**TABLE OF CONTENT**

1. Introduction:
   1. Natural Language Processing (NLP)
   2. TF - IDF
   3. Sentiment Analysis
   4. BERT
2. Objective
3. Literature Survey
4. Proposed methodology
5. Hardware and software requirements
6. TF - IDF
   1. TF - IDF
   2. Term Frequency
   3. Inverse Document Frequency
7. BERT (Bidirectional Encoder Representation from Transformer)
   1. Architecture
   2. Performance
   3. Analysis
8. Training the model
   1. Data identification
   2. Data collection
   3. Data cleaning
   4. Data categorization
   5. Sentimental Analysis
   6. Result comparison
9. Conclusion
10. Observations
11. Reference
12. **INTRODUCTION**
    1. Natural Language Processing (NLP)

The field of natural language processing (NLP) combines linguistics, computer science, and machine learning. NLP is all about teaching computers to comprehend and produce human language. The field focuses on communication between computers and humans in natural language.

NLP blends statistical, machine learning, and deep learning models with computational linguistics—rule-based modeling of human language. With the use of these technologies, computers are now able to interpret human language in the form of text or audio data and fully "understand" what is being said or written, including the speaker's or writer's intentions and mood.

Computer programs that translate text between languages, reply to spoken commands, and quickly summarize vast amounts of text—even in real-time—are all powered by NLP. You've probably used NLP in the form of voice-activated GPS devices, digital assistants, speech-to-text dictation programs, customer service chatbots, and other consumer conveniences. The use of NLP in corporate solutions, however, is expanding as a means of streamlining company operations, boosting worker productivity, and streamlining mission-critical business procedures.

* 1. TF - IDF

Information retrieval and text mining frequently employ the tf-idf weight, which stands for term frequency-inverse document frequency. This weight is a statistical metric used to assess a word's significance to a collection or corpus of documents. The frequency of the term in the corpus offsets the importance, which rises in direct proportion to the frequency of a word in the document. Search engines frequently utilize variations of the tf-idf weighting system as their main score and ranking mechanism for determining how relevant a document is to a user query.

* 1. Sentiment Analysis

Using the technique of sentiment analysis, you may examine a text to ascertain the sentiment it represents. This is accomplished by combining machine learning and natural language processing (NLP).

Using basic Sentiment analysis, a program can understand whether the sentiment behind a piece of text is positive, negative, or neutral.

It is a powerful technique in Artificial intelligence that has important business applications.

* 1. BERT

BERT (Bidirectional Encoder Representations from Transformers) is a recent paper published by researchers at Google AI Language. It has caused a stir in the Machine Learning community by presenting state-of-the-art results in a wide variety of NLP tasks, including Question Answering (SQuAD v1.1), Natural Language Inference (MNLI), and others.

The primary technological advancement of BERT is the application of Transformer's bidirectional training, a well-liked attention model, to language modeling. In contrast, earlier research looked at text sequences from either a left-to-right or a mixed left-to-right and right-to-left training perspective. The study's findings demonstrate that bidirectionally trained language models can comprehend the context and flow of language more deeply than single-direction language models. The authors of the study describe a unique method called Masked LM (MLM), which makes bidirectional training possible in models where it was previously not practicable.

1. **OBJECTIVE**

If someone wants to know about what the political party is doing in their elected year and how they are performing there are two options either they ask the people near them or they do their own research if they ask people near them there can be bias in there result as deep down they are inclined towards a political party and unbeknownst to them, they might give results which have some bias present in them.

To obtain completely unbiased or true neutral results a human’s perception cannot be the base, for this purpose the use of artificial intelligence is a more suitable tool. As the artificial intelligence has no preconceived notion of the topic for which it is being to get some result, one can be sure that if the knowledge base of the AI agent is free from any bias or inclination towards any entity then the results obtained are truly neutral.

The main object of this research was to compare the political ideologies of two major political parties present in the Indian Subcontinent and how they have changed over the last four decades. A political party's ideologies are best reflected in the political manifesto released by them before an election. A manifesto is a written statement of the goals, objectives, or opinions of the issuer, which could be an individual, an organization, a political party, or even the government. A manifesto provides a great insight into the internal working of a political party and what they plan on doing to tackle the challenges given. These are the direct representation of the parties' thinking and ideologies on what matters are more important than others to them.

1. **LITERATURE SURVEY**

The usage of social media in the contemporary digital world is at an all-time high. The majority of individuals use social media sites like Twitter, Facebook, and YouTube to share their thoughts and experiences with other people. These platforms are regularly used by users to discuss incidences or problems they are facing. As opposed to the offline manner of opinion sharing, using these platforms enables the user to communicate their thoughts to a larger audience. Both the government and businesspeople may greatly benefit from an analysis of popular sentiment and opinion. This is one of the reasons why so many media outlets are active during election season and are encouraged to carry out various types of surveys.

In this study, we looked at the political parties manifestos in India.

The paper's main contributions are as follows:

• Making an effective sentiment analysis model for election data.

• Giving the model the ability to comprehend real emotional states.

• Using the transfer learning approach to handle the unsupervised nature of the problem.

• Examining the model's findings in light of the actual election results.

Sentiment analysis has been used for manifestos from a variety of application fields. The Term Frequency Inverse Document Frequency (TF-IDF) method was applied in our machine learning model. For this application, emotion ratings are generated using the textual data of the BJP and INC, two prominent political parties. First, the downloaded manifestos were cleaned up, then they were converted into the matching CSV files.

Utilizing the different political party manifestos from 1999 to 2014, our trained model is tested. People used to share their ideas, opinions, and experiences with the Bharatiya Janata Party (BJP) and the Indian National Congress, two of the biggest political parties at the time, during elections (INC).

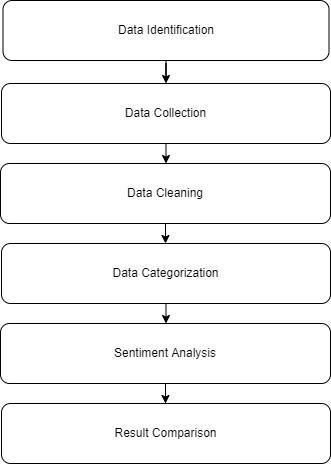
Algorithms for machine learning are employed. For instance, the matrix Term Frequency - Inverse Text Frequency (TF-IDF) offers a word frequency table for a phrase or document. Document Frequency (DF(t)) is a counter for how many times a word appears in documents, whereas Term Frequency (TF(t,d)) is a measure for how often a word appears in document(d). The inverse document frequency (IDF(t)) function estimates a word's relative importance.

The test is performed on a computer with an Intel Core i5 processor, 8GB of RAM, and a 2TB hard drive, with Windows 10 as the operating system. The Python 3.7 Programming Environment and Jupyter Notebook 6.4.3 were used to complete the project. Other tools/libraries that have been adopted are listed below.

Sklearn version 0.22: Scikit-learn is a Python-based machine learning library that includes a number of regression, classification, and clustering algorithms. Matplotlib (version 3.4.3) is a Python plotting library that offers an API for plotting data embedded plots. (This is version 1.3.2.) Pandas is a Python data table handling library that provides methods for quickly and easily performing all data table operations. Pipelines (Version 1.0): Pipelines in Python provide a sequence of data transformation and prediction based on machine learning models that are passed to them. The JSON library can be used to parse JSON from strings or files. JSON Python lists or dictionaries are parsed by the library. JSON strings can also be created from Python dictionaries and lists.

1. **PROPOSED METHODOLOGY**

The methods used for finding the result can be summarized in the following flowchart:



**Figure 1:** The Flow of the Proposed Methodology

Social media use is at an all-time high in the modern digital world. Most people use social media platforms to communicate their ideas and experiences with others (such as Twitter, Facebook, and YouTube). The government gains a lot from studying the attitudes and viewpoints of the populace. This is one of the factors that make a lot of media outlets active during election season and encourages them to conduct different kinds of polls.

In order to analyze the beliefs of Indian political parties for this academic paper, we used a Manifesto from that time period. We created an automatic manifesto analyzer using natural language processing to address the problem's unsupervised nature. Using the Term Frequency Inverse Document Frequency (TF-IDF) approach and the pdfMiner script, we handled the manifesto's textual data in our machine learning model. Additionally, we enhanced the model's capacity to handle manifesto sentiment values.

As previously mentioned, data is first gathered from the party manifestos, then data is extracted to the appropriate CSV files, after which the TF-IDF values are calculated in the manifesto of all the sentences and are divided into five categories: education, agriculture, infrastructure, economy, and health.

These five groups have been chosen because each group's growth may be used to gauge a country's overall development.

The emotion values of each phrase are computed for each of these five categories once the TF-IDF values have been obtained, and they are then separated into five values: negative, semi-negative, neutral, semi-positive, and positive.

In the conclusion, graphs are drawn to allow for reliable comparisons of the collected data.

Since the necessary data was difficult to get and the very last manifesto was chosen for this research, gathering the data was likewise a difficult effort. We picked them for their fair judgment, which was the sole justification.

They give as much information as possible on a political party, which is sufficient to make the most precise judgments.

There were a few challenges encountered:

* Since it was unable to utilize any conventional data extraction techniques, such as APIs or data scraping, we used manifestos as our main source of information.
* The manifestos that we got from the websites of the various political parties include a lot of formatting, such as headers, footers, page numbers, etc., which was causing a mess. To convert each manifesto into raw text, we utilized basic Python packages like nltk, re, and pandas.
* We occasionally get multiple spaces while deleting the header. Once more, we just added space in their place.
* The same is true for an Rs. (Rupees Indication); by condensing the content into a single sentence, the '.' created a muddle. We used "₹" in place of "Rs".

1. **HARDWARE AND SOFTWARE REQUIREMENTS**
   1. Hardware Requirements

Processor: Intel i5 (9th Generation) equivalent or higher

RAM: 8 GB or higher

Operating System: Any operating system that supports Python 3.

GPU: A CPU with compute score of 6.5 or higher

* 1. Software Requirements

Programming Language: Python (3.8 or higher).

IDE: Jupyter Notebook, Visual Studio Code.

Library's: nltk; re; numpy; pandas; pdfminer.six; matplotlib

1. **TF - IDF**
   1. TF-IDF

The tf-idf for each query phrase is added to create one of the most basic ranking functions; many more complex ranking functions are variations of this basic model.

Tf-idf may be utilized effectively for stop-word filtering in a variety of topic areas, such as text categorization and summarization.

The normalized Term Frequency (TF) is calculated as the logarithm of the number of documents in the corpus divided by the number of documents where the specific term appears.

The Inverse Document Frequency (IDF) is computed as the logarithm of the number of documents in the corpus divided by the number of documents where the specific term appears.

* 1. Term Frequency

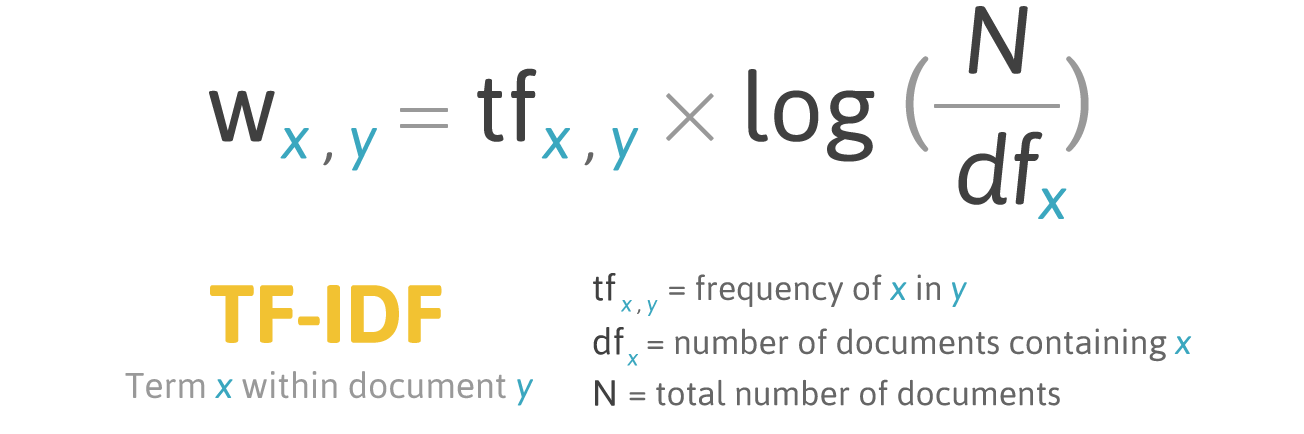
Term Frequency, which measures how frequently a term occurs in a document. Since every document is different in length, it is possible that a term would appear much more often in long documents than in shorter ones. Thus, the term frequency is often divided by the document length (aka. the total number of terms in the document) as a way of normalization:

TF(t) =

* 1. Inverse Document Frequency

Inverse Document Frequency, which measures how important a term is. While computing TF, all terms are considered equally important. However, it is known that certain terms, such as "is", "of", and "that", may appear a lot of times but have little importance. Thus we need to weigh down the frequent terms while scaling up the rare ones, by computing the following:

IDF(t) = loge().



1. **BERT (Bidirectional Encoder Representation from Transformer)**
   1. Architecture

There are three approaches used:

Rule-based approach: Over here, the lexicon method, tokenization, and parsing come in the rule-based. The approach is to count the number of positive and negative words in the given dataset. If the number of positive words is greater than the negative words then the sentiment is positive else vice-versa.

Automatic Approach: This approach works on the machine learning technique. Firstly, the datasets are trained and predictive analysis is done. The next process is the extraction of words from the text. This text extraction can be done using different techniques such as Naive Bayes, Linear Regression, Support Vector, and Deep Learning. This machine learning technique is used.

Hybrid Approach: It is the combination of both the above approaches i.e. rule-based and automatic approaches. The surplus is that the accuracy is high compared to the other two approaches.

People express opinions in complex ways, which makes understanding the subject of human opinions a difficult problem to solve. Rhetorical devices like sarcasm, irony and implied meaning can mislead sentiment analysis, which is why concise and focused opinions like product, book, movie, and music reviews are easier to analyze.

* 1. Performance

The transformer is an attention mechanism that learns the contextual relationships between words (or subwords) in a text and is used by BERT. The transformer's basic design consists of two independent mechanisms: an encoder that reads the text input and a decoder that generates a job prediction. Only the encoder mechanism is required because BERT aims to produce a language model.

The Transformer encoder reads the full sequence of words at once, in contrast to directional models, which read the text input sequentially (from right to left or left to right). Although it would be more appropriate to describe it as non-directional, it is thus thought of as bidirectional. This trait enables the model to understand a word's context depending on all of its surroundings (left and right of the word).

Unquestionably, BERT represents a milestone in machine learning's application to natural language processing. Future practical applications are anticipated to be numerous given how easy it is to use and how quickly they can be fine-tuned. We made an effort to avoid going overboard with the technical specifics in this summary while yet capturing the essence of the work. We strongly suggest reading the complete paper and any related publications that are cited in it for anyone interested in doing more research. The BERT source code and models, which include 103 languages and were kindly made available as open source by the study team, are another helpful resource.

* 1. Analysis

It is still unclear why BERT performs so well on these natural language processing challenges at the cutting edge of technology. Current research has concentrated on understanding the linkages that underlie BERT's output as a result of carefully selected input sequences, examination of internal vector representations by probing classifiers, and the associations implied by attention weights.

1. **TRAINING THE MODEL**
   1. Data identification

There were two approaches to obtain the same result. The mentioned two approaches entail the same result with one resulting in a higher quantity and quality of data.

* The parameters on which the identification and selection of the data were based are as follows:
* The information source should be able to convey the goals and image of the political party comprehensively.
* The data should be credible and its origin must be a very reliable source of information.
* The data should be present in a form that is easily converted into a format that is suitable for further analysis.

Based upon the above parameter the data source which satisfies all of the above requirements was the Political Manifesto which a party releases before the election.

A manifesto is a written statement of the goals, objectives, or opinions of the issuer, which might be an individual, an organization, a political party, or even the government. A manifesto typically supports a new idea with prescriptive ideas for implementing the changes the author believes should be made, accepts a previously published opinion or the consensus, or promotes a new idea. It frequently has a political, social, or aesthetic theme, is occasionally revolutionary, and may even convey a person's attitude on life. Creeds or confessions of faith are common terms used to describe religious manifestos. To summarize this a manifesto lays out the goals and the plan of a political party and its ideologies.

This makes the manifesto a perfect source of information that is required for this research. The consolidation of such data about the political parties and their promises makes the political manifestos a direct image of the party.

* 1. Data Collection

Since the political manifesto for Indian Parties was readily available at a single source so traditional data extraction techniques such as the use of API for data collection or web scraping couldn’t be used.

Each manifesto was manually downloaded from the source which had them available.

The format in which these manifestos were downloaded was Portable Document Format (.pdf). Since there were no other data sources accessible than the manifesto in PDF format on the websites of the political parties, we were constrained to use the pdf files as our primary data source.



Figure 1: Snippet of BJP Manifesto

* 1. Data Cleaning

The PDF file of each manifesto contained a lot of verbiages that is not suitable for further steps of the process.

The extra text present in the documents has to be cleared to make it easier for analysis. To achieve the result a simple Python script was used to clear the aforementioned documents. Since the format of the text written in the file was the same across all the files a single program can be used for cleaning all the files.

Python library “pdfminer.six” was used for reading the PDF file into the programs. The data given by this process was appropriately cleaned by removing the header and footer by using Regular Expressions. The data obtained from the cleaning process was converted into a Pandas DataFrame and then this DataFrame was stored as a CSV(comma separated values) file. CSV format was chosen because it represents data in a tabular form and there are a myriad of libraries available to handle, process, and manipulate CSV file data. The tabular representation of the data makes the understanding and representation of data easier.



Figure 2: A snippet of the python script used to clean the data



Figure 3: After cleaning the data

* 1. Data Categorization

The categories chosen for the categorization of the data that were chosen are Agriculture, Education, Economics, Health, and Infrastructure. These five categories act as the defining pillar for the growth of a country.

Mainly the prevalent topics during an election are directly from these five categories or from one of the many derivatives of these categories.

The data categorization was done by searching each line of the manifesto for words having similar meanings to the categorization keywords. If a matching word was found then the corresponding word’s TF - IDF value was found and stored in a CSV file. TF - IDF values calculate the weight of each word in a sentence thus an ideal way of finding the importance or the emphasis of a word in a sentence.

Any sentences which were devoid of any of these aforementioned words were removed from the database, reducing the size of the useless data significantly.



Figure 4: A snippet of python script on the categorization of data

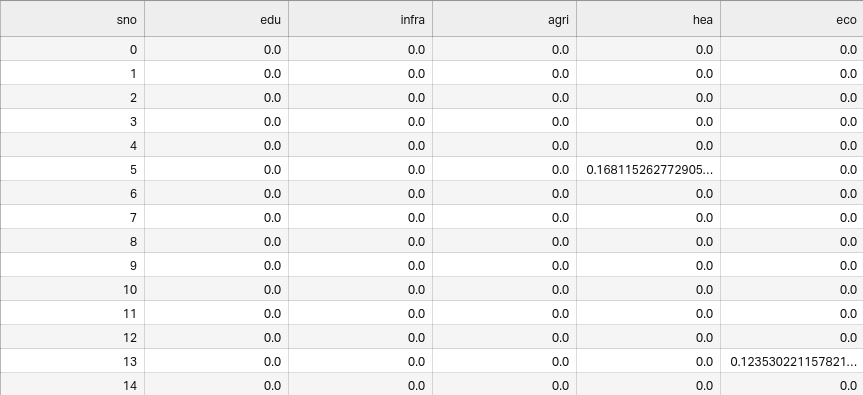


Figure 5: Data after categorization

* 1. Sentimental Analysis

After having the TF-IDF values of each category the data is now sent for sentiment analysis. In sentiment analysis first, the bert model is imported, this bert model is trained on many databases as this model is trained by google on its searches. As seen in figure 6 the python code for sentiment analysis is shown.

After training the model on the categories data we have the sentiment value of each word corresponding to the categories we have as shown in figure 7.

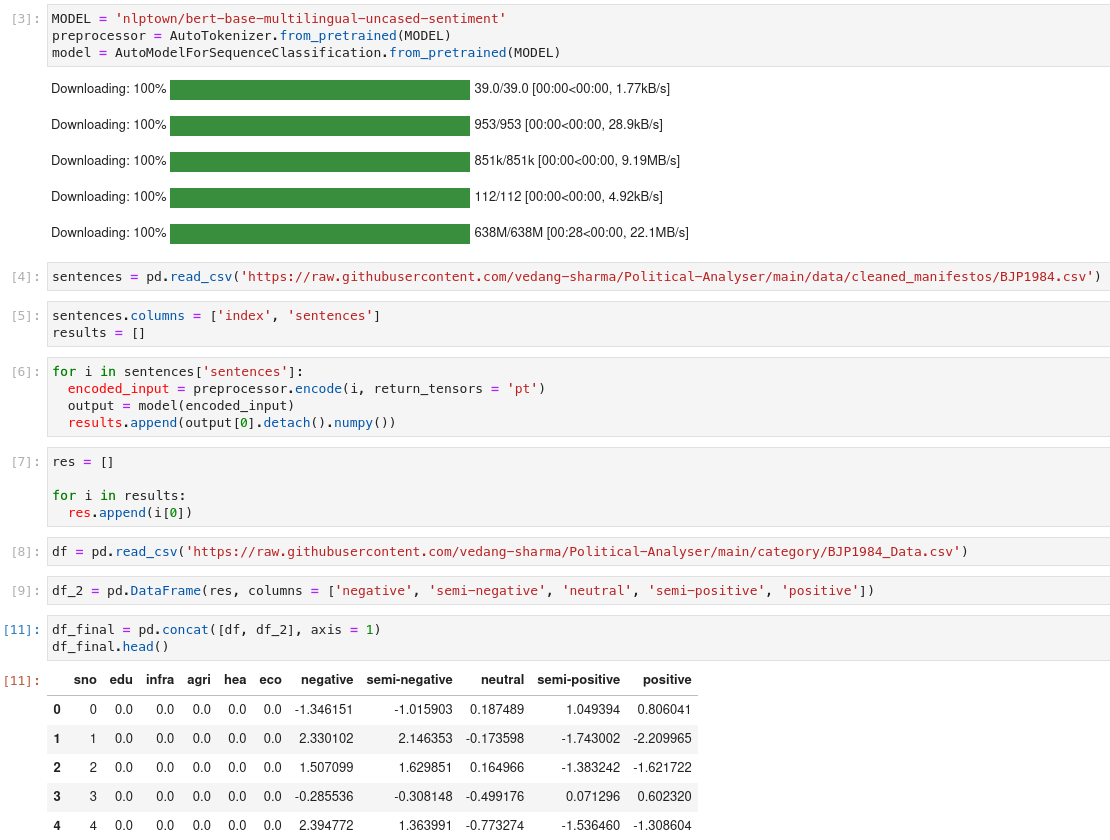


Figure 6: A small snippet of python script to find sentiment analysis of categorized data

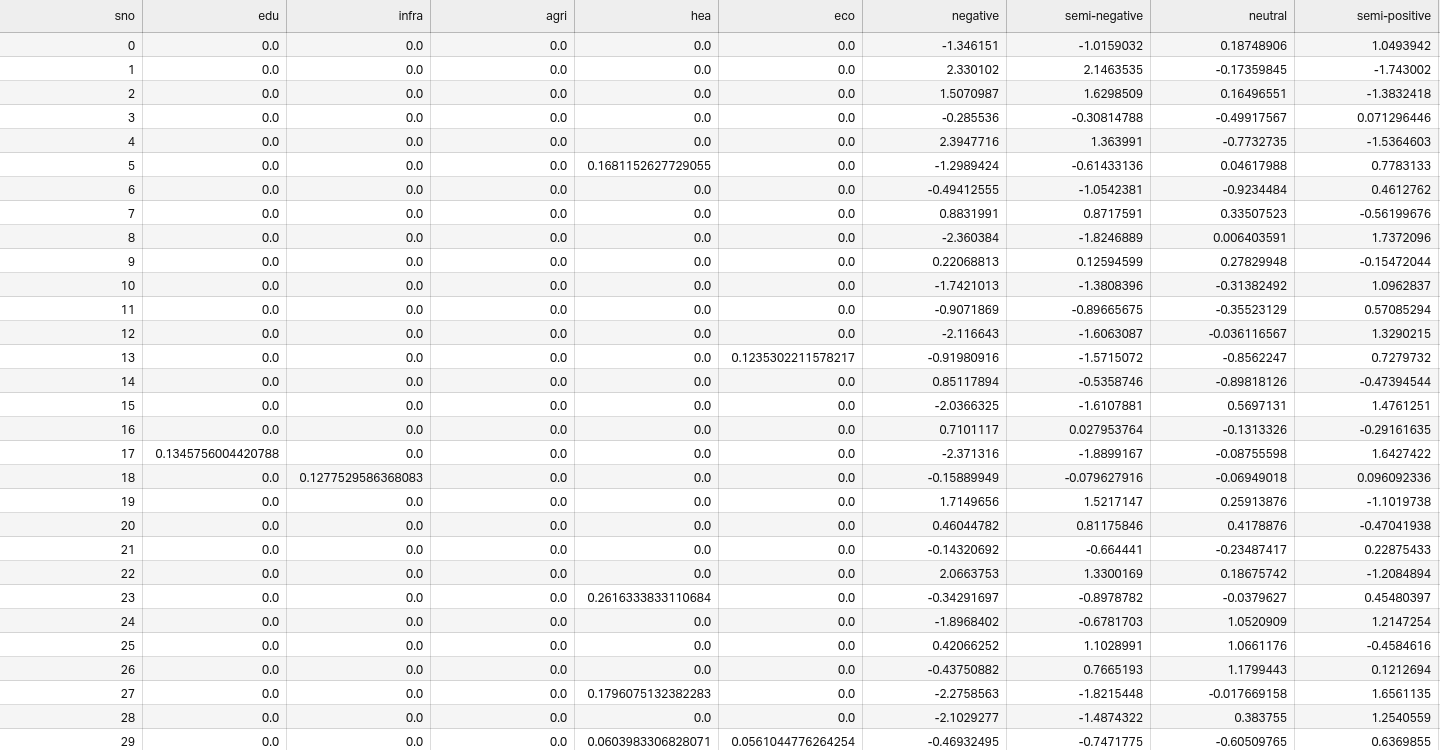


Figure 7: Shown here is the data after sentiment analysis

* 1. Result comparison

After getting the final values for all the manifestos in categories for TF-IDF and sentiment analysis the data is compared.

For that, the average of the values per manifesto was calculated since in a manifesto there are many values of TF-IDF and only those values are only five, so for this reason a python script is been used which average out all the TF-IDF value per category in a manifesto so at the end per manifesto, there is only 5 TF-IDF value on each category.

After this, the graphs are plotted per category comparing the TF-IDF values of the parties which can be seen in the figure.

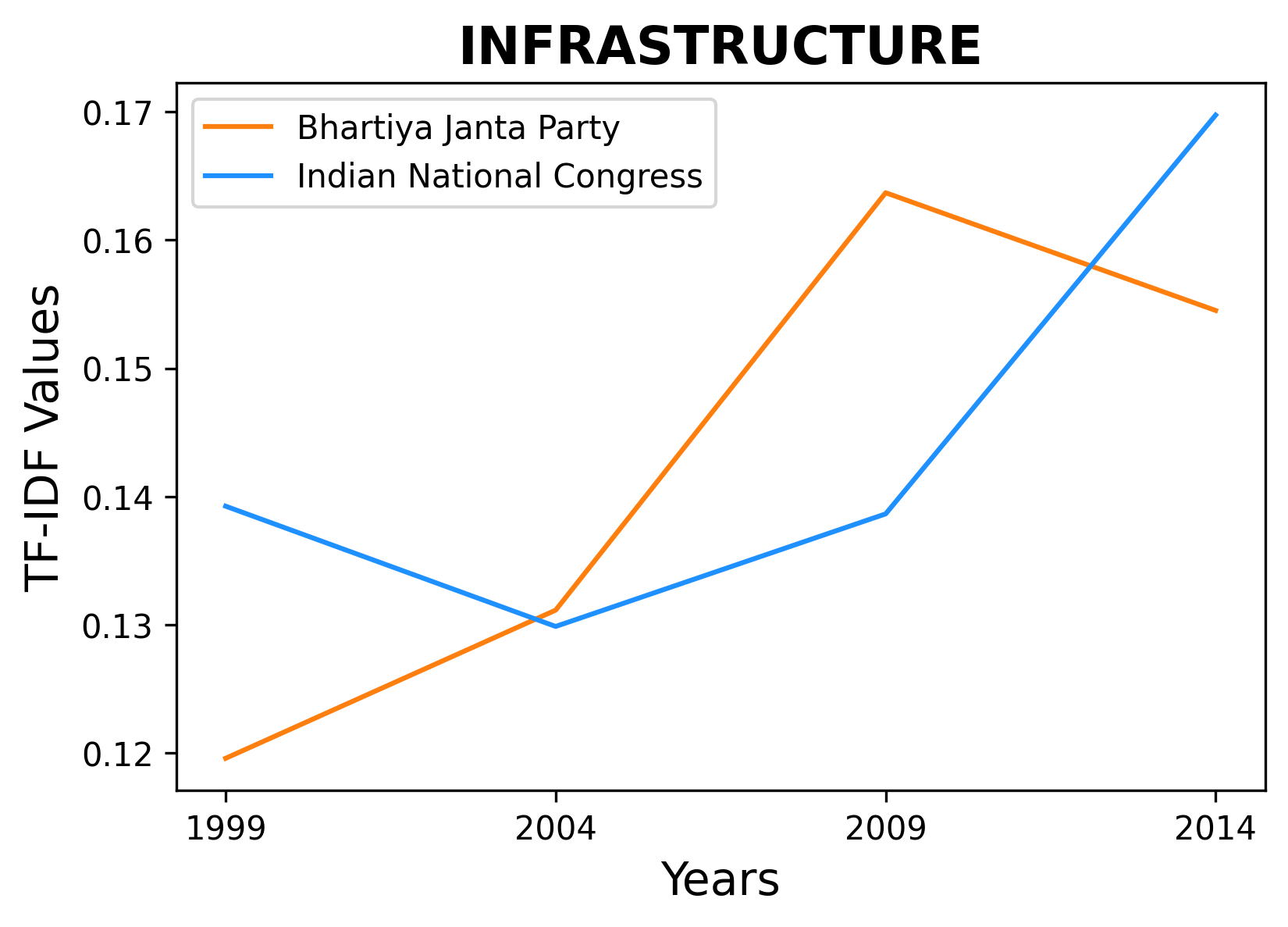


Figure 8:Graph Showing TF-IDF values of each party in the respective year

Now for the sentiment values per manifesto and per category 5 values are negative semi-negative neutral semi-positive and positive again at the need needed values are only one for this a python script is written a snippet is shown in the figure and the graphs that came as output is shown in the figure

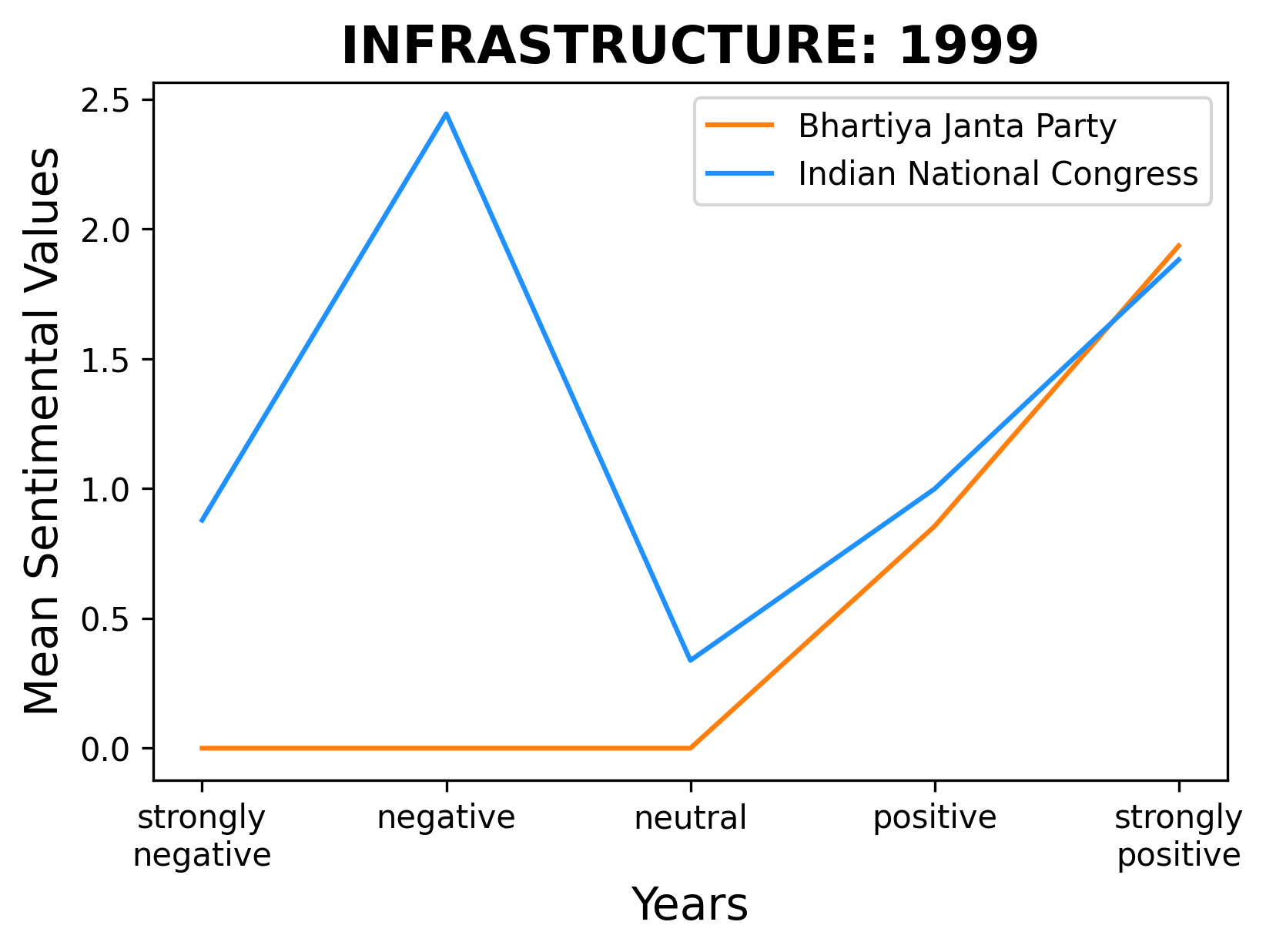


Figure 9: A graph showing infrastructure sentiment values of the party in the year 1999

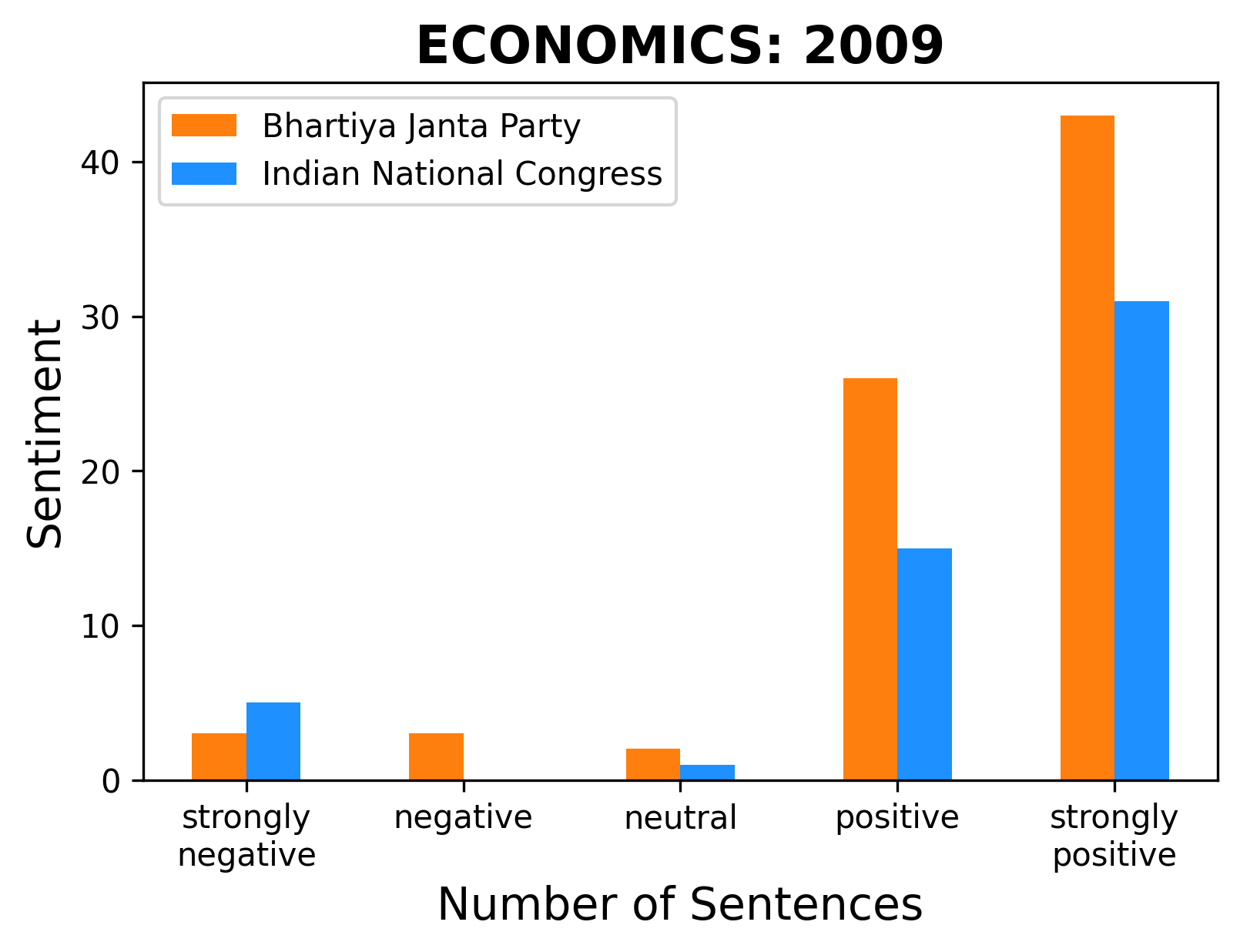
****

Figure 10: A bar graph comparing sentiment values

1. **OBSERVATIONS**

The following observation may be made in light of the data and figure above

* It has been noticed that the manifesto of the ruling party tends to be more upbeat. This might be attributable to the fact that different strategies they employ and their effects are highlighted.
* As an alternative to the aforementioned argument, the main opponent's manifesto is more critical as they work to hold the current administration responsible for its activities.
* In terms of the mood expressed in its manifesto, the BJP has often tended to be more extremist. This is consistent with the party's principles.
* For INC, on the other hand, it is seen as a right neutral or neutral party that is consistent with the views expressed by the party, which are typically more positive.
* According to the following data:

| **Strongly Negative** | **Negative** | **Neutral** | **Positive** | **Strongly Positive** |
| --- | --- | --- | --- | --- |
| 0.3732 | 0.1122 | 0.0651 | 0.0094 | 0.0544 |

* These values represent the difference in every sentiment category between BJP and Congress and it was found that BJP depicts stronger sentiment overall than Congress despite the fact that they are the ruling party at the time or not.
* Among sentiments also BJP shows more negative sentiments as compared to the positive ones. Again affirming the above observations.

1. **CONCLUSION**

According to the research mentioned above, the political party adheres to its ideals as seen by the nation's people as a whole. Whether they are the ruling party or not, they may feel differently, but this still obscures their feelings.

In contrast to Congress, which is the liberal or central-wing party, the BJP, which is the conservative or right-wing party, expresses greater feelings in its manifesto.

When it comes to critiquing the ruling party and praising its accomplishments while in power, Congress has demonstrated negative and positive emotions, respectively. In contrast, the Bharatiya Janata Party (BJP) expresses very strong positive emotion while promoting its accomplishments while the ruling party and is quite negative when criticizing the opposition.

1. **REFERENCE**
2. Prediction of Election Result by Analyzing Twitter Data Via Constituency Model by Ayush Satti and Deepanshu Bhardwaj (<https://www.researchgate.net/publication/360069318_Prediction_of_Election_Result_by_Analyzing_Twitter_Data_Via_Constituency_Model>)
3. Studying Indian Politics with Large-scale Data: Indian Election Data 1961–Today by Francesca R. Jensenius and Gilles Verniers (<https://www.researchgate.net/publication/321175519_Studying_Indian_Politics_with_Large-scale_Data_Indian_Election_Data_1961-Today?enrichId=rgreq-101edebb3a463b8f271f731127b46e9c-XXX&enrichSource=Y292ZXJQYWdlOzMyMTE3NTUxOTtBUzo1NjM1NjAzODAxMDQ3MDRAMTUxMTM3NDY3MzUxNQ%3D%3D&el=1_x_2&_esc=publicationCoverPdf>)
4. Analyzing Political Manifestos through Natural Language Processing and Dashboarding by André Ferreira (<https://towardsdatascience.com/analyzing-political-manifestos-through-natural-language-processing-and-dashboarding-4ad1d62d6b9a>)
5. Sentiment Analysis Using NLP: The Go-To Guide by Sriram Kothandaraman (<https://www.aviso.com/blog/sentiment-analysis-using-nlp>)
6. NLP: Introduction To NLP & Sentiment Analysis by Farhad Malik (<https://medium.com/fintechexplained/sentimental-analysis-an-introduction-7fc21d9b8625>)
7. Natural Language Processing (NLP) by IBM Cloud Education (<https://www.ibm.com/cloud/learn/natural-language-processing>)
8. K. Sparck Jones. "A statistical interpretation of term specificity and its application in retrieval". Journal of Documentation, 28 (1). 1972.
9. G. Salton and Edward Fox and Wu Harry Wu. "Extended Boolean information retrieval". Communications of the ACM, 26 (11). 1983.
10. G. Salton and M. J. McGill. "Introduction to modern information retrieval". 1983
11. G. Salton and C. Buckley. "Term-weighting approaches in automatic text retrieval". Information Processing & Management, 24 (5). 1988.
12. H. Wu and R. Luk and K. Wong and K. Kwok. "Interpreting TF-IDF term weights as making relevance decisions". ACM Transactions on Information Systems, 26 (3). 2008.
13. TF-IDF Vectorizer scikit-learn by Mukesh Chaudhary <https://medium.com/@cmukesh8688/tf-idf-vectorizer-scikit-learn-dbc0244a911a>
14. BERT Explained: State of the art language model for NLP by Rani Horev <https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270>
15. What is Sentiment Analysis? A Complete Guide for Beginners by Manish Shivanandhan <https://www.freecodecamp.org/news/what-is-sentiment-analysis-a-complete-guide-to-for-beginners/>