Title: Polarity and Subjectivity in Indian Election Cycles

Abstract

Elections are the pinnacle of democracy as they shape the governance and the overall future of nations worldwide. With such considerable stakes, prominent political figures have to appeal to the people in whom they will represent and ultimately result in biases. One such place where this occurs is India, a diverse land full of various ideas, people, and parties. In this study, we investigate sentiment analysis in political discourse, specifically focusing on the public sentiment and polarity surrounding BJP's Narendra Modi and Congress's Rahul Gandhi. We aim to dissect these metrics via data from social media posts in order to analyze the effect of these numbers in determining political outcomes. The extent of polarization gives us leverage to discover the nature of such networked election races particularly in regards to location (states, urban/rural), people (caste/minorities/religion), and overall circumstances. The methodology involves employing TextBlob, a natural language processing library, and Pandas, a data manipulation library, to conduct sentiment analysis on public discourse related to Modi and Gandhi. This explores how sentiment analysis tools combined with dataframe analysis can reveal patterns in sentiment and polarity that shape the general nature of public discourse during election cycles.

Introduction

With the advancement of technology and global interconnectedness, it is now easier than ever to analyze public discourse via social media. With the increased use of such mediums in sharing the daily lives of billions of people around the world, an environment is facilitated where data scientists can analyze such communication in relation to the most volatile and subjective portions of human experience, namely election cycles. During election time, citizens often express a vast array of political attitudes and the continued divergence from the “center,” a phenomenon known as polarization. It reflects a world of extremes, specifically a world of posting and uploading the most drastic examples of movements or policy for likes and comments.

This research initiative dives deeper into analyzing the extent of such polarization. It focuses on sentiment analysis within the unique context of Indian elections by specifically examining tweets related to key political figures like Narendra Modi of the Bharatiya Janata Party (BJP) or Rahul Gandhi who has allegiance with the Indian National Congress. Although previous studies regarding sentiment exist especially in the USA, this emphasis on India takes into account a completely different societal dynamic as the analysis sheds light on the impact of factors like religion, caste, and community on public opinion.

The reasoning behind this is to gauge the political strategies and decision making that is utilized through social media. Twitter, in particular, has become a prominent platform for political discussions and expressions of public sentiment. By applying our sentiment analysis approach via Python on these data sets (tweets from key figures), we get to witness first hand the emotion and opinions that serve as a catalyst behind each tweet.

The usefulness and practicality behind these findings are present in the potential to inform political scientists about political strategy and to gauge public sentiment trends in a very different political landscape than the United States. By depicting certain trends and patterns, sentiment across different regions and demographics, the research seeks to bridge existing gaps in regards to the diversity and social media use of India as well as identify the overlying characteristics of its infamous election cycles/campaigns.

Related work:

Here, we will contextualize our research within the broader landscape of related work. Here are some related studies from Google Scholar:

- "Opinion Detection in Thai Political News Columns Based on Subjectivity Analysis": This study by Sukhum, Nitsuwat, and Haruechaiyasak focuses on detecting opinions in Thai political news columns. They propose an opinion mining framework to monitor highly-opinionated news content, particularly in the context of Thai journalism. While our research examines sentiment analysis in the context of Indian political discourse on social media, this study underscores the importance of opinion detection in media monitoring across different cultural/linguistic contexts.
- "Empirical Investigation of Political Party Control and Election Outcomes Using Sentiment Analysis": This abstract depicts the political party control over candidates on election outcomes, employing sentiment analysis of Twitter discussions during the Anambra State gubernatorial election in Nigeria. This is similar to our study as we are also using Twitter as our main database.
- "Boosting Twitter Sentiment Classification Using Different Sentiment Dimensions": This article proposes an approach for enhancing Twitter sentiment classification by incorporating various sentiment dimensions such as opinion strength, emotion, and polarity indicators. This study demonstrates the importance of leveraging different sentiment dimensions for improving sentiment analysis tasks.

Methods

This research aims to quantify the presence of 2 metrics in order to enable us to make sense of the data: Polarity and subjectivity. Polarity for starters is the direction in which emotion is expressed in a given sentence. The intensity of the sentence could be positive, negative, or neutral. Essentially, the polarity score is computed by analyzing the occurrence of certain “trigger” words in relation to their connotation and their association with positive or negative sentiments in a particular text. Textblob comes into play in cultivating these outcomes as it is an open-source library in Python that assists in processing textual data. Textblob employs a NLTK which is a natural language toolkit. Thus, it uses a pre-trained sentiment analysis model that is typically based on a supervised machine learning approach in order to assign a polarity score to each given sentence. Polarity is measured between a domain of [-1,1]. A score of -1 would indicate negative polarity while 1 would indicate the opposite. This numerical approach in analyzing two possible sides of a given structure facilitates our effective analysis.

Subjectivity is another metric. It refers to the extent of opinionization on a given matter. It regards how subjective a piece of information is contrary to how objective/factual it could be. The calculation of this involves analyzing the presence of subjective language elements such as personal pronouns, and emotional expressions within the text. It again uses the NLTK to see already given words that indicate polarity such as “I”, “think,” etc. It takes into account the frequency and nature of such language elements that potentially indicates subjectivity when scoring a text. Unlike polarity, subjectivity’s domain actually ranges from 0 to 1 with 0 indicating factual content to 1 which depicts volatile and opinionated text.

It is important to note that textblob is not 100% accurate and actually has a human to output agreement percentage of 80-85% which is deemed sufficient.

Kolmogorov-Smirnov Test (scipy.stats.ks_2samp).

The Kolmogorov-Smirnov Test is essentially a mechanism to assess whether two distributions originate from the same underlying population distribution. The KS test checks if two distributions, (A and B), are different. It gives a number that shows how much these distributions differ. This number is found by comparing how their cumulative distribution functions (CDFs) differ. In machine learning like in this research project, we use samples from the datasets to calculate these CDFs, which we call eCDFs.

In our scenario, our A and B distributions could be: A, representing the frequency of tweets about Modi in your model training dataset, and B, representing the frequency of tweets about Gandhi. We aimed to determine if the deployed model is encountering a distribution of data (tweets) different from what was used during training. The KS test is nonparametric (does not make any assumptions and measures the central tendency with the median value) statistical test that can be employed to compare the distribution of Modi tweets (A) and Gandhi tweets (B) to discover if they are statistically different. It also ensures that there is statistical significance as we want to ensure that the occurrence of what we are measuring isn't purely by chance.

Results

KS Table for the figures directly from official accounts

	KS Statistic	**P-Value**
Polarity (Gandhi vs Modi)	0.253	0.027
Sentiment (Gandhi vs Modi)	0.133	0.549

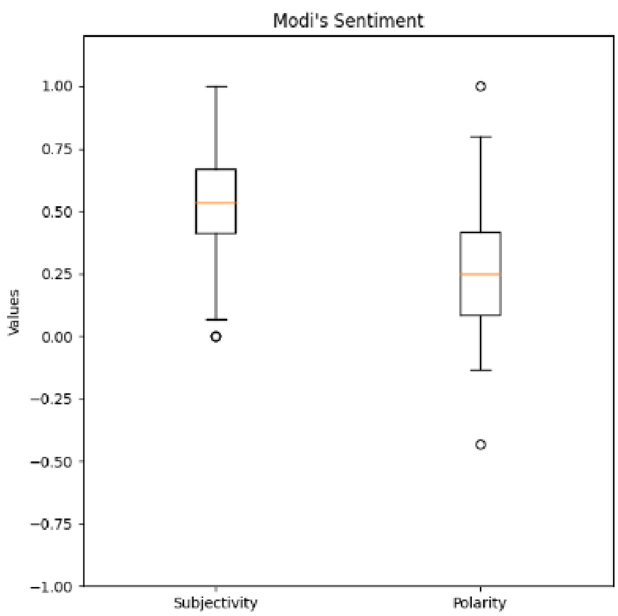
KS Table for tweets revolving the figures (public discourse)

	KS Statistic	**P-Value**
Polarity (Gandhi vs Modi)	0.371	0.012
Sentiment (Gandhi vs Modi)	0.229	0.532

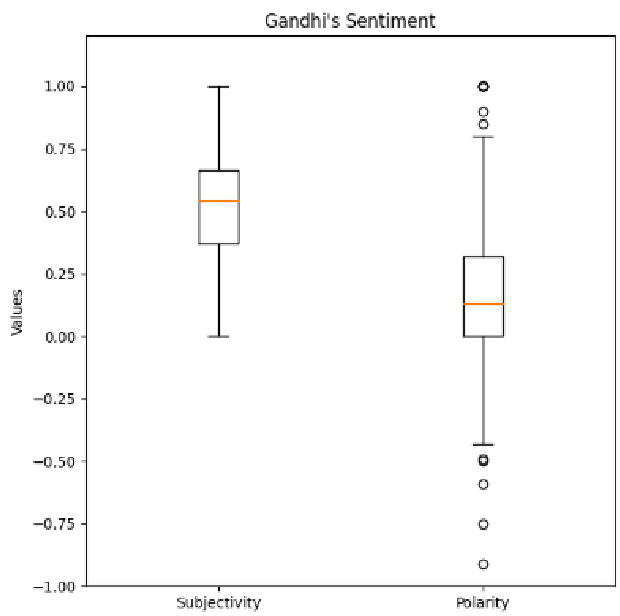
Polarity- KS Statistic (0.253) is a number that shows how different the distributions of sentiment scores are between the two datasets. P-value (0.027) tells us the chance of getting a test statistic as extreme as the one we got if there was no real difference between the sentiment scores of the two groups. Since the p-value is less than 0.05 (the standard we used), the distributions are statistically different and we can reject the null hypothesis. Thus, there's strong proof that the sentiment expressed in tweets about Modi is different from that in tweets about Gandhi. So, people are likely expressing different feelings or opinions about these two figures on Twitter.

Subjectivity- KS Statistic (0.1333) is a number that shows how different the distributions of sentiment scores are between the two datasets. P-value (0.549) tells us the chance of getting a test statistic as extreme as the one we got if there was no real difference between the sentiment scores of the two groups. Since the p-value is greater than 0.05 (the standard we used), the distributions are not statistically different and we actually reject the null hypothesis (It is important to note, when we can't reject the null hypothesis, we can't accept it either). Thus, we don't have enough evidence to say that the distributions of subjectivity scores in tweets about Modi are different from those in tweets about Rahul Gandhi. So, it seems that people are expressing similar levels of subjectivity in tweets about both figures.

The result makes sense for subjectivity because subjectivity, unlike sentiment polarity, is a measure of how opinions or attitudes are expressed rather than whether they are positive, negative, or neutral. This suggests that people are expressing similar levels of subjectivity in tweets about both Modi and Gandhi, indicating that opinions or personal perspectives are similarly present in tweets related to both figures.

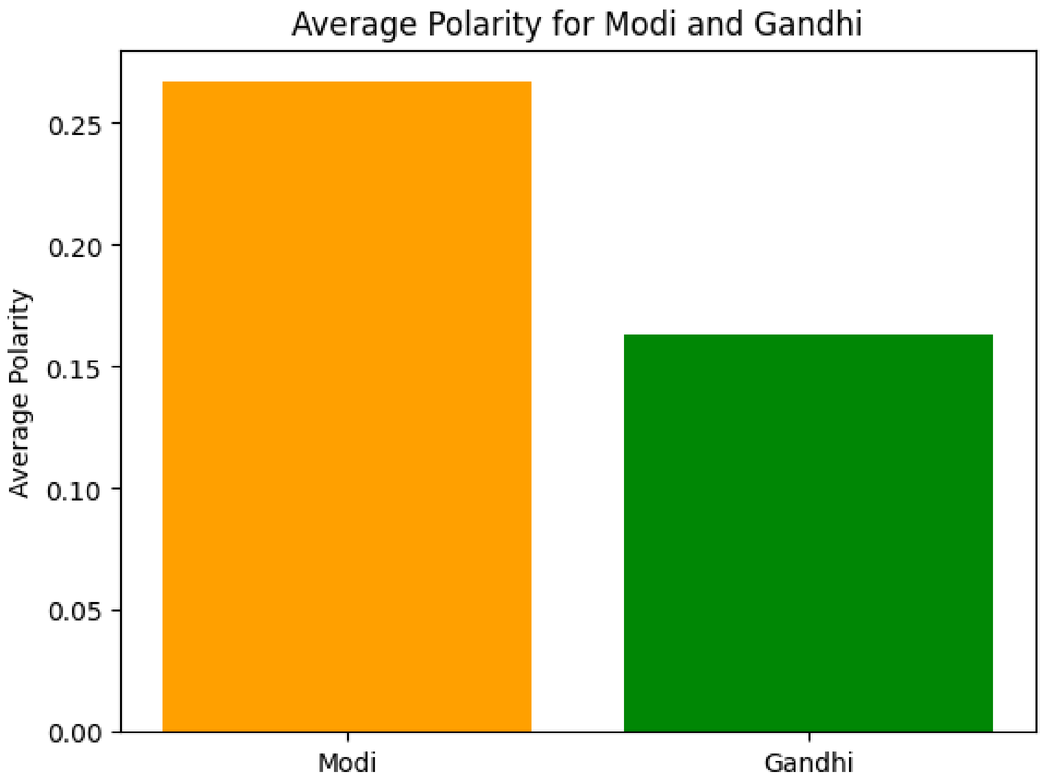


Box and Whisker Plot of Modi's Sentiment

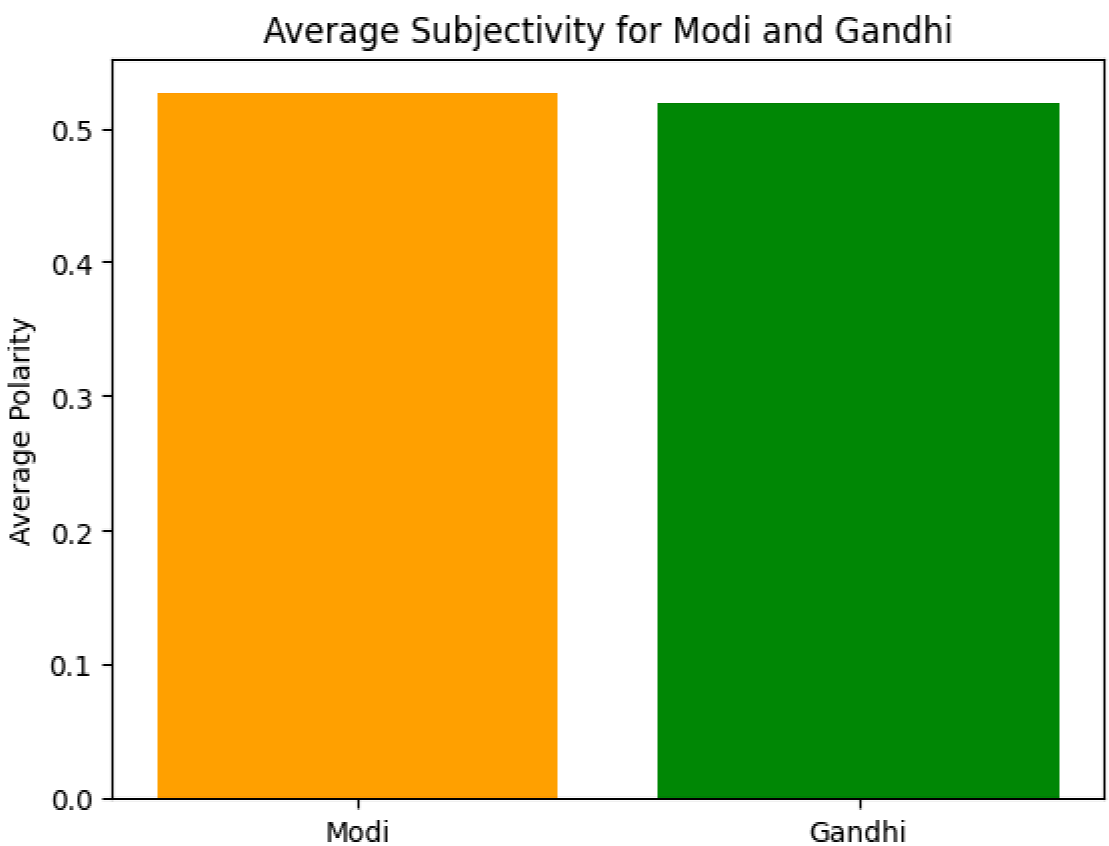


Box and Whisker Plot of Gandhi's Sentiment

Results: Here we can see that Modi's subjectivity and Gandhi's subjectivity are actually pretty similar. This makes sense as subjectivity is an attribute that can be applied to a large extent. However for polarity, Gandhi actually had a greater range and larger Q1 and Q3 values. The dataset also had way more outliers. This discrepancy suggests that Gandhi's statements or actions encompass a wider spectrum of sentiment compared to Modi's. The abundance of outliers indicate instances where his expressions or behaviors deviate notably from the norm.



Average Polarity for Modi and Gandhi
Average Polarity
0.00
0.05
0.10
0.15
0.20
0.25
Modi
Gandhi



Average Subjectivity for Modi and Gandhi
Average Polarity
0.0
0.1
0.2
0.3
0.4
0.5
Modi
Gandhi

There is similar average subjectivity for both Modi and Gandhi. Since subjectivity is a measure of the extent to which language reflects personal opinions, beliefs, or emotions, our findings suggest that both leaders tend to express themselves with comparable degrees of personal perspective rather than objective fact. However, there is a stark difference in terms of the average polarity Modi's tweets convey. This implies that Modi's communication often embodies interesting sentiment - whether positive or negative (less neutrality in expression). It suggests that Modi's communication style may be more polarizing or emotionally charged, evoking stronger reactions or opinions among the audience. This is backed up by many instances of BJP party members and affiliates invoking violence or behaving in a drastic manner in India. Gandhi's lower polarity indicates less pronounced emotional or opinionated undertones. This could imply a more diplomatic or cautious approach to communication (aimed at maintaining broader appeal or avoiding controversy).

Conclusion:

Overall, this study attempts to underscore the role of sentiment analysis in political discourse within the context of Indian elections. By focusing on public sentiment and polarity surrounding key political figures like Narendra Modi and Rahul Gandhi, the research sheds light on the multifaceted nature of political communication in the digital age. Through the application of sentiment analysis tools such as TextBlob and Pandas on social media data (Twitter datasets), the study uncovers patterns and trends that shape public discourse during election cycles. More significantly, this demonstrates the impact of factors like religion, caste, and community on public opinion and the overall implications on such a diverse electorate.

Via statistical evidence of the differences in sentiment scores between tweets about Modi and Gandhi, distinct patterns in public perception of these leaders are established. While subjectivity levels in tweets about both leaders are similar, Modi's tweets exhibit a wider range and higher polarity, suggesting a more polarizing communication style compared to Gandhi's more diplomatic approach. Furthermore, tweets relating to Modi himself (not directly from his Twitter account) also indicate greater polarity levels. Thus, Modi's polarizing nature is supported by the greater polarity showcased in tweets by outsiders ABOUT him.

Overall, this research contributes to our understanding of political strategy, public sentiment trends, and the influence of social media in shaping political discourse in India. The study offers valuable implications for political scientists, policymakers, and those interested in understanding the dynamics of democracy in diverse societies.

Citations:

Baccianella, S., Esuli, A., and Sebastiani, F. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, 2010. Valletta, Malta.

ACE (2012) Parties and candidates in the electoral process—ace, ACE-the electoral knowledge network. http://aceproject.org/ace-en/topics/pc/pcc/pcc06/default. Accessed 4 April 2024

Rudraksh Gohil, Deepa S, Vinay M, Jayapriya J, "Election Forecasting with Machine Learning and Sentiment Analysis: Karnataka 2023", 2023 International Conference on Ambient Intelligence, Knowledge Informatics and Industrial Electronics (AIKIIE)

Remya Panikar, Ramchandra Bhavsar, B. V. Pawar, "Sentiment Analysis: a Cognitive Perspective", 2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS)