issues and Media Influence- The topic highlights societal discussions shaped by media narratives. It delves into modern challenges like technology and its implications.



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Inferences from the LDA

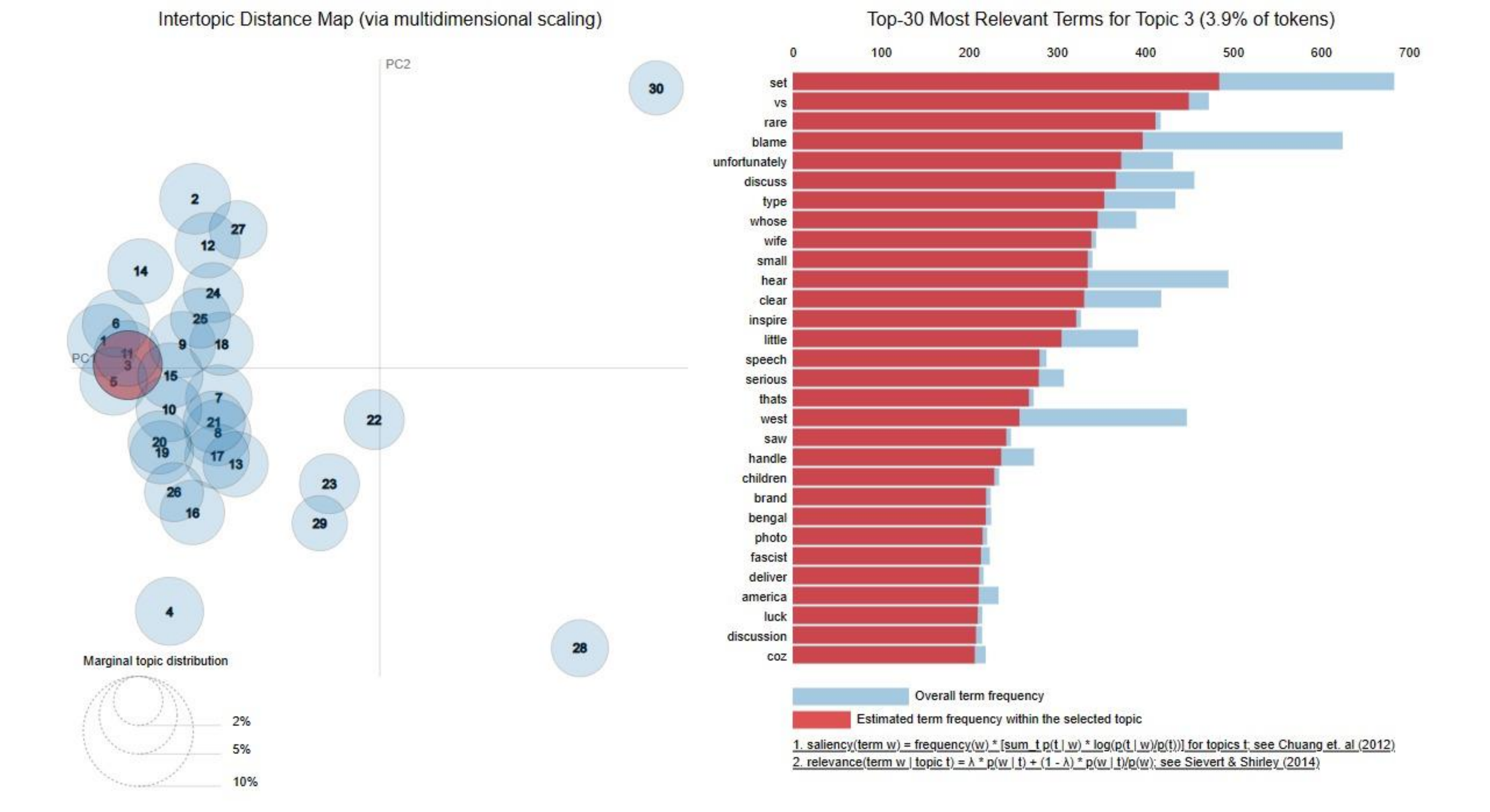
Topic-2: Sports, Events, and National Relations -The conversation is heavily centered around cricket, emphasizing its cultural significance in India. The term 'debate' suggests potential discussions on controversial sports events or decisions.



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

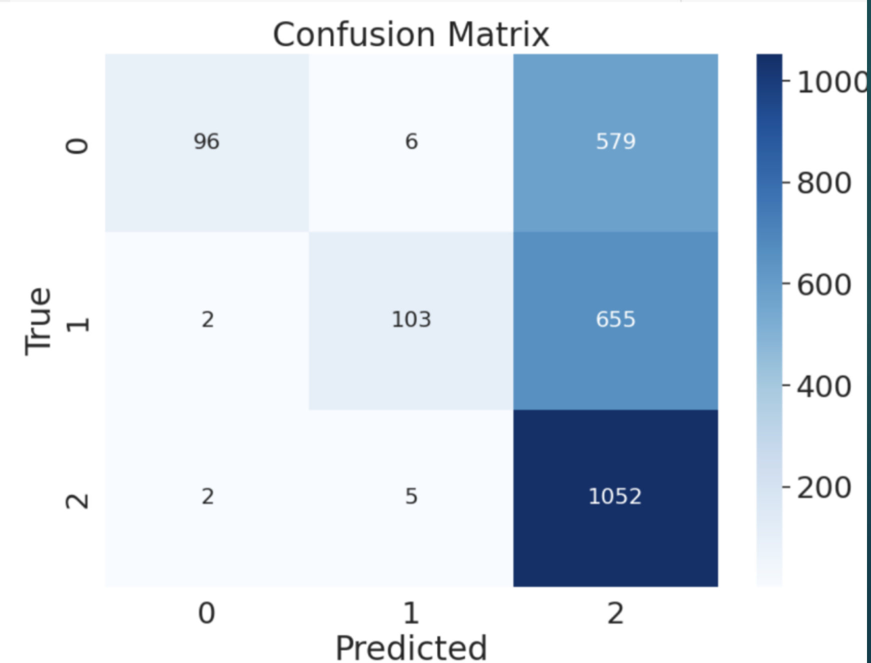
Inferences from the LDA

Topic- 3: Personal and Regional Relations- Discussions on personal relationships intertwine with regional identities and references to international relations



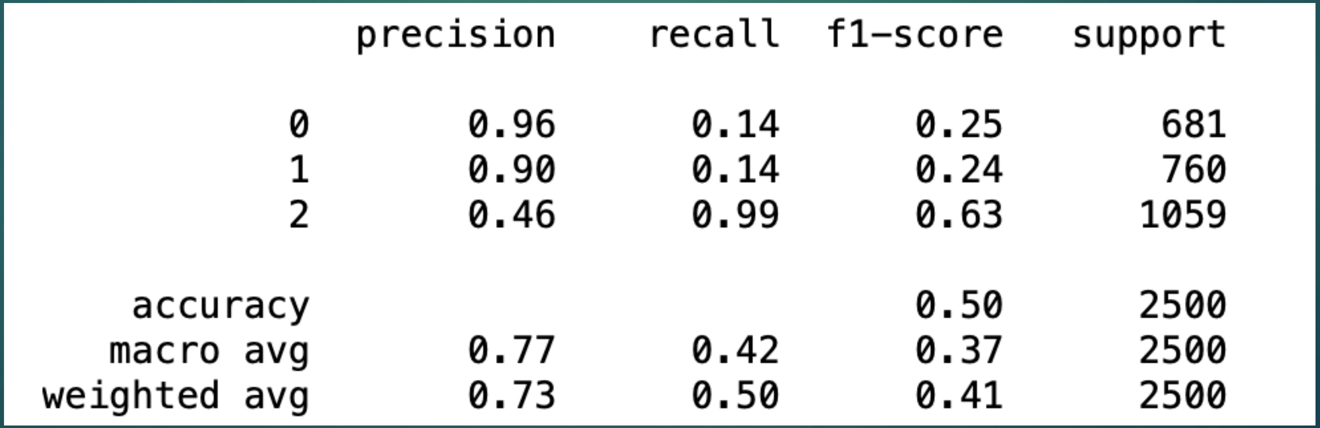
Multinomial Naive Bayes

- Preprocessing text data using tokenization and TF-IDF, sample and split into training and testing sets.
- Training a Multinomial Naive Bayes classifier and calculating a confusion matrix to evaluate the performance
- Amidst upcoming elections, the classifier accurately detected most tweets as positive, with 1052 correct picks.



Multinomial Naive Bayes

- The classifier reliably predicts negative and neutral sentiments but often misses most of them (high precision, low recall for classes 0 and 1).
- It heavily favors positive sentiment, correctly identifying most but with many false positives (high recall, lower precision for class 2).
- The overall accuracy stands at 50%, indicating mixed reliability.
- Given the upcoming elections, the results suggest a skewed perception towards positive sentiment, possibly influenced by dataset composition.

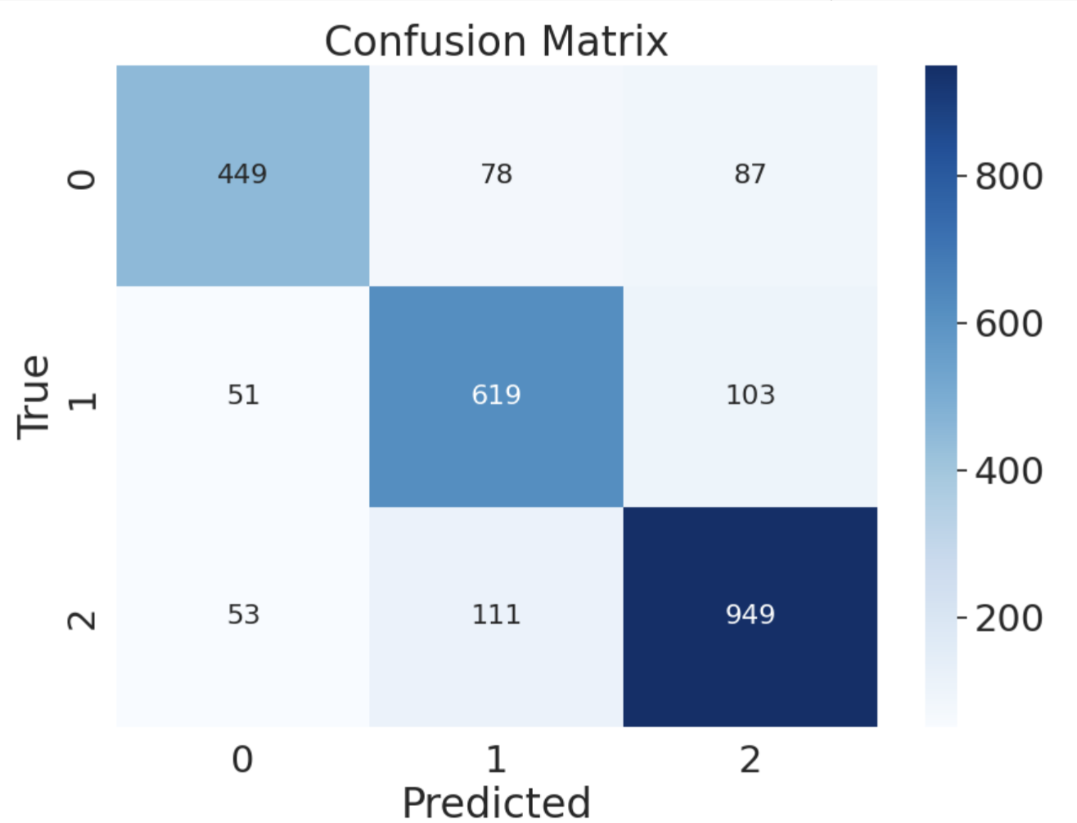


```
              precision    recall  f1-score   support

           0       0.96      0.14      0.25       681
           1       0.90      0.14      0.24       760
           2       0.46      0.99      0.63      1059

    accuracy                           0.50      2500
   macro avg       0.77      0.42      0.37      2500
weighted avg       0.73      0.50      0.41      2500
```

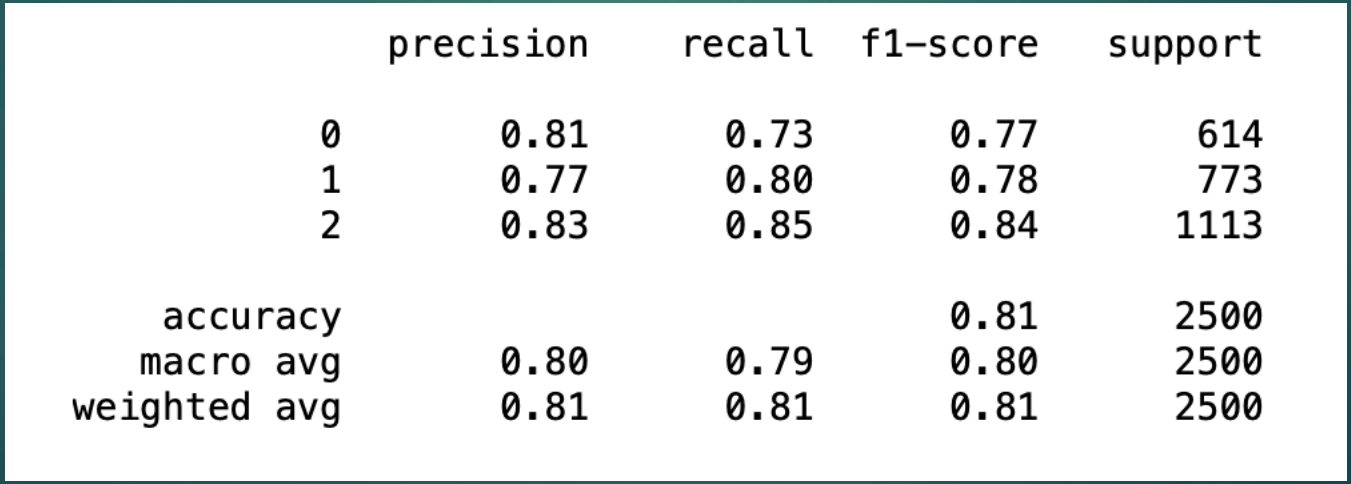
Support Vector Machine Classifier



True \ Predicted	0	1	2
0	449	78	87
1	51	619	103
2	53	111	949

Support Vector Machine Classifier

- The classifier performs consistently well across all sentiment categories, with accuracy at 81%.
- Positive sentiments (class 2) have the highest precision and recall, indicating the classifier's proficiency in detecting them.
- Negative (class 0) and neutral (class 1) sentiments also have comparable precision and recall, showcasing balanced model performance.
- Given the upcoming elections, this balanced classification indicates the dataset likely reflects a diverse set of political opinions and sentiments.



```
              precision    recall  f1-score   support

           0       0.81      0.73      0.77       614
           1       0.77      0.80      0.78       773
           2       0.83      0.85      0.84      1113

    accuracy                           0.81      2500
   macro avg       0.80      0.79      0.80      2500
weighted avg       0.81      0.81      0.81      2500
```

Model Performance Evaluation
Naive Bayes vs SVM

1. **Higher Accuracy**: SVM boasts 81% accuracy, significantly outperforming Naive Bayes at 50%.
2. **Balanced Predictions**: SVM shows consistent precision and recall, whereas Naive Bayes displays a positive sentiment bias.
3. **Better Classification**: SVM's confusion matrix reveals more accurate sentiment categorizations compared to Naive Bayes.
4. **F1-Score**: SVM has a superior F1-score, indicating a harmonized balance between precision and recall, essential for real-world applications.

Key Findings

- The frequent mentions of "BJP", "Congress", and "Modi" highlight their prominence in political discussions, suggesting BJP and Congress as the primary contesting parties with Modi as a key political figure.
- There's a prominent spike in tweet activity between 2 PM and 3 PM
- The repeated mention of "cricket" alongside major political terms underscores cricket's cultural significance in India
- SVM tends to be effective at creating a clear decision boundary, especially when the data is well-separated, while Naive Bayes relies on probabilistic calculations that can be influenced by class imbalances in the training data
- Notably, Rahul Gandhi, the opposition leader in India, has received the most likes and retweets, possibly indicating anti-incumbent sentiment for the upcoming elections.

Future Scope & Managerial Implications

- **Real-time Political Opinion Monitoring:** Developing systems for real-time monitoring of political sentiment can provide insights into public opinion trends during elections, policy changes, and political events. This can be valuable for political parties and government agencies.

- **Election Forecasting: S**entiment analysis can be used to predict election outcomes based on the sentiment expressed in tweets. By analyzing a large volume of political tweets, patterns and trends in public sentiment can be identified and correlated with election results.

- **Sentiment Analysis for Journalism:** Journalists can use sentiment analysis tools to track public sentiment on political issues and use this information to inform their reporting. It can also help in identifying trending topics.

THANK YOU !!!