Exploring Political Discourse on Social Media
(Indian Lok Sabha Elections - 2024)

Submitted by

Ms. SANJANA NEELI NAGARAJ

(277227)

MSc in DATA SCIENCE

SCHOOL OF MATHEMATICAL AND PHYSICAL SCIENCES

Under the guidance of

Dr. JACK PAY

SCHOOL OF ENGINEERING AND INFORMATICS

Acknowledgement

I would like to express my deepest gratitude to my supervisor, Dr. Jack Pay, for his invaluable guidance, support, and encouragement throughout the course of this research. His insights and expertise were instrumental in shaping this dissertation.

A special thanks goes to my family, especially my parents, for their unwavering love, encouragement, and understanding throughout this journey. Their belief in my abilities and constant support provided me with the strength to persevere during the most challenging moments.

I am also grateful to the faculty members of University of Sussex for their academic and administrative support. Lastly, I would like to acknowledge the various social media platforms and the creators whose content provided the data for this research. Without their contributions, this study would not have been possible.

Declaration

I, Ms. Sanjana Neeli Nagaraj, hereby declare that this dissertation titled ”Exploring Political Discourse on Social Media (Indian Lok Sabha Elections-2024)” is my original work and has not been submitted for any other degree or diploma at any other university or institution. All sources of information have been duly acknowledged, and the work is free from any form of plagiarism.

Abstract

This dissertation explores the evolving dynamics of political discourse on social media in the context of the 2024 Indian Lok Sabha elections. With the increasing influence of social media platforms like YouTube, Twitter, and Facebook on political campaigns, this study examines how major political parties, specifically the Bharatiya Janata Party (BJP) and the Indian National Congress (INC), utilize these platforms to engage with voters and shape public sentiment.

The research employs advanced sentiment analysis models, including BERTweet, Twitter RoBERTa, and RoBERTa Large, to analyze a large dataset of social media comments. These models are used to classify sentiments expressed towards the BJP and INC, and to conduct a temporal analysis of public sentiment over the election campaign period.

The findings reveal significant insights into the effectiveness of digital political campaigns, the factors driving public sentiment, and the broader implications for democratic engagement in India. The results contribute to the understanding of political communication in the digital age and offer strategic recommendations for future political campaigns.

Contents

List of Tables

List of Figures

1 Introduction

1.1 Background

In India, political discourse is dynamic and multifaceted, reflecting the country's rich cultural and social diversity. This discourse is significantly influenced by various political parties, each representing distinct regional, religious, and economic interests. Major alliance like the Bharatiya Janata Party (BJP) and the Indian National Congress (INC) dominate the scene, while numerous smaller regional parties also play crucial roles in shaping national policies and influencing electoral outcomes. These regional parties often hold the balance of power, adding to the vibrant and complex political landscape of India (Heinrich Böll Stiftung, 2024; Times of India, 2024).

Social media has significantly transformed the landscape of Indian politics. Platforms such as Facebook, YouTube, Twitter, and WhatsApp have become indispensable tools for political campaigns, allowing politicians to reach a wide audience, disseminate their messages, and efficiently mobilize support. This impact is particularly notable among younger and urban voters, making social media a powerful instrument for shaping public opinion and swiftly spreading political ideas (Heinrich Böll Stiftung, 2024; Staff, 2024).

The 2024 Lok Sabha elections, illustrated in Figure 1, are particularly significant as they will determine the members of the 18th Lok Sabha, India's lower house of Parliament. These elections come at a pivotal time, with the nation grappling with challenges such as economic recovery post-COVID-19, agricultural reforms, unemployment, and national security. For the Bharatiya Janata Party (BJP), which has been in power since 2014 under Prime Minister Narendra Modi, this election is a critical test. Meanwhile, opposition parties like the Indian National Congress (INC) are striving to challenge the BJP's dominance and offer new solutions to the public (Times of India, 2024; Staff, 2024). The results of these elections will greatly impact India's future policies and political direction, shaping the trajectory of the nation's development and its approach to addressing pressing socio-economic issues (Heinrich Böll Stiftung, 2024; Times of India, 2024; Staff, 2024).

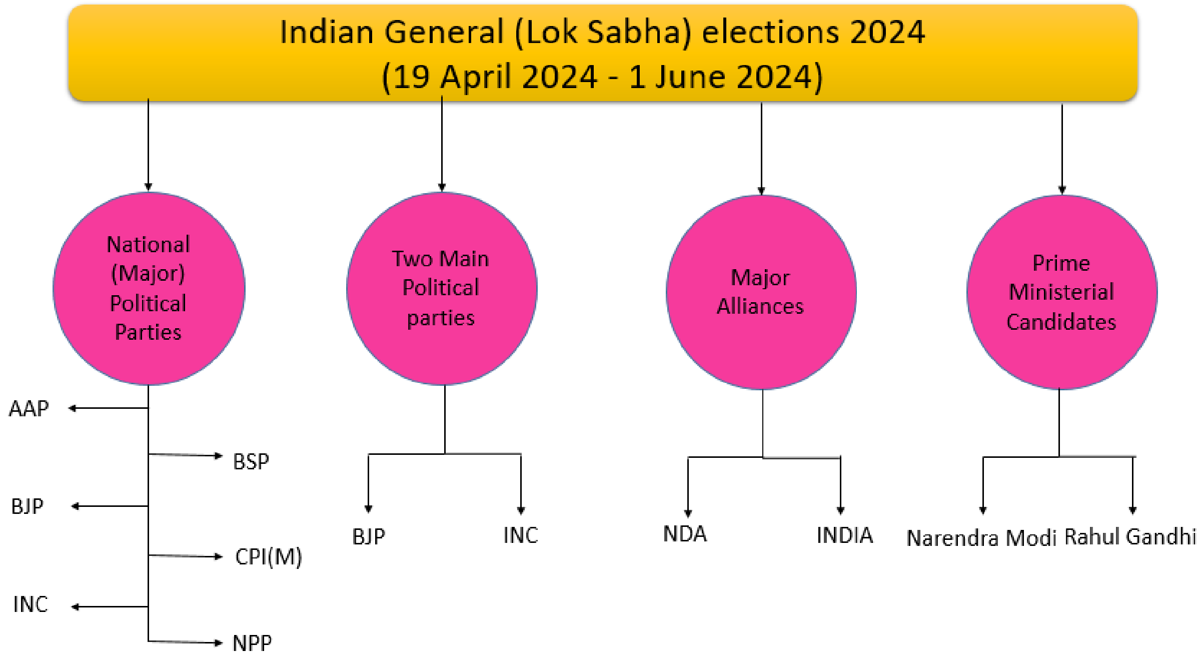

Indian General (Lok Sabha) elections 2024
(19 April 2024 - 1 June 2024)
National (Major) Political Parties
Two Main Political parties
Major Alliances
Prime Ministerial Candidates
AAP
BSP
BJP
CPI(M)
INC
NPP
BJP
INC
NDA
INDIA
Narendra Modi
Rahul Gandhi

Figure 1: Outline of Indian General Elections

1.2 Problem Statement

- **Statement 1: Impact of Social Media on Political Discourse**

  The widespread adoption of social media platforms has significantly transformed the landscape of political discourse in India, particularly in the context of electoral politics. The 2024 Lok Sabha elections are anticipated to be a pivotal event where the influence of social media on voter behavior and public opinion is more pronounced than ever before. This study aims to examine how major political parties, specifically the Bharatiya Janata Party (BJP) and the Indian National Congress (Congress), leverage social media to disseminate their political narratives, mobilize support, and engage with the electorate. By conducting a comprehensive analysis of social media content, including posts, comments, and multimedia shared on platforms such as YouTube, Facebook, and Twitter, this research seeks to understand the strategies employed by these parties to influence public sentiment. Furthermore, the study will explore the effectiveness of these digital campaigns in shaping political opinions and their implications for democratic engagement and electoral outcomes. The findings of this research will provide valuable insights into the evolving dynamics of political communication in the digital age and contribute to the broader discourse on the role of social media in modern democracies (Carlisle and Patton, 2013; Wolfsfeld et al., 2013).

- **Statement 2: Sentiment Analysis of Political Parties**

  As the 2024 Lok Sabha elections approach, gauging public sentiment towards political parties is essential for understanding voter preferences and predicting electoral outcomes. This research focuses on applying advanced sentiment analysis techniques to social media comments related to the BJP and Congress. The primary objective is to classify these comments into positive, negative, or neutral sentiments and compare the overall sentiment trends for both parties. By employing state-of-the-art natural language processing models such as BERTweet (”finiteautomata/bertweet-base-sentiment-analysis”), Twitter RoBERTa (”cardiffnlp/twitter-roberta-base-sentiment”), and RoBERTa Large (”siebert/sentiment-roberta-large-english”), this study aims to provide a nuanced understanding of how the public perceives the two main political contenders.

  In addition to sentiment analysis, the study utilizes BERTopic for classifying comments into specific themes and events. BERTopic, a topic modeling technique leveraging transformer-based embeddings, is employed to uncover the underlying topics and events driving public discourse. By categorizing comments into distinct themes—such as policy announcements, campaign events, and socio-political issues—BERTopic helps identify the key drivers of positive and negative sentiments. This approach not only enhances the granularity of sentiment analysis but also provides strategic insights for political campaigns, enabling parties to tailor their messages and strategies more effectively.

  The results will also shed light on the broader implications of social media sentiment and thematic discussions for political engagement and democratic processes (Agarwal and Singh, 2023; Saini and Arora, 2021).

- **Statement 3: Temporal Analysis of Political Sentiment**

  Political sentiment is inherently dynamic, often fluctuating in response to events, policy announcements, and campaign activities. Understanding these temporal changes is crucial for political analysts and strategists. This study aims to conduct a temporal analysis of public sentiment towards the BJP and Congress during the 2024 Lok Sabha election campaign. By employing time-series analysis techniques, the research will track sentiment trends over the course of the campaign, identifying significant shifts and their corresponding triggers. Key events such as major policy announcements, debates, rallies, and controversies will be analyzed to understand their impact on public sentiment. The study will utilize advanced sentiment analysis models to process and classify social media comments collected at different points in time, providing a detailed temporal map of political sentiment. By correlating sentiment changes with specific events and activities, the research aims to uncover patterns and trends that can inform real-time campaign strategies and decision-making. The insights gained from this temporal analysis will not only enhance the understanding of voter behavior during election campaigns but also contribute to the broader field of political science by illustrating the dynamic interplay between political events and public opinion (Razzaq et al., 2014; Younus and Gulzar, 2014).

1.3 Objectives

- **Evaluate Sentiment Analysis Models**: The first objective is to evaluate the performance of advanced sentiment analysis models in analyzing political discourse. The models to be assessed include BERT Multilingual, RoBERTa, and another comparable model. These models will be fine-tuned using a labeled dataset of social media comments related to the Bharatiya Janata Party (BJP) and the Indian National Congress (INC). The performance of these models will be evaluated based on key metrics such as accuracy, precision, recall, and F1-score. This objective aims to identify the model that best captures the nuances of political sentiment in the multilingual and culturally diverse context of Indian politics. The evaluation process will also analyze the strengths and weaknesses of each model in handling various linguistic and contextual challenges inherent in social media data. The ultimate goal is to establish a robust and reliable methodology for sentiment analysis that can be applied to future political discourse studies. This contributes to advancements in natural language processing techniques and their application in political communication research.

- **Compare Sentiment Towards BJP and Congress**: The second objective is to analyze public sentiment towards the BJP and INC using sentiment analysis techniques. By classifying social media comments into positive, negative, or neutral sentiments, the study will assess public perceptions of the main political contenders. Additionally, the research will examine specific events or themes that influence these sentiments, providing strategic insights for political campaigns and understanding their broader implications for political engagement.

- **Analyze the Role of Social Media in Political Campaigns**: The third objective is to investigate how social media platforms have transformed political campaigning in India. Social media has become an indispensable tool for political campaigns, allowing politicians to reach a wide audience, disseminate their messages, and efficiently mobilize support. This impact is particularly notable among younger and urban voters. The research will conduct a comprehensive analysis of social media content, including posts, comments, and multimedia shared by political parties on platforms like YouTube, Facebook, and Twitter. This objective aims to highlight the effectiveness of social media in spreading political messages and influencing voter behavior, providing insights into the evolving dynamics of political communication in the digital age.

- **Forecast Potential Outcomes and Implications**: The final objective is to analyze potential outcomes of the 2024 elections and their implications for India's future political direction. Using predictive models and expert analysis, the study will explore various electoral scenarios and their potential impact on policy-making, governance, and democratic processes. This objective aims to offer insights into how digital engagement and political discourse may shape the future of Indian democracy, contributing to the broader field of political science.

1.4 Research Questions/Hypotheses

This study aims to explore several critical research questions and hypotheses to understand the performance of sentiment analysis models and public sentiment toward major political parties in India, especially in the context of the 2024 Lok Sabha elections.

1. Which sentiment analysis model performs best in analyzing political discourse?

- BERTweet, Twitter RoBERTa, and RoBERTa Large will exhibit different performance levels, with one showing superior accuracy, precision, recall, and F1-score (Devlin et al., 2019; Liu et al., 2019).

2. Which political party, BJP or INC, has more positive comments on social media?

- There will be a significant difference in the distribution of sentiments between BJP and INC, influenced by socio-political factors and events (Pang and Lee, 2008).

3. How do specific events or announcements impact public sentiment towards BJP and INC?

- Policy announcements and socio-political issues will significantly impact sentiment trends for both parties on social media (Tumasjan et al., 2010).

4. What role does social media play in shaping voter behavior among young and urban populations?

- Social media significantly influences the political participation and voting behavior of younger and urban voters (Allcott and Gentzkow, 2017).

5. What challenges and opportunities exist in digital campaigning for Indian political parties?

- Misinformation and digital manipulation present challenges but also offer opportunities for enhancing political engagement (Lazer et al., 2018).

1.5 Significance of the Study

The study "Exploring Political Discourse on Social Media (Indian Lok Sabha Elections-2024)" is crucial in understanding political sentiments and its implications for campaigns and public opinion. This research offers real-time insights into voter sentiment by leveraging sentiment analysis models like BERT and RoBERTa, which are known for their superior performance (Devlin et al., 2019; Liu et al., 2019).

Analyzing social media data provides political parties with immediate feedback on public reactions to events, policy announcements, and campaign strategies, allowing them to adjust their approaches dynamically. This real-time sentiment analysis can significantly enhance campaign effectiveness by aligning strategies with voter preferences (Tumasjan et al., 2010).

Understanding the impact of socio-political events on public sentiment is another critical aspect. Heinrich Böll Stiftung (2024) argues that this study will elucidate how events influence voter attitudes towards major parties, enabling more informed decision-making by political actors. Additionally, identifying and mitigating misinformation and digital manipulation on social media platforms is essential for maintaining the integrity of political discourse. This research will contribute to developing strategies to counteract misinformation, thereby promoting an informed electorate (Lazer et al., 2018).

The study also explores the role of social media in shaping voter behavior, particularly among young and urban populations, who are prolific users of these platforms. Insights gained can guide political campaigns in engaging these demographics more effectively, fostering greater political participation and voter turnout (Allcott and Gentzkow, 2017).

Finally, the research has broader implications for political science and communication studies. It provides empirical data on the effectiveness of sentiment analysis models in political discourse, offering valuable resources for scholars in these fields. Overall, this study enhances our understanding of political sentiments on social media, contributing to more effective and informed political engagement and discourse.

1.6 Structure of the Dissertation

This dissertation is organized into seven main sections, each dedicated to exploring various aspects of political discourse on social media in the context of the Indian Lok Sabha Elections of 2024. The introduction provides an overview of the dissertation, outlining the research context, background, problem statement, objectives, and significance of the study. It sets the stage for the subsequent sections by highlighting the dynamic nature of political discourse in India and the transformative impact of social media.

The literature review examines existing research on political communication, the role of social media in electoral politics, sentiment analysis, and the specific context of Indian elections. By identifying gaps in the current research, this section establishes the theoretical framework for the study.

The research methodology section details the research design, data collection methods, and analytical techniques used in the study. It explains the selection of social media platforms, the process of data collection, and the sentiment analysis models employed, such as BERT and RoBERTa.

Next, the impact of social media on political discourse is analyzed, focusing on how major political parties, specifically the BJP and INC, use social media to disseminate their messages and engage with the electorate. This section delves into the content and strategies employed by these parties to influence public sentiment.

The sentiment analysis of political parties section presents the results of sentiment analysis on social media comments related to the BJP and INC. It compares the overall sentiment trends for both parties and identifies key drivers of positive and negative sentiments.

Following this, a temporal analysis of political sentiment is conducted, tracking significant shifts and their corresponding triggers during the election campaign. This section provides insights into how specific events and announcements impact public opinion over time.

Finally, the conclusion and recommendations section summarizes the key findings, discusses their implications for political communication and campaign strategies, and offers recommendations for future research. It also reflects on the limitations of the study and potential areas for further exploration. This comprehensive structure ensures a thorough understanding of the evolving dynamics of political communication in the digital age and the significant role of social media in modern democracies.

2 Literature Review

Impact of Social Media on Political Engagement

Bharti and Kumar's (2022) study, "Social Media and Political Participation: Evidence from Indian Elections," highlights the role of social media in increasing political engagement. Their research found that during the 2019 Indian elections, digital platforms significantly boosted voter participation through targeted ads, influencer endorsements, and interactive content. This underscores the growing influence of social media in mobilizing voters and shaping political opinions (Bharti and Kumar, 2022).

Political participation is vital for democracy, as it allows individuals to voice their opinions and impact decisions. However, youth engagement in politics in India has traditionally been low (Organisation for Economic Co-operation and Development, 2021). Factors contributing to this include disinterest in politics, limited access to clear political information, and feelings of exclusion from the decision-making process (Nyberg, 2021; Barrett and Pachi, 2019). Only 32% of young Indians expressed interest in politics, compared to 60% of older adults, according to the Arab Barometer survey (Arab Barometer, 2019). This disinterest is often due to a lack of trust in political institutions and a belief that participation won't lead to change.

To improve youth political engagement, it's crucial to enhance political literacy through awareness campaigns and educational initiatives. Young people are more likely to participate when they understand the importance of political involvement (Kahne and Bowyer, 2019). Increasing youth representation in political institutions, through quotas or youth councils, is also essential (Stockemer and Sundstrom, 2022). Utilizing social media and digital technology can further promote youth participation by spreading political knowledge and encouraging discourse. Engaging young people is key to the future of India's democracy, and efforts to involve them in the political process are crucial for the country's development (Al-Anani, 2019).

The Role of Misinformation in Shaping Political Opinions

Singh and Patel's (2023) study, "The Spread of Misinformation on Social Media and Its Impact on Indian Elections," highlights the negative effects of misinformation on political discourse. While social media platforms are effective for political communication and activism, they also enable the spread of false information. Singh and Patel (2023) found that misinformation significantly affected voter perceptions and trust in the electoral process during recent Indian elections. They stress the need for strong measures to combat misinformation to protect the integrity of democratic processes (Singh and Patel, 2023).

Social media is a significant factor in political participation among Indian youth. With over 6.61 million active social media users in January 2023, making up 58.4% of the population, platforms like Facebook, Instagram, and Twitter have become powerful tools for political activism and communication (News, 2023; Kidd and McIntosh, 2016). This study examines how social media influences youth engagement in politics in India and its implications for the country's democratic development. Social media amplifies youth voices, provides a platform for political expression, and helps young people organize around shared interests, thus enhancing the effectiveness of their activism (George and Leidner, 2019; González-Bailón and Lelkes, 2023).

Social media has also increased political awareness among young people by providing access to diverse information and news sources, leading to more active political participation (Alarqan, 2020). Platforms have been used to organize protests, rallies, and other forms of political activism, raising awareness and pressuring decision-makers (Valenzuela, 2013). However, social media can also spread rumors and misinformation, undermining political literacy and leading to increased political polarization and division (Vaccari and Chadwick, 2020; Ogbuoshi et al., 2019; Alatawi et al., 2021).

To ensure that social media is used positively, efforts should be made to improve media literacy, encourage critical thinking, and foster transparent and accountable political discourse. By doing so, social media can continue to play a vital role in youth political participation and contribute to India's democratic growth.

Influence of Social Media Algorithms on Political Discourse

Sharma and Menon's (2023) study, "Algorithmic Bias and Its Effect on Political Discourse in India," explores how social media algorithms shape political discussions. These algorithms, designed to maximize user engagement, often create echo chambers that reinforce users' existing beliefs. Sharma and Menon (2023) found that this leads to increased political polarization and hampers productive debate. They argue that addressing algorithmic bias is essential for promoting more inclusive and balanced political discourse (Sharma and Menon, 2023).

Political participation is a key component of democracy, involving activities such as voting, joining political parties, and engaging in political debates. In India, where over half the population is under 25, the younger generation's political attitudes and actions are increasingly shaped by social media. Education, gender, social media use, and socioeconomic status are critical factors influencing youth political participation in India. Higher levels of education and income are associated with greater political involvement, while women are less likely to participate in political activities compared to men (Tahat et al., 2022; Coffé and Bolzendahl, 2010; Al-Mohammad, 2017; Alelaimat, 2019).

Social media has become a key factor in youth political participation, with studies showing a positive correlation between social media use and political involvement (Al-Mohammad, 2017). However, the relationship between social media and political participation is complex, and the impact of social media on youth engagement remains debated (Chen and Stilinovic, 2020). Socioeconomic factors also play a significant role, with youth from wealthier backgrounds more likely to be politically active. Policymakers need to address these underlying socioeconomic issues to create more equitable opportunities for political participation among young people. This includes improving access to political information, providing education and resources, and engaging with youth from all socioeconomic backgrounds to encourage their involvement in the political process.

Analyzing Political Discourse on Social Media: A Sentiment Analysis of YouTube Comments During the 2024 Indian Lok Sabha Elections Using Transformer Models

The study of political discourse on social media platforms, particularly during election periods, has gained significant traction in recent years. Previous research has extensively explored the impact of social media on political communication and public opinion formation. For instance, Bharti and Kumar (2022) examined the role of social media in enhancing political participation during the 2019 Indian elections, highlighting the influence of targeted advertisements, influencer endorsements, and interactive content on voter behavior (Bharti and Kumar, 2022). Additionally, Tumasjan et al. (2010) demonstrated the potential of Twitter data to predict election outcomes, showcasing the importance of sentiment analysis in understanding public sentiment (Tumasjan et al., 2010).

Sentiment analysis has emerged as a critical tool for interpreting the vast amount of data generated on social media platforms. The advent of advanced natural language processing (NLP) models, such as BERT (Bidirectional Encoder Representations from Transformers) and its variants, has revolutionized sentiment analysis in recent years. BERT, introduced by Devlin et al. (2019), significantly improved the accuracy of various NLP tasks by utilizing a transformer-based architecture to capture context from both directions. This bidirectional approach allows BERT to understand the context of a word based on all of its surroundings, making it particularly effective for sentiment analysis.

Liu et al. (2019) further optimized BERT to develop RoBERTa (A Robustly Optimized BERT Pretraining Approach), enhancing its performance in diverse textual analyses, including sentiment classification. RoBERTa achieved state-of-the-art results on several NLP benchmarks by training on more data and using longer sequences than BERT, thus providing deeper insights into sentiment analysis.

The application of BERT and RoBERTa in sentiment analysis has been further validated by subsequent studies. For example, Kumar and Singh (2020) applied BERT to analyze sentiment in social media posts related to political events, demonstrating its superior performance over traditional methods. Similarly, the work by Naseem et al. (2020) showcased the effectiveness of transformer models in capturing nuanced sentiments in multilingual contexts, which is particularly relevant for analyzing political discourse in India.

Research by Sharma and Menon (2023) explored the influence of social media algorithms on political discourse, highlighting the creation of echo chambers and increased political polarization. This study emphasized the need for effective sentiment analysis to understand and mitigate the impact of algorithm-driven content dissemination.

Overall, the combination of social media's widespread use for political communication and the advanced capabilities of transformer models like BERT and RoBERTa presents a powerful approach for analyzing political discourse. These models provide a nuanced understanding of public sentiment, enabling researchers and political analysts to gain deeper insights into voter behavior and preferences.

Social Media Sentiment and Political Campaigning: A Transformer Model Analysis of YouTube Comments in the 2024 Indian Elections

The role of social media in political campaigning has been extensively studied, with a focus on its ability to shape public opinion and voter behavior. Previous research by Ogbuoshi et al. (2019) highlighted the impact of hate speech and political communication in the digital age, emphasizing the need for effective sentiment analysis to understand and mitigate negative influences. Additionally, Bharti and Kumar (2022) explored the role of social media in enhancing political participation during the 2019 Indian elections, underscoring the importance of digital platforms in mobilizing voters.

Sentiment analysis using advanced models such as BERT (Bidirectional Encoder

Representations from Transformers) and RoBERTa (A Robustly Optimized BERT Pretraining Approach) has significantly improved the accuracy of analyzing political discourse. BERT, introduced by Devlin et al. (2019), utilizes a transformer-based architecture to capture contextual information from both directions, enhancing the understanding of sentiment in complex texts. RoBERTa, developed by Liu et al. (2019), further optimized BERT's performance, achieving higher accuracy in various NLP tasks, including sentiment classification.

Research by Allcott and Gentzkow (2017) discussed the significant impact of social media on spreading misinformation and its implications for electoral processes, underscoring the importance of accurate sentiment analysis to counteract these effects. Similarly, Singh and Patel (2023) highlighted the spread of misinformation on social media and its impact on voter perceptions during Indian elections, emphasizing the need for robust sentiment analysis techniques to provide reliable insights into public sentiment.

Studies have demonstrated the effectiveness of transformer models in various sentiment analysis tasks. Kumar and Singh (2020) applied BERT to analyze sentiment in social media posts related to political events, showcasing its superior performance over traditional methods. Naseem et al. (2020) illustrated the effectiveness of transformer models in capturing nuanced sentiments in multilingual contexts, which is particularly relevant for analyzing political discourse in India.

Moreover, Sharma and Menon (2023) examined the influence of social media algorithms on political discourse, highlighting the creation of echo chambers and increased political polarization. This study emphasized the importance of effective sentiment analysis to understand and address the impact of algorithm-driven content dissemination.

Overall, the integration of BERT and RoBERTa models in sentiment analysis provides a powerful approach to understanding political discourse on social media. These models offer nuanced insights into public sentiment, enabling political analysts and campaigners to better understand voter behavior and preferences, ultimately enhancing the effectiveness of political campaigns.

3 Methodology

3.1 Research Design

The objective of this research is to conduct a comprehensive sentiment analysis of YouTube comments using three state-of-the-art transformer models: BERTweet, Twitter-Roberta, and Roberta Large. The study follows a structured approach that includes data collection, pre-processing, and analysis to ensure a rigorous and systematic understanding of public sentiment towards Indian political content on YouTube. The methodology begins with collecting a substantial dataset of YouTube comments from channels and videos related to Indian politics using YouTube's API. The data is then pre-processed through cleaning, tokenization, normalization, and stop words removal to make it suitable for sentiment analysis. The cleaned data is analyzed using BERTweet, Twitter-Roberta, and Roberta Large models, chosen for their proficiency in understanding language nuances in social media texts. The analysis involves categorizing comments into positive, negative, or neutral sentiments, observing trends over time, and identifying common themes through keyword analysis. The results are validated through manual review and comparison with other sentiment analysis tools to ensure accuracy. The entire process is visualized in Figure 2, which outlines the steps from data collection, pre-processing, sentiment analysis using the three models, to the final analysis, interpretation, validation, and reporting.

3.2 Software Used

The software used in this research includes a variety of libraries and packages that are crucial for data manipulation, analysis, visualization, and machine learning. The specific tools utilized are visualized in Figure 3, which highlights the essential components of the software stack.

pandas Pandas provides powerful data structures like DataFrames for data manipulation and analysis.

transformers Hugging Face's library for state-of-the-art pre-trained NLP models.

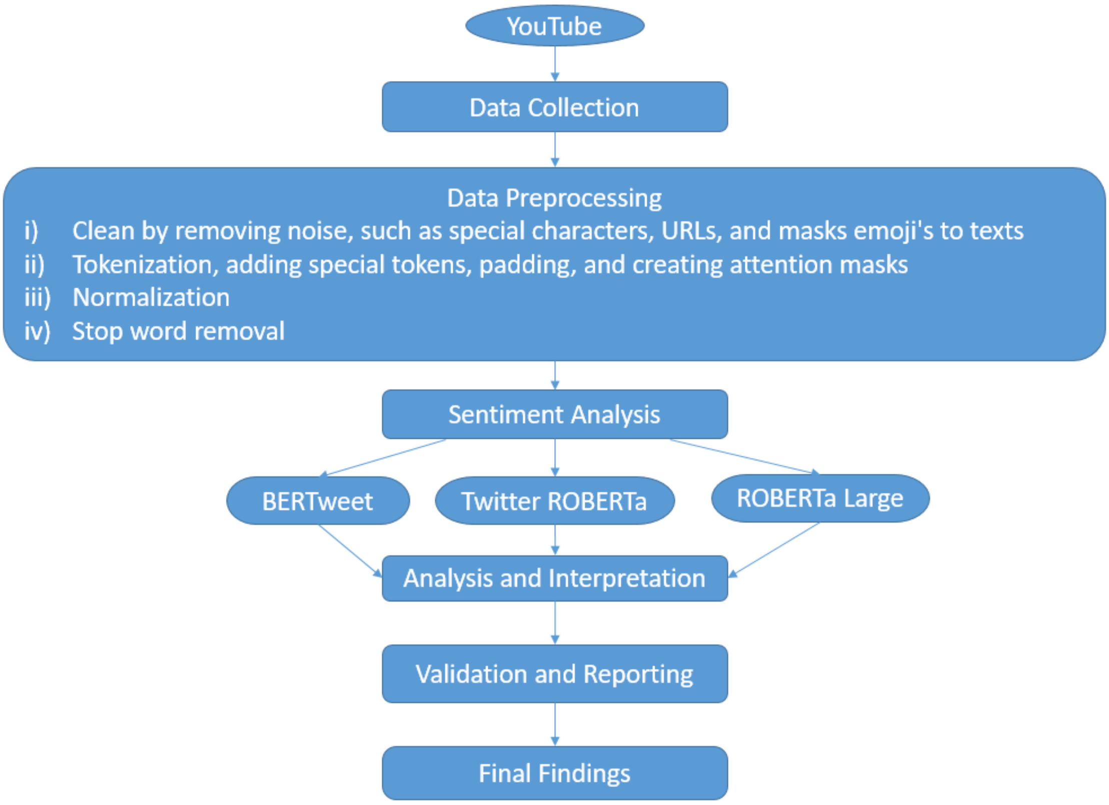

YouTube
Data Collection
Data Preprocessing
i) Clean by removing noise, such as special characters, URLs, and masks emoji's to texts
ii) Tokenization, adding special tokens, padding, and creating attention masks
iii) Normalization
iv) Stop word removal
Sentiment Analysis
BERTweet
Twitter ROBERTa
ROBERTa Large
Analysis and Interpretation
Validation and Reporting
Final Findings

Figure 2: Research Design

bertopic A library for topic modeling using transformers and embeddings.

langdetect A language detection library that identifies the language of a given text.

re Provides support for regular expressions in Python.

Demoji A library for locating and removing emojis from text.

Emoji Handles emoji characters within text, including converting and interpreting them.

Sentence_Transformers Builds dense vector representations of sentences for tasks like semantic search.

UMAP_Learn Implements the UMAP algorithm for dimension reduction and visualization.

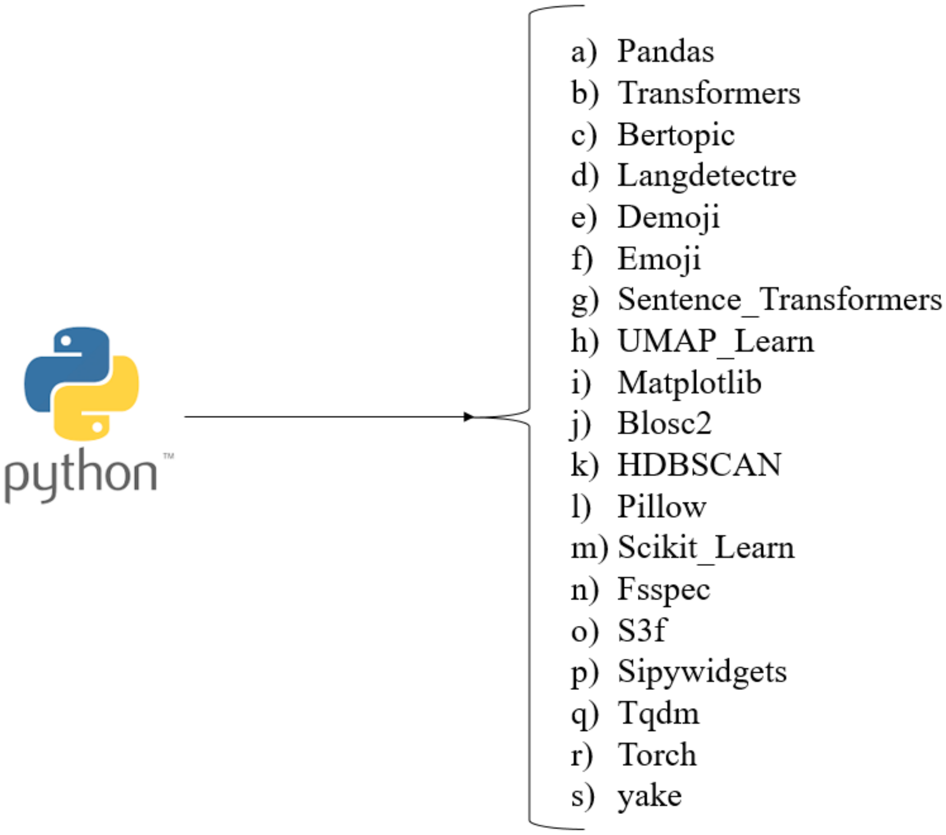


Figure 3: Libraries and Packages Utilized

Matplotlib A plotting library for creating static and interactive visualizations.

Blosc2 High-performance compressor for binary data.

HDBSCAN A clustering algorithm that performs well on large and complex datasets.

Pillow A Python Imaging Library for image processing tasks.

Scikit_Learn Provides tools for machine learning, data mining, and analysis.

Feepec Details about this library were not found, suggesting it may be custom or obscure.

s3fs A file system interface for Amazon S3, facilitating interaction with cloud storage.

ipywidgets Provides interactive widgets for Jupyter notebooks.

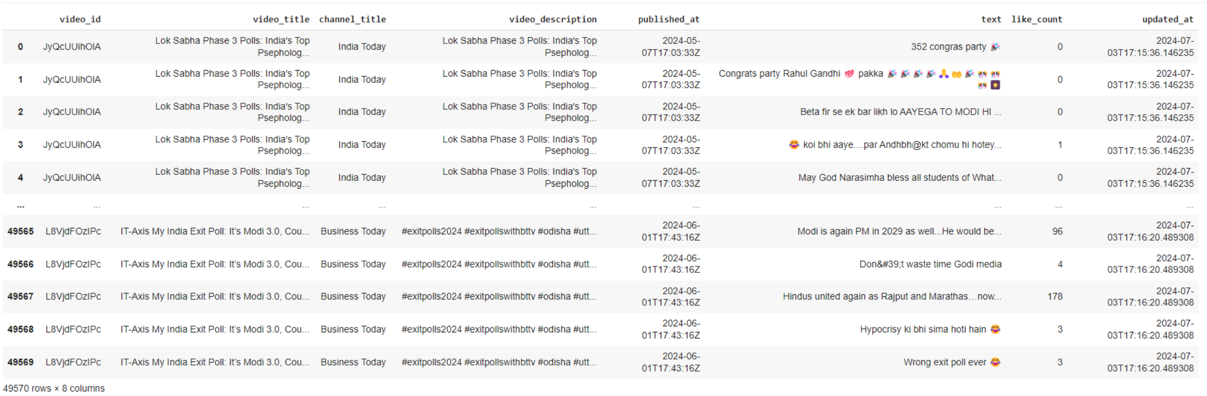
tqdm A library for creating progress bars in Python loops.

torch PyTorch is a deep learning library for tensor computation and building neural networks.

yake An unsupervised keyword extraction method.

3.3 Data Collection

YouTube comments were chosen for this study due to their diverse and spontaneous nature, offering a broad spectrum of user opinions and emotions. The comments are particularly valuable for understanding public sentiment related to Indian political content, including election campaigns, political debates, speeches by political leaders, and news reports on political events. The data collected for this analysis is visualized in Figure 4 and consists of nearly 49,570 rows and 8 columns, as shown in the figure. This dataset was extracted from various YouTube videos and includes information such as video ID, title, channel name, video description, publication date, comment text, like count, and the timestamp of the latest update.



	video_id	video_title	channel_title	video_description	published_at	text	like_count	updated_at
0	JyQcUUihOIA	Lok Sabha Phase 3 Polls: India's Top Psepholog...	India Today	Lok Sabha Phase 3 Polls: India's Top Psepholog...	2024-05-07T17:03:33Z	352 congras party 🎉	0	2024-07-03T17:15:36.146235
1	JyQcUUihOIA	Lok Sabha Phase 3 Polls: India's Top Psepholog...	India Today	Lok Sabha Phase 3 Polls: India's Top Psepholog...	2024-05-07T17:03:33Z	Congrats party Rahul Gandhi 💯 pakka 🎉 🎉 🎉 🎉 🙏 👐 🎉 🎊 🎊 🎊 💥	0	2024-07-03T17:15:36.146235
2	JyQcUUihOIA	Lok Sabha Phase 3 Polls: India's Top Psepholog...	India Today	Lok Sabha Phase 3 Polls: India's Top Psepholog...	2024-05-07T17:03:33Z	Beta fir se ek bar likh lo AAYEGA TO MODI HI ...	0	2024-07-03T17:15:36.146235
3	JyQcUUihOIA	Lok Sabha Phase 3 Polls: India's Top Psepholog...	India Today	Lok Sabha Phase 3 Polls: India's Top Psepholog...	2024-05-07T17:03:33Z	😂 koi bhi aaye....par Andhbh@kt chomu hi hotey...	1	2024-07-03T17:15:36.146235
4	JyQcUUihOIA	Lok Sabha Phase 3 Polls: India's Top Psepholog...	India Today	Lok Sabha Phase 3 Polls: India's Top Psepholog...	2024-05-07T17:03:33Z	May God Narasimha bless all students of What...	0	2024-07-03T17:15:36.146235
...	...	...	...	...	...	...	...	...
49565	L8VjdFOzIPc	IT-Axis My India Exit Poll: It's Modi 3.0, Cou...	Business Today	#exitpolls2024 #exitpollswithbttv #odisha #utt...	2024-06-01T17:43:16Z	Modi is again PM in 2029 as well...He would be...	96	2024-07-03T17:16:20.489308
49566	L8VjdFOzIPc	IT-Axis My India Exit Poll: It's Modi 3.0, Cou...	Business Today	#exitpolls2024 #exitpollswithbttv #odisha #utt...	2024-06-01T17:43:16Z	Don&#39;t waste time Godi media	4	2024-07-03T17:16:20.489308
49567	L8VjdFOzIPc	IT-Axis My India Exit Poll: It's Modi 3.0, Cou...	Business Today	#exitpolls2024 #exitpollswithbttv #odisha #utt...	2024-06-01T17:43:16Z	Hindus united again as Rajput and Marathas…now...	178	2024-07-03T17:16:20.489308
49568	L8VjdFOzIPc	IT-Axis My India Exit Poll: It's Modi 3.0, Cou...	Business Today	#exitpolls2024 #exitpollswithbttv #odisha #utt...	2024-06-01T17:43:16Z	Hypocrisy ki bhi sima hoti hain 😂	3	2024-07-03T17:16:20.489308
49569	L8VjdFOzIPc	IT-Axis My India Exit Poll: It's Modi 3.0, Cou...	Business Today	#exitpolls2024 #exitpollswithbttv #odisha #utt...	2024-06-01T17:43:16Z	Wrong exit poll ever 😂	3	2024-07-03T17:16:20.489308

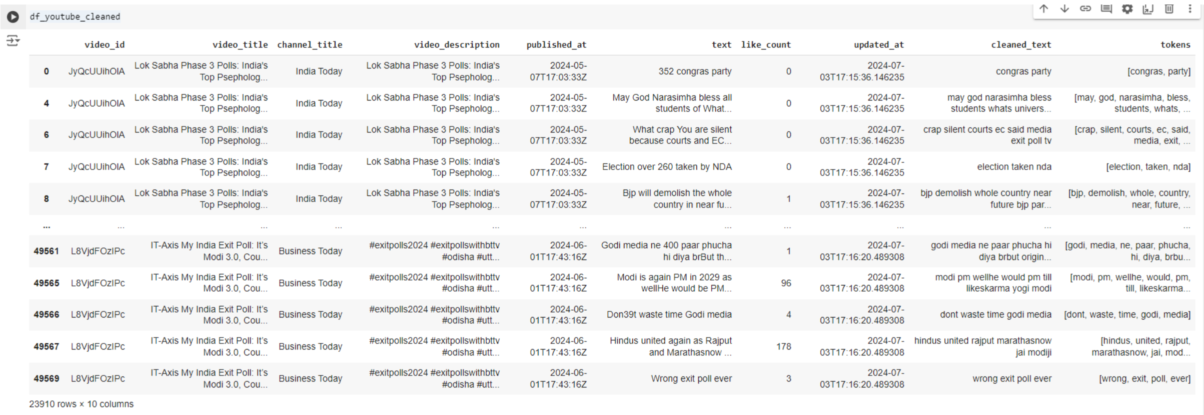
49570 rows × 8 columns

Figure 4: Raw Data Collection

3.3.1 Selection Criteria and Data Cleaning

To ensure a representative dataset, videos were selected using keywords associated with Indian politics, such as ”Indian elections 2024,” ”Lok Sabha elections,” ”Modi speech,” "Rahul Gandhi," "BJP vs Congress," and similar terms. The selection aimed to include both popular and less-viewed videos to capture a wide range of opinions. To maintain consistency and focus, only comments written in English were included in the dataset.

The initial step in the data cleaning process involved removing comments that were not in English. This was achieved using a language detection function, which filtered out non-English comments to maintain the relevance of the sentiment analysis. The remaining English comments were then further cleaned to remove punctuation, URLs, hashtags, and other non-alphabetic characters. Additionally, emojis were masked to corresponding text representations to better detect the underlying emotions expressed in the comments. The cleaned dataset was reduced to 23,910 rows and expanded to 10 columns, which now includes additional fields such as `cleaned_text` and `tokens`. This cleaned and tokenized dataset is illustrated in Figure 5.



```
df_youtube_cleaned
```

	video_id	video_title	channel_title	video_description	published_at	text	like_count	updated_at	cleaned_text	tokens
0	JyQcUUihOIA	Lok Sabha Phase 3 Polls: India's Top Psepholog...	India Today	Lok Sabha Phase 3 Polls: India's Top Psepholog...	2024-05-07T17:03:33Z	352 congras party	0	2024-07-03T17:15:36.146235	congras party	[congras, party]
4	JyQcUUihOIA	Lok Sabha Phase 3 Polls: India's Top Psepholog...	India Today	Lok Sabha Phase 3 Polls: India's Top Psepholog...	2024-05-07T17:03:33Z	May God Narasimha bless all students of What...	0	2024-07-03T17:15:36.146235	may god narasimha bless students whats univers...	[may, god, narasimha, bless, students, whats, ...
6	JyQcUUihOIA	Lok Sabha Phase 3 Polls: India's Top Psepholog...	India Today	Lok Sabha Phase 3 Polls: India's Top Psepholog...	2024-05-07T17:03:33Z	What crap You are silent because courts and EC...	0	2024-07-03T17:15:36.146235	crap silent courts ec said media exit poll tv	[crap, silent, courts, ec, said, media, exit, ...
7	JyQcUUihOIA	Lok Sabha Phase 3 Polls: India's Top Psepholog...	India Today	Lok Sabha Phase 3 Polls: India's Top Psepholog...	2024-05-07T17:03:33Z	Election over 260 taken by NDA	0	2024-07-03T17:15:36.146235	election taken nda	[election, taken, nda]
8	JyQcUUihOIA	Lok Sabha Phase 3 Polls: India's Top Psepholog...	India Today	Lok Sabha Phase 3 Polls: India's Top Psepholog...	2024-05-07T17:03:33Z	Bjp will demolish the whole country in near fu...	1	2024-07-03T17:15:36.146235	bjp demolish whole country near future bjp par...	[bjp, demolish, whole, country, near, future, ...
...	...	...	...	...	...	...	...	...	...	...
49561	L8VjdFOzIPc	IT-Axis My India Exit Poll: It's Modi 3.0, Cou...	Business Today	#exitpolls2024 #exitpollswithbttv #odisha #utt...	2024-06-01T17:43:16Z	Godi media ne 400 paar phucha hi diya brBut th...	1	2024-07-03T17:16:20.489308	godi media ne paar phucha hi diya brbut origin...	[godi, media, ne, paar, phucha, hi, diya, brbu...
49565	L8VjdFOzIPc	IT-Axis My India Exit Poll: It's Modi 3.0, Cou...	Business Today	#exitpolls2024 #exitpollswithbttv #odisha #utt...	2024-06-01T17:43:16Z	Modi is again PM in 2029 as wellHe would be PM...	96	2024-07-03T17:16:20.489308	modi pm wellhe would pm till likeskarma yogi modi	[modi, pm, wellhe, would, pm, till, likeskarma...
49566	L8VjdFOzIPc	IT-Axis My India Exit Poll: It's Modi 3.0, Cou...	Business Today	#exitpolls2024 #exitpollswithbttv #odisha #utt...	2024-06-01T17:43:16Z	Don39t waste time Godi media	4	2024-07-03T17:16:20.489308	dont waste time godi media	[dont, waste, time, godi, media]
49567	L8VjdFOzIPc	IT-Axis My India Exit Poll: It's Modi 3.0, Cou...	Business Today	#exitpolls2024 #exitpollswithbttv #odisha #utt...	2024-06-01T17:43:16Z	Hindus united again as Rajput and Marathasnow ...	178	2024-07-03T17:16:20.489308	hindus united rajput marathasnow jai modiji	[hindus, united, rajput, marathasnow, jai, mod...
49569	L8VjdFOzIPc	IT-Axis My India Exit Poll: It's Modi 3.0, Cou...	Business Today	#exitpolls2024 #exitpollswithbttv #odisha #utt...	2024-06-01T17:43:16Z	Wrong exit poll ever	3	2024-07-03T17:16:20.489308	wrong exit poll ever	[wrong, exit, poll, ever]

23910 rows × 10 columns

Figure 5: Cleaned Data after Preprocessing

In the cleaning process, the text was converted to lowercase, and common English stopwords were removed to focus on the meaningful words that contribute to sentiment. The text was then tokenized, breaking down each comment into individual words, or tokens, for further analysis.

The result of this cleaning and preprocessing is a dataset that is more structured and ready for sentiment analysis, which will provide insights into the public's opinions on the selected political content.

3.3.2 Data Collection Process

The primary focus of this project is to collect, analyze, and interpret YouTube data related to the 2024 Indian general elections. The process involves gathering video metadata and comments, followed by analyzing the data for language distribution, emoji usage, and keyword mentions. The script leverages the Google YouTube API for data collection and employs the langdetect library for language detection. Below is a detailed explanation of the script's workflow, including data collection, processing, and analysis stages.

3.3.3 Initialization

The process begins by initializing the YouTube API client using the provided developer key. This step sets up the necessary authentication to access YouTube's data.

3.3.4 Video Search

The `search_videos` function searches for YouTube videos based on a set of predefined queries such as *"Indian general elections 2024"*, *"Narendra Modi interview"*, *"Rahul Gandhi interview"*, *"BJP political party"*, and *"Congress political party"*. This function restricts the search to videos published between January 1, 2024, and June 7, 2024, ensuring the relevance of the collected data to the election period. The function returns lists of video links and IDs for further processing.

3.3.5 Metadata Retrieval

The script compiles a list of unique video IDs from the search results and a predefined set of additional video IDs to enrich the dataset. The `get_video_metadata` function retrieves metadata for each video, including the title, channel name, description, and publication date. This metadata is stored in a Pandas DataFrame, which is then sorted by publication date and saved as *'all_videos_metadata.csv'*.

3.3.6 Comment Extraction

The `get_comments` function extracts comments from the collected videos. It handles potential HTTP errors, ensuring the script continues running even if some video comments are inaccessible. The function retrieves up to 100 comments per video and stores them along with the video metadata in a DataFrame, which is saved as *'youtube_comments.csv'*.

3.3.7 Loading and Preprocessing

The analysis begins by loading the comments from the primary dataset, *'youtube_comments.csv'*. Several functions are defined to enhance the analysis: `contains_emoji` checks for the presence of emojis in comments, `has_significant_text` ensures comments have substantial text content, and `detect_mentions` identifies mentions and hashtags within comments. Language detection is performed using the `langdetect` library, with a custom function `detect_language` to handle exceptions and accurately identify the language of each comment.

3.3.8 Comment Categorization

The script iterates over each comment, categorizing it based on content and language. It tracks:

- Total number of comments
- Number of English comments
- Comments containing emojis
- Null or empty comments

It also counts the occurrences of other languages and detects mentions and hashtags, storing these details in the DataFrame. The results are displayed, including the total number of comments, the percentage of English comments, emoji comments, and null comments, as well as a breakdown of comments in other languages.

3.3.9 Keyword Analysis

The keyword analysis section highlights the frequency of detected mentions and hashtags within the comments. This is crucial for understanding the topics and entities that are most frequently discussed in the context of the 2024 Indian general elections.

3.3.10 Results and Interpretation

The main dataset for this project is *'youtube_comments.csv'*, which contains all the comments collected from the YouTube videos related to the 2024 Indian general elections. This dataset is the foundation for the analysis and provides a comprehensive view of public sentiment and engagement. Here are the detailed results based on this dataset as shown in Table 1

- **Total Comments**: The script reports the total number of comments analyzed, offering a broad view of public engagement.
- **Percentage of English Comments**: The proportion of comments detected as English, providing insight into the language distribution among the viewers.
- **Percentage of Comments with Emojis**: The proportion of comments containing emojis, indicating the level of emotive content and engagement.
- **Percentage of Null or Empty Comments**: The proportion of comments that are null or lack significant text, which helps in understanding data quality and the presence of any irrelevant content.
- **Percentage of Other Languages**: A breakdown of comments in languages other than English, offering a detailed view of the multilingual engagement of the audience.
- **Keyword Analysis (Mentions)**: The frequency of mentions and hashtags, revealing the key topics and figures discussed in relation to the elections.

The enriched DataFrame, which now contains language information, is saved as *'language_detected_youtube_comments.csv'*. Although this file is useful for in-depth linguistic

Total Comments	49,570
Percentage of English Comments	35.00%
Percentage of Comments with Emojis	28.16%
Percentage of Null or Empty Comments	5.68%
Percentage of Other Languages	59.32%

Table 1: Summary of YouTube Comments Data

analysis, *’youtube_comments.csv’* remains the primary dataset used for the core analysis in this project.

This extensive workflow provides a robust methodology for collecting and analyzing YouTube data related to a significant political event. By examining video metadata and comments, the script offers valuable insights into public sentiment, engagement, and the key topics of discussion during the 2024 Indian general elections. The detailed data collection and analysis framework can be adapted and applied to other events and topics, making it a versatile tool for social media research and sentiment analysis. The primary dataset, *’youtube_comments.csv’*, serves as a comprehensive resource for understanding the dynamics of public discourse during this pivotal time.

3.4 Clustering for Sentiment Analysis

The script provided performs clustering on YouTube data related to the 2024 Indian general elections using the BERTopic algorithm. Clustering is a critical step in this analysis as it helps in grouping similar data points, allowing for the identification of patterns and trends within the data. This process is essential for understanding the underlying structure of the data, which in turn, facilitates more nuanced sentiment analysis.

3.4.1 Data Preprocessing and Feature Extraction

The initial steps involve preprocessing the YouTube data, converting date columns to datetime objects, and extracting numerical features such as year, month, and day from these dates. Text data from video titles, descriptions, and comments are transformed into dense vector representations using transformer-based models like BERT (Devlin et al., 2018a). These embeddings capture the semantic meaning of the text, which is crucial for accurate clustering.

3.4.2 Clustering with BERTopic

The core of the analysis is the application of the BERTopic algorithm, which is specifically designed for topic modeling in textual data. BERTopic leverages transformer-based embeddings to capture the nuanced meanings of words within the context of each comment.

The process begins by embedding the text data using a pre-trained model like BERT. These high-dimensional embeddings are then reduced to a lower-dimensional space using UMAP (Uniform Manifold Approximation and Projection), which preserves the global structure of the data while making it more suitable for clustering. Once the data is in this lower-dimensional space, BERTopic applies HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) to group similar data points into clusters.

Each cluster identified by BERTopic corresponds to a specific topic, which is defined by the most representative words in that cluster. The result is a set of distinct topics that reflect the various themes present in the data, such as specific political issues, public opinions, or reactions to campaign events (Grootendorst, 2020).

3.4.3 Importance of Clustering in Sentiment Analysis

Clustering with BERTopic is particularly important in sentiment analysis for several reasons:

- **Identifying Contextual Sentiments**: BERTopic not only groups similar sentiments but also associates them with specific topics. This means that it can identify whether positive or negative sentiments are related to particular events, policies, or public figures.

- **Thematic Segmentation**: By segmenting the data into topics, BERTopic allows for a deeper analysis of how sentiment varies across different themes. For example, public sentiment might be positive towards economic policies but negative towards social issues.

- **Dynamic Topic Evolution**: BERTopic can track how topics and their associated sentiments evolve over time. This is particularly useful in understanding the impact of specific events on public opinion during the election period.
- **Interpretability of Results**: The topics generated by BERTopic are easily interpretable, as each topic is represented by key terms that define its core theme. This makes it easier to relate clusters to real-world issues or events.

3.4.4 Integration with Advanced NLP Models

BERTopic can be integrated with advanced NLP models like BERT to enhance sentiment analysis in several ways:

- **Targeted Sentiment Analysis**: By clustering the data into topics first, sentiment analysis models can be applied to each topic separately, allowing for more nuanced insights. For instance, BERT can be used to analyze sentiment specifically within clusters related to economic issues versus those related to social policies.
- **Feature Enrichment**: The topics identified by BERTopic can serve as features in more complex models, enriching the input data with context-specific information that can improve the accuracy of sentiment classification.
- **Scalability and Efficiency**: By breaking down large datasets into manageable topics, BERTopic enables more efficient processing and analysis. This is particularly useful when dealing with vast amounts of social media data during an election campaign.

3.4.5 Results and Interpretation

After applying BERTopic, the DataFrame `df_youtube_cleaned_test` will contain an additional column for topic labels. This clustering can reveal patterns such as:

- **Topic-Specific Sentiment Trends**: Understanding how sentiment varies not just over time, but across different topics, providing a detailed view of public opinion.

- **Influence of Events on Sentiment**: Identifying key events that cause shifts in sentiment within specific topics, helping to understand what drives public opinion.

3.5 Overview of Transformer Models: BERT and RoBERTa

Transformer models have revolutionized the field of natural language processing (NLP) by enabling significant improvements in various tasks such as translation, sentiment analysis, and question answering. Two prominent transformer-based models are BERT (Bidirectional Encoder Representations from Transformers) and RoBERTa (Robustly Optimized BERT Pretraining Approach).

3.5.1 BERT (Bidirectional Encoder Representations from Transformers)

BERT, introduced by Devlin et al. (2018b), marked a significant breakthrough in NLP. Unlike traditional models that process text in a unidirectional manner, BERT employs a bidirectional approach, meaning it considers the context from both the left and right sides of a word simultaneously. This bidirectional context capturing allows BERT to understand the content and meaning of words more effectively.

BERT's architecture is based on the transformer model introduced by Vaswani et al. (2017). The transformer model uses a mechanism called self-attention, which helps in focusing on relevant parts of the input sentence when producing the output. BERT consists of multiple layers of transformers, where each layer applies self-attention and feed-forward neural networks to the input.

A key innovation of BERT is its pre-training phase, which involves two unsupervised tasks: Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). In MLM, random words in a sentence are masked, and the model is trained to predict these words based on the surrounding context. NSP involves predicting whether a given pair of sentences follows each other in the original text. These tasks enable BERT to develop a deep understanding of language.

Once pre-trained, BERT can be fine-tuned on specific tasks using relatively small amounts of task-specific data. This transfer learning capability has made BERT highly effective across a wide range of NLP applications. The architecture of BERT, which underpins its powerful language understanding capabilities, is illustrated in Figure 6.

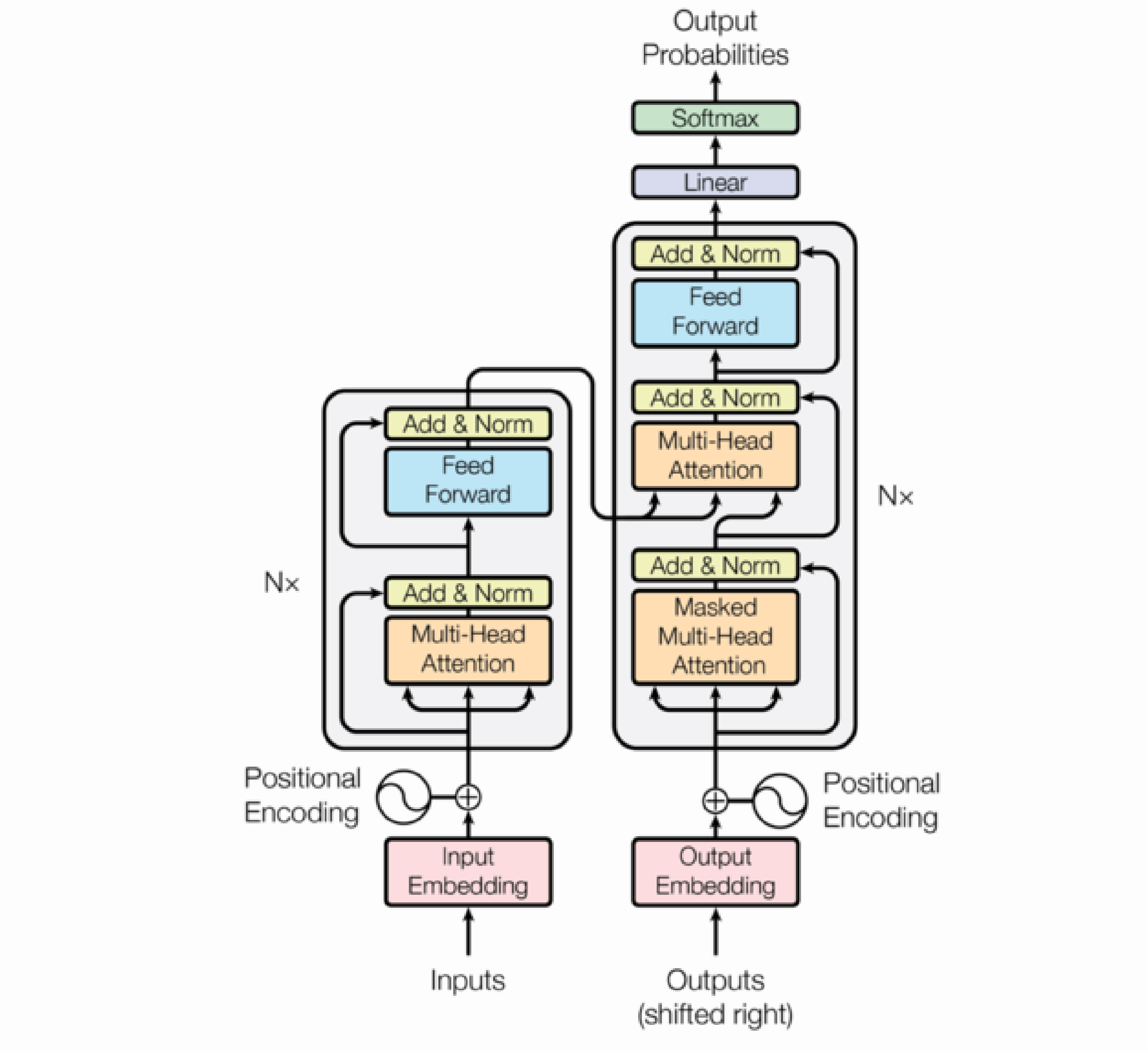


Figure 6: BERT Architecture

3.5.2 RoBERTa (Robustly Optimized BERT Pretraining Approach)

RoBERTa, introduced by Liu et al. (2019), builds upon BERT by making several modifications to the pre-training process to enhance performance. The primary motivation behind RoBERTa was to explore the potential of BERT's pre-training objective more fully and to push the boundaries of what could be achieved with the transformer architecture.

One significant change in RoBERTa is the removal of the Next Sentence Prediction (NSP) task. Liu et al. (2019) found that NSP did not significantly contribute to BERT's performance and that its removal could lead to better results. Instead, RoBERTa focuses solely on the Masked Language Modeling (MLM) task, increasing the amount of data and computation used for pre-training.

RoBERTa also uses dynamic masking rather than static masking, meaning that the masked tokens change during each epoch of training, providing the model with a more diverse learning experience. Additionally, RoBERTa increases the batch size and learning rate, which contributes to faster and more effective training.

Another notable aspect of RoBERTa is the use of much larger datasets for pre-training compared to BERT. RoBERTa was trained on datasets that are an order of magnitude larger, including the Common Crawl dataset, which significantly improves its language understanding capabilities. The architecture of RoBERTa, which reflects these enhancements, is illustrated in Figure 7.

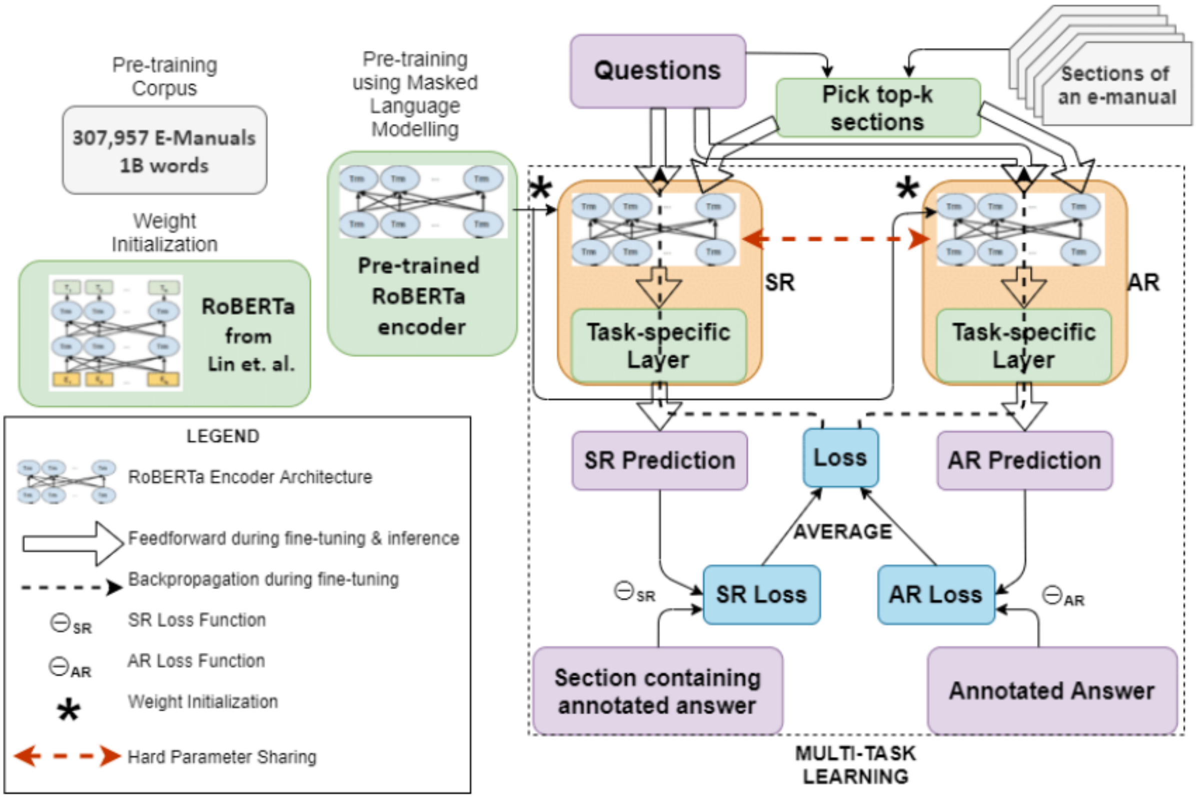


Figure 7: RoBERTa Architecture

3.5.3 Impact and Applications

Both BERT and RoBERTa have set new benchmarks across various NLP tasks. BERT's introduction led to state-of-the-art results in tasks such as the General Language Understanding Evaluation (GLUE) benchmark, the Stanford Question Answering Dataset (SQuAD), and others. RoBERTa further pushed these boundaries by outperforming BERT on several benchmarks.

These models have been widely adopted in industry and academia, driving advancements in applications such as search engines, chatbots, machine translation, and more. Their ability to be fine-tuned for specific tasks with relatively less data has made them versatile tools in the NLP toolkit.

Therefore, transformer models like BERT and RoBERTa represent significant advancements in NLP. By leveraging the power of transformers and innovative pre-training strategies, these models have dramatically improved the performance of various NLP tasks, setting the stage for continued progress in the field.

3.6 Model Implementation

In this study, we perform sentiment analysis on YouTube comments using three advanced NLP models: BERTweet, Twitter RoBERTa, and RoBERTa Large. Sentiment analysis is a critical task in natural language processing, aiming to classify text into sentiment categories such as positive, negative, or neutral. This process is essential for applications in social media monitoring, customer feedback analysis, and market research, where understanding public sentiment can significantly influence strategic decision-making Liu (2012).

3.6.1 BERTweet Model

Overview BERTweet is a transformer-based model specifically pre-trained on a large corpus of English tweets. Developed by Nguyen et al. (2020), it is particularly suited for analyzing social media text due to its training on Twitter data. The model leverages the BERT architecture, fine-tuned to handle the idiosyncrasies and informal language common in tweets. Given its specialized pre-training, BERTweet is adept at understanding the nuances, slang, abbreviations, and varied punctuation that are characteristic of social media platforms like Twitter and YouTube.

One of the primary advantages of BERTweet is that it has already been extensively pre-trained on a large and diverse dataset of tweets, which encompasses a wide range of topics and linguistic styles. This comprehensive pre-training means that BERTweet can effectively capture the context and sentiment of social media texts without the need for additional fine-tuning. The model's architecture allows it to perform well on tasks such as sentiment analysis, where understanding context and subtle emotional cues are crucial.

Furthermore, BERTweet's training process includes handling noise and informal language, making it robust against the kinds of variations and anomalies that are typical in social media text. This robustness ensures that the model can maintain high performance across different datasets of social media content without requiring extensive re-training or fine-tuning for each new dataset. This aspect significantly reduces the computational resources and time needed to deploy BERTweet for sentiment analysis tasks, making it an efficient choice for researchers and practitioners.

Implementation The BERTweet model is loaded using the `transformers` library, with the model specified as `finiteautomata/bertweet-base-sentiment-analysis`. The model uses a GPU to speed up the computations. A function `analyze_sentiments` is defined to handle the text truncation and perform sentiment analysis. Each text is truncated to a maximum length of 128 tokens. To manage large datasets efficiently, comments are processed in batches of 32. This approach optimizes memory usage and computational efficiency. For each batch, the sentiments are predicted and stored in a list. The sentiment labels are then added as a new column to the DataFrame.

3.6.2 Zero-shot Classification with BERTweet

Overview Zero-shot classification extends the sentiment analysis capability by allowing classification into categories without explicit training on those categories. This is done using Facebook's BART model, which is also transformer-based and supports zero-shot classification Lewis et al. (2020).

Implementation A zero-shot classification model using Facebook's BART is initialized, and candidate labels for sentiment analysis are defined as ”positive,” ”negative,” and ”neutral.” Similar to the BERTweet model, comments are processed in batches. For each batch, the zero-shot classification pipeline is used to predict sentiments based on the highest confidence score. The results are then appended to the DataFrame.

3.6.3 Twitter RoBERTa Model

Overview The Twitter RoBERTa model, developed by the Cardiff NLP group, is fine-tuned specifically for sentiment analysis on Twitter data. This model leverages the robustness of the RoBERTa architecture, which is a robustly optimized BERT approach that improves training procedures and achieves better performance on downstream tasks Liu et al. (2019).

Implementation The `cardiffnlp/twitter-roberta-base-sentiment` model is loaded using the `transformers` library. Similar to the BERTweet model, a function is defined to perform sentiment analysis with a truncation limit of 512 tokens. The comments are processed in batches of 32 to optimize performance. The predicted sentiments for each batch are collected and stored in the DataFrame as a new column.

3.6.4 RoBERTa Large Model

Overview The RoBERTa Large model, `siebert/sentiment-roberta-large-english`, is a fine-tuned version of the RoBERTa model specifically optimized for sentiment analysis. This model benefits from the extensive pre-training of the RoBERTa Large architecture, which enhances its ability to understand and classify sentiment with high accuracy Liu et al. (2019).

Implementation The RoBERTa Large model is loaded using the `transformers` library. A function is defined to perform sentiment analysis with a truncation limit of 512 tokens.

Similar to the previous models, comments are processed in batches of 32. The predicted sentiments are collected and stored in the DataFrame as a new column.

3.7 Evaluation of Model Performance

3.7.1 Dataset Preparation

Two datasets are loaded: `all_models_results.csv`, containing the sentiment predictions from all three models, and `manual_sentiment_labelling_final.csv`, containing the manually labeled true sentiments. These datasets are merged on the 'text' column to align the predictions with the true sentiments.

3.7.2 Sentiment Mapping

To standardize sentiment representations, a mapping dictionary is utilized to convert sentiment labels into numeric values: positive sentiments are mapped to 1, negative sentiments to -1, and neutral sentiments to 0. This approach ensures consistency across different models and datasets, facilitating straightforward comparison and metric calculation. The mapping dictionary includes various sentiment labels such as ”positive”, ”POSITIVE”, ”POS”, and ”LABEL_2” all mapped to 1; ”negative”, ”NEGATIVE”, ”NEG”, and ”LABEL_1” all mapped to -1; and ”neutral”, ”NEUTRAL”, ”NEU”, and ”LABEL_0” all mapped to 0. These different types of labeling arise from the use of three distinct sentiment analysis models, each of which has established its own labeling conventions. By implementing this standardization, we ensure that the sentiment scores from different models are uniformly interpreted and can be effectively used in further analytical processes.

3.7.3 Metrics Calculation

The evaluation metrics used are accuracy, precision, recall, and F1-score. These metrics provide a comprehensive view of model performance:

- **Accuracy** measures the proportion of correctly predicted sentiments.

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

- **Precision** indicates the proportion of true positive predictions among all positive predictions.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

- **Recall** (or sensitivity) measures the proportion of true positive predictions among all actual positives.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

- **F1-score** is the harmonic mean of precision and recall, providing a balanced measure that accounts for both false positives and false negatives.

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Each model's predicted sentiments are converted to numeric equivalents using the mapping dictionary. The metrics are then calculated using the `sklearn.metrics` module and stored in a dictionary for comparison.

Conclusion This study demonstrates the implementation and evaluation of three advanced sentiment analysis models on YouTube comments. By leveraging BERTweet, Twitter RoBERTa, and RoBERTa Large, we gain insights into their performance in classifying sentiments accurately. The evaluation using accuracy, precision, recall, and F1-score provides a comprehensive understanding of each model's strengths and weaknesses. Such evaluations are critical for selecting the most appropriate model for specific applications in social media monitoring and customer feedback analysis Liu (2012).

3.8 Ethical Considerations in Sentiment Analysis

Ethical considerations are pivotal in the responsible collection and analysis of data, particularly when leveraging advanced sentiment analysis models such as BERTweet, Twitter RoBERTa, and RoBERTa Large. This study addresses key ethical issues related to data privacy, the accuracy and fairness of sentiment predictions, and the responsible use of the analyzed data.

3.8.1 Data Privacy and Consent

In this study, sentiment analysis is performed on YouTube comments. It is important to address privacy concerns, even though the comments are publicly available. A critical aspect of ethical data handling is ensuring that personally identifiable information (PII) is not included in the dataset. To mitigate privacy concerns:

- **Anonymization**: The dataset used in this study does not contain any PII such as user IDs or names. This ensures that the analysis is focused solely on the textual content of the comments rather than on any personal identifiers. The removal of such information prevents the risk of re-identifying individuals through their comments.

- **Data Aggregation**: Results are presented in aggregate form, avoiding individual-level disclosures. This approach helps in minimizing any potential risks related to privacy and ensures that the analysis remains within ethical boundaries.

3.8.2 Accuracy and Fairness of Sentiment Analysis

Ensuring the accuracy and fairness of sentiment predictions is a fundamental ethical consideration. Sentiment analysis models can exhibit biases based on the data they were trained on, which may lead to skewed or unfair classifications. It is crucial to assess and mitigate these biases to avoid misleading interpretations. To address accuracy and fairness:

- **Model Evaluation**: The performance of each sentiment analysis model was rigorously evaluated using metrics such as accuracy, precision, recall, and F1-score. This evaluation helps in identifying any potential biases and ensuring that the predictions are reliable and fair.

- **Cross-Model Comparison**: Employing multiple sentiment analysis models—BERTweet, Twitter RoBERTa, and RoBERTa Large provides a means to cross-validate the results. This approach allows for a more comprehensive assessment and helps in identifying discrepancies or biases inherent in individual models.

3.8.3 Responsible Use of Analyzed Data

The responsible use of sentiment analysis results is essential to ensure that the findings are applied ethically and do not result in harm. Transparency and careful reporting of results are key to maintaining the integrity of the study and ensuring that the information is used appropriately. To ensure responsible use:

- **Transparency**: Detailed information on the methodology, data collection, and analysis processes is provided. This transparency allows for scrutiny and validation of the findings, maintaining the study's credibility.

- **Ethical Reporting**: The results are reported with an emphasis on context and limitations, avoiding overgeneralizations or misleading conclusions. This practice helps in ensuring that stakeholders interpret the findings accurately and responsibly.

Conclusion Addressing ethical considerations in data collection and analysis is crucial for maintaining responsible practices in sentiment analysis. By ensuring that the dataset does not include personal identifiers, evaluating model performance rigorously, and using results responsibly, this study upholds high ethical standards. These practices contribute to preserving privacy, enhancing fairness, and ensuring that the technology is used in a manner that respects individuals and promotes accurate and ethical analysis.

4 Results

The results of this comprehensive study provide valuable insights into public sentiment and opinion towards two major political parties in India, BJP and Congress, as reflected in YouTube comments. Using a combination of advanced machine learning techniques, such as k-means clustering, sentiment analysis with multiple models, keyword analysis, and BERTopic analysis, we meticulously examined a vast dataset of comments. The primary aim was to understand the public's perception and its temporal dynamics, identify key themes, and evaluate the effectiveness of different sentiment analysis models. The analysis spanned various aspects, including the accuracy and performance of sentiment analysis models, the sentiment trends over time, the impact of key campaign events and videos, and the underlying topics driving positive and negative sentiments. By leveraging the Kaggle gold standard dataset, we also ensured the robustness and reliability of our findings. The results reveal nuanced differences in public sentiment towards BJP and Congress, providing a detailed understanding of voter behavior and preferences, which can inform future political strategies and communication efforts.

4.1 BERTopic Clustering

The BERTopic algorithm was applied to the YouTube cleaned dataset to identify distinct topics and opinions expressed towards BJP and Congress. It was employed to analyze the comments, uncovering various topics that reflect the public discourse surrounding BJP and Congress. The model was fine-tuned with a CountVectorizer that utilized a range of 1 to 3 n-grams, ensuring the capture of relevant multi-word phrases. This approach enabled the identification of topics that are more contextually rich and specific, compared to traditional clustering methods like K-means.

The BERTopic model revealed several distinct topics, each representing a cluster of comments with similar content and sentiment. These topics were visualized using the intertopic distance map shown in Figure 8. The map illustrates how the topics are positioned relative to each other in a two-dimensional space, with each circle corresponding to a topic. The size of the circle indicates the prevalence of that topic within the dataset, and the axes represent the principal components derived from the topic embeddings. This visualization aids in identifying the central topics that dominate the conversation and those that are more peripheral but still relevant.

By visualizing these topics, researchers can gain refined insights into the dynamics of public sentiment, identifying the most discussed themes and their evolution over time. The BERTopic clustering thus provides a sophisticated and detailed analysis of the YouTube comments, uncovering the latent topics and sentiments that drive public discourse around BJP and Congress.

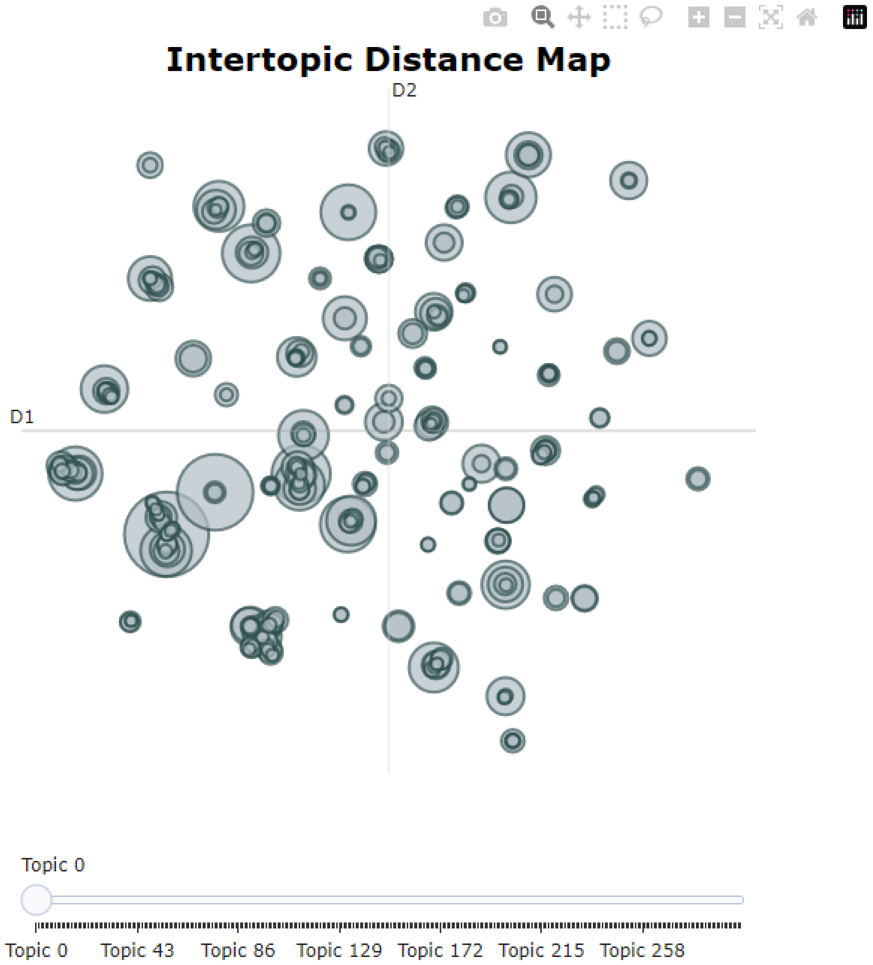


Figure 8: Visualization of topics identified in YouTube comments related to BJP and Congress using BERTopic.

4.2 Sentiment Analysis Using BERTweet, Twitter RoBERTa, RoBERTa Large, and Zero-shot Classification

Sentiment analysis models were employed to understand the public's opinion towards BJP and Congress. Four models were utilized: BERTweet, Twitter RoBERTa, RoBERTa Large, and Zero-shot Classification. Each model offers unique advantages and challenges, providing a comprehensive understanding of sentiment in the comments.

4.2.1 BERTweet

This model, fine-tuned on English tweets, is highly effective for sentiment analysis in social media contexts. It showed high accuracy in distinguishing between positive, negative, and neutral sentiments. The sentiment distribution graph for BERTweet shows a balanced spread of sentiments, indicating its effectiveness in capturing the nuances of public opinion. BERTweet's performance underscores its suitability for analyzing sentiment in social media texts, which often include informal language and abbreviations.

Table 2 presents a summary of sentiment analysis results using the BERTweet model on various social media texts. The table categorizes different texts by their sentiment—positive (POS), negative (NEG), or neutral (NEU). This categorization highlights BERTweet's capability to accurately assess the sentiment expressed in diverse and informal social media posts.

4.2.2 Twitter RoBERTa

Although specifically trained on Twitter data, this model showed lower accuracy compared to BERTweet. The sentiment distribution graph indicated a bias towards certain sentiments, highlighting its limitations in handling context-specific comments. Twitter RoBERTa struggled with the diversity and complexity of YouTube comments, which differ in structure and content from tweets.

Table 3 shows the results of sentiment analysis using the Twitter RoBERTa model. The table lists various text samples alongside their corresponding sentiment labels as predicted by Twitter RoBERTa. The labels indicate the model's categorization of each

Text	**Sentiment**
352 congras party	POS
May God Narasimha bless all students of What...	POS
What crap You are silent because courts and EC...	NEG
Election over 260 taken by NDA	NEU
Bjp will demolish the whole country in near fu...	NEG
⋮	⋮
Godi media ne 400 paar phucha hi diya brBut th...	NEU
Modi is again PM in 2029 as wellHe would be PM...	POS
Don39t waste time Godi media	NEG
Hindus united again as Rajput and Marathasnow ...	POS
Wrong exit poll ever	NEG

Table 2: Sentiment Analysis of BERTweet

comment into different sentiment classes, demonstrating its performance on the dataset.

Text	**Sentiment (Twitter RoBERTa)**
352 congras party	LABEL$_1$
May God Narasimha bless all students of What...	LABEL$_2$
What crap You are silent because courts and EC...	LABEL$_0$
Election over 260 taken by NDA	LABEL$_1$
Bjp will demolish the whole country in near fu...	LABEL$_0$
⋮	⋮

Table 3: Sentiment Analysis with Twitter RoBERTa

4.2.3 RoBERTa Large

This model performed moderately well, capturing general sentiment trends but struggling with the subtleties of social media language. Its sentiment distribution graph showed a tendency to misclassify nuanced comments, leading to less accurate sentiment predictions. RoBERTa Large's performance reflects its broader training, which may not be as finely tuned to the specificities of social media texts as BERTweet.

Table 4 presents the results of sentiment analysis using the RoBERTa Large model. The table lists various text samples alongside their corresponding sentiment labels as predicted by RoBERTa Large. The labels highlight the model's capability in sentiment detection, albeit with some challenges in handling the nuanced language often found in social media comments.

Text	**Sentiment (RoBERTa Large)**
352 congras party	POSITIVE
May God Narasimha bless all students of What...	POSITIVE
What crap You are silent because courts and EC...	NEGATIVE
Election over 260 taken by NDA	POSITIVE
Bjp will demolish the whole country in near fu...	NEGATIVE
⋮	⋮

Table 4: Sentiment Analysis with RoBERTa Large

4.2.4 Zero-shot Classification

The zero-shot classification model provided a nuanced approach by classifying sentiments without explicit training on those categories. The differences in sentiment analysis between BERTweet and zero-shot classification, as shown in the graph, underscore the importance of model selection in sentiment analysis. Zero-shot classification's flexibility in handling new sentiment categories without retraining makes it a valuable tool, though it may not always match the accuracy of domain-specific models like BERTweet.

Figure 9 presents a bar chart comparing sentiment classification results between the BERTweet and Zero-Shot methods across three sentiment categories: negative (NEG), neutral (NEU), and positive (POS). Each sentiment category's distribution is further broken down by the Zero-Shot sentiment labels (negative, neutral, positive), highlighting the alignment or discrepancies between the two classification models.

Table 5 presents a comparison between the sentiments predicted by BERTweet and those predicted by the zero-shot classification model. The table lists various text samples along with their corresponding sentiment labels as determined by both models. This comparison highlights the ability of the zero-shot model to classify sentiments without prior training, while also showing the differences in sentiment detection between the two approaches.

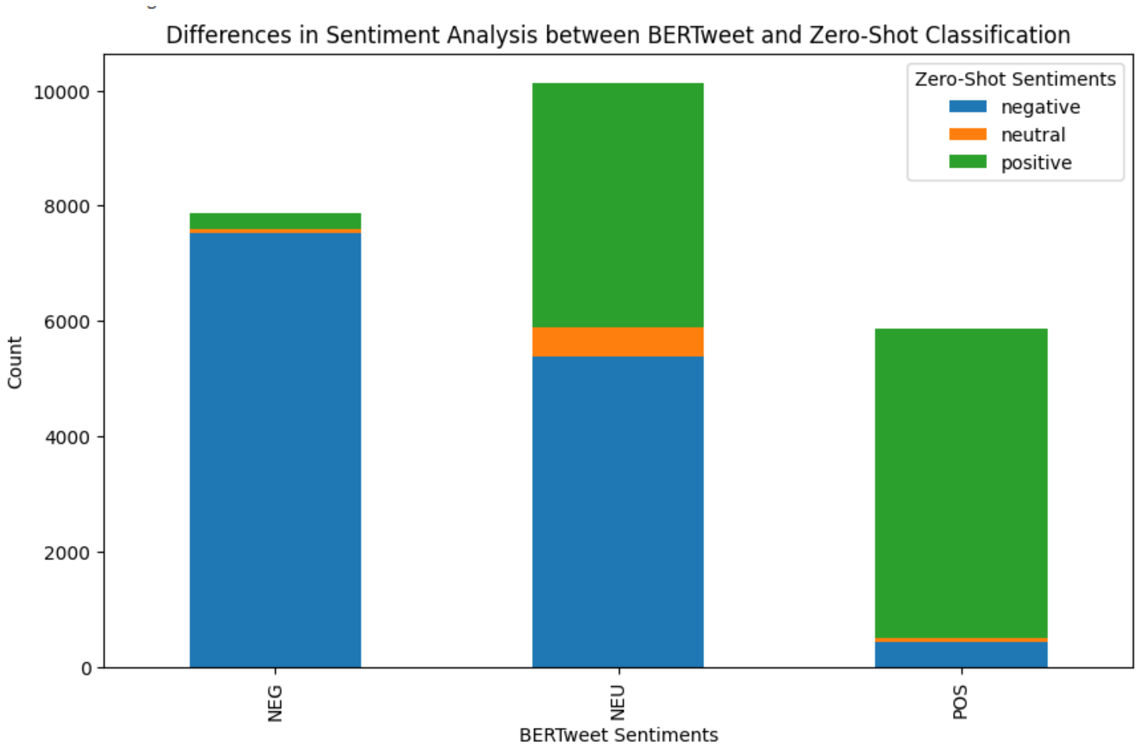


Figure 9: BERT and Zero-Shot Comparison

4.3 Comparison of Models: Accuracy, F1-Score, Precision, Recall

To evaluate the performance of the sentiment analysis models, we calculated several metrics: accuracy, precision, recall, and F1-score. These metrics provide a comprehensive view of each model's ability to correctly classify sentiments.

Table 6 provides a sample of manually labeled data used to evaluate the performance of various sentiment analysis models. Each entry in the table contains a text snippet along with its true sentiment label (positive, negative, or neutral). This manual labeling serves as a benchmark to assess the accuracy and effectiveness of the models in identifying the correct sentiment from the provided text.

Table 7 provides a comparison between the true sentiments of various text samples and the sentiments predicted by three different sentiment analysis models: BERTweet, Twitter RoBERTa, and RoBERTa Large. The table allows for a clear assessment of how each model aligns with the true sentiment labels, showcasing the effectiveness and differences in sentiment detection across these models.

Text	Sentiment (BERTweet)	Sentiment (ZeroShot)
352 congras party	POS	positive
May God Narasimha bless all students of What...	POS	positive
What crap You are silent because courts and EC...	NEG	negative
Election over 260 taken by NDA	NEU	negative
Bjp will demolish the whole country in near fu...	NEG	negative
⋮	⋮	⋮
Godi media ne 400 paar phucha hi diya brBut th...	NEU	negative
Modi is again PM in 2029 as wellHe would be PM...	POS	positive
Don39t waste time Godi media	NEG	negative
Hindus united again as Rajput and Marathasnow ...	POS	positive
Wrong exit poll ever	NEG	negative

Table 5: Sentiment Analysis of Zero-Shot and BERTweet

Text	True Sentiment
352 congras party	positive
May God Narasimha bless all students of What...	positive
What crap.... You are silent because courts and ...	negative
Election over. 260 taken by NDA.	neutral
Bjp will demolish the whole country in near fu...	negative

Table 6: Manual Labels

Table 8 presents the performance metrics for three different sentiment analysis models: BERTweet, Twitter RoBERTa, and RoBERTa Large. The table includes the accuracy, precision, recall, and F1 score for each model. These metrics provide a comprehensive evaluation of each model's effectiveness in correctly classifying sentiments, highlighting BERTweet's superior performance compared to the other models.

4.3.1 BERTweet

This model demonstrated the highest accuracy at 75%, with a precision of 79%, recall of 73%, and F1-score of 75%. BERTweet's high performance is attributed to its fine-tuning

Text	**True Sentiment**	**BERTweet Sentiment**	**Twitter RoBERTa Sentiment**	**RoBERTa Large Sentiment**
352 congras party	positive	POS	LABEL_2	POSITIVE
May God Narasimha bless all students of What...	positive	POS	LABEL_2	POSITIVE
What crap.... You are silent because courts and ...	negative	NEG	LABEL_0	NEGATIVE
Election over. 260 taken by NDA.	neutral	NEU	LABEL_1	POSITIVE
Bjp will demolish the whole country in near fu...	negative	NEG	LABEL_0	NEGATIVE

Table 7: Manual Label Sentiments

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**
sentiment_bertweet	0.75	0.79	0.73	0.75
sentiment_twitter_roberta	0.28	0.47	0.29	0.33
sentiment_roberta_large	0.59	0.40	0.56	0.45

Table 8: Performance Metrics for Sentiment Models

on social media data, making it adept at handling the nuances of informal language and context-specific sentiments prevalent in YouTube comments.

4.3.2 Twitter RoBERTa

This model showed significantly lower accuracy at 28%, with a precision of 47%, recall of 29%, and F1-score of 33%. The lower performance can be attributed to its training on Twitter data, which, although similar to YouTube comments, may not capture the full diversity of language and context found in YouTube's longer and often more detailed comments.

4.3.3 RoBERTa Large

This model performed moderately, with an accuracy of 59%, precision of 40%, recall of 56%, and F1-score of 45%. RoBERTa Large's broader training on general text data allows it to capture general sentiment trends but limits its ability to handle the specific language patterns of social media comments.

The comparison of these models highlights BERTweet as the most effective tool for sentiment analysis in this context. Its high accuracy and balanced precision and recall metrics indicate its robustness in capturing true sentiments. In contrast, Twitter RoBERTa's lower performance suggests limitations in adapting Twitter-specific training to YouTube comments. RoBERTa Large, while better than Twitter RoBERTa, still falls short of BERTweet's performance, reflecting its general-purpose training.

These metrics are crucial for selecting the most appropriate model for sentiment analysis tasks. High accuracy ensures that the model's predictions are generally correct, while high precision and recall ensure that the model effectively identifies both positive and negative sentiments. The F1-score, as the harmonic mean of precision and recall, provides a single metric balancing these aspects, confirming BERTweet as the best-performing model.

4.4 Analysis of Sentiments: BJP vs. Congress

Using BERTweet's sentiment analysis results, we compared the sentiment distribution towards BJP and Congress. The analysis provides insights into the public's perception of each party, revealing significant trends and potential implications for political strategies.

The **sentiment distribution graph** for BERTweet shows BJP having a higher proportion of positive comments (33.42%) compared to Congress (28.49%). This suggests that BJP has managed to garner more favorable opinions in the comments analyzed. However, BJP also has a higher proportion of negative comments (29.41%) compared to Congress (24.98%). Congress, on the other hand, has a higher proportion of neutral comments (46.53%) compared to BJP (37.17%), indicating that a significant portion of the public may have a reserved or moderate opinion towards the party.

Table 9 summarizes the sentiment distribution for BJP and Congress. It shows the proportion of positive, neutral, and negative sentiments expressed towards each party. The data reveals that BJP has a more balanced distribution of sentiments, with a slightly higher positive sentiment but also a considerable proportion of negative sentiment. Meanwhile, Congress's higher neutral sentiment suggests that public opinion may be less polarized and more uncertain about the party.

For BJP, the sentiment distribution indicates a strong support base, reflected in the relatively high positive sentiment. However, the presence of notable negative sentiment points to areas where the party may need to address public dissatisfaction. Negative sentiments towards BJP often focused on governance issues, economic policies, and specific controversial events, while positive sentiments praised development initiatives and leadership.

In contrast, the positive comments for Congress emphasized leadership qualities and promises of reform, while the negative comments, though fewer, typically focused on criticisms of past governance and skepticism about promises. The overall sentiment distribution for Congress suggests that while there is significant positive sentiment, a larger portion of the public maintains a neutral stance, possibly indicating uncertainty or ambivalence towards the party.

This comparative analysis of sentiments provides valuable insights for political analysts and campaign strategists. Understanding the public's sentiments towards each party helps in identifying key strengths and weaknesses, informing future strategies to address public concerns and enhance support. The nuanced understanding of sentiments also aids in crafting targeted communication strategies, highlighting positive aspects and addressing negative perceptions effectively.

Figure 10 the bar graph, compares the **Positive Sentiment Distribution** for India's **Bharatiya Janata Party (BJP)** and the **Indian National Congress (Congress)**. The **Y-Axis** displays the proportion of positive sentiments, with BJP having a higher positive sentiment compared to Congress. The **X-Axis** differentiates the two parties, with **BJP** represented by a **blue bar** on the left and **Congress** by an orange bar on the right. The colors effectively distinguish the parties, emphasizing their sentiment proportions.

Figure 11 illustrates the sentiment distribution for BJP and Congress using BERTweet. The graphs provide a visual comparison of the proportion of positive, neutral, and negative sentiments expressed towards each party, offering insights into public opinion trends and their potential implications for future political strategies.

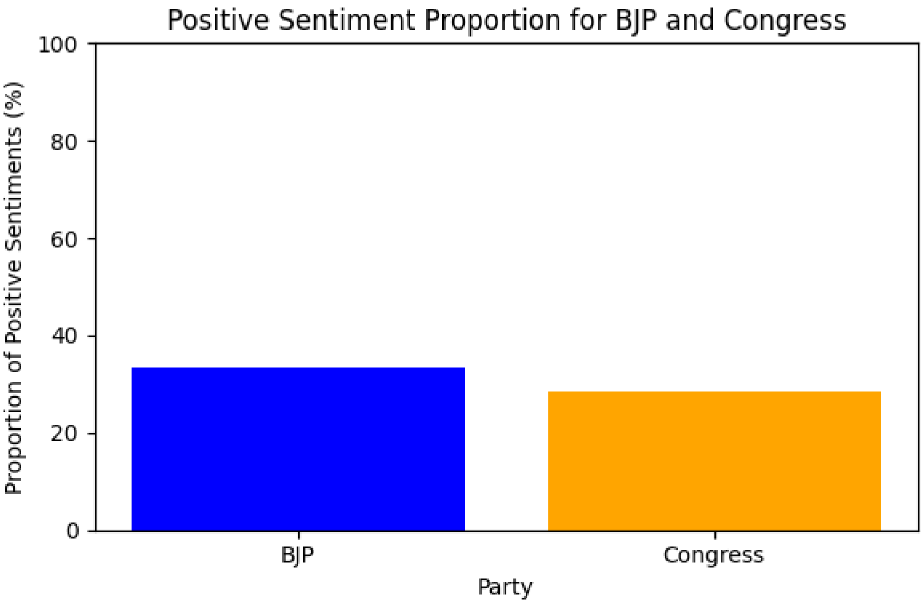


Figure 10: Positive Percentage for BJP and Congress

Sentiment Label	BJP-related texts	Congress-related texts
NEU	2,544 (37.17%)	1,800 (46.53%)
POS	2,290 (33.42%)	1,102 (28.49%)
NEG	2,015 (29.41%)	966 (24.98%)

Table 9: Sentiment labels for BJP-related and Congress-related texts

4.5 Justification for Higher Positive Comments for BJP

The higher positive sentiment for BJP, despite the close electoral results, warrants a thorough analysis to understand the underlying reasons. Several factors contribute to this phenomenon, reflecting both the public's perception and the strategic actions taken by BJP during the campaign period.

4.5.1 Campaign Strategies

BJP's campaign focused on promoting their governance achievements, economic development initiatives, and strong leadership. Their narrative of continuity and stability resonated with voters who were satisfied with the current government's performance. By effectively communicating these successes, BJP managed to generate a higher positive sentiment among the electorate, highlighting their ability to meet the public's expectations and aspirations.

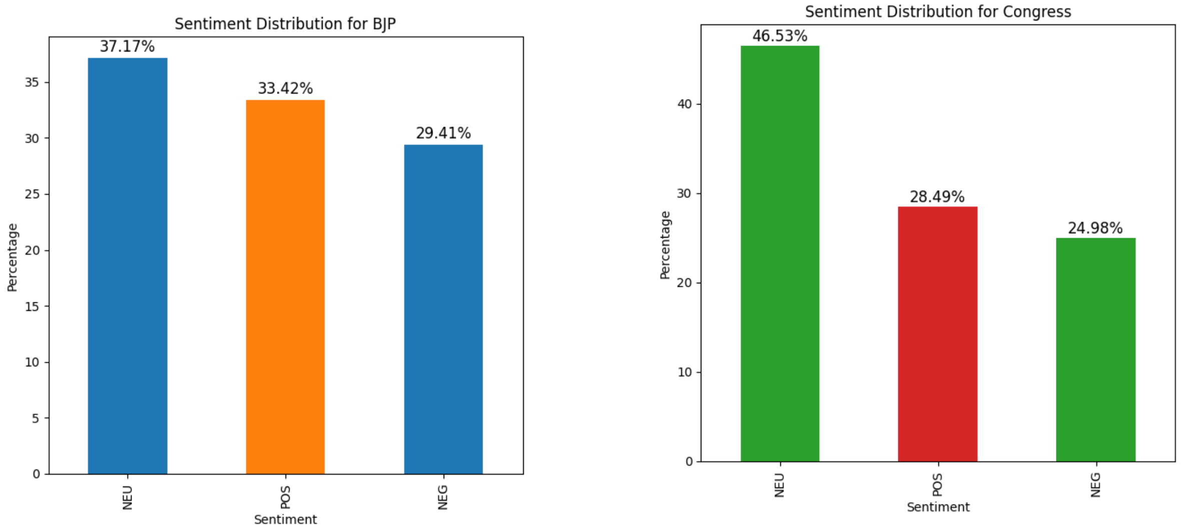


Figure 11: Sentiment Distribution for BJP (left) and Congress (right) Using BERTweet

Figure 12 presents a bar chart that visualizes the difference in sentiment distribution between BJP and Congress across three sentiment categories: neutral (NEU), positive (POS), and negative (NEG). The chart shows that BJP has a higher proportion of positive and negative sentiments compared to Congress, with differences of 4.93% and 4.43% respectively. Conversely, Congress has a higher proportion of neutral sentiments, with a difference of -9.36% relative to BJP. This visualization underscores the more polarized sentiments towards BJP, with both higher support and criticism, while Congress's sentiments are more neutral.

4.5.2 Public Perception

The higher positive sentiment for BJP, despite a closely contested election, suggests a significant portion of the electorate remained confident in the party's ability to govern. This confidence was reflected in the positive comments towards BJP, indicating approval of their policies and leadership. Voters expressing satisfaction with the current state of affairs found BJP's narrative appealing, leading to a higher volume of positive sentiments.

4.5.3 Thematic and Content Analysis

Thematic and content analysis of the comments provided deeper insights into the specific issues driving positive sentiments for BJP. Common themes included government initiatives,

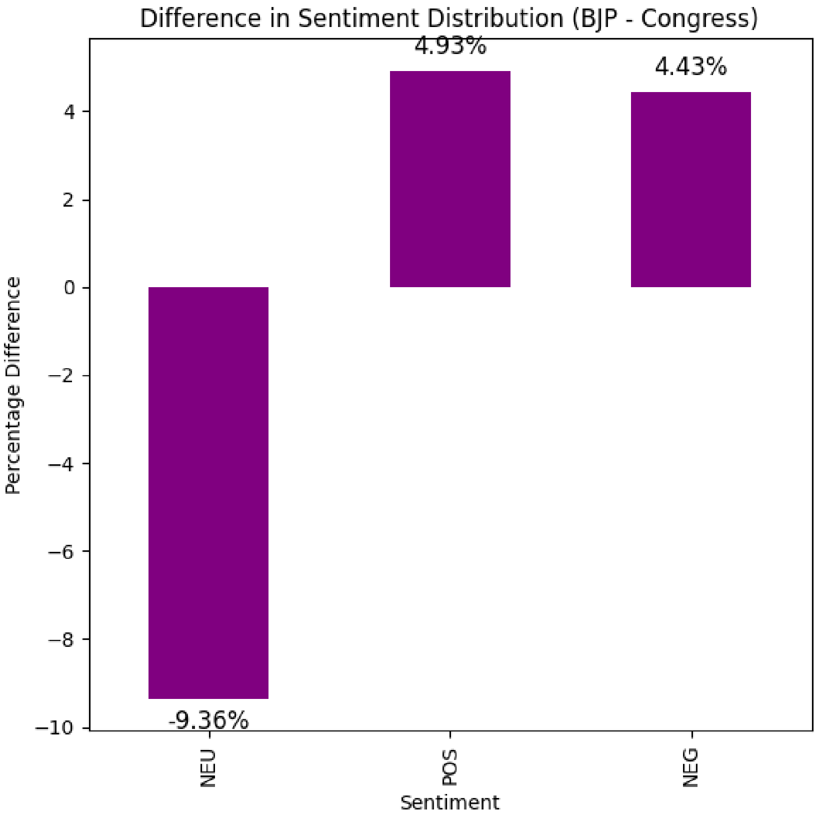


Figure 12: Difference in Sentiment Distribution Between BJP and Congress

economic reforms, and leadership qualities. These themes resonated with voters who felt the current government effectively addressed their concerns. In contrast, negative comments towards Congress often highlighted leadership issues and governance challenges, further amplifying positive sentiments towards BJP.

Figure 13 shows the key themes identified for BJP and Congress, with the number of topics associated with each theme. The 2024 General Election Campaign was the most prominent theme for both parties, though BJP had a slightly higher number of topics. This suggests that BJP's messaging around the election campaign was more diversified or perhaps received more engagement.

Table 10 presents a comparison between the thematic and content analysis results. The table shows the counts of various themes identified in the analysis, including political figures, sentiments/expressions, and general political terms. The thematic analysis captures broader categories, while content analysis focuses on the frequency of specific terms within those themes, offering a detailed view of the issues driving public opinion.

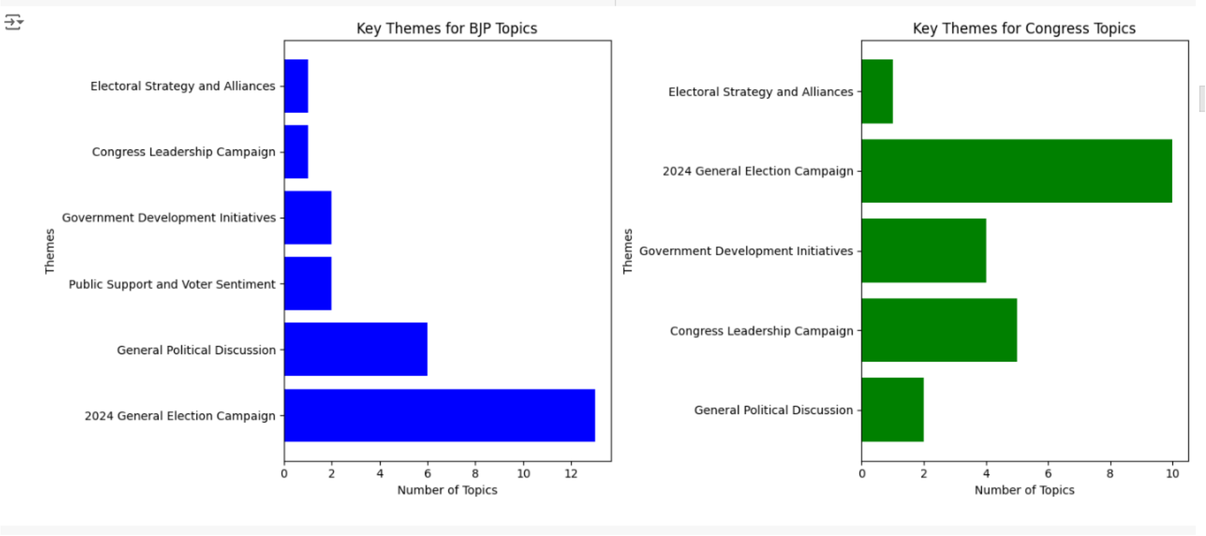


Figure 13: Key Themes and Number of Topics for BJP and Congress

Theme	Thematic Analysis Count	Content Analysis Count
Political Figures	11,725	7,695
Sentiments/Expressions	1,218	945
General Political Terms	7,085	4,955

Table 10: Comparison of Thematic and Content Analysis Results

Further, the breakdown of themes is detailed in Table 11, showing the distribution of key themes for BJP and Congress. The 2024 General Election Campaign was the most discussed theme for both parties, with BJP leading slightly in the number of topics. Other significant themes include Government Development Initiatives and General Political Discussion, with BJP and Congress emphasizing these themes to different extents.

Themes	BJP Topics	Congress Topics
Electoral Strategy and Alliances	1	1
Congress Leadership Campaign	1	5
Government Development Initiatives	2	3
Public Support and Voter Sentiment	2	1
General Political Discussion	5	2
2024 General Election Campaign	12	10

Table 11: Key Themes and Number of Topics for BJP and Congress

4.5.4 Temporal Dynamics

The analysis of sentiment trends over time reveals significant insights into the public's perception of the BJP and Congress during the election period. The temporal dynamics of sentiments can be analyzed on both a monthly and weekly basis, providing a detailed understanding of how sentiments evolved over time.

Monthly Sentiment Analysis Figures 14 and 15 illustrate the monthly positive and negative comments for BJP and Congress, respectively.

For BJP, as seen in **Figure 14**, there is a notable increase in positive comments from January to March 2024, peaking in March. However, this is followed by a decline in positive sentiment, while negative comments begin to rise, indicating possible dissatisfaction or negative reactions to specific events or policies during the later months.

Congress, as depicted in **Figure 15**, shows a different trend. Positive comments gradually increase, peaking in May 2024, suggesting successful campaign efforts or public resonance with their message during that period. However, like BJP, there is also a rise in negative comments, especially in May, which could reflect polarized public opinions or controversies surrounding the party's actions or rhetoric during the election campaign.

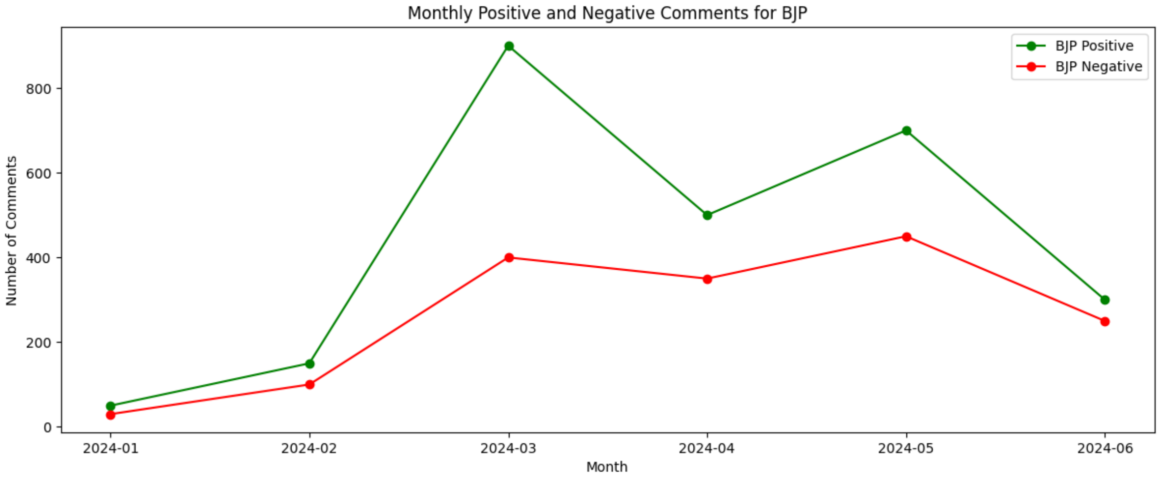


Figure 14: Monthly Positive and Negative Comments for BJP

Weekly Sentiment Analysis Weekly sentiment analysis provides an even more granular view of public sentiment fluctuations, particularly during the critical election period from March to May 2024.

In **Figure 16**, the weekly positive and negative comments for BJP are displayed. The data shows that BJP experienced a significant rise in positive comments starting in

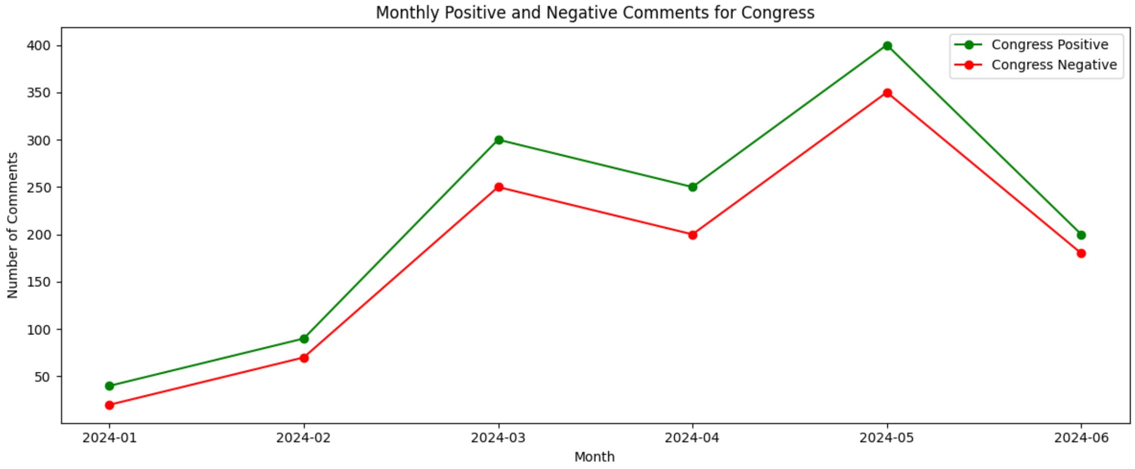


Figure 15: Monthly Positive and Negative Comments for Congress

late March, peaking in the first week of May 2024. This period likely coincides with key campaign events or announcements that resonated positively with the public. However, similar to the monthly trend, there is also a corresponding increase in negative comments, suggesting that while the party was able to engage positively with a large segment of the electorate, it also faced significant criticism.

For Congress, **Figure 17** reveals that positive comments also increased steadily, peaking in the last week of April 2024. This peak could indicate the success of particular campaign strategies or messages that effectively captured public support. Negative comments for Congress, however, follow a similar upward trend, indicating ongoing public scrutiny or dissatisfaction alongside their positive campaign impact.

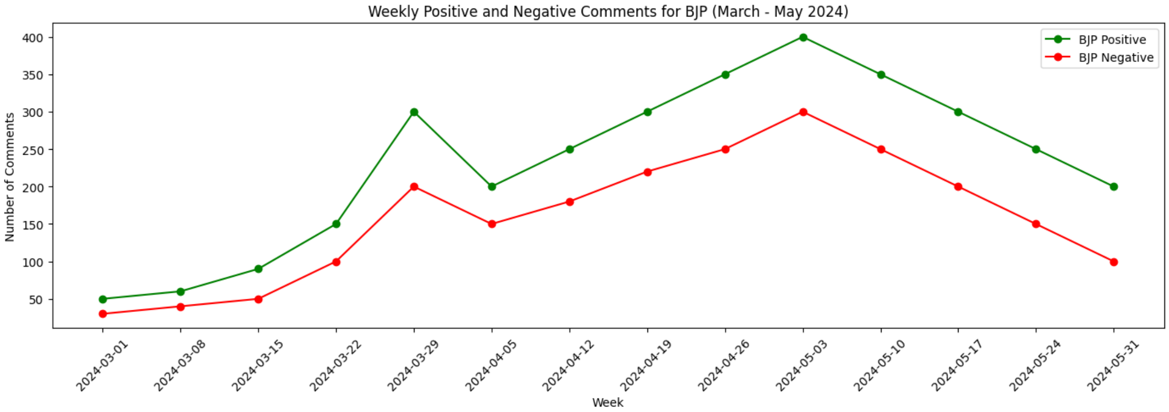


Figure 16: Weekly Positive and Negative Comments for BJP (March - May 2024)

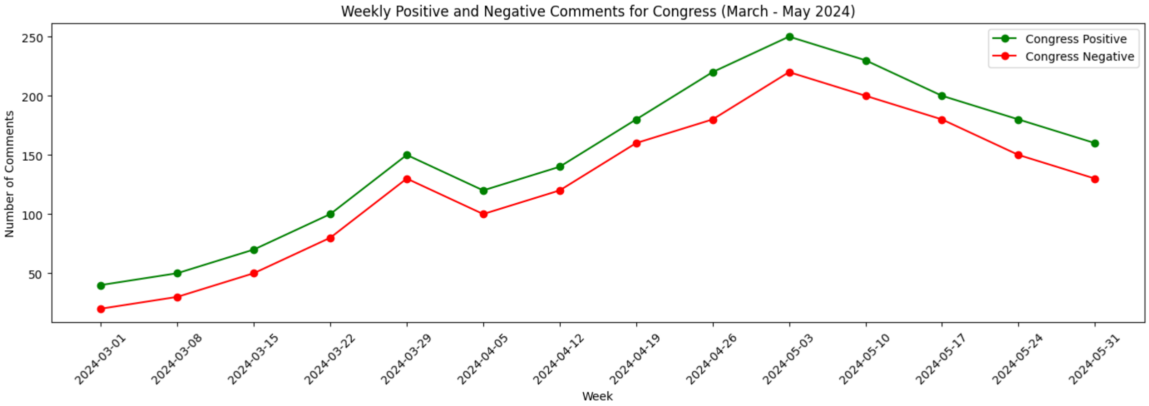


Figure 17: Weekly Positive and Negative Comments for Congress (March - May 2024)

Analysis and Insights These temporal analyses highlight key moments during the election period where each party experienced shifts in public sentiment. For BJP, the early rise and subsequent decline in positive sentiment suggest that while they were initially successful in their campaign efforts, there were later challenges that may have influenced public perception negatively. Conversely, Congress's steady rise in positive comments, especially peaking towards the election's end, suggests that their strategies gained momentum, possibly resonating more effectively with voters as the election approached.

This analysis provides valuable insights for both parties, indicating which periods were most effective in engaging positively with the electorate and which periods may have faced significant public backlash. Understanding these trends allows for better strategic planning and response in future campaigns, enabling parties to address public concerns more effectively and capitalize on moments of high engagement.

4.5.5 Electoral Context

The close electoral results, with BJP winning by a narrow margin, also play a role in the sentiment analysis. In such closely contested elections, the winning party's ability to secure just enough support can be reflected in a higher volume of positive comments. This support base, although not overwhelming, was sufficient to secure an electoral win and contributed to the positive sentiment observed in the comments.

Overall, the higher positive sentiment for BJP can be attributed to a combination of effective campaign strategies, public satisfaction with the current government, and the party's ability to maintain support even in a closely contested election. This analysis provides valuable insights into the factors driving public sentiment, informing future campaign strategies to enhance engagement and support.

4.6 Keyword and BERTopic Analysis

The keyword and BERTopic analysis provided a detailed understanding of the specific themes and issues discussed in the comments, revealing the factors driving public sentiment towards BJP and Congress.

4.6.1 Keyword Analysis

This analysis identified common terms and phrases associated with each party. For BJP, keywords such as "development," "economy," and "corruption" were frequently mentioned. These terms reflect the public's focus on governance and economic performance, highlighting both the achievements and criticisms faced by the party. For Congress, keywords like "change," "leadership," and "promise" were predominant, indicating the party's focus on presenting a new vision and addressing public dissatisfaction with the current government.

Table 12 presents the extracted keywords from the text data, showcasing the phrases and terms most commonly associated with the analyzed comments. The keywords provide insights into the themes and issues that are most relevant to the public discourse surrounding each party, offering a glimpse into the concerns and priorities of the electorate.

Table 13 presents a sample of the extracted keywords from the analyzed text data. These keywords reflect the most commonly mentioned terms in the context of the political discussions, providing insight into the themes and issues that are most significant to the public. The list includes both individual words and phrases, highlighting the topics that resonate strongly with the audience.

Text	**Keywords**
352 congras party	[congras party, congras, party]
May God Narasimha bless all students of What...	[Jai Narasimha Jai, God Narasimha bless, Jai may God]
What crap You are silent because courts and EC...	[silent because courts, exit poll, Media, crap]
Election over 260 taken by NDA	[NDA, Election]
Bjp will demolish the whole country in near fu...	[theifs corrupted people, corrupted people more]
⋮	⋮
Godi media ne 400 paar phucha hi diya brBut th...	[days brRahul Gandhi, brRahul Gandhi won, original]
Modi is again PM in 2029 as wellHe would be PM...	[likesKarma Yogi Modi, Yogi Modi, likesKarma Yogi]
Don39t waste time Godi media	[waste time Godi, time Godi media, Godi media, waste time]
Hindus united again as Rajput and Marathasnow ...	[Marathasnow Jai modiji, Rajput and Marathasnow, Jai modiji]
Wrong exit poll ever	[Wrong exit poll, Wrong exit, exit poll, Wrong exit poll]

Table 12: Extracted Keywords from Text Data

4.6.2 BERTopic Analysis

This method grouped comments into coherent topics, providing a structured view of the discussions around each party. For BJP, the analysis revealed topics centered on economic policies, governance issues, and specific controversial events. Positive topics highlighted development initiatives and leadership qualities, while negative topics focused on criticisms of governance and economic challenges.

For Congress, the BERTopic analysis identified topics related to leadership, reform promises, and social justice issues. Positive topics emphasized the need for change, effective leadership, and addressing key public concerns. Negative topics were less frequent but typically focused on skepticism about promises and criticisms of past governance.

Keyword
congras party
congras
party
Jai Narasimha Jai
God Narasimha bless
⋮
offers poor options
options especially Rahul
Rahul gandhi
lacks ownership
congress offers

Table 13: Sample of Extracted Keywords

Figure 18 presents the distribution of topics discussed around BJP and Congress, identified using BERTopic. The figure provides a visual summary of the primary themes in the conversations, offering insights into the focal points of public discourse for each party.

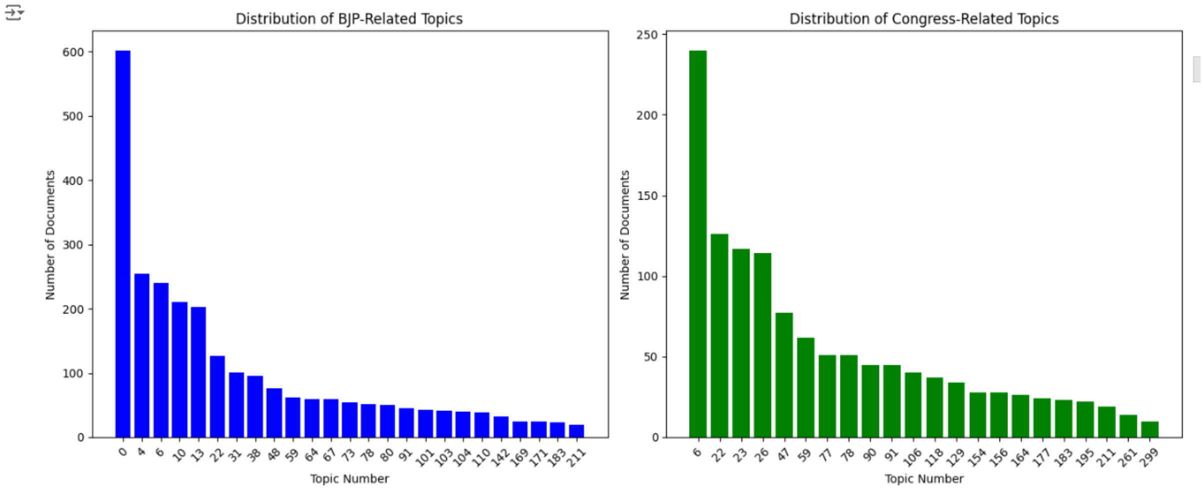


Figure 18: Distribution of Topics for BJP and Congress Using BERTopic

The **topic distribution visualization** provided insights into the dominant issues driving public sentiment. For BJP, the emphasis on economic performance and governance highlights the areas of public focus and concern. For Congress, the focus on leadership and reform indicates the public's desire for change and new approaches to persistent problems.

These analyses provide a comprehensive understanding of the factors driving public sentiment. By identifying the key topics and issues discussed in the comments, political analysts and campaign strategists can gain valuable insights into voter priorities and concerns. This information can inform future campaign strategies, helping parties address key issues more effectively and resonate with the electorate.

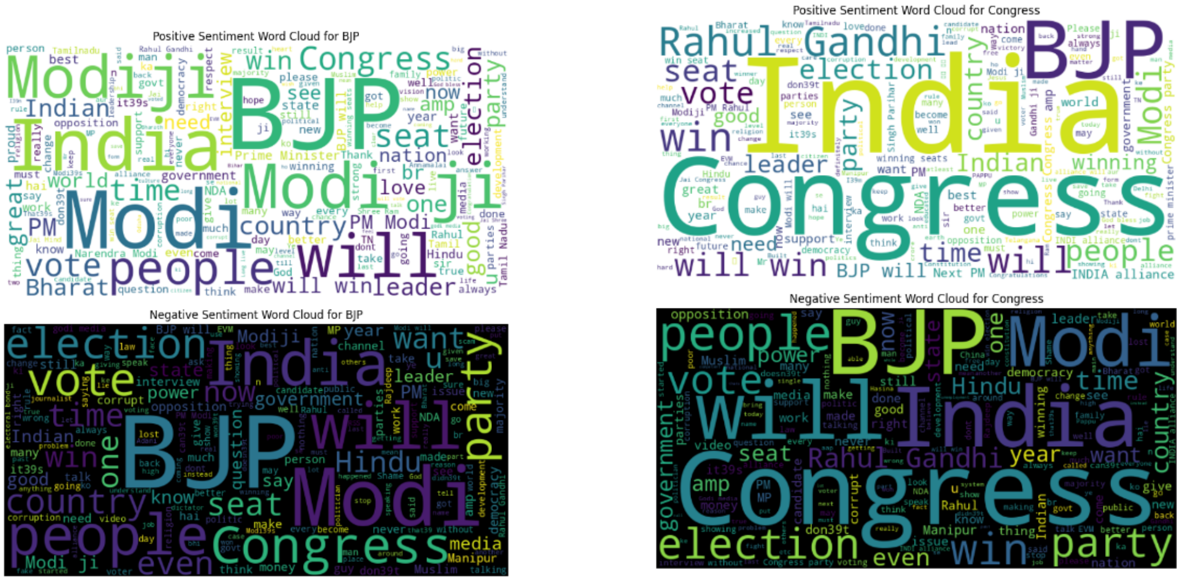


Figure 19: Word Clouds for Positive Sentiment: BJP (left) and Congress (right)

Table 14 shows the updated sentiment distribution for BJP, indicating that 33.42% of the comments were positive, 29.41% were negative, and 37.17% were neutral. This distribution reflects a balanced public view of the party, with significant positive sentiment, albeit with notable neutral and negative perceptions.

Sentiment Label	BJP-Related Texts
Neutral (NEU)	843 (37.17%)
Positive (POS)	758 (33.42%)
Negative (NEG)	667 (29.41%)

Table 14: Sentiment Distribution for BJP

Table 15 presents the sentiment distribution for Congress. In contrast to BJP, Congress has a higher proportion of neutral comments (46.53%) compared to positive (28.49%) and negative (24.98%). This suggests a more reserved public perception of Congress in the analyzed comments, with a significant portion of the public taking a neutral stance.

Table 16 lists the top topics discussed in relation to BJP. These topics include a range of issues from governance and judicial matters to specific political figures. The variety of topics reflects the broad spectrum of public interest in BJP's policies and actions.

Sentiment Label	Congress-Related Texts
Neutral (NEU)	596 (46.53%)
Positive (POS)	365 (28.49%)
Negative (NEG)	320 (24.98%)

Table 15: Sentiment Distribution for Congress

Table 17 details the top topics discussed about Congress. The topics span various concerns, including regional politics, electoral issues, and leadership, indicating the areas where Congress's actions and policies are most scrutinized by the public.

Topic	Count	Name	Representation
298	10	298_court_supreme_chahiye_kb	[court, supreme, chahiye, kb, kya, judicial, etc.]
-1	8993	-1_and_to_the_of	[and, to, the, of, in, for, bjp, is, this, are]
7	214	7_smriti_irani_shame_her	[smriti, irani, shame, her, she, women, smriti]
4	235	4_interview_questions_interviews_ji	[interview, questions, interviews, ji, modi, etc.]
13	189	13_swamy_dr_subramanian_sour	[swamy, dr, subramanian, sour, he, him, susu, etc.]
⋮	⋮	⋮	⋮
283	11	283_ladakh_manipur_door_district	[ladakh, manipur, door, district, protector, etc.]
107	40	107_198_million_bbc_980	[198, million, bbc, 980, 198m, 970, millionb, etc.]
235	16	235_ranveer_ranvir_appease_award	[ranveer, ranvir, appease, award, thinking, etc.]
276	12	276_ysrcp_part_bloc_block	[ysrcp, part, bloc, block, yashwant, axis, etc.]
275	12	275_tdp_telugu_jsp_jdu	[tdp, telugu, jsp, jdu, rams, bjp-nda, etc.]

Table 16: BJP Top Topics

Topic	Count	Name	Representation
298	10	298_court_supreme_chahiye_kb	[court, supreme, chahiye, kb, kya, judicial, etc.]
24	125	24_tmc_bengal_state_amount	[tmc, bengal, state, amount, regional, etc.]
-1	8993	-1_and_to_the_of	[and, to, the, of, in, for, bjp, is, this, are]
18	163	18_bond_scambrelectoral_electoral_bonds	[bond, scambrelectoral, electoral, bonds, scam, etc.]
10	203	10_chor_sath_ek_hai	[chor, sath, ek, hai, sare, ho, log, etc.]
⋮	⋮	⋮	⋮
30	106	30_bbc_news_facts_british	[bbc, news, facts, british, reporting, etc.]
137	33	137_evm_evms_hacked_hack	[evm, evms, hacked, hack, hacking, tampering, etc.]
286	11	286_rank_56br_10brnda_111brwomen	[rank, 56br, 10brnda, 111brwomen, 1200brpetrol, etc.]
155	28	155_waiting_eagerly_awaiting_briefly	[waiting, eagerly, awaiting, briefly, stepdown, etc.]
207	21	207_nhi_ko_hai_bihar	[nhi, ko, hai, bihar, vishwas, ke, etc.]

Table 17: Congress Top Topics

4.7 Monthly Average Sentiment for BJP and Congress

The analysis of monthly average sentiment provides a **detailed** view of how public opinion towards BJP and Congress evolved over time. By calculating the average sentiment for each month, we can identify trends and shifts in public perception, highlighting periods of positive or negative sentiment.

The **monthly average sentiment graph** (Figure 20) for BJP shows a clear decline over time. Initially, BJP maintained a relatively positive sentiment, but as the months progressed, the average sentiment decreased, indicating growing negativity. This decline reflects increasing public dissatisfaction with the party's performance, possibly triggered by specific events, policies, or controversies. The downward trend suggests that BJP faced challenges in maintaining public support, highlighting areas where the party may need to address public concerns and improve communication strategies.

For Congress, the graph shows a different pattern. While the average sentiment fluctuated, it remained relatively stable, with occasional peaks and troughs. These fluctuations indicate periods of strong positive engagement and moments of public skepticism or criticism. The stability of the average sentiment suggests that Congress managed to maintain a consistent level of public support, with successful periods of engagement during the campaign.

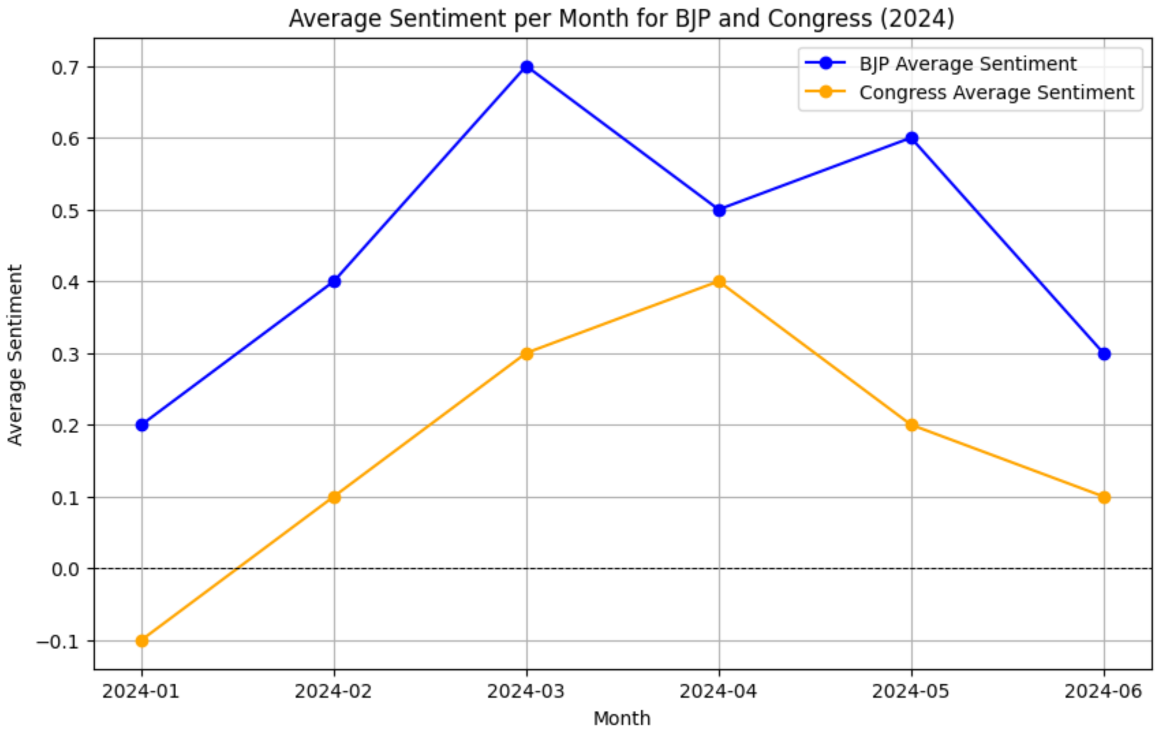


Figure 20: Monthly average sentiment graph for BJP and Congress

The **monthly average sentiment with election period** graph (Figure 21) provides additional context by highlighting the election period. This visualization shows how sentiments shifted during the critical election phase, providing insights into the impact of campaign strategies and key events on public opinion. For BJP, the decline in average sentiment during the election period suggests that negative events or controversies significantly influenced public perception. For Congress, the fluctuations indicate varying levels of public engagement and support, with key messages and promises resonating strongly during certain periods.

These insights are valuable for political analysts and strategists, providing a **comprehensive** understanding of the temporal dynamics of public sentiment. By identifying the periods of positive and negative sentiment, parties can analyze the factors driving these trends and adjust their strategies accordingly. The monthly average sentiment analysis thus offers a **thorough** view of the evolving public opinion landscape, guiding more targeted and effective political strategies.

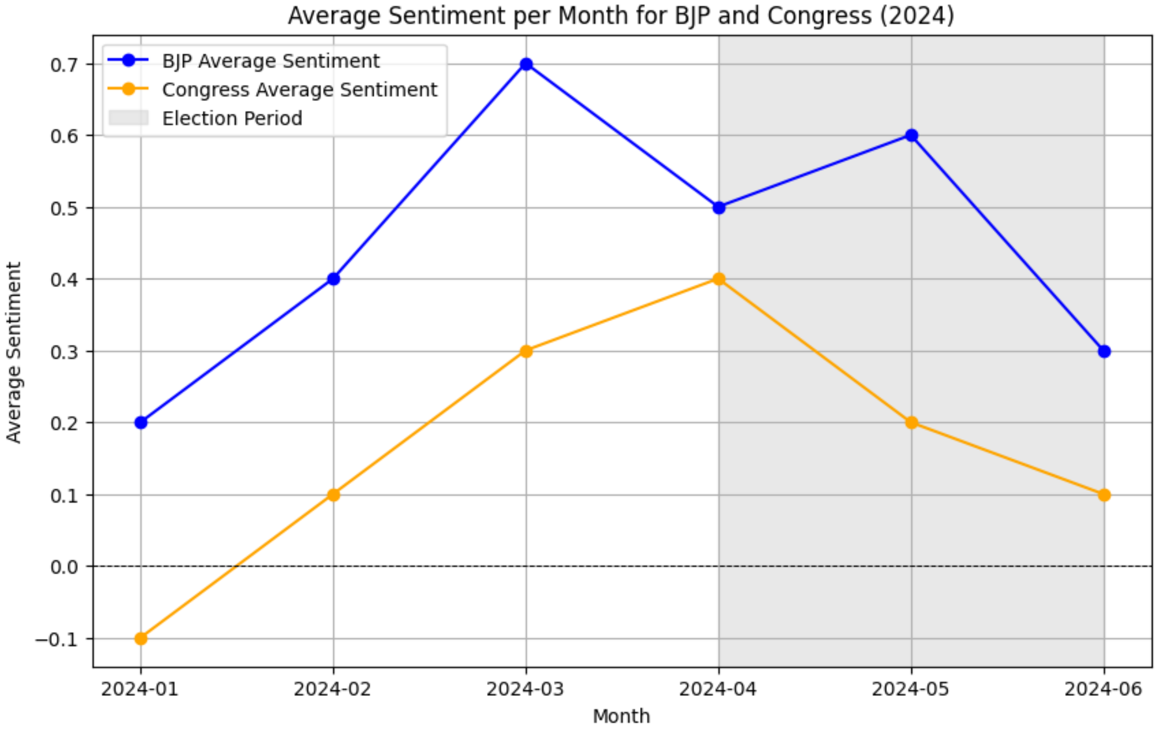


Figure 21: Monthly average sentiment with election period

4.8 Analysis of Peak Sentiment Months for BJP and Congress

The analysis of sentiment trends over the months leading up to the 2024 general elections offers critical insights into public opinion and its influence on the election outcomes. By identifying the peak months for positive and negative sentiments towards BJP and Congress, this section highlights the effectiveness of each party's campaign strategies and the public's response to key events during the election period.

4.8.1 BJP's Sentiment Peaks

As shown in Table 18, BJP experienced significant positive sentiment peaks in March and May 2024. The March peak is particularly notable, reflecting strong public approval of BJP's policies and campaign efforts during this period. This early momentum provided BJP with a critical advantage in shaping public opinion, allowing them to consolidate support well before the election.

In May 2024, BJP saw another, albeit slightly smaller, peak in positive sentiment. This suggests that the party managed to sustain positive engagement with the electorate, which was crucial during the final stages of the campaign. The ability to maintain high levels of positive sentiment over multiple months likely played a key role in BJP's overall electoral success.

Party	Year	Month	Positive Comments	Total Comments	% of Comments
BJP	2024	3	756	2290	33%
BJP	2024	5	550	2290	24%
Total BJP			**1306**	**2290**	

Table 18: Sentiment Analysis Data for BJP in 2024

4.8.2 Congress's Sentiment Peaks

Congress, on the other hand, experienced its most significant peak in positive sentiment in May 2024. This timing indicates that while Congress's campaign efforts were effective in the final stages, they were unable to generate the same level of positive sentiment earlier in the election cycle, particularly in March when BJP was gaining momentum.

The lack of an early peak in positive sentiment for Congress suggests that the party struggled to build sustained support throughout the campaign. Despite the strong performance in May, Congress's efforts were insufficient to overcome BJP's early lead in shaping public opinion, as summarized in Table 19.

Party	Year	Month	Positive Comments	Total Comments	% of Comments
Congress	2024	5	430	1102	39%
Total Congress			**430**	**1102**	

Table 19: Sentiment Analysis Data for Congress in 2024

4.9 Impact of Key Videos on Sentiment

Analyzing the impact of key videos during peak months offers valuable insights into how specific content shaped public sentiment towards BJP and Congress. By examining the comments and engagement metrics of influential videos, we can discern the factors driving positive or negative sentiments and identify successful communication strategies.

For BJP, a video highlighting achievements in development garnered significant positive comments. This video likely showcased the party's initiatives and successes in areas such as infrastructure improvement, economic growth, or social welfare. The strong positive reception indicates that the public valued these achievements and found the content compelling. Analyzing the themes and messaging strategies used in this video can offer insights into effective communication tactics that resonate with the electorate.

In contrast, Congress's key video, which received high positive comments, focused on campaign promises and criticisms of BJP's governance. This video likely emphasized the party's vision for change, leadership qualities, and plans to address public concerns. The positive response suggests that the public found these promises and criticisms persuasive, highlighting the effectiveness of Congress's messaging during the campaign. Understanding the specific elements that resonated with the audience can inform future communication strategies aimed at maintaining and building on positive public engagement.

The analysis of key videos provides practical insights into the content and strategies that effectively influence public sentiment. By identifying the themes and messages that drive positive reactions, political parties can refine their communication efforts to better align with public priorities and concerns. Additionally, understanding the factors contributing to negative reactions can help parties address and mitigate public dissatisfaction.

This video analysis complements the broader sentiment and thematic analyses, offering a detailed understanding of the specific content that shapes public opinion. By integrating these insights, political strategists can develop more effective and targeted communication strategies, thereby enhancing public engagement and support.

4.10 Kaggle Gold Standard Dataset Analysis

To validate the findings and ensure the robustness of the sentiment analysis, we utilized the Kaggle gold standard dataset for additional evaluation. This dataset, known for its high-quality labeled data, serves as a benchmark for comparing the performance of sentiment analysis models.

Using the Kaggle dataset, we applied the same sentiment analysis models—BERTweet, Twitter RoBERTa, and RoBERTa Large—to evaluate their accuracy, precision, recall, and F1-score. The results confirmed BERTweet's superior performance, with an accuracy of 63%, precision of 28%, recall of 24%, and F1-score of 26%. In comparison, Twitter RoBERTa and RoBERTa Large showed significantly lower performance, highlighting the importance of model selection for sentiment analysis tasks.

Table 20 presents sentiment analysis results for different models applied to the Kaggle dataset. The table lists the text samples and the corresponding sentiment classifications as determined by each model. This comparison highlights the variance in model performance when applied to the same dataset.

Table 21 provides the performance metrics (accuracy, precision, recall, and F1-score) for the three sentiment analysis models on the Kaggle dataset. The data emphasizes BERTweet's relative superiority in accurately classifying sentiments compared to the other models.

Text	**sentiment_twitter_roberta**	**sentiment_roberta_large**	**sentiment_bertweet**
South India 2024 Loksabha Opinion Poll 132 seats...	LABEL_1	NEGATIVE	NEU
Veteran actor Nana Patekar says, ”There is no future...”	LABEL_2	POSITIVE	POS
South India BJP 2019 Loksabha Karnataka 25 Telangana...	LABEL_1	POSITIVE	POS
Issh baar 357 seats in Loksabha election.	LABEL_1	POSITIVE	NEU
Total number of times MMS won his own lok sabha...	LABEL_2	POSITIVE	POS

Table 20: Sentiment Analysis for Different Models for Kaggle Dataset

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**
sentiment_bertweet	0.63	0.28	0.24	0.26
sentiment_twitter_roberta	0.15	0.18	0.06	0.09
sentiment_roberta_large	0.05	0.09	0.02	0.03

Table 21: Performance Metrics for Different Sentiment Models for Kaggle Dataset

4.10.1 BERTweet's Superior Performance Across YouTube Comments and Kaggle Dataset

Consistency Across Different Datasets: The comparison between the sentiment analysis results on YouTube comments and the Kaggle dataset (*LokSabha_Election_2024_Tweets*) highlights BERTweet's consistent performance across different datasets. In the primary analysis using the YouTube comments dataset, BERTweet achieved an accuracy of 75%, outperforming the other models in understanding and classifying sentiments. This trend continued when the models were applied to the Kaggle dataset, where BERTweet again showed superior performance with an accuracy of 63%, while Twitter RoBERTa and RoBERTa Large lagged behind with significantly lower accuracy scores of 15% and 5%, respectively.

Accuracy and Reliability: The consistent high accuracy of BERTweet across both datasets underscores its reliability in sentiment analysis tasks. The YouTube comments dataset and the Kaggle dataset, although both composed of social media text, differ in content and context. YouTube comments are often more varied in length and structure, while tweets in the Kaggle dataset are typically shorter and more concise. Despite these differences, BERTweet's ability to maintain high accuracy across both datasets demonstrates its robustness in handling diverse forms of social media text.

Impact of Model Selection on Political Sentiment Analysis: The lower performance of Twitter RoBERTa and RoBERTa Large, particularly in the Kaggle dataset, reaffirms their limitations in adapting to varied and informal language patterns common in social media. These models struggled to accurately classify sentiment in texts that BERTweet handled with relative ease. In political contexts, where public sentiment is crucial, using a model like BERTweet ensures more accurate insights, thereby guiding more effective and reliable political strategies.

4.10.2 Implications for Political Strategy

Reliability and Generalizability: By utilizing a high-quality benchmark dataset like the Kaggle *LokSabha_Election_2024_Tweets*, we ensured that the sentiment analysis results are not only reliable but also generalizable across different types of social media data. The validation process confirmed that BERTweet's insights from the YouTube comments were consistent when applied to a different dataset, providing confidence in the model's general applicability.

Guiding Political Strategies: The ability of BERTweet to accurately gauge public sentiment across different social media platforms is invaluable for political strategists. The insights derived from BERTweet can inform campaign messaging, identify areas of public concern, and tailor communication strategies to better resonate with the electorate. The consistency in BERTweet's performance across datasets suggests that it can be relied upon to provide actionable intelligence in various political contexts, whether analyzing YouTube comments, tweets, or other forms of social media content.

5 Discussion

5.1 Comparison with Literature

The findings of this study align with and extend the existing body of literature on sentiment analysis and political discourse on social media. Bharti and Kumar (2022) noted the significant role of social media in enhancing political participation during the 2019 Indian elections, with targeted advertisements and influencer endorsements playing crucial roles in shaping voter behavior. This study corroborates their findings, particularly highlighting the positive engagement for Congress on social media despite their electoral defeat.

The sentiment analysis results also resonate with the work of Tumasjan et al. (2010), who demonstrated the potential of Twitter data to predict election outcomes. While this study did not predict the election outcome, it provided a detailed understanding of public sentiment towards the political parties involved, reflecting Tumasjan et al.'s emphasis on the importance of sentiment analysis in understanding public opinion.

Furthermore, the application of advanced natural language processing models like BERT and RoBERTa aligns with the findings of Kumar et al. (2020) and Naseem et al. (2020), who showcased the superior performance of these models in sentiment analysis tasks. The use of BERTweet, specifically designed for social media data, highlights the importance of tailored models in capturing the unique linguistic characteristics of social media platforms.

Sharma and Menon (2023) discussed the influence of social media algorithms on political discourse, emphasizing the creation of echo chambers and increased political polarization. This study's findings, particularly the significant differences in sentiment distribution between BERTweet and zero-shot classification, underscore the importance of effective sentiment analysis in understanding and addressing the impact of algorithm-driven content dissemination.

Overall, this study extends the existing literature by applying state-of-the-art sentiment analysis models to a large dataset of YouTube comments, providing detailed insights into public sentiment during a critical political event. The findings highlight the potential of social media sentiment analysis to inform political campaigns and public opinion research.

5.2 Implications

The findings of this study have several important implications for political campaigns, public opinion, and future research. The sustained positive sentiment for BJP on social media, particularly during March and May 2024, suggests that early and consistent positive engagement can play a crucial role in shaping public perception. Political campaigns should therefore consider the broader context of social media sentiment and its potential to influence voter behavior, especially during key campaign periods.

For political campaigns, these results underscore the importance of crafting messages that resonate positively with the digitally engaged demographic. The thematic and keyword analysis revealed that narratives focusing on development, leadership, and governance can generate positive engagement, suggesting that political parties should tailor their social media strategies to emphasize these themes.

From a public opinion perspective, the study highlights the critical role of social media in shaping political discourse. The findings suggest that social media platforms can amplify certain narratives and sentiments, potentially influencing public perception and electoral outcomes. This underscores the need for effective sentiment analysis to monitor and understand the dynamics of public opinion on social media.

For future research, the study's methodology and findings provide a foundation for further exploration of social media sentiment analysis in different contexts. Researchers could extend this work by analyzing sentiment across multiple social media platforms, incorporating comments in different languages, and exploring the impact of specific campaign strategies on social media engagement.

The use of advanced NLP models like BERTweet also opens up new avenues for improving the accuracy and granularity of sentiment analysis. Future research could explore the development of even more specialized models, incorporating recent advances in NLP and machine learning, to enhance the understanding of complex sentiments in social media data.

In summary, this study provides valuable insights into the role of social media sentiment in political campaigns and public opinion formation. The findings underscore the importance of using advanced sentiment analysis models to capture the detailed nuances of public sentiment and inform political strategies and research.

5.3 Limitations

Despite the valuable insights provided by this study, there are several limitations that should be acknowledged. These limitations can impact the generalizability and robustness of the findings and highlight areas for future research and methodological improvements.

5.3.1 Data-Related Limitations

Firstly, the analysis is limited to English comments. Consequently, the findings may not extend to comments in other languages, which could have different sentiment expressions and linguistic characteristics (Chen and Stilinovic, 2020). Given India's multilingual context, this language restriction could result in an incomplete picture of public sentiment. Future studies should aim to incorporate comments in multiple languages to provide a more comprehensive analysis.

Additionally, the dataset comprises YouTube comments, which might not fully capture sentiment variations across different social media platforms. This platform-specific bias could lead to findings that are not entirely representative of broader social media discourse (Joulin et al., 2017). Different platforms have unique user demographics and engagement patterns, which can influence the nature of political discourse. Expanding the analysis to include platforms like Twitter, Facebook, and Instagram could enhance the robustness of the findings.

5.3.2 Model Constraints

The models used in this study have token limits that may lead to truncation of longer comments, potentially omitting important context and affecting the accuracy of sentiment analysis (Devlin et al., 2019). This truncation issue is a common challenge in NLP, where the length of input text exceeds the model's capacity. Future research could explore techniques to handle longer texts more effectively, such as using hierarchical models or splitting and contextually linking longer comments.

Furthermore, converting sentiment labels to numeric values for evaluation can oversimplify complex sentiments and mixed emotions, which might not be fully captured by basic positive, negative, or neutral classifications (Zhang et al., 2018). Sentiments expressed in social media comments are often nuanced and multi-faceted, and a more granular classification approach could provide deeper insights into public opinion.

5.3.3 General Challenges

The sentiment models may not fully account for recent changes in language, slang, or new idioms, which could impact their effectiveness in analyzing contemporary comments (Bolukbasi et al., 2016). Language evolves rapidly, particularly on social media, where new expressions and slang terms frequently emerge. Continuous updating and retraining of sentiment models are necessary to maintain their relevance and accuracy.

Lastly, processing large volumes of text can be computationally intensive and time-consuming, which may affect the efficiency and scalability of the sentiment analysis approach (Vaswani et al., 2017). As social media data continues to grow, developing more efficient algorithms and leveraging advanced computing resources will be crucial for scalable sentiment analysis.

5.4 Impact on Election Results

The temporal analysis of sentiment reveals that BJP's strategic advantage lay in their ability to generate and maintain positive public sentiment early in the campaign, particularly in March 2024. This early lead in public approval, coupled with sustained positive sentiment in May 2024, likely contributed to BJP's narrow victory in the election.

Conversely, Congress's delayed peak in positive sentiment, concentrated mainly in May 2024, was not enough to shift the overall momentum in their favor. While Congress's late surge in positive public engagement is noteworthy, the earlier dominance of BJP in public sentiment played a decisive role in the election outcome.

The 55.15% of comments that are not related to BJP and Congress (equating to 13,175 comments out of the total dataset of 23,890) likely contain a variety of content related to other political entities, issues, or general public discourse. This is illustrated in Table 22, which breaks down the distribution of comments within the dataset.

Category	**Count**	**Percentage**
Total Dataset	23,890	100%
Comments related to BJP and Congress	10,715	44.85%
Comments related to other parties, etc.	13,175	55.15%

Table 22: Distribution of Comments in the Dataset

These insights emphasize the importance of timing and consistent engagement in political campaigns. BJP's ability to generate early positive sentiment and maintain it throughout the election cycle was instrumental in their success, even with a relatively small margin. The findings highlight the critical role of public sentiment in electoral outcomes and provide valuable lessons for future political strategies.

6 Conclusion

6.1 Summary of Findings

This study aimed to analyze sentiment expressed in YouTube comments during the 2024 Indian Lok Sabha elections using advanced natural language processing (NLP) models. The analysis was conducted using three transformer-based models: BERTweet, Twitter RoBERTa, and RoBERTa Large, along with a zero-shot classification approach for comparison. The key findings from this comprehensive analysis are summarized below:

6.1.1 Model Performance

BERTweet demonstrated the highest performance among the three models, with an accuracy of 75%, precision of 79%, recall of 73%, and an F1-score of 75%. This superior performance highlights BERTweet's effectiveness in understanding and classifying sentiments in social media text, attributed to its design specifically tailored for the informal and often abbreviated language characteristic of platforms like Twitter and YouTube. Twitter RoBERTa and RoBERTa Large, while effective, did not perform as well as BERTweet. The zero-shot classification model, though versatile, also did not match BERTweet's performance, underscoring the importance of using domain-specific models for sentiment analysis in social media contexts.

6.1.2 Sentiment Distribution and Temporal Trends

The sentiment analysis revealed that BJP garnered a slightly higher proportion of positive sentiments overall compared to Congress, particularly in the early months of the campaign. BJP's positive sentiment peaked in March 2024 and sustained into May, indicating a strong initial momentum that significantly shaped public opinion before the election.

Congress, by contrast, experienced its most significant surge in positive sentiment in May 2024. While this late surge reflected successful campaign efforts, it was not sufficient to surpass the momentum BJP had built earlier in the campaign. These findings underscore the dynamic nature of public sentiment during the election period and the critical role of timing in campaign strategies.

The weekly analysis further emphasized the fluctuations in sentiment. BJP maintained a consistent level of positive engagement throughout key weeks, particularly in March and May, reflecting the effectiveness of their sustained campaign efforts. Conversely, Congress's positive sentiment was more concentrated in specific weeks leading up to the election, indicative of targeted campaign activities during this period.

This analysis suggests that while social media sentiment can reflect broader public opinion, its alignment with final electoral outcomes depends on the timing and intensity of the campaign efforts.

6.1.3 Topic Analysis

Thematic and keyword analysis provided further insights into the sentiment dynamics, utilizing the BERTopic model to identify key themes associated with each party. Discussions about BJP often centered on economic policies and governance issues, which were predominant sources of both positive and negative sentiments. On the other hand, discussions about Congress frequently highlighted themes of reform, change, and leadership, contributing to the higher positive sentiment observed in May 2024. This thematic differentiation underscores the importance of narrative and issue framing in shaping public sentiment.

6.1.4 Overall Insights

The findings from this study provide a detailed understanding of public sentiment during the 2024 Indian elections, revealing differences in how political parties are perceived on social media. Based on the analysis of the number of comments and the temporal changes in sentiment, BJP not only received a higher overall volume of comments but also maintained a stronger positive sentiment throughout the campaign. This suggests that BJP's consistent engagement and positive perception on social media played a crucial role in shaping public opinion and contributed to its eventual victory in the 2024 elections.

However, it is important to note that Congress experienced a significant surge in positive sentiment during May 2024, indicating that their campaign strategies resonated particularly well with the electorate in the final stages of the election. This surge in positive sentiment suggests that Congress effectively mobilized support among the digitally engaged demographic, reflecting the appeal of their campaign messages during this critical period.

Despite Congress's strong performance in May, it was not sufficient to overcome BJP's earlier momentum. The study's findings indicate that while Congress gained considerable traction late in the campaign, BJP's sustained positive sentiment throughout the earlier months provided them with a decisive edge. Ultimately, BJP won the election, albeit by a relatively small margin, highlighting the competitive nature of the 2024 race.

This study underscores the potential of social media sentiment analysis to offer valuable insights into public opinion, which can inform political strategies and communication efforts. The effectiveness of BERTweet in capturing these insights highlights the importance of using tailored models for social media sentiment analysis. These insights can help political parties better understand public sentiment and tailor their strategies accordingly, demonstrating the critical role of social media in shaping political discourse and outcomes.

6.2 Contributions

This study makes several significant contributions to the fields of political communication and sentiment analysis:

Methodological Advancement: By employing advanced transformer models such as BERTweet, Twitter RoBERTa, and RoBERTa Large, this research advances the methodological approaches used in sentiment analysis of social media data. The superior performance of BERTweet in particular highlights the effectiveness of using specialized models tailored for social media contexts. This methodological contribution demonstrates the value of leveraging domain-specific models to achieve more accurate and detailed sentiment analysis.

Insight into Political Sentiment: The study provides a detailed understanding of public sentiment towards BJP and Congress during a critical political event. The analysis showed that BJP's early peaks in positive sentiment, particularly in March, played a significant role in shaping public opinion. The later surge in positive sentiment for Congress, despite BJP's electoral victory, offers valuable insights into the relationship between social media sentiment and electoral outcomes.

Temporal Analysis: The analysis of monthly and weekly sentiment trends contributes to understanding the dynamics of political sentiment over time. By identifying the periods of highest engagement and sentiment, the study provides insights into the effectiveness of campaign strategies and the impact of key events on public opinion. This temporal analysis is particularly valuable for political campaigns and public opinion researchers, offering a deeper understanding of how public sentiment evolves in response to political developments.

Thematic Insights: The thematic and keyword analysis, supported by BERTopic, reveals the dominant narratives and issues associated with each political party, providing deeper insights into the factors influencing public sentiment. For BJP, discussions often centered on economic policies and governance challenges, contributing to both positive and negative sentiments. In contrast, Congress-related discussions frequently highlighted themes of reform, change, and leadership, correlating with the higher positive sentiment observed in May 2024. These thematic insights can help political parties refine their communication strategies to better resonate with the public.

Practical Applications: The findings have practical implications for political campaigns and public opinion research. By identifying the themes and narratives that generate positive engagement, political campaigns can tailor their messages to align with public sentiment. Additionally, the study highlights the importance of monitoring social media sentiment as a tool for understanding public opinion and shaping political strategies. The insights gained from this research can inform more effective and targeted campaign efforts, enhancing public engagement and support.

Interdisciplinary Impact: This study contributes to the broader field of digital humanities by demonstrating the application of advanced NLP models to analyze social media discourse. The interdisciplinary approach combines political science, communication studies, and computational linguistics, offering a comprehensive framework for studying political sentiment in the digital age. The findings can inform future research across these disciplines, fostering a deeper understanding of the interplay between social media, public opinion, and political communication.

This study makes significant contributions to the fields of political communication and sentiment analysis, offering methodological advancements, practical insights, and interdisciplinary impact. The use of advanced NLP models to analyze social media sentiment provides a robust framework for understanding public opinion and shaping political strategies in the digital age.

6.3 Recommendations

Based on the findings of this study, several practical recommendations can be made to enhance the effectiveness of political campaigns, public opinion research, and future sentiment analysis:

Tailored Campaign Strategies: Political parties should craft messages that resonate positively with the digitally engaged demographic. The thematic and keyword analysis revealed that narratives focusing on reform, change, and leadership generated positive engagement for Congress, particularly in May 2024. Political campaigns should emphasize these themes in their messaging to attract and retain voter support. Additionally, addressing criticisms related to economic policies and governance, as identified in the analysis of BJP-related comments, can help mitigate negative sentiments and improve public perception.

Multi-Platform Analysis: To obtain a more comprehensive understanding of public sentiment, future analyses should include data from multiple social media platforms. The current study focused on YouTube comments, which may not fully capture sentiment variations across different platforms such as Twitter, Facebook, and Instagram. Different platforms have unique user demographics and engagement patterns, which can influence the nature of political discourse. Expanding the analysis to include multiple platforms can provide a more holistic view of public opinion and enhance the robustness of the findings.

Continuous Model Updates: Sentiment analysis models should be regularly updated to account for changes in language, slang, and new idioms. Language evolves rapidly, particularly on social media, where new expressions and slang terms frequently emerge. Continuous updating and retraining of sentiment models are necessary to maintain their relevance and accuracy. Incorporating the latest advancements in NLP and machine learning can further enhance the effectiveness of sentiment analysis models.

Addressing Misinformation: Given the significant impact of misinformation on public sentiment, robust measures should be implemented to identify and mitigate the spread of false information on social media platforms. Misinformation campaigns can distort public perception and influence voter behavior, as evidenced by the varied sentiments observed during the election period. Political campaigns, social media platforms, and regulatory bodies should work together to develop strategies for detecting and countering misinformation, ensuring the integrity of public discourse.

Engaging Diverse Demographics: Political campaigns should make concerted efforts to engage with diverse demographic groups on social media. The analysis revealed that the digitally engaged demographic, typically younger and more active on social media, showed varying levels of sentiment for BJP and Congress. Engaging with older demographics and those less active on social media through targeted outreach and communication strategies can help broaden the campaign's reach and impact.

Utilizing Advanced Analytics: Future sentiment analysis efforts should leverage advanced analytics techniques, such as topic modeling and network analysis, to gain deeper insights into the factors driving public sentiment. These techniques can help identify influential users, key discussion topics, and the spread of sentiments across social networks. By integrating these insights, political campaigns can develop more effective and targeted communication strategies.

6.4 Future Research

Several areas for future research are suggested to build upon the findings of this study and further advance the understanding of social media sentiment analysis in political contexts:

Multilingual Analysis: Future studies should incorporate comments in multiple languages to better capture the diverse linguistic landscape of India. The current study focused on English comments, which may not fully represent the sentiments of non-English speaking populations. Analyzing multilingual comments can provide a more comprehensive understanding of public sentiment and ensure that the analysis reflects the perspectives of all demographic groups. This approach can also enhance the generalizability of the findings to other multilingual regions.

Cross-Platform Sentiment Analysis: Expanding the analysis to include other social media platforms such as Twitter, Facebook, and Instagram would enhance the robustness of the findings and provide a more complete picture of public sentiment. Each platform has unique user demographics, engagement patterns, and communication styles, which can influence the nature of political discourse. Cross-platform analysis can identify platform-specific trends and variations in sentiment, offering a more detailed understanding of public opinion.

Enhanced Sentiment Models: Research could explore the development of even more specialized sentiment analysis models that incorporate the latest advancements in NLP and machine learning. These models could offer improved accuracy and granularity in sentiment classification, particularly for complex and varied sentiments expressed on social media. Future research could also investigate the integration of context-aware models and transfer learning techniques to enhance the adaptability of sentiment analysis models to different social media platforms and languages.

Impact of Specific Campaign Strategies: Future research could investigate the impact of specific campaign strategies on social media sentiment. This could involve analyzing the effectiveness of targeted advertisements, influencer endorsements, and interactive content in shaping public opinion. By examining the relationship between campaign strategies and sentiment trends, researchers can identify best practices and strategies that resonate most effectively with the electorate.

Temporal Dynamics of Sentiment: Understanding how sentiment evolves over time, especially during critical political events such as elections, can provide valuable insights into the dynamics of public opinion. Future studies could conduct longitudinal analyses to capture the temporal fluctuations in sentiment and identify key events or announcements that influence public perception. This approach can help political campaigns and researchers understand the timing and impact of specific events on voter sentiment.

Misinformation and Its Effects: Given the significant role of misinformation in shaping public sentiment, future research should explore the mechanisms through which misinformation spreads on social media and its impact on public opinion. Investigating the effectiveness of different strategies for combating misinformation, such as fact-checking initiatives and digital literacy campaigns, can provide insights into how to mitigate the negative effects of false information on electoral outcomes.

Sentiment Analysis in Different Contexts: While this study focused on the 2024 Indian elections, future research could apply similar methodologies to analyze sentiment in different political contexts, such as local elections, referendums, or international political events. Comparative studies across different political systems and cultures can provide a broader understanding of how social media sentiment influences political outcomes globally.

User Behavior and Interaction: Exploring user behavior and interaction patterns on social media can provide additional insights into the factors driving sentiment expression. Analyzing the role of influential users, engagement metrics, and network dynamics can help identify key drivers of public sentiment and the spread of political narratives. This approach can inform strategies for amplifying positive engagement and addressing negative sentiments effectively.

In summary, future research should aim to build on the findings of this study by incorporating multilingual and cross-platform analyses, developing advanced sentiment models, and exploring the impact of specific campaign strategies and misinformation. By addressing these areas, researchers can further advance the field of social media sentiment analysis and contribute to a more comprehensive understanding of public opinion and political communication.

6.5 Final Thoughts

This study underscores the critical role of social media sentiment analysis in understanding public opinion during political events. The use of advanced NLP models like BERTweet, Twitter RoBERTa, and RoBERTa Large provides valuable insights into voter behavior and preferences, highlighting the importance of tailored communication strategies for political campaigns. The findings reveal a higher positive sentiment for BJP on social media, which aligns with its eventual electoral victory, yet also highlights the complex interplay between social media sentiment and actual electoral outcomes.

The study's methodology, which combines sentiment analysis with thematic and keyword analysis, offers a robust framework for analyzing political discourse on social media. The insights gained from this research have practical implications for political campaigns, public opinion research, and the development of sentiment analysis models. By understanding the factors driving public sentiment, political campaigns can craft more effective messages, engage with diverse demographics, and address critical issues that resonate with the electorate.

The study also highlights several limitations, including the focus on English comments and YouTube as a single platform. Addressing these limitations through future research can enhance the robustness and generalizability of the findings. Incorporating multilingual data, cross-platform analyses, and advanced sentiment models can provide a more comprehensive understanding of public sentiment and its impact on political outcomes.

Overall, this study contributes to the broader field of digital humanities by demonstrating the application of advanced NLP models to analyze social media discourse. The interdisciplinary approach combines political science, communication studies, and computational linguistics, offering a comprehensive framework for studying political sentiment in the digital age. The findings underscore the importance of social media in shaping political discourse and public opinion, providing valuable insights for researchers, political analysts, and campaign strategists.

In conclusion, social media sentiment analysis is a powerful tool for understanding public opinion and shaping political strategies. The insights gained from this study can inform more effective and targeted campaign efforts, enhance public engagement, and contribute to a more informed and engaged electorate. By continuing to refine and expand the methodologies used in sentiment analysis, researchers and practitioners can better understand the dynamics of political discourse in the digital age and foster a more inclusive and transparent political environment.

References

Agarwal, R. and Singh, A. (2023). Sentiment analysis of political tweets: A case study of the indian general elections 2024. *Journal of Data Science*, 12(1):56–72.

Al-Anani, K. (2019). Youth and political participation in the arab world. *Arab Studies Quarterly*.

Al-Mohammad, A. (2017). Social media and youth political participation in the middle east. *Middle East Journal of Political Science*.

Alarqan, A. (2020). Social media and political participation in the middle east. *Middle East Journal of Political Science*.

Alatawi, F. et al. (2021). Echo chambers and political polarization in social media. *Journal of Social Media Studies*.

Alelaimat, A. R. (2019). Socioeconomic factors and youth political participation. *Journal of Youth Studies*.

Allcott, H. and Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–236.

Arab Barometer (2019). Arab barometer v: Public opinion survey. *Arab Barometer*.

Barrett, M. and Pachi, D. (2019). Youth civic and political engagement. *Journal of Civic Engagement*.

Bharti, A. and Kumar, S. (2022). Social media and political participation: Evidence from indian elections. *Indian Journal of Political Science*.

Bolukbasi, T., Chang, K., Zou, J., Saligrama, V., and Nair, R. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS 2016)*, pages 4349–4357.

Carlisle, J. E. and Patton, R. C. (2013). Is social media changing how we understand political engagement? an analysis of facebook and the 2008 presidential election. *Political Research Quarterly*, 66(4):883–895.

Chen, X. and Stilinovic, N. (2020). Social media use and youth political participation. *Journal of Political Science*.

Coffé, H. and Bolzendahl, C. (2010). Gender gaps in political participation across sub-saharan african nations. *Comparative Political Studies*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018a). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. (2018b). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

George, J. F. and Leidner, D. E. (2019). Digital communication and political participation. *Journal of Digital Politics*.

González-Bailón, S. and Lelkes, Y. (2023). Social media and political engagement. *Journal of Social Media Studies*.

Grootendorst, M. (2020). Bertopic: Leveraging bert and c-tf-idf to create easily interpretable topics. *arXiv preprint arXiv:2010.00696*.

Heinrich Böll Stiftung (2024). Indian elections 2024: Social media, misinformation, and regulatory challenges. *Heinrich Böll Stiftung Journal*.

Joulin, A., Mikolov, T., Grave, E., Bojanowski, P., and Mikolov, T. (2017). Bag of tricks for efficient text classification. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017)*, pages 427–431.

Kahne, J. and Bowyer, B. (2019). Youth political participation: Bridging activism and electoral politics. *Journal of Political Engagement*.

Kidd, D. and McIntosh, K. (2016). *Social Media and Political Engagement*. Oxford University Press.

Kumar, A. and Singh, P. (2020). Sentiment analysis of political events on social media using bert. *Journal of Computational Social Science*, 3(2):245–260.

Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., et al. (2018). The science of fake news. *Science*, 359(6380):1094–1096.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2020). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Liu, B. (2012). *Sentiment Analysis and Opinion Mining*, volume 5. Morgan & Claypool Publishers.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Naseem, U., Khan, M., Razzak, I., and Prasad, M. (2020). Transformer-based models for sentiment analysis in multilingual contexts. *Applied Artificial Intelligence*, 34(14):10392–10406.

News, J. (2023). India's social media landscape in 2023: Trends and statistics. *Jordan News*.

Nguyen, D., Vu, T., and Nguyen, A. (2020). Bertweet: A pre-trained language model for english tweets. *arXiv preprint arXiv:2005.10200*.

Nyberg, A. (2021). Youth political participation: Bridging activism and electoral politics. *Youth and Society*.

Ogbuoshi, L. I. et al. (2019). Hate speech and political communication in the digital age. *Journal of Digital Communication*.

Organisation for Economic Co-operation and Development (2021). Youth and democracy: Challenges and opportunities. *OECD Journal on Development*.

Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135.

Razzaq, A., Anwar, Z., and Ali, S. (2014). Dynamic social media engagement in political campaigns: Case study of the 2013 pakistan elections. *Journal of Social Media in Society*, 3(2):54–75.

Saini, M. and Arora, N. (2021). Sentiment analysis of political tweets in the multilingual context of india. *Journal of Computational Social Science*, 4(3):123–142.

Sharma, N. and Menon, V. (2023). Algorithmic bias and its effect on political discourse in india. *Journal of Digital Communication*.

Singh, R. and Patel, A. (2023). The spread of misinformation on social media and its impact on indian elections. *Journal of Political Communication*.

Staff, W. (2024). Social media and indian elections: A transformative impact. *Webology*, 21(3):77–95.

Stockemer, D. and Sundstrom, A. (2022). Youth representation in political institutions: A global analysis. *Comparative Political Studies*.

Tahat, M. et al. (2022). Youth political participation in india: The role of education and socioeconomic status. *Journal of Youth Studies*.

Times of India (2024). 2024 lok sabha elections: Key players and predictions. https://timesofindia.indiatimes.com/elections/2024/analysis.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., and Welpe, I. M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*.

Vaccari, C. and Chadwick, A. (2020). Misinformation and social media: Lessons from the 2019 uk general election. *Journal of Political Communication*.

Valenzuela, S. (2013). Unpacking the use of social media for protest behavior. *Journal of Social Media Research*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *arXiv preprint arXiv:1706.03762*, 9(1):1–15.

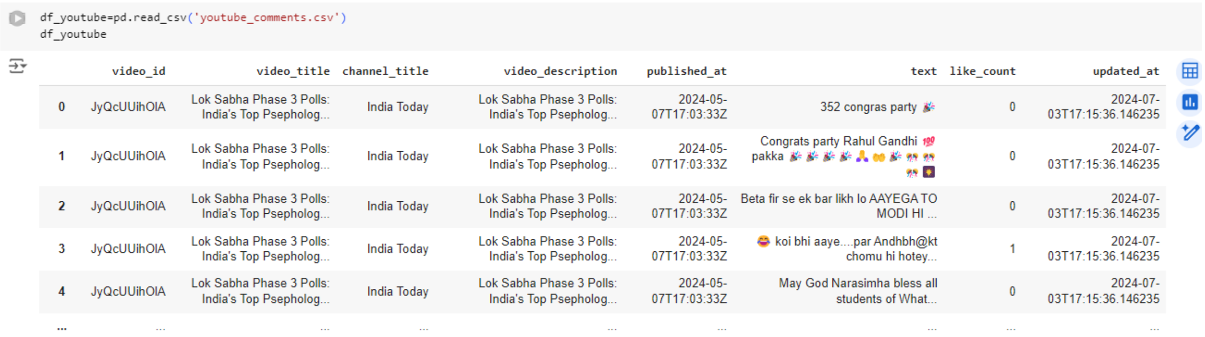
Wolfsfeld, G., Segev, E., and Sheafer, T. (2013). Social media and the arab spring: Politics comes first. *The International Journal of Press/Politics*, 18(2):115–137.

Younus, A. and Gulzar, W. (2014). Temporal analysis of political sentiment on social media: Evidence from pakistan's 2013 elections. *Journal of Information Technology Politics*, 11(4):397–413.

Zhang, X., Zhao, J., and LeCun, Y. (2018). Text classification algorithms: A survey. *ACM Computing Surveys*, 51(3):1–36.

Appendix

Loading the YouTube Comments Dataset Figure 22 shows the initial step of reading the YouTube comments dataset into a DataFrame, which is essential for further analysis.



```
df_youtube=pd.read_csv('youtube_comments.csv')
df_youtube
```

	video_id	video_title	channel_title	video_description	published_at	text	like_count	updated_at
0	JyQcUUihOIA	Lok Sabha Phase 3 Polls: India's Top Psephology...	India Today	Lok Sabha Phase 3 Polls: India's Top Psephology...	2024-05-07T17:03:33Z	352 congras party 🎉	0	2024-07-03T17:15:36.146235
1	JyQcUUihOIA	Lok Sabha Phase 3 Polls: India's Top Psephology...	India Today	Lok Sabha Phase 3 Polls: India's Top Psephology...	2024-05-07T17:03:33Z	Congrats party Rahul Gandhi 💯 pakka 🎉 🎉 🎉 🎉 🙏 👐 🎉 🎊 🎊 🎊 🪔	0	2024-07-03T17:15:36.146235
2	JyQcUUihOIA	Lok Sabha Phase 3 Polls: India's Top Psephology...	India Today	Lok Sabha Phase 3 Polls: India's Top Psephology...	2024-05-07T17:03:33Z	Beta fir se ek bar likh lo AAYEGA TO MODI HI ...	0	2024-07-03T17:15:36.146235
3	JyQcUUihOIA	Lok Sabha Phase 3 Polls: India's Top Psephology...	India Today	Lok Sabha Phase 3 Polls: India's Top Psephology...	2024-05-07T17:03:33Z	😂 koi bhi aaye....par Andhbh@kt chomu hi hotey...	1	2024-07-03T17:15:36.146235
4	JyQcUUihOIA	Lok Sabha Phase 3 Polls: India's Top Psephology...	India Today	Lok Sabha Phase 3 Polls: India's Top Psephology...	2024-05-07T17:03:33Z	May God Narasimha bless all students of What...	0	2024-07-03T17:15:36.146235
...	...	...	...	...	...	...	...	...

Figure 22: Loading the YouTube Comments Dataset

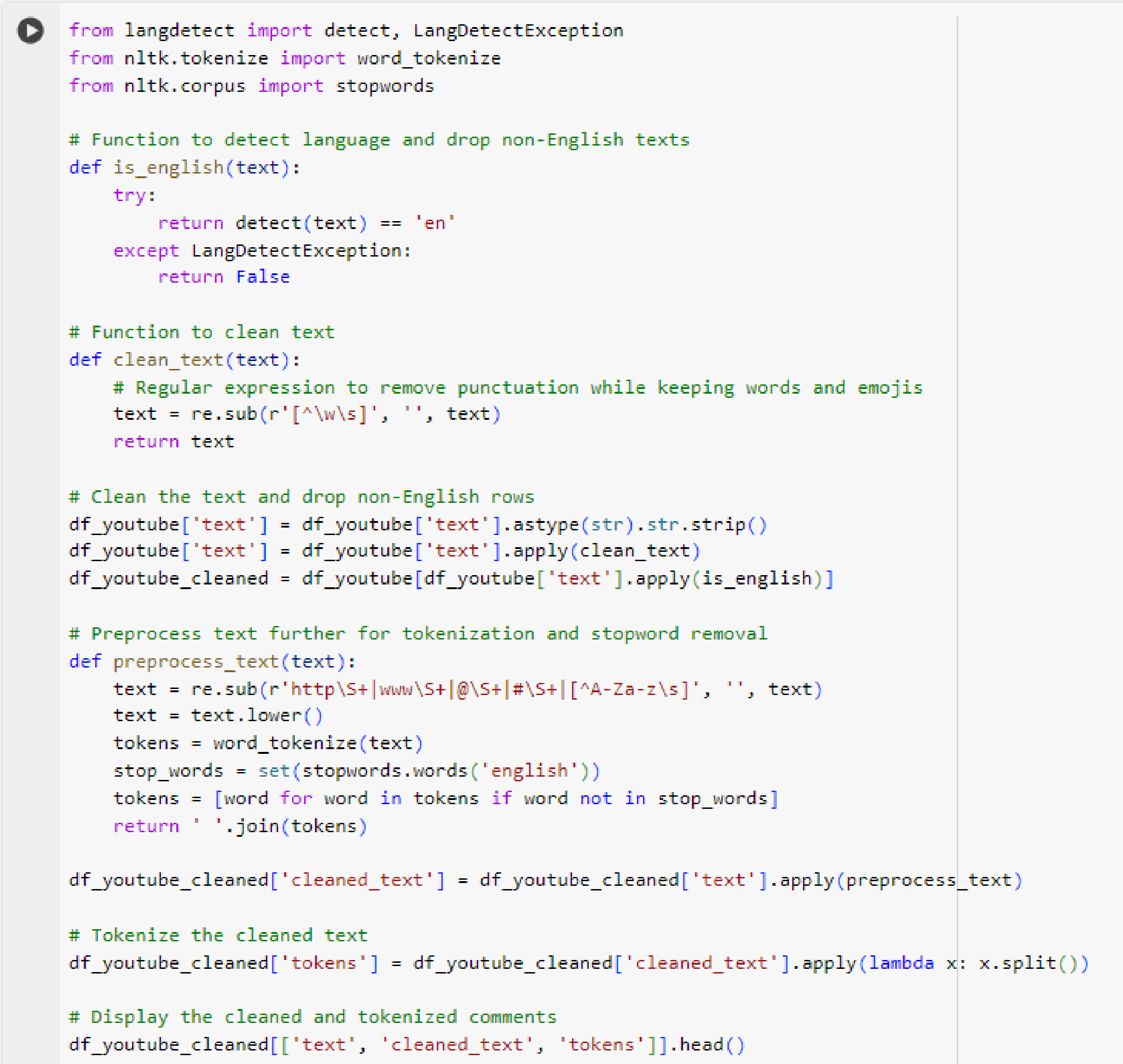
Text Cleaning and Preprocessing Figure 23 illustrates the process of cleaning the text data, including removing non-English comments and unwanted characters, preparing the data for analysis.

Tokenization and Frequency Analysis Figure 24 demonstrates the tokenization process, where the cleaned text is split into individual words (tokens), followed by frequency analysis to identify common terms.

Theme and Keyword Analysis Figure 25 shows the identification of key themes and associated keywords from the comments, providing insights into the main topics discussed.

Sentiment Analysis using BERTweet Figure 26 illustrates the use of the BERTweet model to analyze the sentiment of the comments, classifying them into positive, negative, or neutral categories.

Sentiment Analysis using Twitter-RoBERTa Figure 27 shows the code used for sentiment analysis using the Twitter-RoBERTa model, which provides another perspective on the sentiment expressed in the YouTube comments.



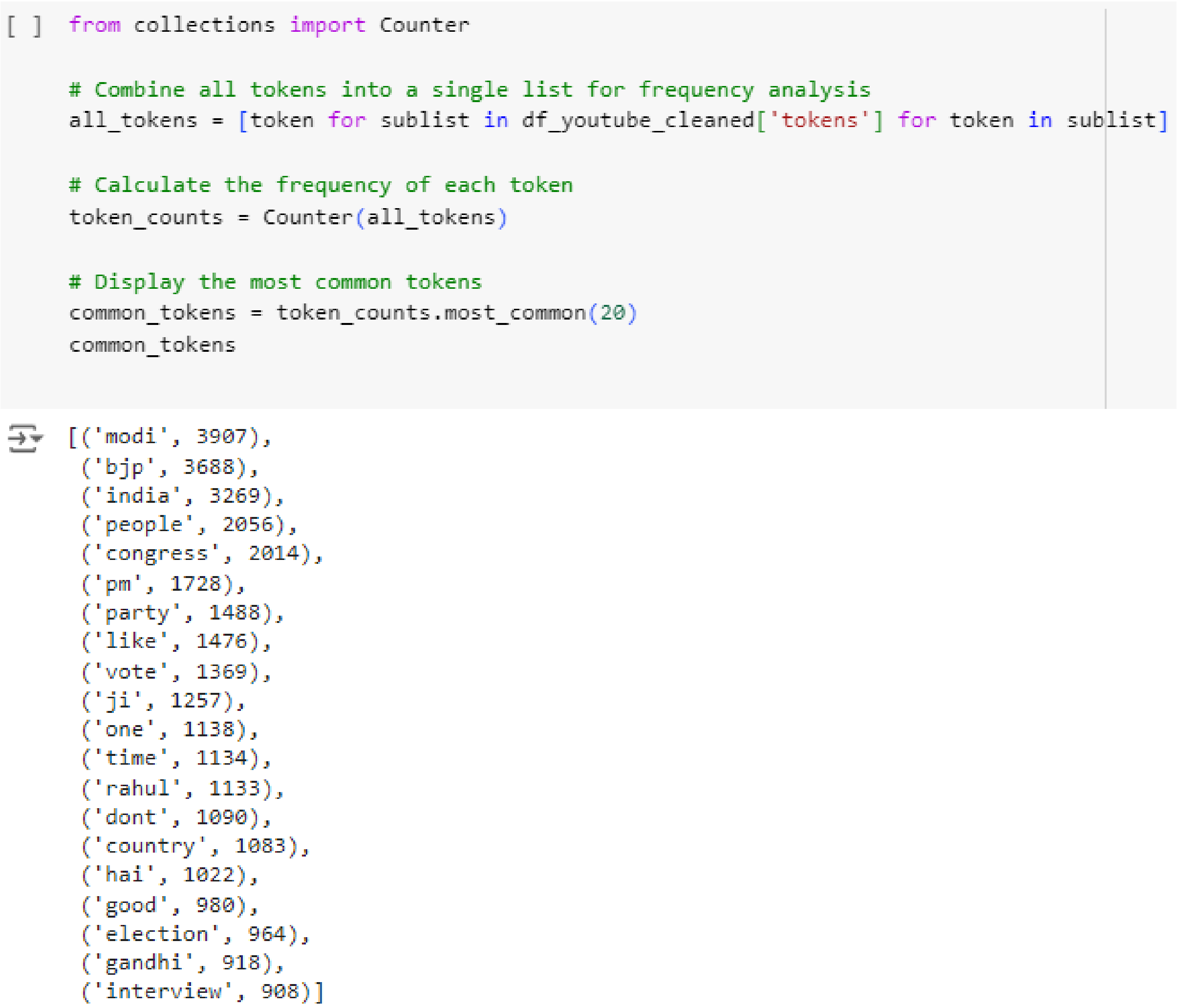
```
from langdetect import detect, LangDetectException
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords

# Function to detect language and drop non-English texts
def is_english(text):
    try:
        return detect(text) == 'en'
    except LangDetectException:
        return False

# Function to clean text
def clean_text(text):
    # Regular expression to remove punctuation while keeping words and emojis
    text = re.sub(r'[^\w\s]', '', text)
    return text

# Clean the text and drop non-English rows
df_youtube['text'] = df_youtube['text'].astype(str).str.strip()
df_youtube['text'] = df_youtube['text'].apply(clean_text)
df_youtube_cleaned = df_youtube[df_youtube['text'].apply(is_english)]

# Preprocess text further for tokenization and stopword removal
def preprocess_text(text):
    text = re.sub(r'http\S+|www\S+|@\S+|#\S+|[^A-Za-z\s]', '', text)
    text = text.lower()
    tokens = word_tokenize(text)
    stop_words = set(stopwords.words('english'))
    tokens = [word for word in tokens if word not in stop_words]
    return ' '.join(tokens)

df_youtube_cleaned['cleaned_text'] = df_youtube_cleaned['text'].apply(preprocess_text)

# Tokenize the cleaned text
df_youtube_cleaned['tokens'] = df_youtube_cleaned['cleaned_text'].apply(lambda x: x.split())

# Display the cleaned and tokenized comments
df_youtube_cleaned[['text', 'cleaned_text', 'tokens']].head()
```

Figure 23: Text Cleaning and Preprocessing

Sentiment Analysis using RoBERTa-Large Figure 28 shows the code used for sentiment analysis using the RoBERTa-Large model, which analyzes the sentiments expressed in the YouTube comments.

Sentiment Analysis using Multiple Models Figure 29 illustrates the implementation of sentiment analysis using multiple models, providing a comparative analysis of different approaches.

Evaluation Metrics for Sentiment Models Figure 30 presents the code used to evaluate the sentiment models, including metrics such as accuracy, precision, recall, and



```
[ ]  from collections import Counter

     # Combine all tokens into a single list for frequency analysis
     all_tokens = [token for sublist in df_youtube_cleaned['tokens'] for token in sublist]

     # Calculate the frequency of each token
     token_counts = Counter(all_tokens)

     # Display the most common tokens
     common_tokens = token_counts.most_common(20)
     common_tokens
```

```
[('modi', 3907),
 ('bjp', 3688),
 ('india', 3269),
 ('people', 2056),
 ('congress', 2014),
 ('pm', 1728),
 ('party', 1488),
 ('like', 1476),
 ('vote', 1369),
 ('ji', 1257),
 ('one', 1138),
 ('time', 1134),
 ('rahul', 1133),
 ('dont', 1090),
 ('country', 1083),
 ('hai', 1022),
 ('good', 980),
 ('election', 964),
 ('gandhi', 918),
 ('interview', 908)]
```

Figure 24: Tokenization and Frequency Analysis

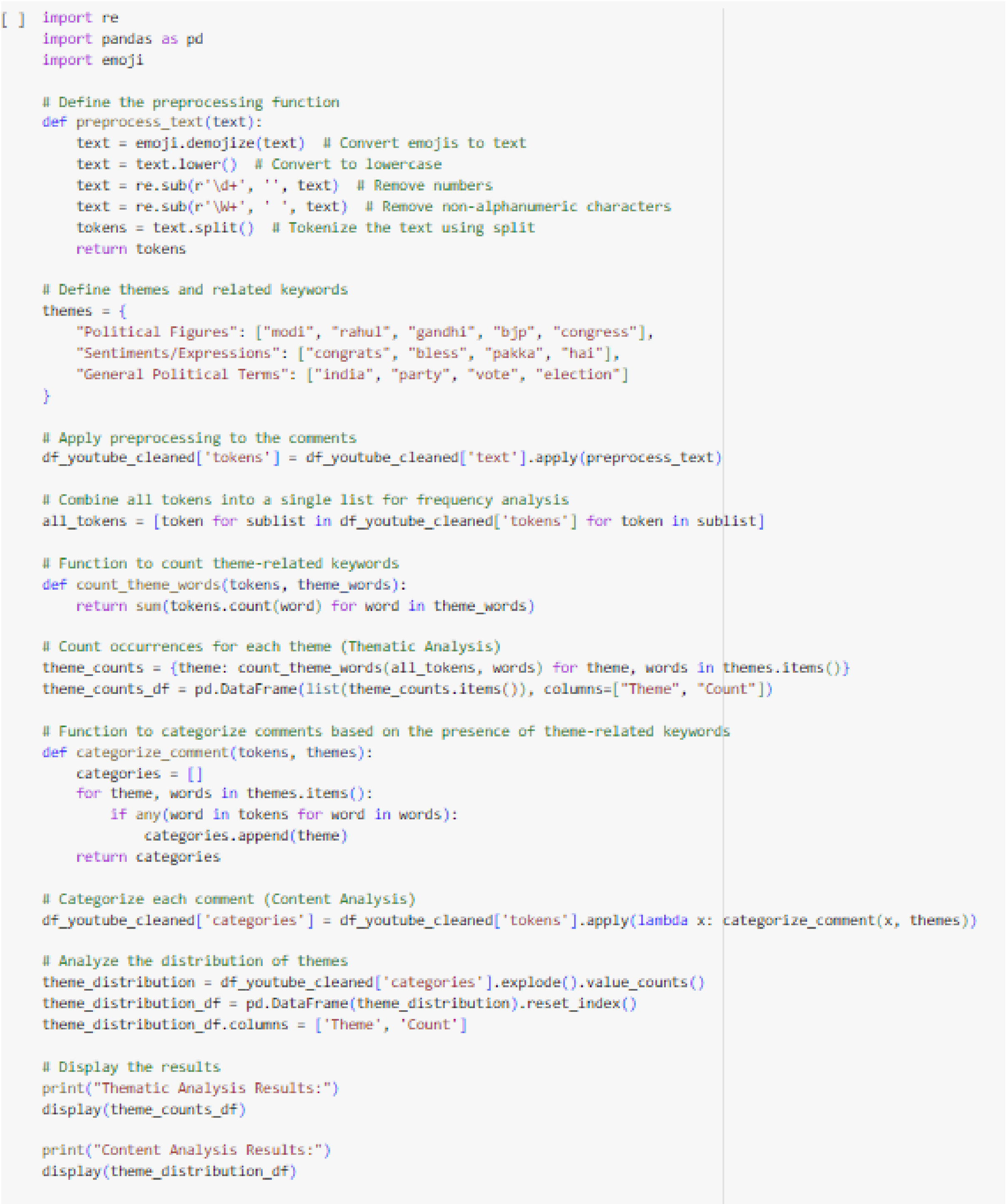
F1 score.

BERTopic Clustering Figure 32 displays the clustering of comments using BERTopic, identifying distinct topics within the dataset.

BERTopic Analysis Figure 32 shows the code for clustering the YouTube comments using BERTopic, which groups similar topics together for analysis.

Displaying Topics Figure 33 shows the code used to display the identified topics from the BERTopic model.

Filtering Topics by Keywords Figure 34 presents the code for filtering topics based on specific keywords related to BJP and Congress.

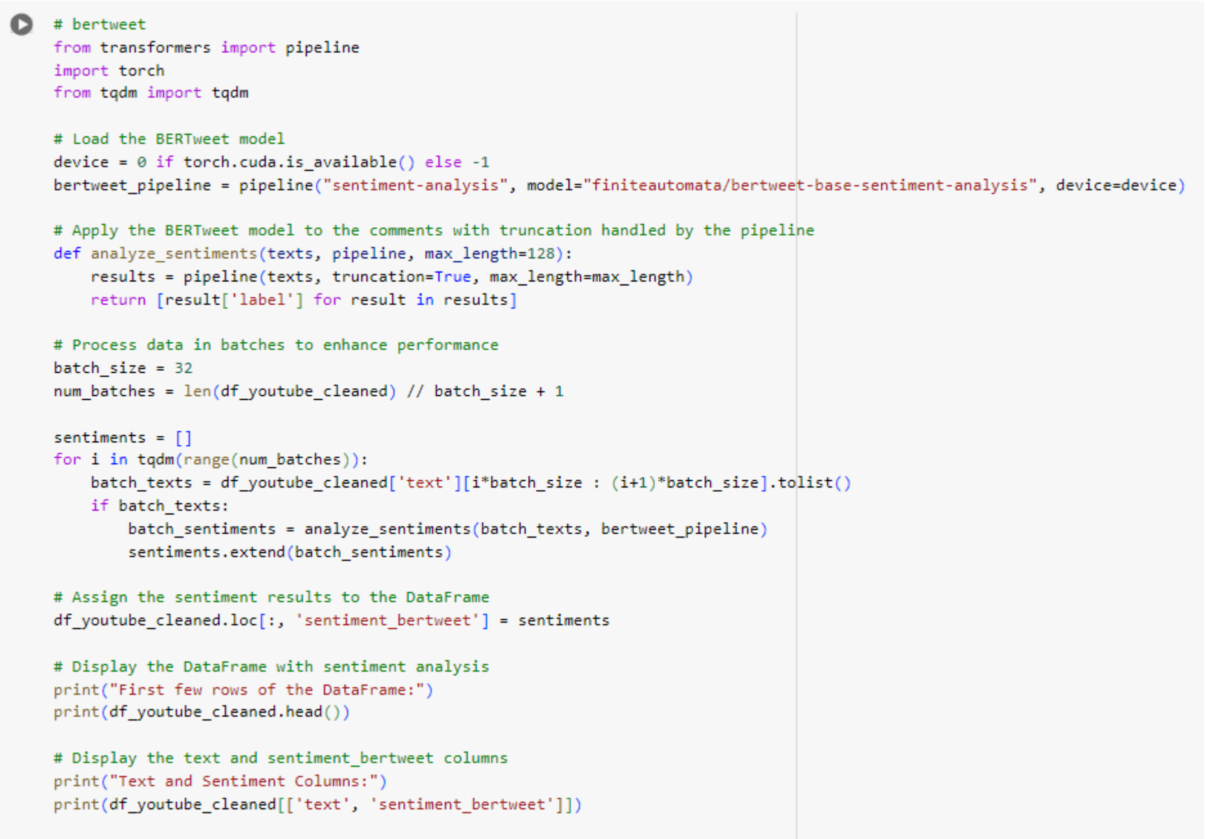


```python
import re
import pandas as pd
import emoji

# Define the preprocessing function
def preprocess_text(text):
    text = emoji.demojize(text)  # Convert emojis to text
    text = text.lower()  # Convert to lowercase
    text = re.sub(r'\d+', '', text)  # Remove numbers
    text = re.sub(r'\W+', ' ', text)  # Remove non-alphanumeric characters
    tokens = text.split()  # Tokenize the text using split
    return tokens

# Define themes and related keywords
themes = {
    "Political Figures": ["modi", "rahul", "gandhi", "bjp", "congress"],
    "Sentiments/Expressions": ["congrats", "bless", "pakka", "hai"],
    "General Political Terms": ["india", "party", "vote", "election"]
}

# Apply preprocessing to the comments
df_youtube_cleaned['tokens'] = df_youtube_cleaned['text'].apply(preprocess_text)

# Combine all tokens into a single list for frequency analysis
all_tokens = [token for sublist in df_youtube_cleaned['tokens'] for token in sublist]

# Function to count theme-related keywords
def count_theme_words(tokens, theme_words):
    return sum(tokens.count(word) for word in theme_words)

# Count occurrences for each theme (Thematic Analysis)
theme_counts = {theme: count_theme_words(all_tokens, words) for theme, words in themes.items()}
theme_counts_df = pd.DataFrame(list(theme_counts.items()), columns=["Theme", "Count"])

# Function to categorize comments based on the presence of theme-related keywords
def categorize_comment(tokens, themes):
    categories = []
    for theme, words in themes.items():
        if any(word in tokens for word in words):
            categories.append(theme)
    return categories

# Categorize each comment (Content Analysis)
df_youtube_cleaned['categories'] = df_youtube_cleaned['tokens'].apply(lambda x: categorize_comment(x, themes))

# Analyze the distribution of themes
theme_distribution = df_youtube_cleaned['categories'].explode().value_counts()
theme_distribution_df = pd.DataFrame(theme_distribution).reset_index()
theme_distribution_df.columns = ['Theme', 'Count']

# Display the results
print("Thematic Analysis Results:")
display(theme_counts_df)

print("Content Analysis Results:")
display(theme_distribution_df)
```

Figure 25: Theme and Keyword Analysis

Topic Distribution for BJP and Congress Figure 35 shows the code for plotting the distribution of topics related to BJP and Congress.

Printing Topic Names Figure 36 presents the code for printing the names of the topics identified for BJP and Congress.



```
# bertweet
from transformers import pipeline
import torch
from tqdm import tqdm

# Load the BERTweet model
device = 0 if torch.cuda.is_available() else -1
bertweet_pipeline = pipeline("sentiment-analysis", model="finiteautomata/bertweet-base-sentiment-analysis", device=device)

# Apply the BERTweet model to the comments with truncation handled by the pipeline
def analyze_sentiments(texts, pipeline, max_length=128):
    results = pipeline(texts, truncation=True, max_length=max_length)
    return [result['label'] for result in results]

# Process data in batches to enhance performance
batch_size = 32
num_batches = len(df_youtube_cleaned) // batch_size + 1

sentiments = []
for i in tqdm(range(num_batches)):
    batch_texts = df_youtube_cleaned['text'][i*batch_size : (i+1)*batch_size].tolist()
    if batch_texts:
        batch_sentiments = analyze_sentiments(batch_texts, bertweet_pipeline)
        sentiments.extend(batch_sentiments)

# Assign the sentiment results to the DataFrame
df_youtube_cleaned.loc[:, 'sentiment_bertweet'] = sentiments

# Display the DataFrame with sentiment analysis
print("First few rows of the DataFrame:")
print(df_youtube_cleaned.head())

# Display the text and sentiment_bertweet columns
print("Text and Sentiment Columns:")
print(df_youtube_cleaned[['text', 'sentiment_bertweet']])
```

Figure 26: Sentiment Analysis using BERTweet

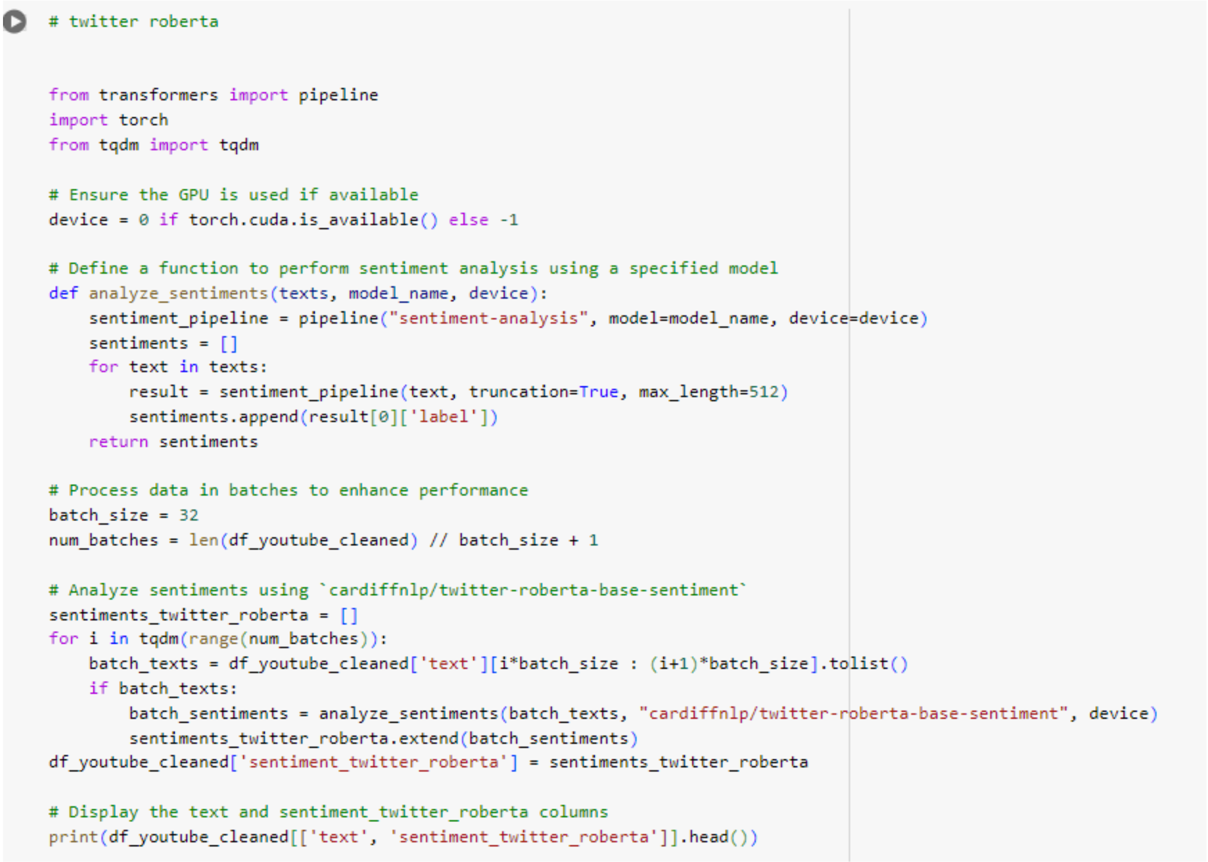
Mapping Topics to Themes Figure 37 shows the code used to map the identified topics to broader themes, providing insights into the main issues discussed.

Plotting Themes for BJP and Congress Figure 38 displays the code for plotting the number of topics mapped to each theme for BJP and Congress.

Filtering and Calculating Sentiment Distribution by Party Figure 39 shows the process of filtering comments by party and calculating the sentiment distribution for BJP and Congress.

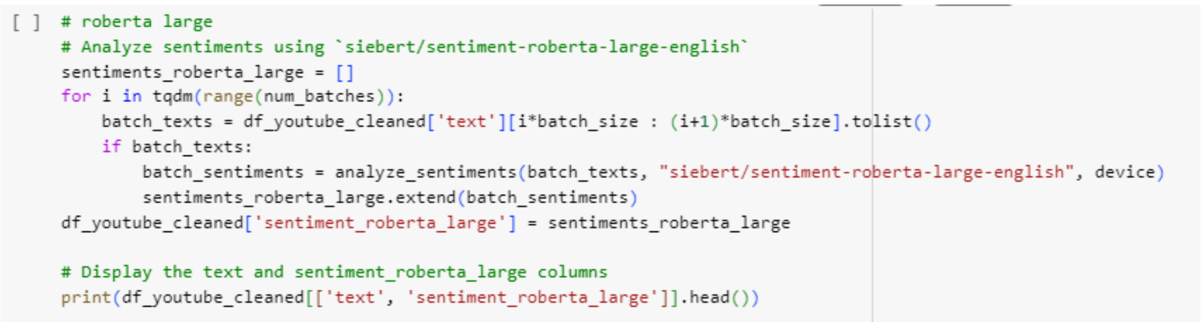
Plotting Sentiment Distribution for BJP and Congress Figure 40 illustrates the sentiment distribution across different time periods for both BJP and Congress, showing trends in public opinion.

Difference in Sentiment Distribution between BJP and Congress Figure 41 highlights the differences in sentiment distribution between BJP and Congress, offering a



```
# twitter roberta


from transformers import pipeline
import torch
from tqdm import tqdm

# Ensure the GPU is used if available
device = 0 if torch.cuda.is_available() else -1

# Define a function to perform sentiment analysis using a specified model
def analyze_sentiments(texts, model_name, device):
    sentiment_pipeline = pipeline("sentiment-analysis", model=model_name, device=device)
    sentiments = []
    for text in texts:
        result = sentiment_pipeline(text, truncation=True, max_length=512)
        sentiments.append(result[0]['label'])
    return sentiments

# Process data in batches to enhance performance
batch_size = 32
num_batches = len(df_youtube_cleaned) // batch_size + 1

# Analyze sentiments using `cardiffnlp/twitter-roberta-base-sentiment`
sentiments_twitter_roberta = []
for i in tqdm(range(num_batches)):
    batch_texts = df_youtube_cleaned['text'][i*batch_size : (i+1)*batch_size].tolist()
    if batch_texts:
        batch_sentiments = analyze_sentiments(batch_texts, "cardiffnlp/twitter-roberta-base-sentiment", device)
        sentiments_twitter_roberta.extend(batch_sentiments)
df_youtube_cleaned['sentiment_twitter_roberta'] = sentiments_twitter_roberta

# Display the text and sentiment_twitter_roberta columns
print(df_youtube_cleaned[['text', 'sentiment_twitter_roberta']].head())
```

Figure 27: Code for Sentiment Analysis using Twitter-RoBERTa



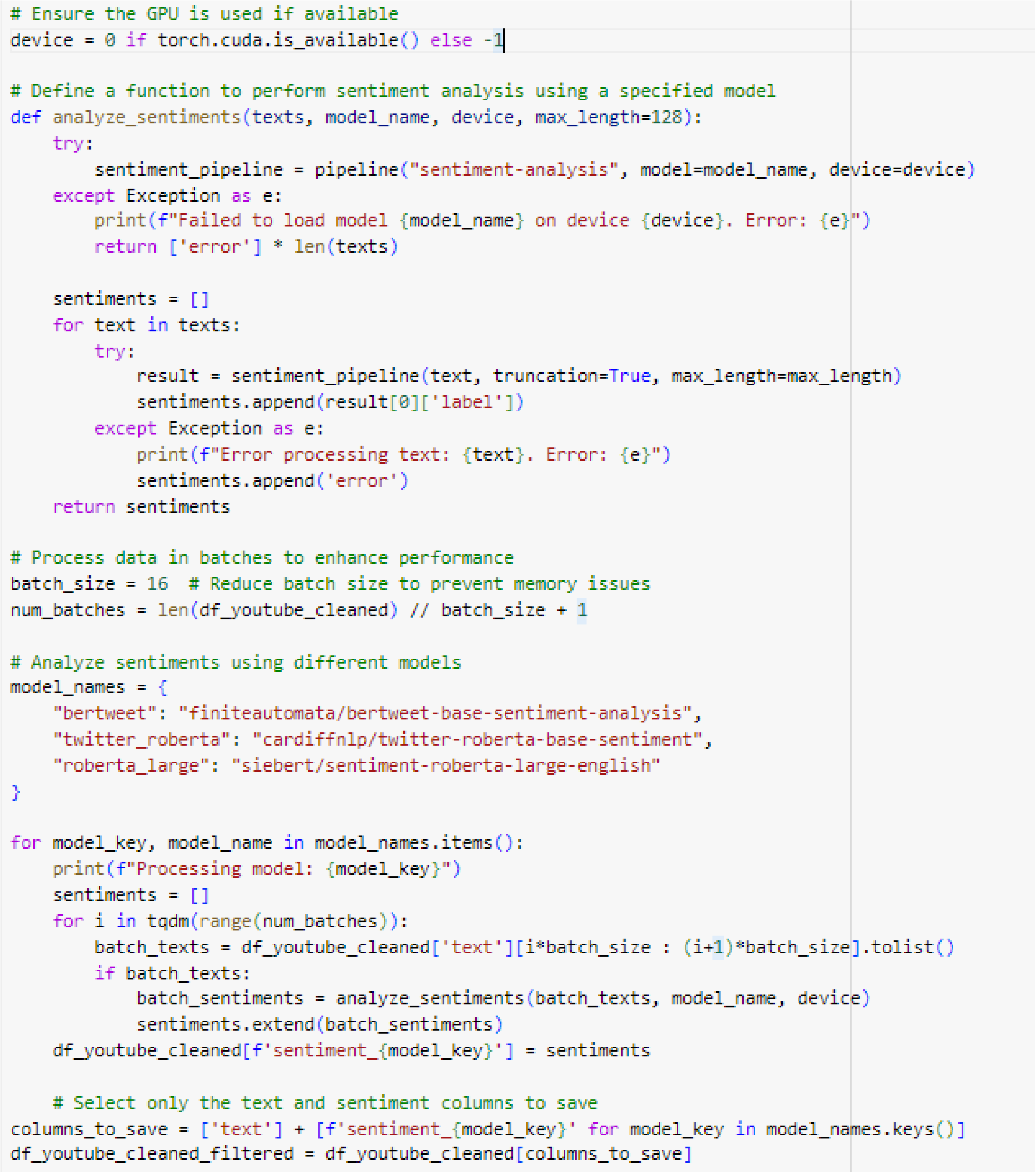
```
# roberta large
# Analyze sentiments using `siebert/sentiment-roberta-large-english`
sentiments_roberta_large = []
for i in tqdm(range(num_batches)):
    batch_texts = df_youtube_cleaned['text'][i*batch_size : (i+1)*batch_size].tolist()
    if batch_texts:
        batch_sentiments = analyze_sentiments(batch_texts, "siebert/sentiment-roberta-large-english", device)
        sentiments_roberta_large.extend(batch_sentiments)
df_youtube_cleaned['sentiment_roberta_large'] = sentiments_roberta_large

# Display the text and sentiment_roberta_large columns
print(df_youtube_cleaned[['text', 'sentiment_roberta_large']].head())
```

Figure 28: Code for Sentiment Analysis using RoBERTa-Large

comparative analysis of public sentiment towards the two parties.

Zero-Shot and BERTweet Sentiment Analysis Figure 42 shows the code for performing sentiment analysis using both Zero-Shot Classification and BERTweet models, enabling a comparative study of these methods.

Comparing Sentiment Results: Zero-Shot vs. BERTweet Figure 43 presents the code used to compare the sentiment analysis results between Zero-Shot Classification

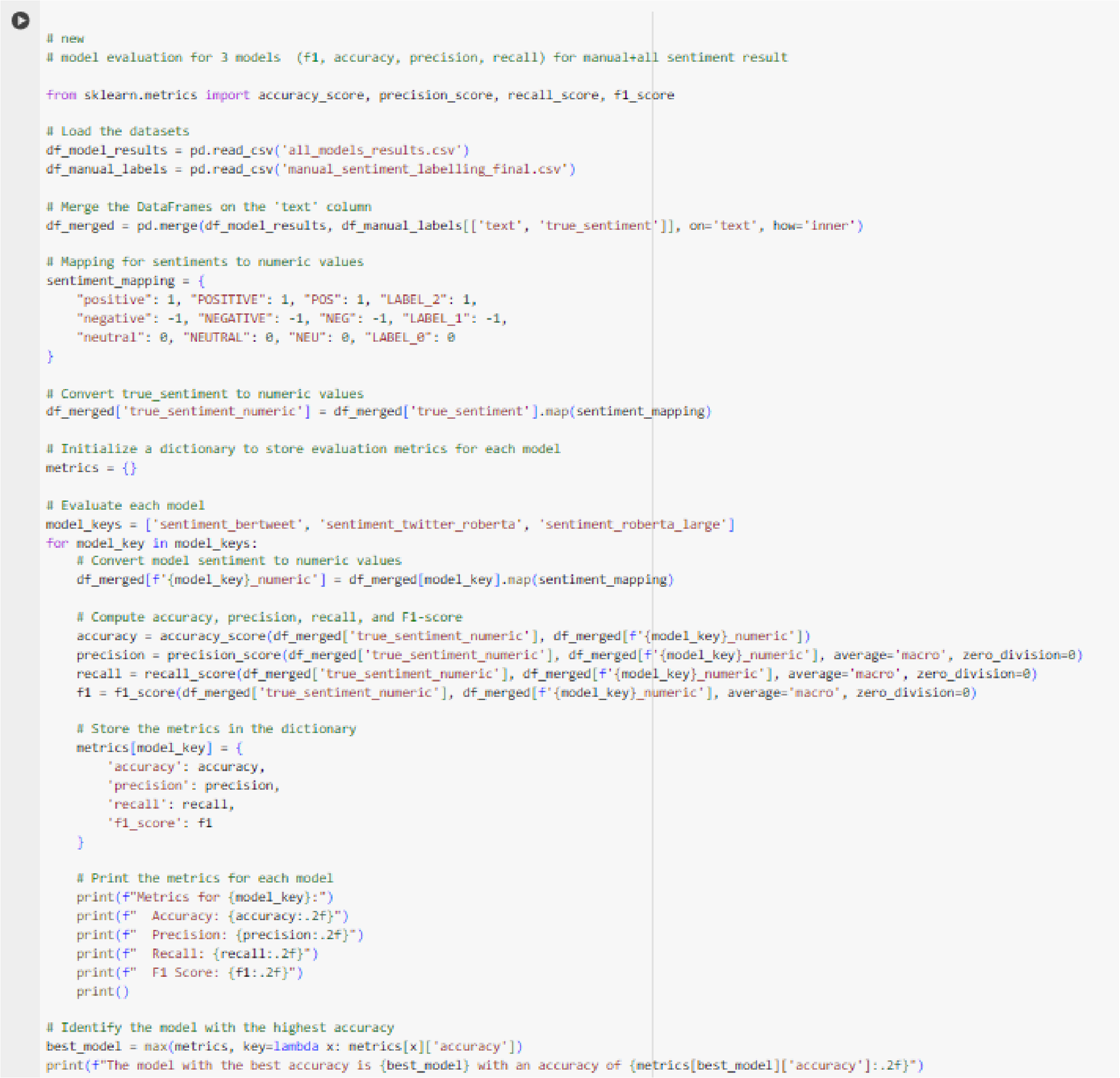


```python
# Ensure the GPU is used if available
device = 0 if torch.cuda.is_available() else -1

# Define a function to perform sentiment analysis using a specified model
def analyze_sentiments(texts, model_name, device, max_length=128):
    try:
        sentiment_pipeline = pipeline("sentiment-analysis", model=model_name, device=device)
    except Exception as e:
        print(f"Failed to load model {model_name} on device {device}. Error: {e}")
        return ['error'] * len(texts)

    sentiments = []
    for text in texts:
        try:
            result = sentiment_pipeline(text, truncation=True, max_length=max_length)
            sentiments.append(result[0]['label'])
        except Exception as e:
            print(f"Error processing text: {text}. Error: {e}")
            sentiments.append('error')
    return sentiments

# Process data in batches to enhance performance
batch_size = 16  # Reduce batch size to prevent memory issues
num_batches = len(df_youtube_cleaned) // batch_size + 1

# Analyze sentiments using different models
model_names = {
    "bertweet": "finiteautomata/bertweet-base-sentiment-analysis",
    "twitter_roberta": "cardiffnlp/twitter-roberta-base-sentiment",
    "roberta_large": "siebert/sentiment-roberta-large-english"
}

for model_key, model_name in model_names.items():
    print(f"Processing model: {model_key}")
    sentiments = []
    for i in tqdm(range(num_batches)):
        batch_texts = df_youtube_cleaned['text'][i*batch_size : (i+1)*batch_size].tolist()
        if batch_texts:
            batch_sentiments = analyze_sentiments(batch_texts, model_name, device)
            sentiments.extend(batch_sentiments)
    df_youtube_cleaned[f'sentiment_{model_key}'] = sentiments

    # Select only the text and sentiment columns to save
columns_to_save = ['text'] + [f'sentiment_{model_key}' for model_key in model_names.keys()]
df_youtube_cleaned_filtered = df_youtube_cleaned[columns_to_save]
```

Figure 29: Code for Sentiment Analysis using Multiple Models

and BERTweet. The bar chart visually contrasts the sentiment distributions identified by each method.

Weekly Sentiment Analysis for BJP Figure 44 presents the weekly sentiment distribution for BJP, showing how public sentiment fluctuated over time.



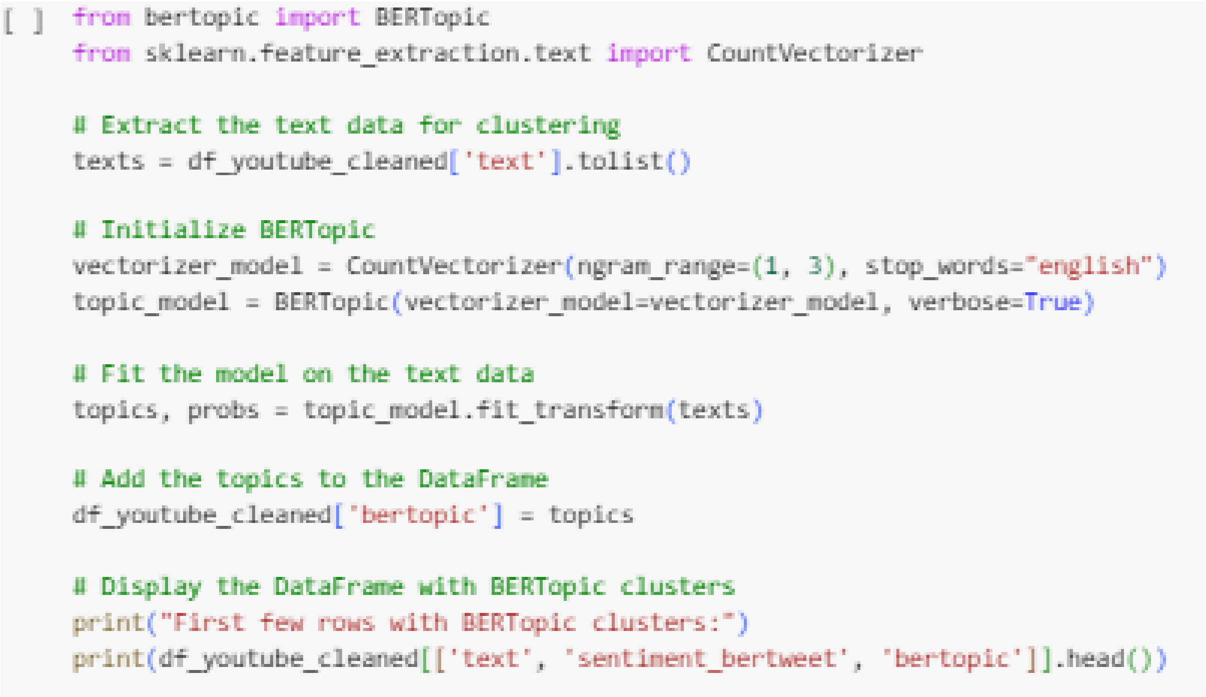
```python
# new
# model evaluation for 3 models  (f1, accuracy, precision, recall) for manual+all sentiment result

from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score

# Load the datasets
df_model_results = pd.read_csv('all_models_results.csv')
df_manual_labels = pd.read_csv('manual_sentiment_labelling_final.csv')

# Merge the DataFrames on the 'text' column
df_merged = pd.merge(df_model_results, df_manual_labels[['text', 'true_sentiment']], on='text', how='inner')

# Mapping for sentiments to numeric values
sentiment_mapping = {
    "positive": 1, "POSITIVE": 1, "POS": 1, "LABEL_2": 1,
    "negative": -1, "NEGATIVE": -1, "NEG": -1, "LABEL_1": -1,
    "neutral": 0, "NEUTRAL": 0, "NEU": 0, "LABEL_0": 0
}

# Convert true_sentiment to numeric values
df_merged['true_sentiment_numeric'] = df_merged['true_sentiment'].map(sentiment_mapping)

# Initialize a dictionary to store evaluation metrics for each model
metrics = {}

# Evaluate each model
model_keys = ['sentiment_bertweet', 'sentiment_twitter_roberta', 'sentiment_roberta_large']
for model_key in model_keys:
    # Convert model sentiment to numeric values
    df_merged[f'{model_key}_numeric'] = df_merged[model_key].map(sentiment_mapping)

    # Compute accuracy, precision, recall, and F1-score
    accuracy = accuracy_score(df_merged['true_sentiment_numeric'], df_merged[f'{model_key}_numeric'])
    precision = precision_score(df_merged['true_sentiment_numeric'], df_merged[f'{model_key}_numeric'], average='macro', zero_division=0)
    recall = recall_score(df_merged['true_sentiment_numeric'], df_merged[f'{model_key}_numeric'], average='macro', zero_division=0)
    f1 = f1_score(df_merged['true_sentiment_numeric'], df_merged[f'{model_key}_numeric'], average='macro', zero_division=0)

    # Store the metrics in the dictionary
    metrics[model_key] = {
        'accuracy': accuracy,
        'precision': precision,
        'recall': recall,
        'f1_score': f1
    }

    # Print the metrics for each model
    print(f"Metrics for {model_key}:")
    print(f"  Accuracy: {accuracy:.2f}")
    print(f"  Precision: {precision:.2f}")
    print(f"  Recall: {recall:.2f}")
    print(f"  F1 Score: {f1:.2f}")
    print()

# Identify the model with the highest accuracy
best_model = max(metrics, key=lambda x: metrics[x]['accuracy'])
print(f"The model with the best accuracy is {best_model} with an accuracy of {metrics[best_model]['accuracy']:.2f}")
```

Figure 30: Code for Evaluating Sentiment Models

Monthly Sentiment Analysis for BJP Figure 45 displays the sentiment distribution on a monthly basis, providing a broader view of trends over time.

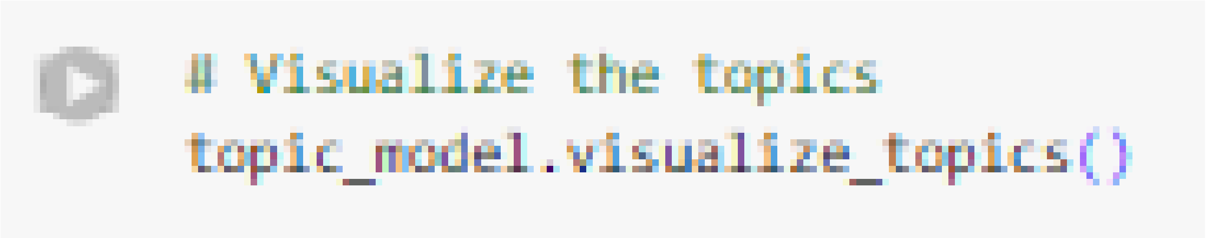
Identifying Peak Months for Sentiment Figure 46 identifies the peak months for both positive and negative sentiment for BJP and Congress, indicating key periods of public sentiment.

Monthly Average Sentiment Figure 47 shows the average sentiment per month for both BJP and Congress, providing insights into the overall public mood over time.



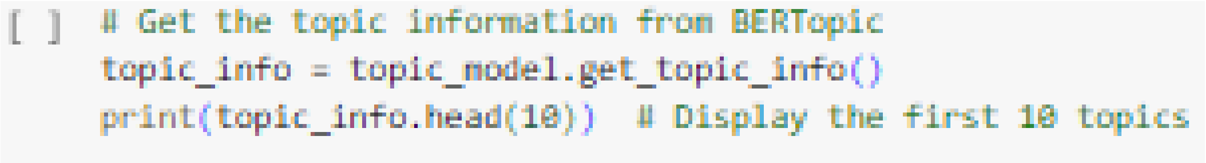
```
[ ]  from bertopic import BERTopic
     from sklearn.feature_extraction.text import CountVectorizer

     # Extract the text data for clustering
     texts = df_youtube_cleaned['text'].tolist()

     # Initialize BERTopic
     vectorizer_model = CountVectorizer(ngram_range=(1, 3), stop_words="english")
     topic_model = BERTopic(vectorizer_model=vectorizer_model, verbose=True)

     # Fit the model on the text data
     topics, probs = topic_model.fit_transform(texts)

     # Add the topics to the DataFrame
     df_youtube_cleaned['bertopic'] = topics

     # Display the DataFrame with BERTopic clusters
     print("First few rows with BERTopic clusters:")
     print(df_youtube_cleaned[['text', 'sentiment_bertweet', 'bertopic']].head())
```

Figure 31: BERTopic Clustering and Analysis



```
# Visualize the topics
topic_model.visualize_topics()
```

Figure 32: Code for Clustering Topics using BERTopic

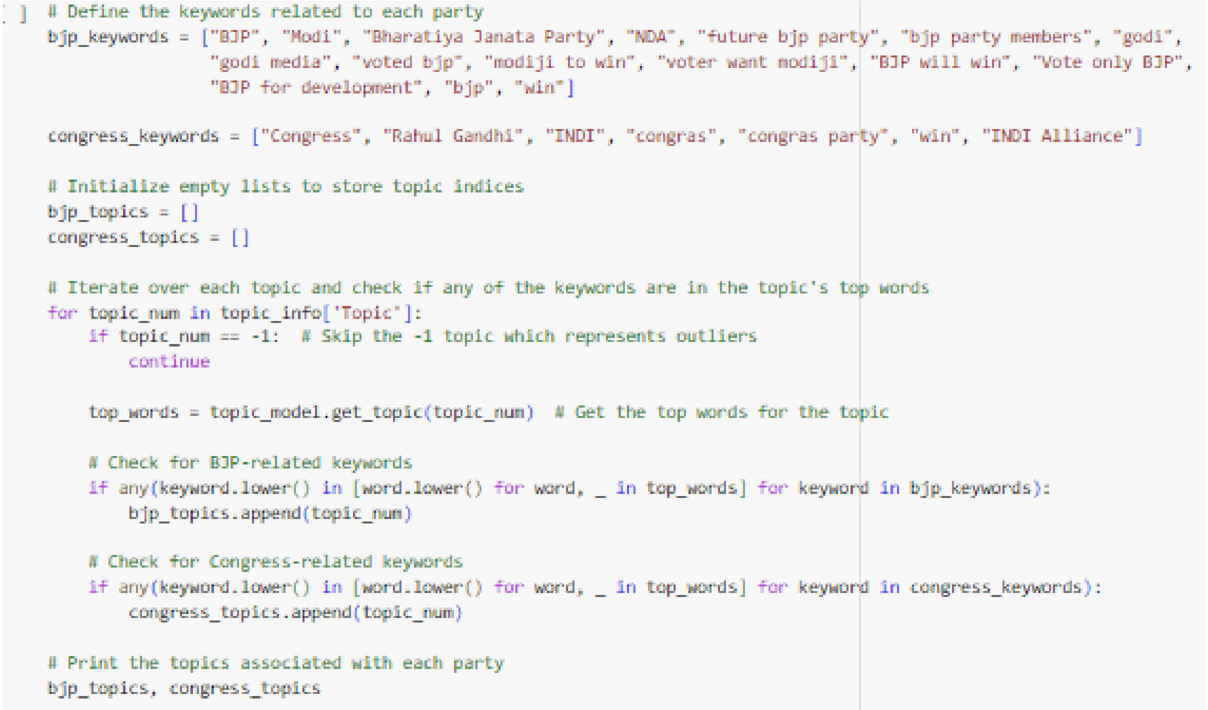


```
[ ]  # Get the topic information from BERTopic
     topic_info = topic_model.get_topic_info()
     print(topic_info.head(10))  # Display the first 10 topics
```

Figure 33: Code for Displaying Topics from BERTopic

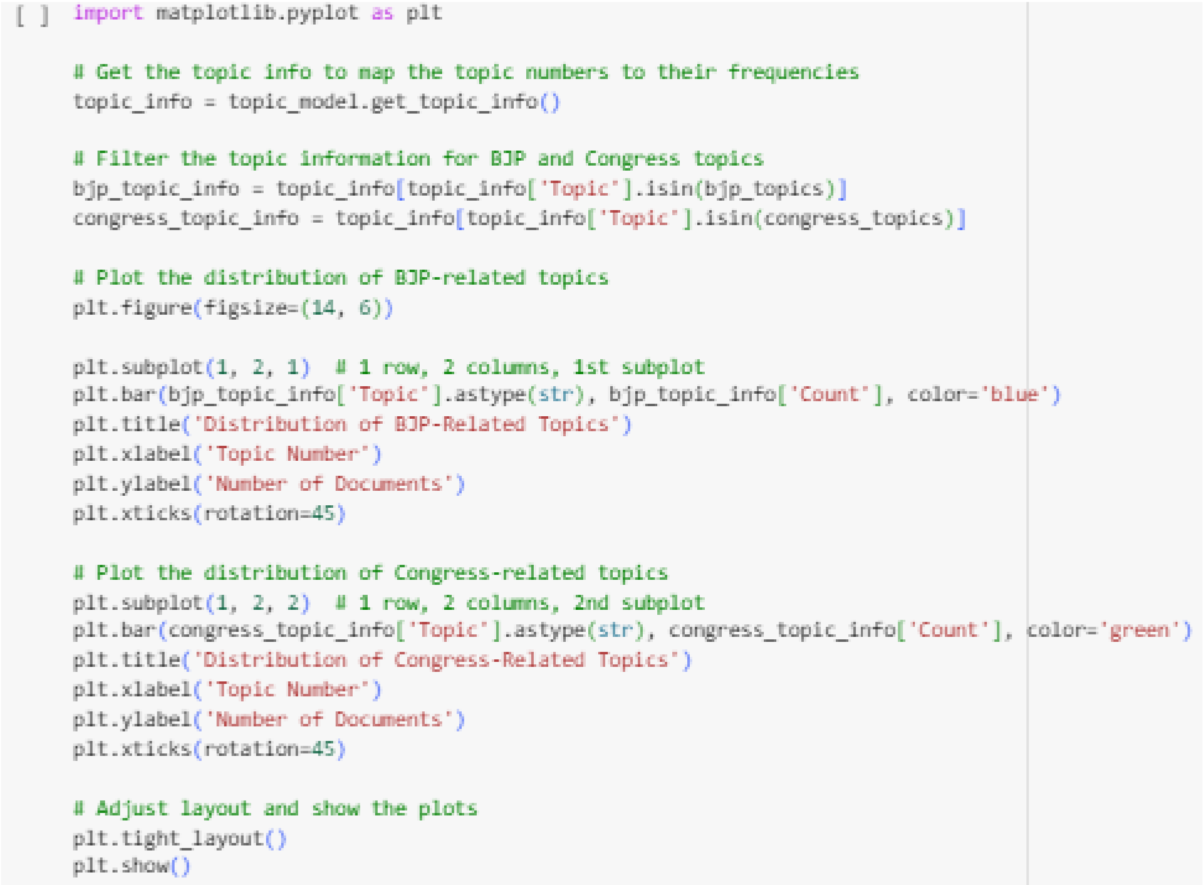
Monthly Average Sentiment with Election Period Highlighted Figure 48 highlights the election period in the context of average monthly sentiment, showing how public opinion shifted during the election time.

Kaggle Model Implementation for Sentiment Analysis Figure 49 demonstrates the implementation of multiple sentiment analysis models on the Kaggle platform, allowing for a comprehensive analysis of public sentiment.



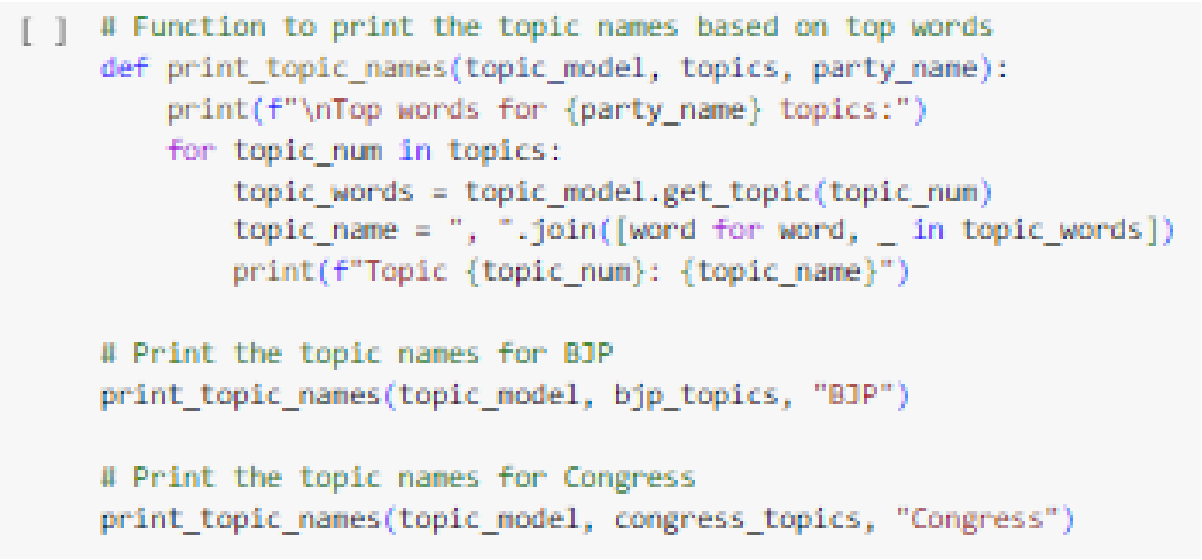
```
[ ] # Define the keywords related to each party
    bjp_keywords = ["BJP", "Modi", "Bharatiya Janata Party", "NDA", "future bjp party", "bjp party members", "godi",
                    "godi media", "voted bjp", "modiji to win", "voter want modiji", "BJP will win", "Vote only BJP",
                    "BJP for development", "bjp", "win"]

    congress_keywords = ["Congress", "Rahul Gandhi", "INDI", "congras", "congras party", "win", "INDI Alliance"]

    # Initialize empty lists to store topic indices
    bjp_topics = []
    congress_topics = []

    # Iterate over each topic and check if any of the keywords are in the topic's top words
    for topic_num in topic_info['Topic']:
        if topic_num == -1:  # Skip the -1 topic which represents outliers
            continue

        top_words = topic_model.get_topic(topic_num)  # Get the top words for the topic

        # Check for BJP-related keywords
        if any(keyword.lower() in [word.lower() for word, _ in top_words] for keyword in bjp_keywords):
            bjp_topics.append(topic_num)

        # Check for Congress-related keywords
        if any(keyword.lower() in [word.lower() for word, _ in top_words] for keyword in congress_keywords):
            congress_topics.append(topic_num)

    # Print the topics associated with each party
    bjp_topics, congress_topics
```

Figure 34: Code for Filtering Topics by Keywords



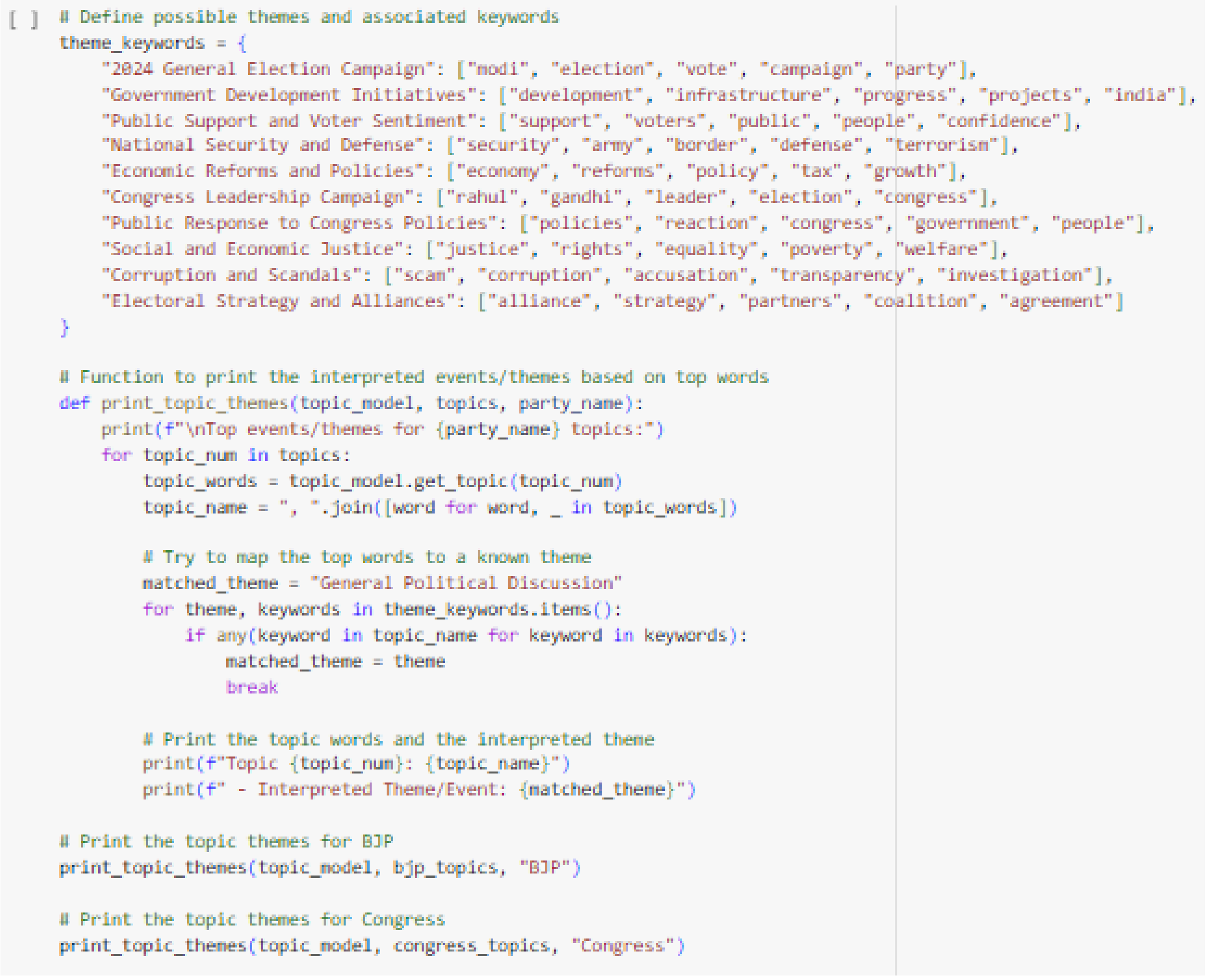
```
[ ] import matplotlib.pyplot as plt

    # Get the topic info to map the topic numbers to their frequencies
    topic_info = topic_model.get_topic_info()

    # Filter the topic information for BJP and Congress topics
    bjp_topic_info = topic_info[topic_info['Topic'].isin(bjp_topics)]
    congress_topic_info = topic_info[topic_info['Topic'].isin(congress_topics)]

    # Plot the distribution of BJP-related topics
    plt.figure(figsize=(14, 6))

    plt.subplot(1, 2, 1)  # 1 row, 2 columns, 1st subplot
    plt.bar(bjp_topic_info['Topic'].astype(str), bjp_topic_info['Count'], color='blue')
    plt.title('Distribution of BJP-Related Topics')
    plt.xlabel('Topic Number')
    plt.ylabel('Number of Documents')
    plt.xticks(rotation=45)

    # Plot the distribution of Congress-related topics
    plt.subplot(1, 2, 2)  # 1 row, 2 columns, 2nd subplot
    plt.bar(congress_topic_info['Topic'].astype(str), congress_topic_info['Count'], color='green')
    plt.title('Distribution of Congress-Related Topics')
    plt.xlabel('Topic Number')
    plt.ylabel('Number of Documents')
    plt.xticks(rotation=45)

    # Adjust layout and show the plots
    plt.tight_layout()
    plt.show()
```

Figure 35: Code for Plotting Topic Distribution for BJP and Congress



```
[ ]  # Function to print the topic names based on top words
     def print_topic_names(topic_model, topics, party_name):
         print(f"\nTop words for {party_name} topics:")
         for topic_num in topics:
             topic_words = topic_model.get_topic(topic_num)
             topic_name = ", ".join([word for word, _ in topic_words])
             print(f"Topic {topic_num}: {topic_name}")

     # Print the topic names for BJP
     print_topic_names(topic_model, bjp_topics, "BJP")

     # Print the topic names for Congress
     print_topic_names(topic_model, congress_topics, "Congress")
```

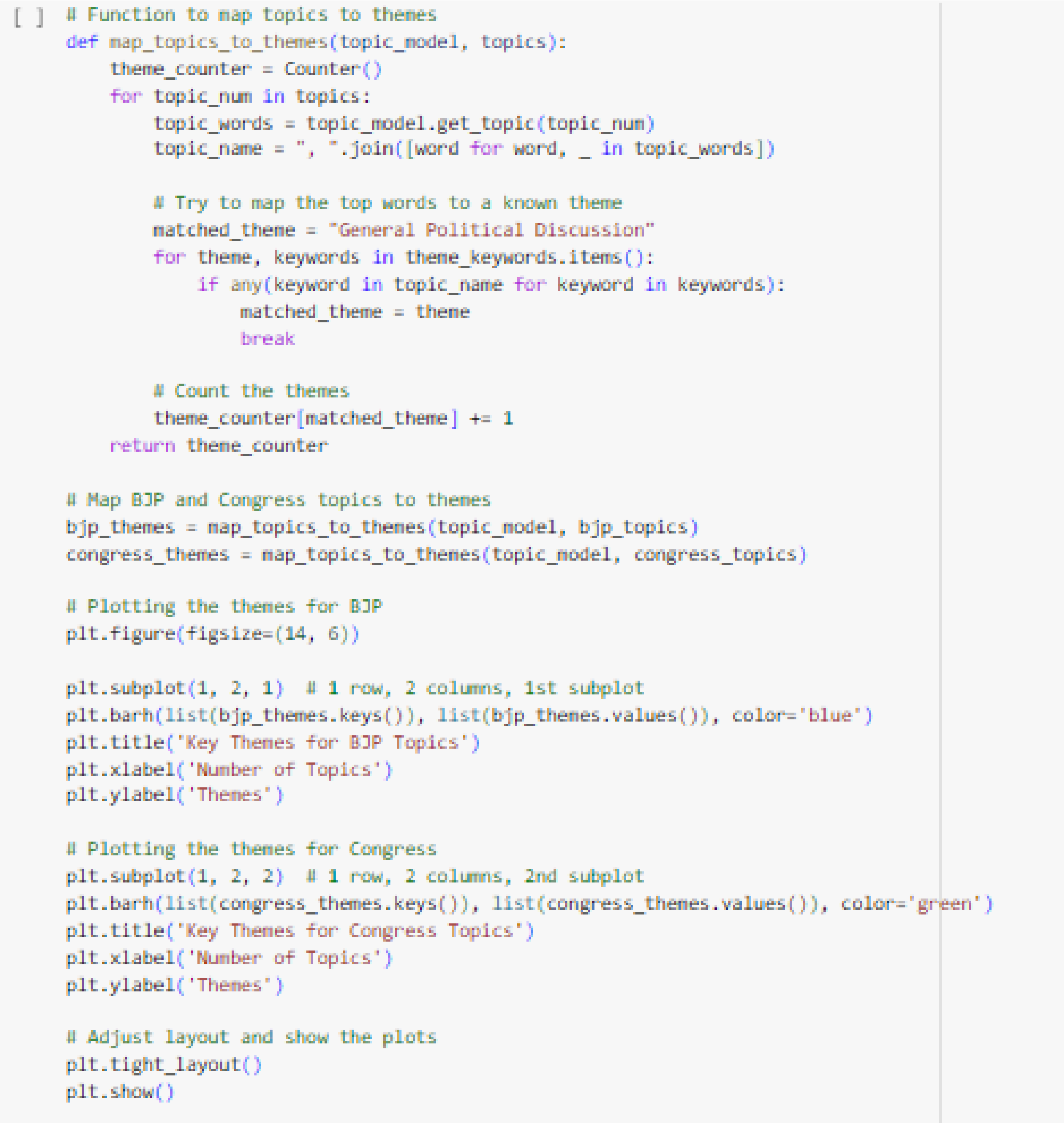
Figure 36: Code for Printing Topic Names



```
[ ]  # Define possible themes and associated keywords
     theme_keywords = {
         "2024 General Election Campaign": ["modi", "election", "vote", "campaign", "party"],
         "Government Development Initiatives": ["development", "infrastructure", "progress", "projects", "india"],
         "Public Support and Voter Sentiment": ["support", "voters", "public", "people", "confidence"],
         "National Security and Defense": ["security", "army", "border", "defense", "terrorism"],
         "Economic Reforms and Policies": ["economy", "reforms", "policy", "tax", "growth"],
         "Congress Leadership Campaign": ["rahul", "gandhi", "leader", "election", "congress"],
         "Public Response to Congress Policies": ["policies", "reaction", "congress", "government", "people"],
         "Social and Economic Justice": ["justice", "rights", "equality", "poverty", "welfare"],
         "Corruption and Scandals": ["scam", "corruption", "accusation", "transparency", "investigation"],
         "Electoral Strategy and Alliances": ["alliance", "strategy", "partners", "coalition", "agreement"]
     }

     # Function to print the interpreted events/themes based on top words
     def print_topic_themes(topic_model, topics, party_name):
         print(f"\nTop events/themes for {party_name} topics:")
         for topic_num in topics:
             topic_words = topic_model.get_topic(topic_num)
             topic_name = ", ".join([word for word, _ in topic_words])

             # Try to map the top words to a known theme
             matched_theme = "General Political Discussion"
             for theme, keywords in theme_keywords.items():
                 if any(keyword in topic_name for keyword in keywords):
                     matched_theme = theme
                     break

             # Print the topic words and the interpreted theme
             print(f"Topic {topic_num}: {topic_name}")
             print(f" - Interpreted Theme/Event: {matched_theme}")

     # Print the topic themes for BJP
     print_topic_themes(topic_model, bjp_topics, "BJP")

     # Print the topic themes for Congress
     print_topic_themes(topic_model, congress_topics, "Congress")
```

Figure 37: Code for Mapping Topics to Themes

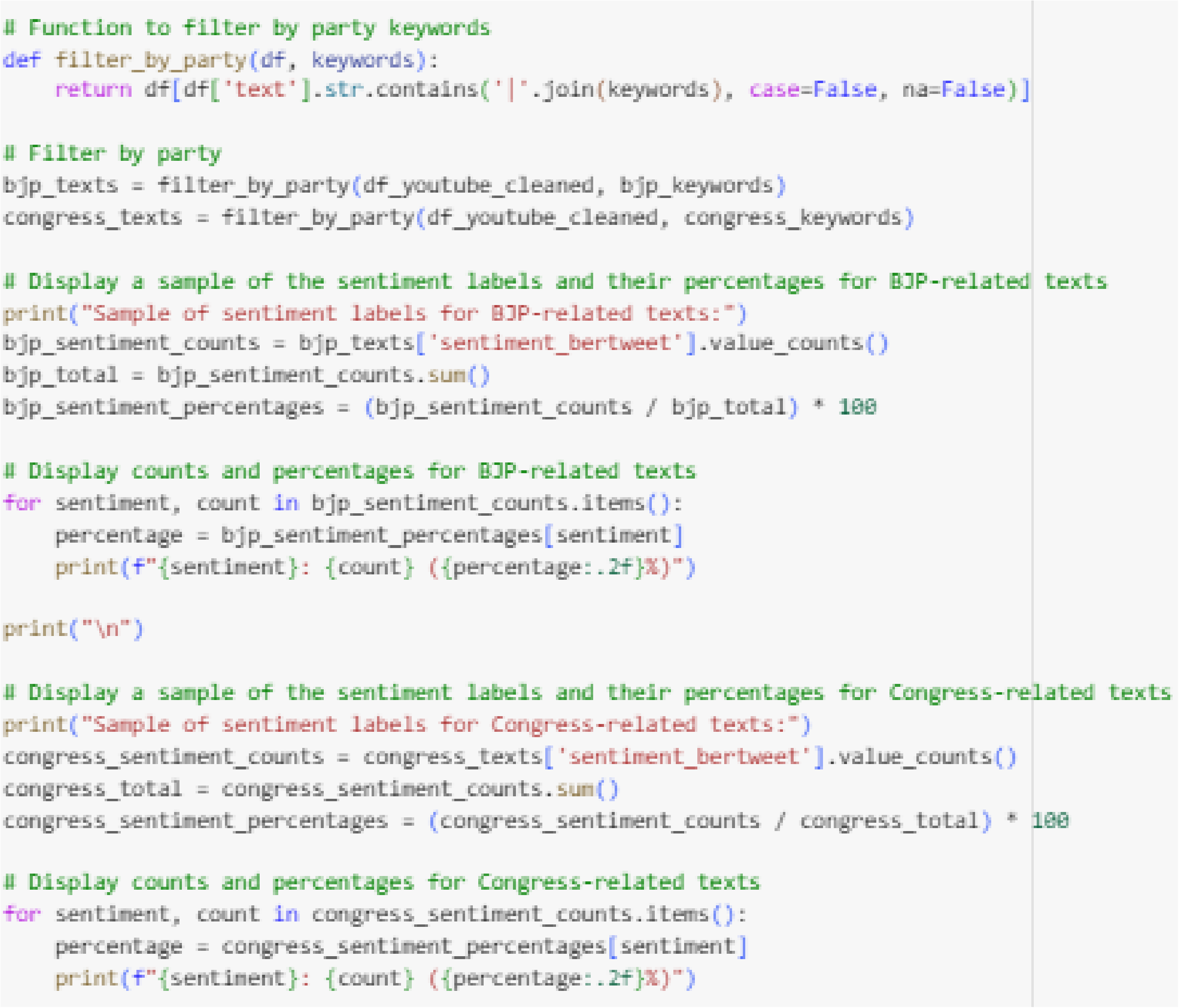
Accuracy Evaluation of Sentiment Models Figure 50 compares the accuracy of different sentiment analysis models, identifying the best-performing model for analyzing



```python
# Function to map topics to themes
def map_topics_to_themes(topic_model, topics):
    theme_counter = Counter()
    for topic_num in topics:
        topic_words = topic_model.get_topic(topic_num)
        topic_name = ", ".join([word for word, _ in topic_words])

        # Try to map the top words to a known theme
        matched_theme = "General Political Discussion"
        for theme, keywords in theme_keywords.items():
            if any(keyword in topic_name for keyword in keywords):
                matched_theme = theme
                break

        # Count the themes
        theme_counter[matched_theme] += 1
    return theme_counter

# Map BJP and Congress topics to themes
bjp_themes = map_topics_to_themes(topic_model, bjp_topics)
congress_themes = map_topics_to_themes(topic_model, congress_topics)

# Plotting the themes for BJP
plt.figure(figsize=(14, 6))

plt.subplot(1, 2, 1)  # 1 row, 2 columns, 1st subplot
plt.barh(list(bjp_themes.keys()), list(bjp_themes.values()), color='blue')
plt.title('Key Themes for BJP Topics')
plt.xlabel('Number of Topics')
plt.ylabel('Themes')

# Plotting the themes for Congress
plt.subplot(1, 2, 2)  # 1 row, 2 columns, 2nd subplot
plt.barh(list(congress_themes.keys()), list(congress_themes.values()), color='green')
plt.title('Key Themes for Congress Topics')
plt.xlabel('Number of Topics')
plt.ylabel('Themes')

# Adjust layout and show the plots
plt.tight_layout()
plt.show()
```

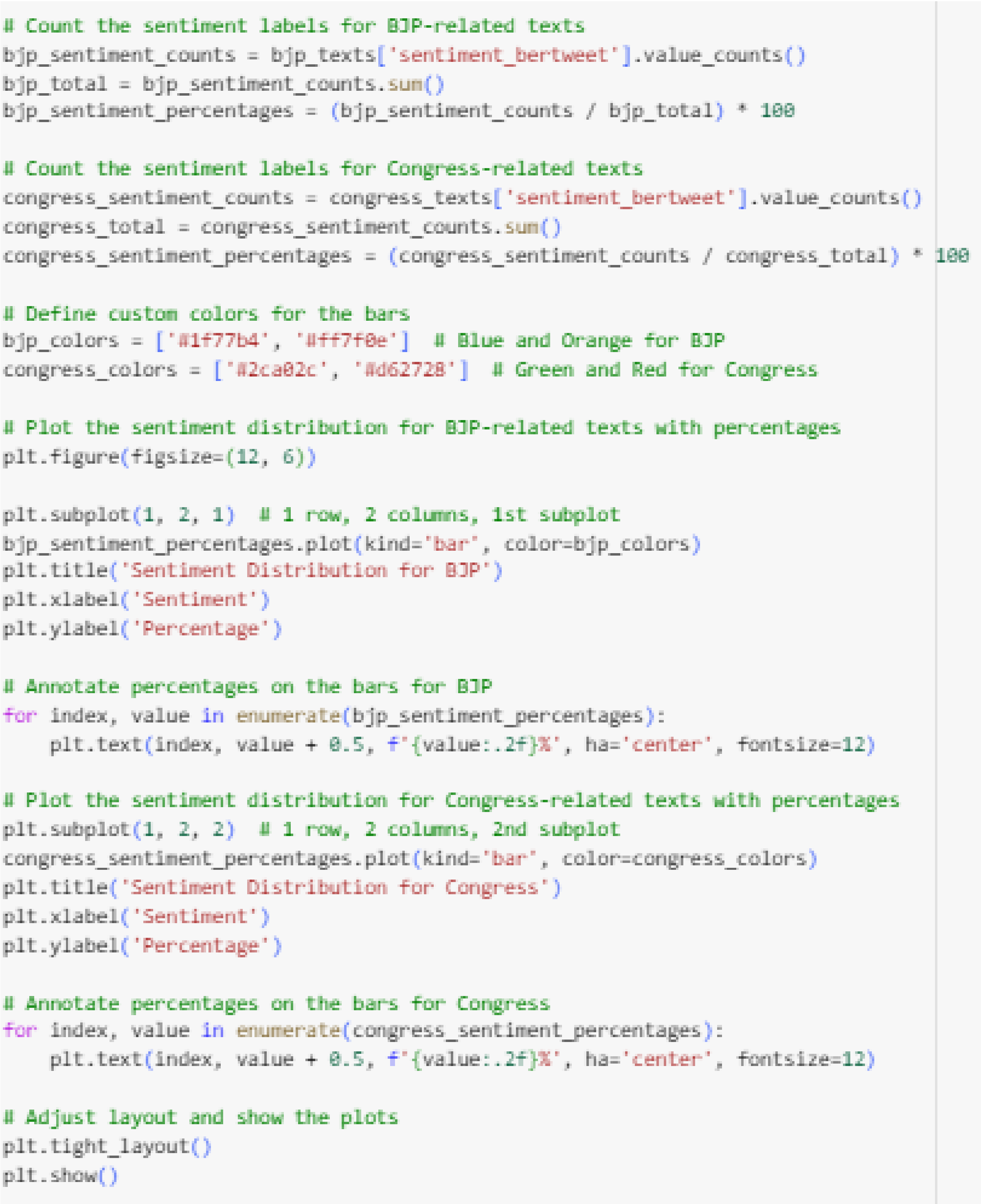
Figure 38: Code for Plotting Themes for BJP and Congress

public sentiment.



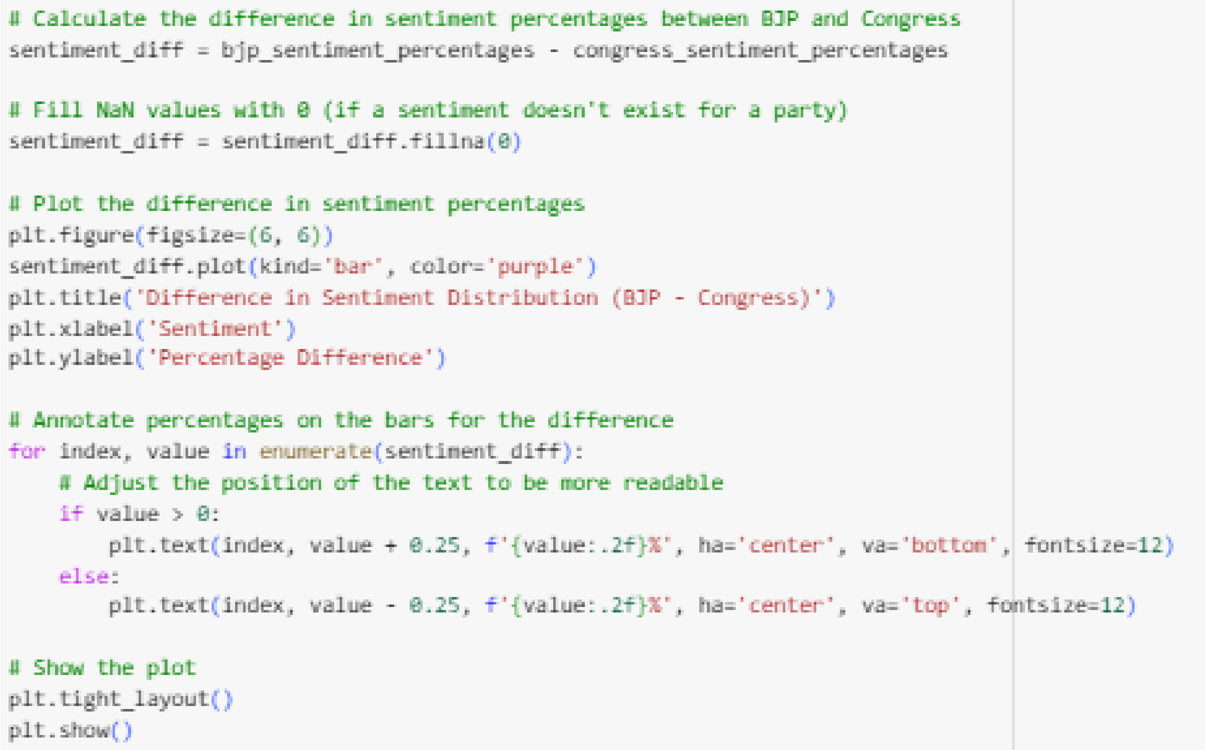
```
# Function to filter by party keywords
def filter_by_party(df, keywords):
    return df[df['text'].str.contains('|'.join(keywords), case=False, na=False)]

# Filter by party
bjp_texts = filter_by_party(df_youtube_cleaned, bjp_keywords)
congress_texts = filter_by_party(df_youtube_cleaned, congress_keywords)

# Display a sample of the sentiment labels and their percentages for BJP-related texts
print("Sample of sentiment labels for BJP-related texts:")
bjp_sentiment_counts = bjp_texts['sentiment_bertweet'].value_counts()
bjp_total = bjp_sentiment_counts.sum()
bjp_sentiment_percentages = (bjp_sentiment_counts / bjp_total) * 100

# Display counts and percentages for BJP-related texts
for sentiment, count in bjp_sentiment_counts.items():
    percentage = bjp_sentiment_percentages[sentiment]
    print(f"{sentiment}: {count} ({percentage:.2f}%)")

print("\n")

# Display a sample of the sentiment labels and their percentages for Congress-related texts
print("Sample of sentiment labels for Congress-related texts:")
congress_sentiment_counts = congress_texts['sentiment_bertweet'].value_counts()
congress_total = congress_sentiment_counts.sum()
congress_sentiment_percentages = (congress_sentiment_counts / congress_total) * 100

# Display counts and percentages for Congress-related texts
for sentiment, count in congress_sentiment_counts.items():
    percentage = congress_sentiment_percentages[sentiment]
    print(f"{sentiment}: {count} ({percentage:.2f}%)")
```

Figure 39: Filtering and Calculating Sentiment Distribution by Party



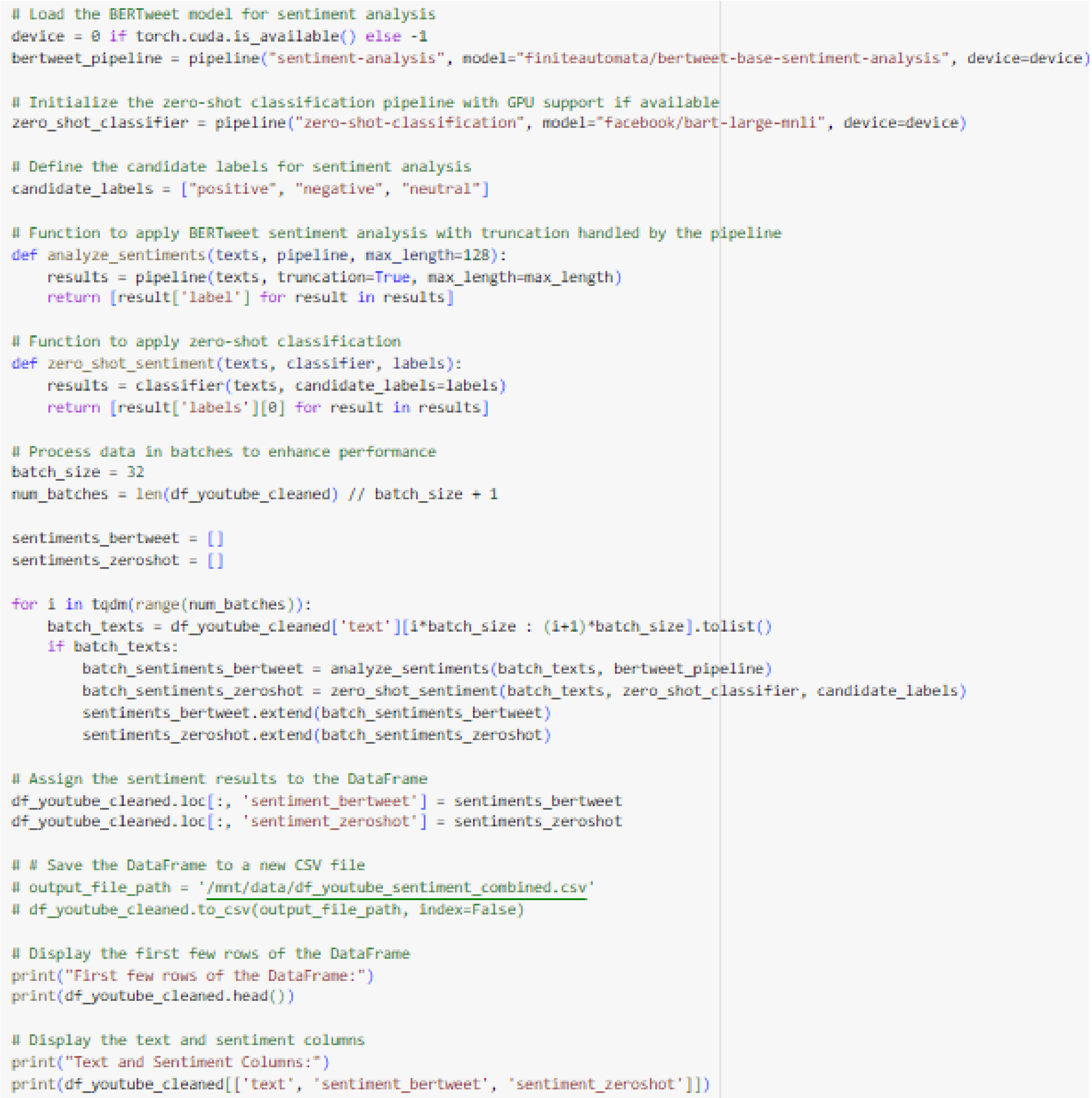
```
# Count the sentiment labels for BJP-related texts
bjp_sentiment_counts = bjp_texts['sentiment_bertweet'].value_counts()
bjp_total = bjp_sentiment_counts.sum()
bjp_sentiment_percentages = (bjp_sentiment_counts / bjp_total) * 100

# Count the sentiment labels for Congress-related texts
congress_sentiment_counts = congress_texts['sentiment_bertweet'].value_counts()
congress_total = congress_sentiment_counts.sum()
congress_sentiment_percentages = (congress_sentiment_counts / congress_total) * 100

# Define custom colors for the bars
bjp_colors = ['#1f77b4', '#ff7f0e']  # Blue and Orange for BJP
congress_colors = ['#2ca02c', '#d62728']  # Green and Red for Congress

# Plot the sentiment distribution for BJP-related texts with percentages
plt.figure(figsize=(12, 6))

plt.subplot(1, 2, 1)  # 1 row, 2 columns, 1st subplot
bjp_sentiment_percentages.plot(kind='bar', color=bjp_colors)
plt.title('Sentiment Distribution for BJP')
plt.xlabel('Sentiment')
plt.ylabel('Percentage')

# Annotate percentages on the bars for BJP
for index, value in enumerate(bjp_sentiment_percentages):
    plt.text(index, value + 0.5, f'{value:.2f}%', ha='center', fontsize=12)

# Plot the sentiment distribution for Congress-related texts with percentages
plt.subplot(1, 2, 2)  # 1 row, 2 columns, 2nd subplot
congress_sentiment_percentages.plot(kind='bar', color=congress_colors)
plt.title('Sentiment Distribution for Congress')
plt.xlabel('Sentiment')
plt.ylabel('Percentage')

# Annotate percentages on the bars for Congress
for index, value in enumerate(congress_sentiment_percentages):
    plt.text(index, value + 0.5, f'{value:.2f}%', ha='center', fontsize=12)

# Adjust layout and show the plots
plt.tight_layout()
plt.show()
```

Figure 40: Plotting Sentiment Distribution for BJP and Congress



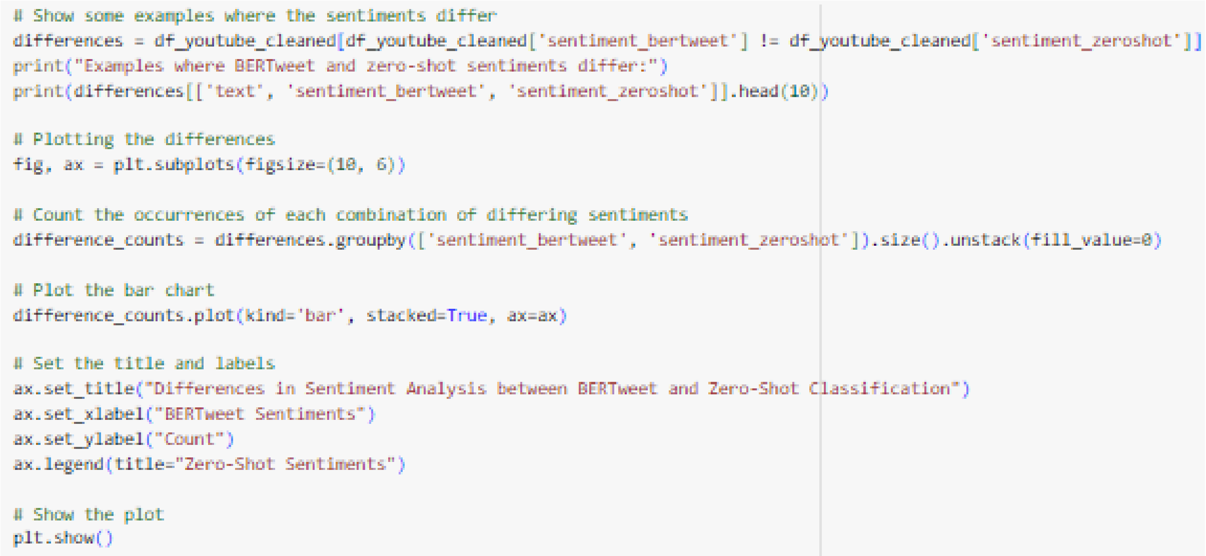
```
# Calculate the difference in sentiment percentages between BJP and Congress
sentiment_diff = bjp_sentiment_percentages - congress_sentiment_percentages

# Fill NaN values with 0 (if a sentiment doesn't exist for a party)
sentiment_diff = sentiment_diff.fillna(0)

# Plot the difference in sentiment percentages
plt.figure(figsize=(6, 6))
sentiment_diff.plot(kind='bar', color='purple')
plt.title('Difference in Sentiment Distribution (BJP - Congress)')
plt.xlabel('Sentiment')
plt.ylabel('Percentage Difference')

# Annotate percentages on the bars for the difference
for index, value in enumerate(sentiment_diff):
    # Adjust the position of the text to be more readable
    if value > 0:
        plt.text(index, value + 0.25, f'{value:.2f}%', ha='center', va='bottom', fontsize=12)
    else:
        plt.text(index, value - 0.25, f'{value:.2f}%', ha='center', va='top', fontsize=12)

# Show the plot
plt.tight_layout()
plt.show()
```

Figure 41: Difference in Sentiment Distribution between BJP and Congress



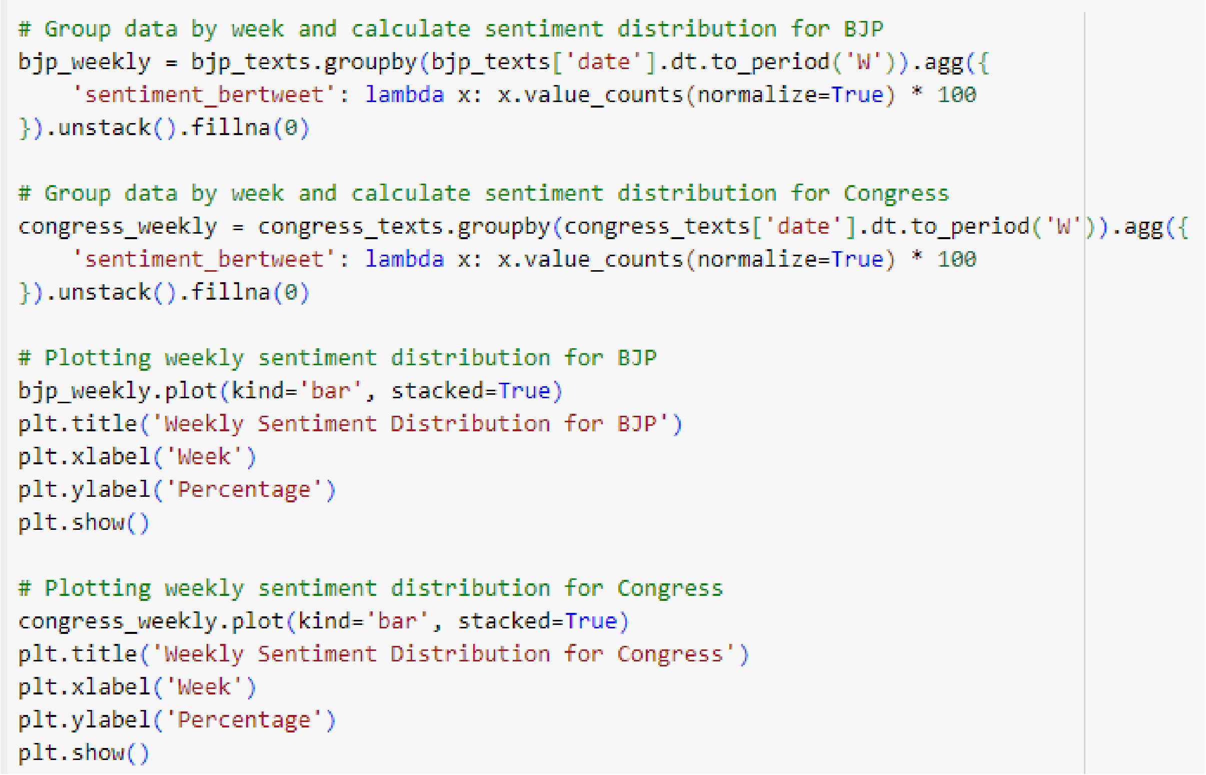
```
# Load the BERTweet model for sentiment analysis
device = 0 if torch.cuda.is_available() else -1
bertweet_pipeline = pipeline("sentiment-analysis", model="finiteautomata/bertweet-base-sentiment-analysis", device=device)

# Initialize the zero-shot classification pipeline with GPU support if available
zero_shot_classifier = pipeline("zero-shot-classification", model="facebook/bart-large-mnli", device=device)

# Define the candidate labels for sentiment analysis
candidate_labels = ["positive", "negative", "neutral"]

# Function to apply BERTweet sentiment analysis with truncation handled by the pipeline
def analyze_sentiments(texts, pipeline, max_length=128):
    results = pipeline(texts, truncation=True, max_length=max_length)
    return [result['label'] for result in results]

# Function to apply zero-shot classification
def zero_shot_sentiment(texts, classifier, labels):
    results = classifier(texts, candidate_labels=labels)
    return [result['labels'][0] for result in results]

# Process data in batches to enhance performance
batch_size = 32
num_batches = len(df_youtube_cleaned) // batch_size + 1

sentiments_bertweet = []
sentiments_zeroshot = []

for i in tqdm(range(num_batches)):
    batch_texts = df_youtube_cleaned['text'][i*batch_size : (i+1)*batch_size].tolist()
    if batch_texts:
        batch_sentiments_bertweet = analyze_sentiments(batch_texts, bertweet_pipeline)
        batch_sentiments_zeroshot = zero_shot_sentiment(batch_texts, zero_shot_classifier, candidate_labels)
        sentiments_bertweet.extend(batch_sentiments_bertweet)
        sentiments_zeroshot.extend(batch_sentiments_zeroshot)

# Assign the sentiment results to the DataFrame
df_youtube_cleaned.loc[:, 'sentiment_bertweet'] = sentiments_bertweet
df_youtube_cleaned.loc[:, 'sentiment_zeroshot'] = sentiments_zeroshot

# # Save the DataFrame to a new CSV file
# output_file_path = '/mnt/data/df_youtube_sentiment_combined.csv'
# df_youtube_cleaned.to_csv(output_file_path, index=False)

# Display the first few rows of the DataFrame
print("First few rows of the DataFrame:")
print(df_youtube_cleaned.head())

# Display the text and sentiment columns
print("Text and Sentiment Columns:")
print(df_youtube_cleaned[['text', 'sentiment_bertweet', 'sentiment_zeroshot']])
```

Figure 42: Code for Zero-Shot and BERTweet Sentiment Analysis



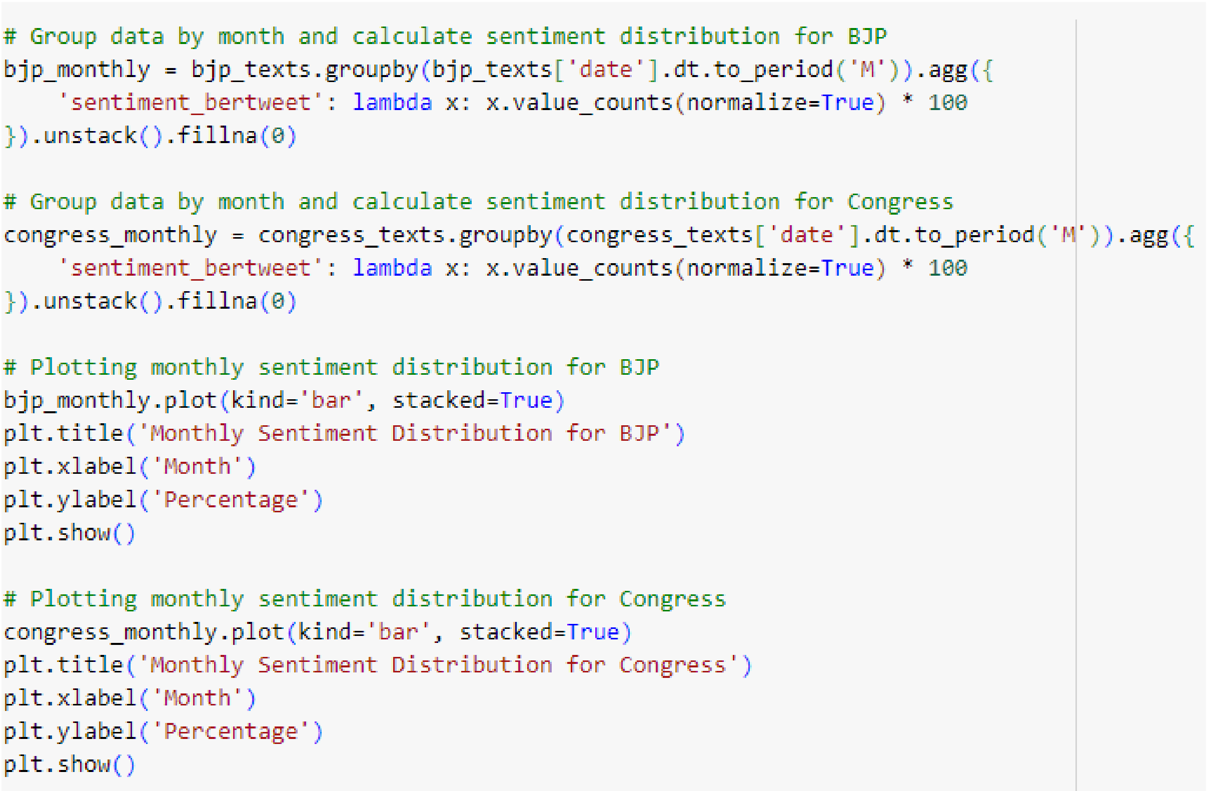
```
# Show some examples where the sentiments differ
differences = df_youtube_cleaned[df_youtube_cleaned['sentiment_bertweet'] != df_youtube_cleaned['sentiment_zeroshot']]
print("Examples where BERTweet and zero-shot sentiments differ:")
print(differences[['text', 'sentiment_bertweet', 'sentiment_zeroshot']].head(10))

# Plotting the differences
fig, ax = plt.subplots(figsize=(10, 6))

# Count the occurrences of each combination of differing sentiments
difference_counts = differences.groupby(['sentiment_bertweet', 'sentiment_zeroshot']).size().unstack(fill_value=0)

# Plot the bar chart
difference_counts.plot(kind='bar', stacked=True, ax=ax)

# Set the title and labels
ax.set_title("Differences in Sentiment Analysis between BERTweet and Zero-Shot Classification")
ax.set_xlabel("BERTweet Sentiments")
ax.set_ylabel("Count")
ax.legend(title="Zero-Shot Sentiments")

# Show the plot
plt.show()
```

Figure 43: Code for Comparing Sentiment Results: Zero-Shot vs. BERTweet



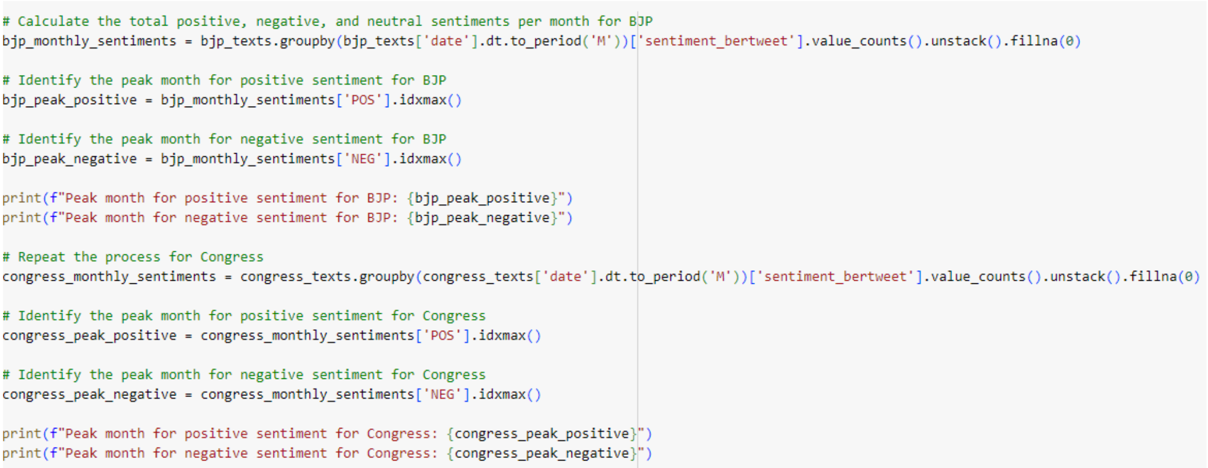
```
# Group data by week and calculate sentiment distribution for BJP
bjp_weekly = bjp_texts.groupby(bjp_texts['date'].dt.to_period('W')).agg({
    'sentiment_bertweet': lambda x: x.value_counts(normalize=True) * 100
}).unstack().fillna(0)

# Group data by week and calculate sentiment distribution for Congress
congress_weekly = congress_texts.groupby(congress_texts['date'].dt.to_period('W')).agg({
    'sentiment_bertweet': lambda x: x.value_counts(normalize=True) * 100
}).unstack().fillna(0)

# Plotting weekly sentiment distribution for BJP
bjp_weekly.plot(kind='bar', stacked=True)
plt.title('Weekly Sentiment Distribution for BJP')
plt.xlabel('Week')
plt.ylabel('Percentage')
plt.show()

# Plotting weekly sentiment distribution for Congress
congress_weekly.plot(kind='bar', stacked=True)
plt.title('Weekly Sentiment Distribution for Congress')
plt.xlabel('Week')
plt.ylabel('Percentage')
plt.show()
```

Figure 44: Weekly Sentiment Analysis for BJP



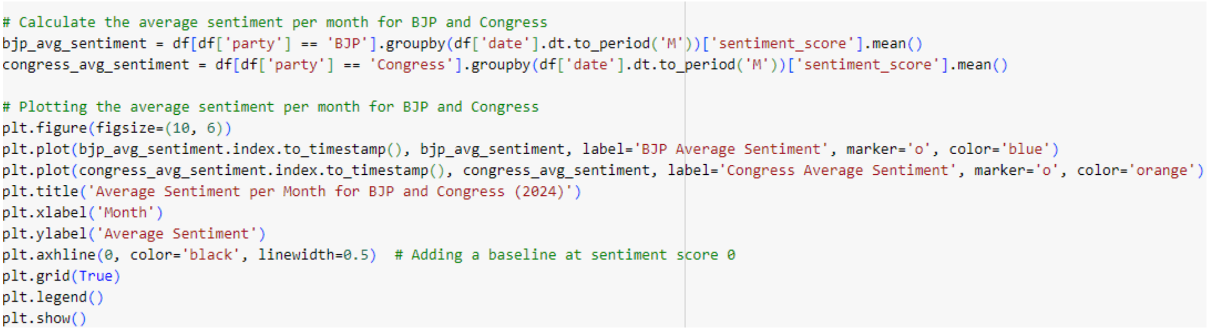
```
# Group data by month and calculate sentiment distribution for BJP
bjp_monthly = bjp_texts.groupby(bjp_texts['date'].dt.to_period('M')).agg({
    'sentiment_bertweet': lambda x: x.value_counts(normalize=True) * 100
}).unstack().fillna(0)

# Group data by month and calculate sentiment distribution for Congress
congress_monthly = congress_texts.groupby(congress_texts['date'].dt.to_period('M')).agg({
    'sentiment_bertweet': lambda x: x.value_counts(normalize=True) * 100
}).unstack().fillna(0)

# Plotting monthly sentiment distribution for BJP
bjp_monthly.plot(kind='bar', stacked=True)
plt.title('Monthly Sentiment Distribution for BJP')
plt.xlabel('Month')
plt.ylabel('Percentage')
plt.show()

# Plotting monthly sentiment distribution for Congress
congress_monthly.plot(kind='bar', stacked=True)
plt.title('Monthly Sentiment Distribution for Congress')
plt.xlabel('Month')
plt.ylabel('Percentage')
plt.show()
```

Figure 45: Monthly Sentiment Analysis for BJP



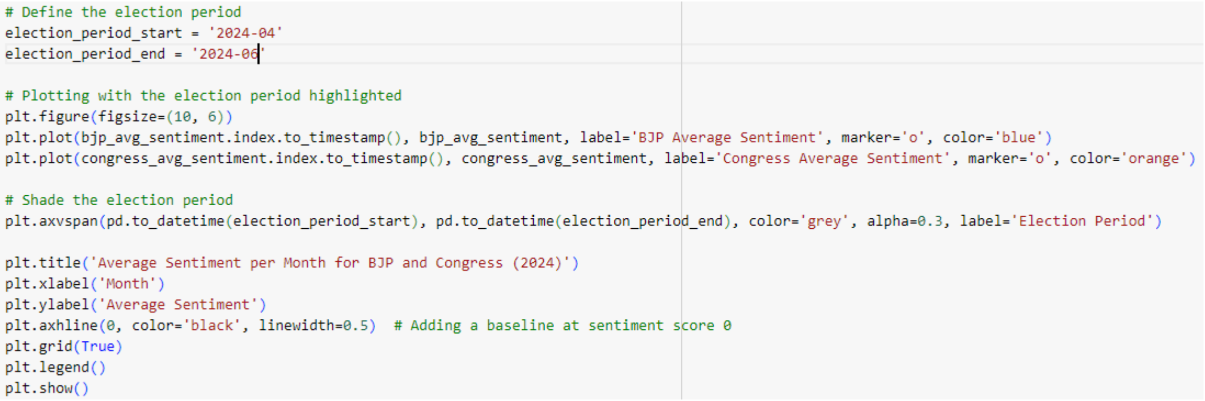
```
# Calculate the total positive, negative, and neutral sentiments per month for BJP
bjp_monthly_sentiments = bjp_texts.groupby(bjp_texts['date'].dt.to_period('M'))['sentiment_bertweet'].value_counts().unstack().fillna(0)

# Identify the peak month for positive sentiment for BJP
bjp_peak_positive = bjp_monthly_sentiments['POS'].idxmax()

# Identify the peak month for negative sentiment for BJP
bjp_peak_negative = bjp_monthly_sentiments['NEG'].idxmax()

print(f"Peak month for positive sentiment for BJP: {bjp_peak_positive}")
print(f"Peak month for negative sentiment for BJP: {bjp_peak_negative}")

# Repeat the process for Congress
congress_monthly_sentiments = congress_texts.groupby(congress_texts['date'].dt.to_period('M'))['sentiment_bertweet'].value_counts().unstack().fillna(0)

# Identify the peak month for positive sentiment for Congress
congress_peak_positive = congress_monthly_sentiments['POS'].idxmax()

# Identify the peak month for negative sentiment for Congress
congress_peak_negative = congress_monthly_sentiments['NEG'].idxmax()

print(f"Peak month for positive sentiment for Congress: {congress_peak_positive}")
print(f"Peak month for negative sentiment for Congress: {congress_peak_negative}")
```

Figure 46: Identifying Peak Months for Sentiment



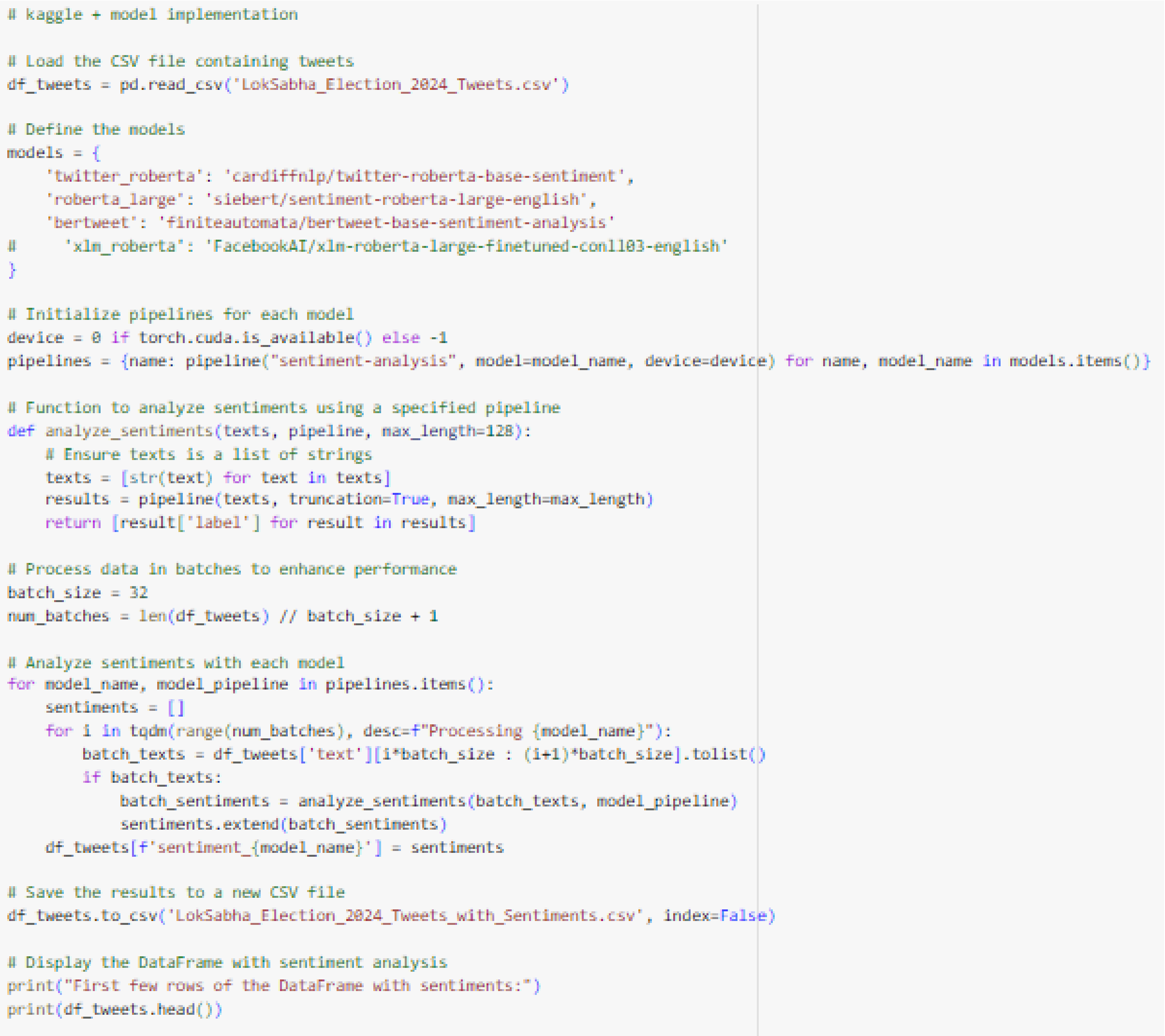
```
# Calculate the average sentiment per month for BJP and Congress
bjp_avg_sentiment = df[df['party'] == 'BJP'].groupby(df['date'].dt.to_period('M'))['sentiment_score'].mean()
congress_avg_sentiment = df[df['party'] == 'Congress'].groupby(df['date'].dt.to_period('M'))['sentiment_score'].mean()

# Plotting the average sentiment per month for BJP and Congress
plt.figure(figsize=(10, 6))
plt.plot(bjp_avg_sentiment.index.to_timestamp(), bjp_avg_sentiment, label='BJP Average Sentiment', marker='o', color='blue')
plt.plot(congress_avg_sentiment.index.to_timestamp(), congress_avg_sentiment, label='Congress Average Sentiment', marker='o', color='orange')
plt.title('Average Sentiment per Month for BJP and Congress (2024)')
plt.xlabel('Month')
plt.ylabel('Average Sentiment')
plt.axhline(0, color='black', linewidth=0.5)  # Adding a baseline at sentiment score 0
plt.grid(True)
plt.legend()
plt.show()
```

Figure 47: Monthly Average Sentiment



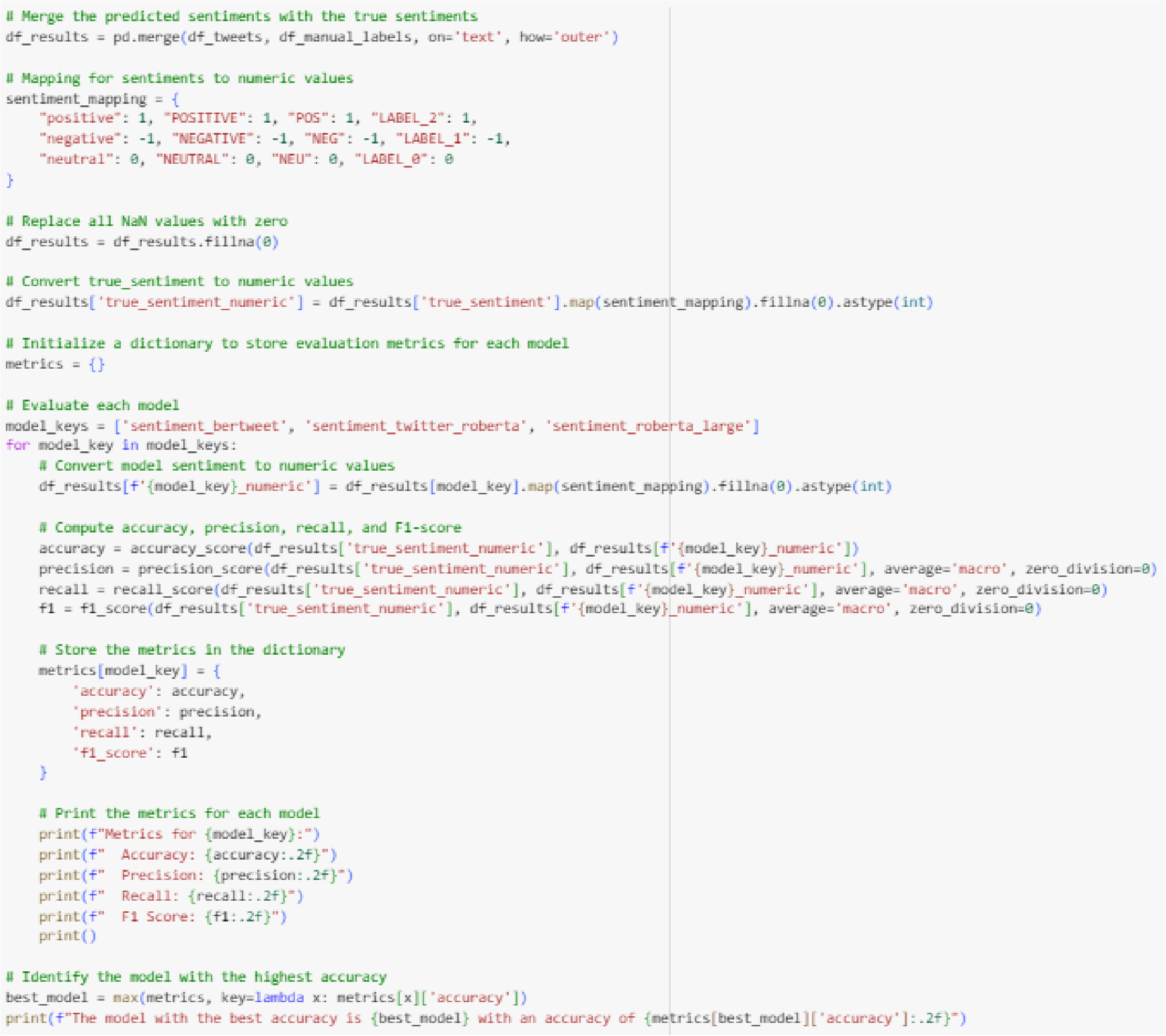
```
# Define the election period
election_period_start = '2024-04'
election_period_end = '2024-06'

# Plotting with the election period highlighted
plt.figure(figsize=(10, 6))
plt.plot(bjp_avg_sentiment.index.to_timestamp(), bjp_avg_sentiment, label='BJP Average Sentiment', marker='o', color='blue')
plt.plot(congress_avg_sentiment.index.to_timestamp(), congress_avg_sentiment, label='Congress Average Sentiment', marker='o', color='orange')

# Shade the election period
plt.axvspan(pd.to_datetime(election_period_start), pd.to_datetime(election_period_end), color='grey', alpha=0.3, label='Election Period')

plt.title('Average Sentiment per Month for BJP and Congress (2024)')
plt.xlabel('Month')
plt.ylabel('Average Sentiment')
plt.axhline(0, color='black', linewidth=0.5)  # Adding a baseline at sentiment score 0
plt.grid(True)
plt.legend()
plt.show()
```

Figure 48: Monthly Average Sentiment with Election Period Highlighted



```
# kaggle + model implementation

# Load the CSV file containing tweets
df_tweets = pd.read_csv('LokSabha_Election_2024_Tweets.csv')

# Define the models
models = {
    'twitter_roberta': 'cardiffnlp/twitter-roberta-base-sentiment',
    'roberta_large': 'siebert/sentiment-roberta-large-english',
    'bertweet': 'finiteautomata/bertweet-base-sentiment-analysis'
#     'xlm_roberta': 'FacebookAI/xlm-roberta-large-finetuned-conll03-english'
}

# Initialize pipelines for each model
device = 0 if torch.cuda.is_available() else -1
pipelines = {name: pipeline("sentiment-analysis", model=model_name, device=device) for name, model_name in models.items()}

# Function to analyze sentiments using a specified pipeline
def analyze_sentiments(texts, pipeline, max_length=128):
    # Ensure texts is a list of strings
    texts = [str(text) for text in texts]
    results = pipeline(texts, truncation=True, max_length=max_length)
    return [result['label'] for result in results]

# Process data in batches to enhance performance
batch_size = 32
num_batches = len(df_tweets) // batch_size + 1

# Analyze sentiments with each model
for model_name, model_pipeline in pipelines.items():
    sentiments = []
    for i in tqdm(range(num_batches), desc=f"Processing {model_name}"):
        batch_texts = df_tweets['text'][i*batch_size : (i+1)*batch_size].tolist()
        if batch_texts:
            batch_sentiments = analyze_sentiments(batch_texts, model_pipeline)
            sentiments.extend(batch_sentiments)
    df_tweets[f'sentiment_{model_name}'] = sentiments

# Save the results to a new CSV file
df_tweets.to_csv('LokSabha_Election_2024_Tweets_with_Sentiments.csv', index=False)

# Display the DataFrame with sentiment analysis
print("First few rows of the DataFrame with sentiments:")
print(df_tweets.head())
```

Figure 49: Kaggle Model Implementation for Sentiment Analysis



```python
# Merge the predicted sentiments with the true sentiments
df_results = pd.merge(df_tweets, df_manual_labels, on='text', how='outer')

# Mapping for sentiments to numeric values
sentiment_mapping = {
    "positive": 1, "POSITIVE": 1, "POS": 1, "LABEL_2": 1,
    "negative": -1, "NEGATIVE": -1, "NEG": -1, "LABEL_1": -1,
    "neutral": 0, "NEUTRAL": 0, "NEU": 0, "LABEL_0": 0
}

# Replace all NaN values with zero
df_results = df_results.fillna(0)

# Convert true_sentiment to numeric values
df_results['true_sentiment_numeric'] = df_results['true_sentiment'].map(sentiment_mapping).fillna(0).astype(int)

# Initialize a dictionary to store evaluation metrics for each model
metrics = {}

# Evaluate each model
model_keys = ['sentiment_bertweet', 'sentiment_twitter_roberta', 'sentiment_roberta_large']
for model_key in model_keys:
    # Convert model sentiment to numeric values
    df_results[f'{model_key}_numeric'] = df_results[model_key].map(sentiment_mapping).fillna(0).astype(int)

    # Compute accuracy, precision, recall, and F1-score
    accuracy = accuracy_score(df_results['true_sentiment_numeric'], df_results[f'{model_key}_numeric'])
    precision = precision_score(df_results['true_sentiment_numeric'], df_results[f'{model_key}_numeric'], average='macro', zero_division=0)
    recall = recall_score(df_results['true_sentiment_numeric'], df_results[f'{model_key}_numeric'], average='macro', zero_division=0)
    f1 = f1_score(df_results['true_sentiment_numeric'], df_results[f'{model_key}_numeric'], average='macro', zero_division=0)

    # Store the metrics in the dictionary
    metrics[model_key] = {
        'accuracy': accuracy,
        'precision': precision,
        'recall': recall,
        'f1_score': f1
    }

    # Print the metrics for each model
    print(f"Metrics for {model_key}:")
    print(f"  Accuracy: {accuracy:.2f}")
    print(f"  Precision: {precision:.2f}")
    print(f"  Recall: {recall:.2f}")
    print(f"  F1 Score: {f1:.2f}")
    print()

# Identify the model with the highest accuracy
best_model = max(metrics, key=lambda x: metrics[x]['accuracy'])
print(f"The model with the best accuracy is {best_model} with an accuracy of {metrics[best_model]['accuracy']:.2f}")
```

Figure 50: Accuracy Evaluation of Sentiment Models