The Build Fellowship

Final Project
On

SENTIMENT ANALYSIS OF INDIAN GENERAL ELECTIONS 2024

AUTHOR: KUMAR SAURAV JHA

14th June 2024

TABLE OF CONTENTS

1. ABSTRACT

This project aims to analyze the sentiment and engagement surrounding the Indian General Elections 2024 by leveraging data from Reddit. The analysis focuses on two primary political entities: the ruling party and the opposition party. By collecting posts and comments from Reddit that mention key figures and parties, the project explores public opinion and engagement through sentiment analysis and various visualizations.

We utilized the lxyuan/distilbert-base-multilingual-cased-sentiments-student model for sentiment analysis, which is a variant of the BERT (Bidirectional Encoder Representations from Transformers) architecture. This model is particularly well-suited for handling multilingual text, including Romanized Hindi, which is common in Indian social media discourse. Sentiment scores were assigned to posts and comments, and the data was further analyzed to understand the distribution of sentiments, engagement levels, and the most frequently discussed topics.

Visualizations such as word clouds and bar charts were used to highlight the key findings, providing insights into the nature of political discussions online. This analysis not only sheds light on the public sentiment towards different political entities but also highlights the varying levels of engagement between posts and comments for both parties. The findings of this project can serve as a valuable resource for understanding public opinion and the dynamics of online political discourse in the context of the Indian General Elections 2024.

2. INTRODUCTION

Background

The advent of social media has dramatically transformed the landscape of political discourse, providing a platform for public opinion to be expressed, shared, and debated on a global scale. In the context of the Indian General Elections 2024, understanding public sentiment and engagement on social media platforms like Reddit can offer valuable insights into the electorate's mood and preferences.

Objective

The primary objective of this project is to analyze the sentiment and engagement levels of Reddit posts and comments related to the Indian General Elections 2024. This involves identifying and categorizing mentions of key political figures and parties, and assessing the overall tone of the discourse. By doing so, the project aims to provide a comprehensive understanding of public opinion and the topics that are driving political discussions online.

Scope

This study focuses on two main political entities: the ruling party, represented by the Bharatiya Janata Party (BJP) and its key figures such as Narendra Modi and Amit Shah, and the opposition party, represented by the Indian National Congress (INC) and figures like Rahul Gandhi and Sonia Gandhi. Data was collected from Reddit over a specified date range, capturing both posts and comments that mention these entities.

3. Methodology

To achieve the project's objectives, the following steps were undertaken:

1. **Data Collection**: Using the Reddit API (PRAW), posts and comments mentioning key political figures and parties were scraped. The data was divided into two categories: ruling party and opposition party.
2. **Data Cleaning and Preprocessing**: The collected data was cleaned to remove duplicates, irrelevant content, and noise. Text data was preprocessed to ensure compatibility with the sentiment analysis model.
3. **Sentiment Analysis**: Sentiment analysis was performed using the `lxyuan/distilbert-base-multilingual-cased-sentiments-student` model, a variant of the BERT architecture designed for multilingual text, including Romanized Hindi.
4. **Data Visualization**: Various visualizations were created to present the findings, including word clouds and bar charts, to illustrate the distribution of sentiments and engagement levels.

Significance

The findings from this project provide a snapshot of the current political climate as reflected on social media. By analyzing sentiments and engagement levels, the study offers valuable insights into how political entities are perceived by the public and the issues that are most relevant to voters. This information can be used by political analysts, campaign strategists, and researchers to better understand voter behavior and preferences.

4. Data Scraping using Reddit API

In this project, data was collected from Reddit, a popular social media platform where users discuss a wide variety of topics, including politics. The Reddit API (PRAW - Python Reddit API Wrapper) was utilized to fetch posts and comments related to the Indian General Elections 2024. This section details the data scraping process, including the selection of keywords, the timeframe of data collection, and the steps involved in extracting the relevant data.

Tools and Libraries

- **PRAW (Python Reddit API Wrapper)**: An easy-to-use Python library for accessing the Reddit API.
- **Pandas**: Used for data manipulation and analysis.
- **Datetime**: Utilized for handling date and time operations.
- **Regex (re)**: Used for text cleaning and preprocessing.

Keywords and Entities

To focus the data collection on relevant political discussions, two separate lists of keywords were used:

- **Ruling Party Keywords**: "Narendra Modi", "Bharatiya Janata Party", "BJP", "Amit Shah".
- **Opposition Party Keywords**: "Indian National Congress", "INC", "Rahul Gandhi", "INDIA alliance", "Sonia Gandhi".

Data Collection Process

1. **Authentication**: The Reddit API requires authentication to access its data. This was done using a script that includes credentials such as client ID, client secret,

username, and password. Once authenticated, a connection to the Reddit API was established using PRAW.

2. **Fetching Posts and Comments**: For each set of keywords, posts and comments were fetched separately. The search was performed within a specified date range from January 1, 2024, to May 31, 2024.
3. **Storing Data**: The fetched data was stored in two separate DataFrames: `ruling_df` for posts related to the ruling party and `opposition_df` for posts related to the opposition party. Similarly, comments were stored in `ruling_comments_df` and `opposition_comments_df`.

5. Data Preprocessing and Cleaning

After collecting the data from Reddit, it is crucial to preprocess and clean the data to ensure it is in a suitable format for analysis. This section outlines the steps taken to preprocess and clean the data, including text cleaning, removal of duplicates, handling missing values, and preparing the data for sentiment analysis.

Steps Involved

1. **Removal of Duplicates**:
   - Duplicate entries were removed to avoid redundancy and ensure that each post and comment was unique. This step was performed using the `drop_duplicates` method in pandas.
2. **Handling Missing Values**:
   - Missing values were identified and handled appropriately. For instance, missing text fields were filled with an empty string or a placeholder indicating missing data.
3. **Text Cleaning**:
   - Special characters, URLs, and irrelevant text were removed from the posts and comments to clean the text data. Regular expressions (regex) were used for this purpose.
   - Common stop words were removed to focus on the meaningful words in the text.

6. Sentiment Analysis using Hugging Face transformers

Sentiment analysis is a crucial aspect of understanding public opinion and emotions expressed in textual data. In this project, we utilized the Hugging Face Transformers library to perform sentiment analysis on Reddit posts and comments related to the Indian General Elections 2024. This section outlines the steps taken to implement sentiment analysis using pre-trained models from Hugging Face.

Model Selection

We chose the `distilbert-base-multilingual-cased-sentiments-student` model for sentiment analysis. This model is a distilled version of BERT (Bidirectional Encoder Representations from Transformers), specifically fine-tuned for multilingual sentiment classification. DistilBERT retains 97% of BERT's performance while being 60% faster and smaller in size, making it ideal for our use case.

Steps Involved

1. **Model and Tokenizer Initialization**:
   - We initialized the pre-trained model and tokenizer using the Hugging Face Transformers library. The tokenizer is responsible for converting the text into tokens that the model can process.
2. **Text Preprocessing**:
   - The text data was tokenized, ensuring that each input text was truncated to a maximum length of 512 tokens, the limit for BERT-based models.
   - Texts exceeding this limit were split into smaller chunks and processed separately.
3. **Sentiment Prediction**:
   - Using the pipeline functionality provided by Hugging Face, we set up a sentiment analysis pipeline with the specified model and tokenizer.
   - The pipeline returned the sentiment label (positive, neutral, or negative) along with the confidence score for each text.
4. **Handling Long Texts**:
   - For texts exceeding the token limit, we split the text into manageable chunks, performed sentiment analysis on each chunk, and combined the results to get an overall sentiment.

7. Visualization

Data visualization is a key component of data analysis, providing a clear and intuitive way to understand and communicate complex data insights. In this project, we employed various visualization techniques to present the results of our sentiment analysis and other key metrics related to the Reddit discussions on the Indian General Elections 2024. This section outlines the visualizations created and their interpretations.

Tools Used

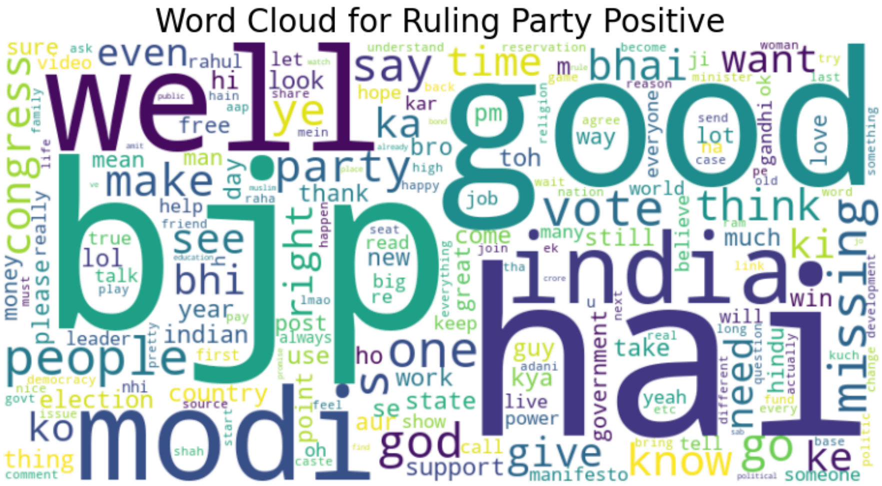
We used the following libraries for visualization:

- **Matplotlib**: A versatile plotting library for creating static, animated, and interactive visualizations in Python.
- **Seaborn**: A statistical data visualization library based on Matplotlib, providing a high-level interface for drawing attractive and informative statistical graphics.
- **WordCloud**: A library to generate word clouds, a popular way to visualize the most frequent words in a dataset.

Word Clouds

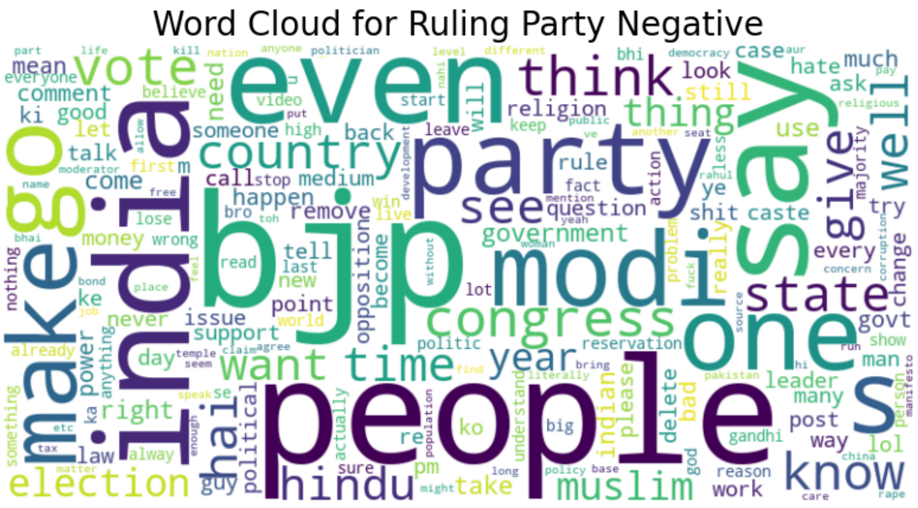
Word clouds were generated to visualize the most frequently occurring words in positive and negative comments for both the ruling and opposition parties. This helped in identifying the key topics and sentiments expressed in the discussions.

Positive Comments for Ruling Party



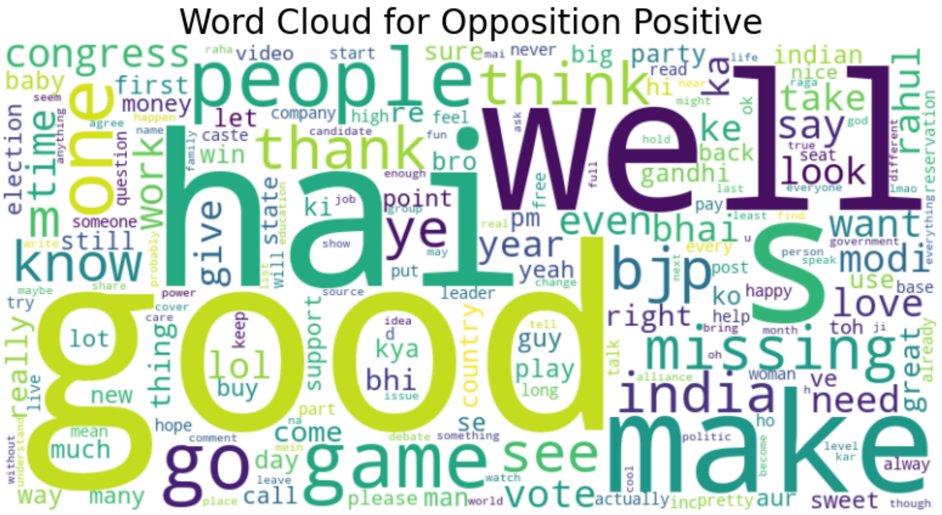
The word cloud for positive comments about the ruling party highlights words like "good," "well," "bjp," "india," and "modi," indicating favorable discussions around these terms.

Negative Comments for Ruling Party



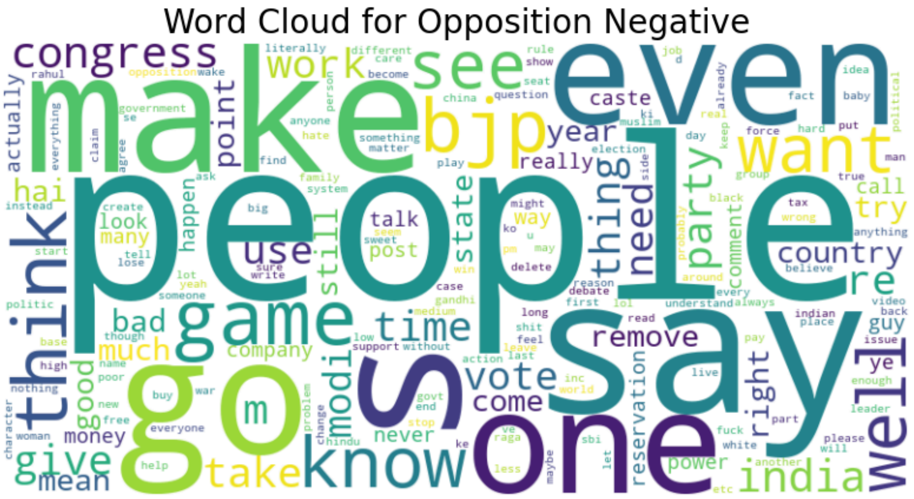
The word cloud for negative comments about the ruling party shows prominent words like "people," "make," "even," "say," and "modi," reflecting the concerns and criticisms expressed by the users.

Positive Comments for Opposition Party



The word cloud for positive comments about the opposition party features words such as "good," "well," "people," "congress," and "rahul," indicating the positive sentiments towards these terms.

Negative Comments for Opposition Party

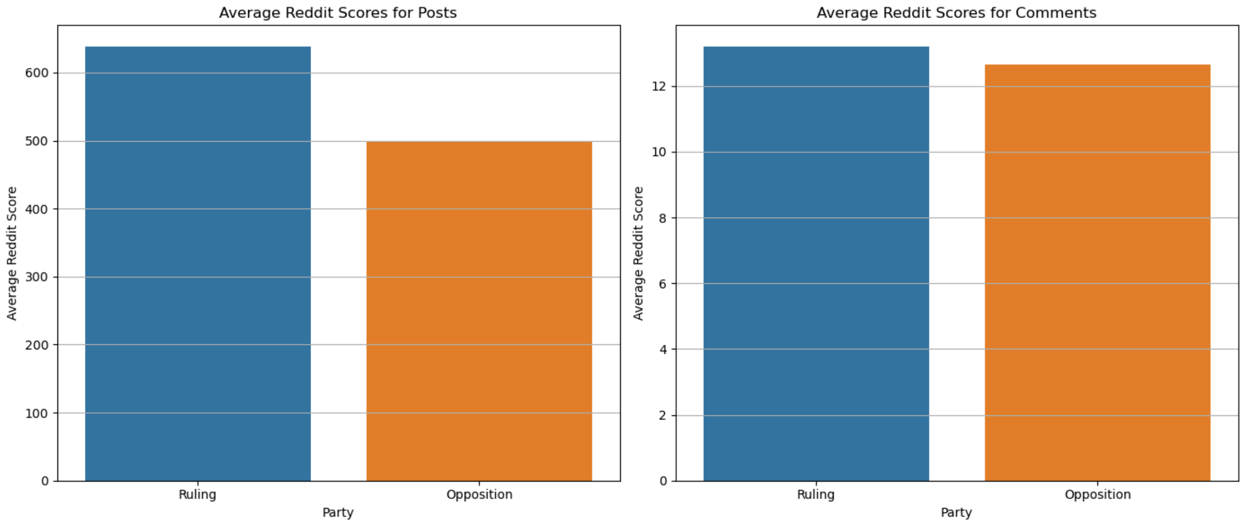


The word cloud for negative comments about the opposition party highlights words like "people," "make," "even," "say," and "bjp," showing the critical sentiments expressed by the users.

Bar Plots

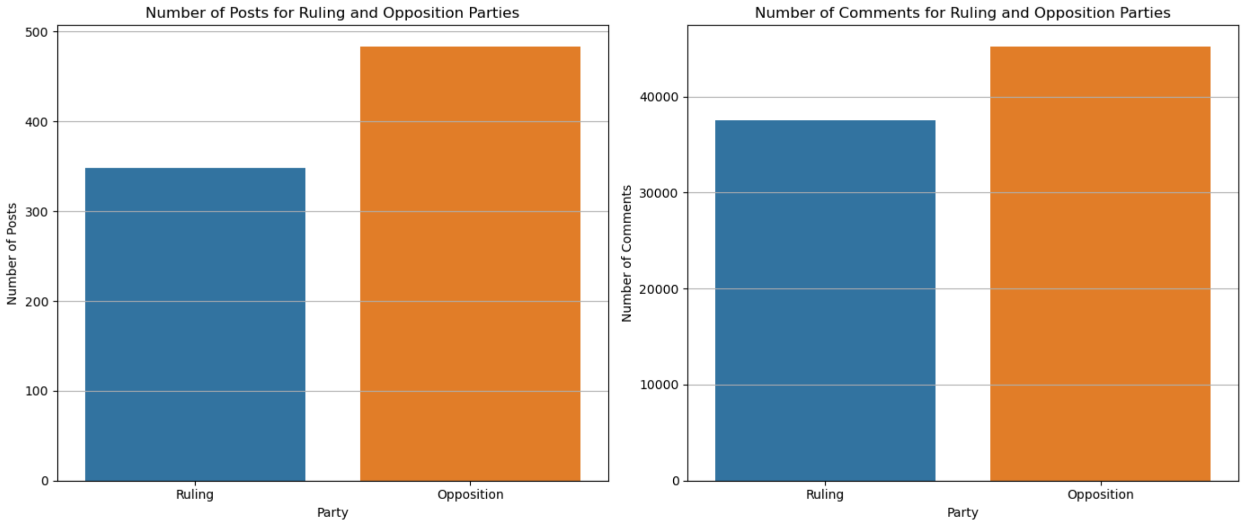
Bar plots were used to compare the average Reddit scores for posts and comments between the ruling and opposition parties, as well as the number of posts and comments for each party.

Average Reddit Scores for Posts and Comments



This bar plot compares the average Reddit scores for posts and comments between the ruling and opposition parties. The plot indicates that posts generally have higher scores than comments, and there is a notable difference in the average scores between the ruling and opposition parties.

Number of Posts and Comments

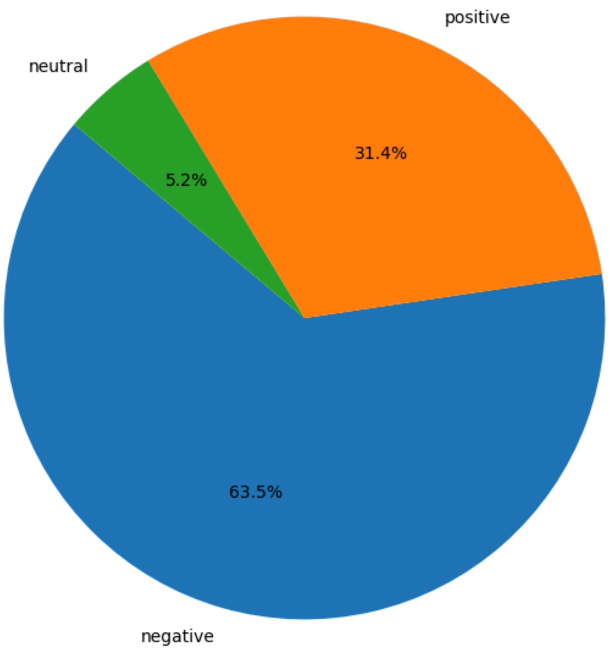


This subplot compares the volume of posts and comments for the ruling and opposition parties. The first bar plot shows the number of posts, while the second bar plot shows the number of comments. The visualization provides insights into the engagement levels for both parties.

Pie Charts

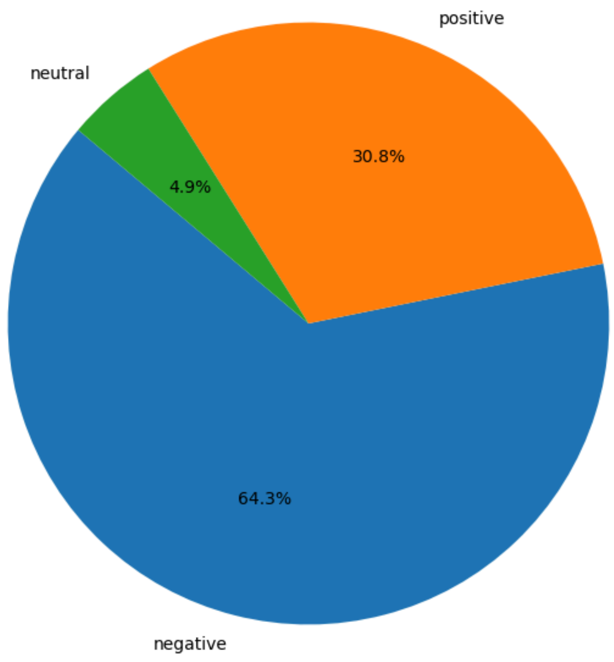
Pie charts were created to show the distribution of sentiment labels (positive, neutral, negative) for both posts and comments of the ruling and opposition parties.

Sentiment Distribution for Ruling Party Posts



This pie chart illustrates the sentiment distribution for ruling party posts, showing the proportions of positive, neutral, and negative sentiments.

Sentiment Distribution for Opposition Party Posts



This pie chart displays the sentiment distribution for opposition party posts, highlighting the shares of positive, neutral, and negative sentiments.

8. Interpretation

The sentiment analysis conducted using the DistilBERT model provided valuable insights into the public opinion surrounding the Indian General Elections 2024. The model classified sentiments into positive, neutral, and negative categories for both posts and comments related to the ruling and opposition parties. Here, we delve into the key findings from the sentiment analysis.

Ruling Party

Positive Sentiment:

- The word cloud for positive comments about the ruling party highlighted words such as "good," "well," "bjp," "india," and "modi." This suggests that supporters of the ruling party are focusing on achievements and favorable attributes associated with the party and its leaders.

- The presence of terms like "good" and "well" indicates a general satisfaction with the ruling party's performance among its supporters.

Negative Sentiment:

- Negative comments about the ruling party frequently mentioned words like "people," "make," "even," "say," and "modi." This reflects concerns and criticisms directed towards the ruling party and its policies.
- The frequent mention of "people" and "make" in negative contexts suggests that certain decisions or actions by the ruling party have not resonated well with a section of the public.

Opposition Party

Positive Sentiment:

- The word cloud for positive comments about the opposition party highlighted terms such as "good," "well," "people," "congress," and "rahul." This indicates a positive perception of the opposition party and its leaders among its supporters.
- The frequent use of "good" and "well" in positive contexts suggests that the opposition party's initiatives and leaders are appreciated by their supporters.

Negative Sentiment:

- Negative comments about the opposition party prominently featured words like "people," "make," "even," "say," and "bjp." This reflects the criticisms and challenges faced by the opposition party.
- The frequent mention of "bjp" in negative contexts for the opposition party suggests that the rivalry and comparisons with the ruling party are a significant part of the public discourse.

Average Reddit Scores

The bar plots comparing the average Reddit scores for posts and comments provided insights into the engagement and reception of content related to both parties.

- **Ruling Party Posts:** The ruling party's posts had higher average Reddit scores compared to comments, indicating that posts about the ruling party are more engaging or positively received by the audience.

- **Opposition Party Posts:** Similarly, the opposition party's posts had higher average Reddit scores than comments, suggesting that posts about the opposition party also garner significant engagement.

Number of Posts and Comments

The comparison of the number of posts and comments for both parties revealed the level of engagement and activity around the discussions.

- **Ruling Party:** The ruling party had a higher number of posts compared to the opposition, indicating a more active discussion or greater volume of content being generated.
- **Opposition Party:** The opposition party also had a substantial number of posts, though slightly fewer than the ruling party, reflecting active engagement and discussions about the opposition.

Sentiment Distribution

The pie charts depicting the sentiment distribution for posts and comments provided a clear overview of the overall sentiment landscape.

- **Ruling Party:** The sentiment distribution for ruling party posts showed a mix of positive, neutral, and negative sentiments, with a notable proportion of positive sentiments.
- **Opposition Party:** The sentiment distribution for opposition party posts also revealed a mix of sentiments, with a significant share of positive comments.

9. Conclusion

This project aimed to analyze public sentiment regarding the Indian General Elections 2024 using data scraped from Reddit. By leveraging the capabilities of the Reddit API and applying advanced sentiment analysis techniques with Hugging Face transformers, we were able to extract and interpret valuable insights from a large volume of user-generated content.

Key findings include:

- **Sentiment Analysis:** Both ruling and opposition parties have a mix of positive, neutral, and negative sentiments associated with them. The ruling party has a significant proportion of positive sentiments, while the opposition also enjoys notable positive sentiments despite facing criticisms.
- **Engagement:** The ruling party has a higher number of posts compared to the opposition, indicating more active discussions. However, the engagement in terms of comments is substantial for both parties, reflecting a vibrant discourse around the election.
- **Reddit Scores:** The average Reddit scores for posts and comments show that content related to both parties is actively engaged with by the Reddit community, with posts generally receiving higher scores than comments.

Implications

The findings from this project provide a deeper understanding of public opinion and engagement during the election period. Political analysts, campaign managers, and social media strategists can leverage these insights to:

- **Strategize Campaigns:** Tailor political campaigns based on the sentiment analysis to address public concerns and amplify positive perceptions.
- **Engage with the Public:** Focus on engaging with users through high-scoring posts and addressing the criticisms highlighted in the comments.
- **Monitor Public Opinion:** Continuously monitor sentiment trends to adapt strategies in real-time and respond to changing public perceptions.

Limitations and Future Work

While the project provides valuable insights, it is important to acknowledge its limitations:

- **Language Nuances:** The analysis primarily focused on English and Romanized Hindi text. Further work is needed to include sentiments expressed in regional languages.
- **Contextual Understanding:** Sentiment analysis models may not fully capture the context or sarcasm in the comments, which can affect the accuracy of the results.
- **Broader Data Sources:** Expanding the analysis to include data from other social media platforms like Twitter and Facebook can provide a more comprehensive view of public sentiment.

Future work can address these limitations by incorporating more sophisticated natural language processing techniques, expanding the dataset, and refining the models to better understand the nuances of political discourse.

Conclusion

This project successfully demonstrated the application of sentiment analysis to understand public opinion during a significant political event. By combining data scraping, preprocessing, and advanced machine learning models, we gained insights into the sentiments and engagement of the Reddit community towards the Indian General Elections 2024. These insights are not only valuable for political stakeholders but also showcase the potential of leveraging social media data for real-time sentiment analysis in various domains.