International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-3, September 2019

NLP Based Text Analytics and Visualization of Political Speeches

Paritosh D. Katre

Abstract: ***This paper presents a case study of implementing computational methods like Natural Language Processing (NLP) to perform Text Analytics and Visualization on political speech transcripts. The speech transcripts are published on websites, social media, and documents in large volumes and multiple languages. These transcripts are available in unstructured textual format and thus they are a part of big-data requiring analytics to derive insights from it. In this experiment, a significantly large volume of speech transcripts are analyzed and graphical visualizations are generated such as Lexical Dispersion Plot, Time Series Plot, WordCloud, Bar-Graphs using various Python libraries. The study has been useful in identifying issues highlighted across a large number of speech transcripts. So far, the linguists have tried to perform analysis using manual linguistic approaches which are extremely time-consuming and complex to understand the Political Discourse. Our experiment of applying NLP based text analytics proves to be a very efficient technique for Political Discourse Analysis (PDA).***

Keywords: natural language processing (NLP), big data, political discourse analysis (PDA), text analytics and visualization

I. INTRODUCTION

Candidates from different political parties give speeches at different events daily. The transcripts of the speeches given by leading or influential political candidates are generally available for free access over the internet thereby making a large pool of unstructured big data. In the age of big-data challenges, this data can be a reliable source to perform Political Discourse Analysis (PDA) [1]. Political Discourse Analysis includes transcripts of speeches, debates, articles and other forms of documentation of political views. This paper presents a case study in which Natural Language Processing (NLP) and visualization techniques have been applied for analyzing the issues raised and the emphasis given to them in the speeches. English transcripts of the speeches are considered in the purview of this case study.

II. RELATED WORK AND NEED

Data science is a fine blend of statistics and computational thinking (CT); it is a useful methodology for learning disciplinary concepts in a variety of fields, from science to social Science [2], and from political science and to humanities. Political speeches are a popular concept of study in the area of linguistics. Previously, a lot of work has been done on analyzing political speeches, pre-election speeches and the critical discourse [3] by linguists using manual linguistic approaches. Junling Wang has explored to analyze presidential speeches, which focused on relationships between language, ideology, and power [4]. Another linguistic study by Chang Pu was based on a discourse analysis of presidential speeches. It aimed at uncovering implicit meanings hidden in the speech [5]. The transcripts of such political speeches form a large amount of textual information. But text is unstructured data that requires processing to extract useful information from the unstructured data [6]. Therefore, to perform political attitude study and election campaign study [7] based on the numerous transcripts of speeches available over the internet, the use of advanced big-data, web-mining as well as semantic web techniques seems to be a promising direction [8]. Extracting knowledge from unstructured transcripts of speeches is limited by the ability of computers to understand the meaning of human language [9]. Therefore, there is a need to analyze campaign data being a niche business [10]. However, NLP based approaches need to be further developed for effectively catering to this requirement.

III. MOTIVATION

Although, a lot of work has been published to analyze political discourse by using manual linguistic approaches, analyzing a few hundreds of speeches with manual methods is a complex and time-consuming process. Therefore, it is decided to conduct this experiment of analyzing the speech transcripts of political candidates using computational methods like Natural Language Processing (NLP) and Visualization.

IV. BIG DATA POTENTIAL

A. Properties of the Data

On most occasions, political candidates give multiple speeches per day. These speeches run over long lengths whilst covering a myriad of diverse issues and topics. In India, such speech transcriptions are also found in various regional languages. Therefore, the problem of Political Discourse Analysis covers the three defining properties of big-data which are – Volume, Variety, and Velocity. Volume refers to the amount of data (length and number of speech transcripts), Variety refers to multilingualism and speeches delivered by a large number of candidates belonging to different parties and Velocity which refers to the speed at which this data is being collected and published.

Manuscript published on 30 September 2019
* Correspondence Author
Paritosh D. Katre*, Computer Engineering Department, Vishwakarma Institute of Information Technology, Pune, India. Email: katreparitosh@gmail.com

The textual transcripts are typically unstructured forms of information requiring NLP processing. Thus, it is observed that the speech transcripts have all the characteristics of big-data.

B. Sources of the Data

Generally, the offices of leading political candidates create official speech transcripts and make them freely available on their official websites. Such speech transcriptions are usually published as a legal protection from misquoting by the newspapers and the subsequent litigations. Sometimes, the event organizers also create the transcripts for their documentation and pool it online. These speeches which are available in the form of website content (web-pages) or word documents published on the websites can be scraped using Web Scraping Tools. Such content is available from hundreds of regional and national party websites, event websites or institutional websites.

V. ELEMENTS OF POLITICAL SPEECHES

Before jumping into applying NLP techniques on the speeches, understanding the basic elements of a political speech is important. The elements help us define the relevant and irrelevant parts of the speech from an NLP perspective. The removal of irrelevant content from the speech reduces the dimensionality of the data to be processed. Although elements of the speech may vary slightly, the identification of most predominant elements can help in defining the NLP rules.

The introduction of a speech contains citing the event where the speech is being delivered, greetings and finally making opening statements. The main body of the speech contains various social issues and concerns, criticism and ideological points. The speech also contains good emphasis on promises and assurances on reforms and schemes for public benefit. Understanding these common elements can be helpful in text analytics.

VI. METHODOLGY

This section defines the methods and technologies implemented in this case study. This section will encompass methods right from the collection of data (speeches) to the analytical models used. Fig. 1 illustrates the process flow diagram. The sources of data have been concealed for an objective focus on the technical aspects of this study.

A. Technologies Used

Table- I: List of Technologies Used

Process Steps	Technologies/Libraries Used
Scraping	Selenium Webdriver, requests, re, time
Data Extraction	nltk.corpus.PlainTextCorpus Reader, glob, re, pandas, datetime, stirng, collections
Managing, Storing and Cleaning	pandas, nltk, stopwords, re, string
NLP based Analytics and Visualizations	pandas, numpy, nltk, matplotlib, WordCloud, Counter

In this project, Python has been chosen considering it is an open-source, high-level, interpreted programming language that provides excellent functionality for processing linguistic data. Also, it supports a wide range of scientific libraries for computing [11]. With the base of python, Natural Language Toolkit (NLTK) is used for the processing of textual corpora throughout the experiment. The entire process code is done in the Jupyter Notebook environment.

To name a few, ‘re’, a python library is used to work with Regular Expressions to extract meaningful metadata from the text files and accumulate it in the dataframe. In this experiment, the multi-file corpus consisted of over 259 speech transcripts, therefore the usage of the Plain Text Corpus Reader module of NLTK enabled us to access and read the text files.

B. Scraping

The speech transcriptions are scraped from the official websites to ensure authenticity. To do that, Selenium Web Driver is used, which is a Web Automation Tool. The path of the website is provided through the python block of code and scraping functions are built to customize the scraping of speeches. The data scraped is then segregated into different batches according to the year of speech and stored into different folders.

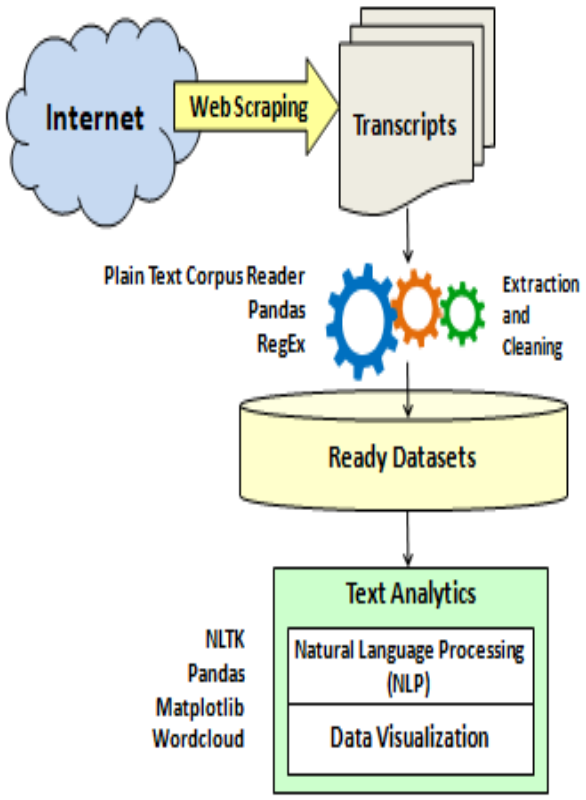

Fig. 1.Process Flow Diagram

C. Extraction of Data

To extract meaningful metadata like date of the speech and title of the speech, PlainTextCorpusReader is used to access the folder of .txt files and read the files. RegEx pattern matching technique is used to extract the date and title of the speech and store them into subsequent dataframe.

Once extracted, the curated dataframe is exported to the 'pickle' file for storage and further use. This procedure is carried out for all the speech transcripts in a year-wise manner.

The date and place information of the speech must be captured to interpret the emphasis given on the issues relevant to the local audiences. The date may signify the impact of political events around that period and its reflection in the speech.

D. Managing, Storing and Cleaning of Data

Various dataframes are consolidated into one single dataframe and the data is further cleaned using 'Pandas' methods, 'string' methods, regular expression, etc. Using 'nltk' and 'string' libraries, the speeches are stripped from punctuation, converted to lowercase, tokenized, stopwords (English) are dropped, the date is formatted as DateTime object, etc.

It is also essential to remove the HTML tags, if present, from the extracted web-pages to get plain text of the speech transcriptions. Such removal of noise and unwanted tags helps in preparing the data for further analysis.

VII. DETAILED ANALYTICS

The detailed analysis section is a vital part of the experiment because the results from this section throw light on the final results and give us a direction to interpret the results concisely.

The analytical section includes results in the form of quantitative values and visuals based on data. Word clouds are generated to observe the frequency and occurrence of different words being used in the speeches. Word clouds can be used to highlight significant textual data points from the data.

A. General Overview of Speech Transcripts

Table- II provides a general overview of speech transcripts and statistics about how many samples are used, what is the average length of speeches, number of speeches per day/week/month/year, etc. The variance in the statistical values indicates that as the audience, event, time, and location change, the quantitative parameters of the speeches change. Furthermore, the subjects of the speech affect the quantitative parameters to a large extent because certain issues require an extended elaboration while some may not.

Table- II: Statistics of Speeches Collected

Statistical Item	Speeches
Number of Speeches	259
Number of days in which at least one speech is delivered	211
Maximum number of Speeches given in a day	4
Average number of words in speech transcripts	1256

Fig. 2 shows a statistical graph of the speeches extracted from the official source to characterize the dataset used in this experiment.

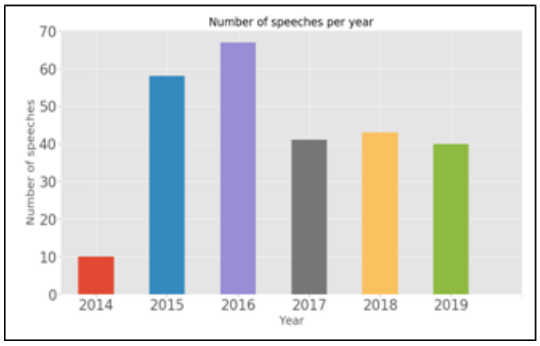

Fig. 2.Speeches Year-Wise

From Fig. 2, it can be inferred that the corpus contains a maximum number of samples from the year 2016. Similarly, graphical visualizations are generated to analyze the trend of speeches per day in a week, speeches per week as well as speeches per month. The comparison of these bar-graphs within different candidates leads to comparative analysis.

B. Lexical Dispersion Plot

The importance of a word/ token can be estimated by its dispersion in the corpus. The tokens vary in their distribution throughout the text which tells when or where certain tokens are used in the text. Lexical dispersion is a measure of the word's homogeneity across the corpus. Word distributions can be generated to get an overall sense of topics, spread of the topics and topic shifts. A lexical dispersion plot depicts the occurrences of the word and the frequency with which they appear from the beginning of the corpus. Thus, lexical dispersion plots are useful in identifying patterns.

The x-axis of lexical dispersion plot shows the word offset i.e. the word occurrence and frequency throughout the speeches and the y-axis shows the specific issues.

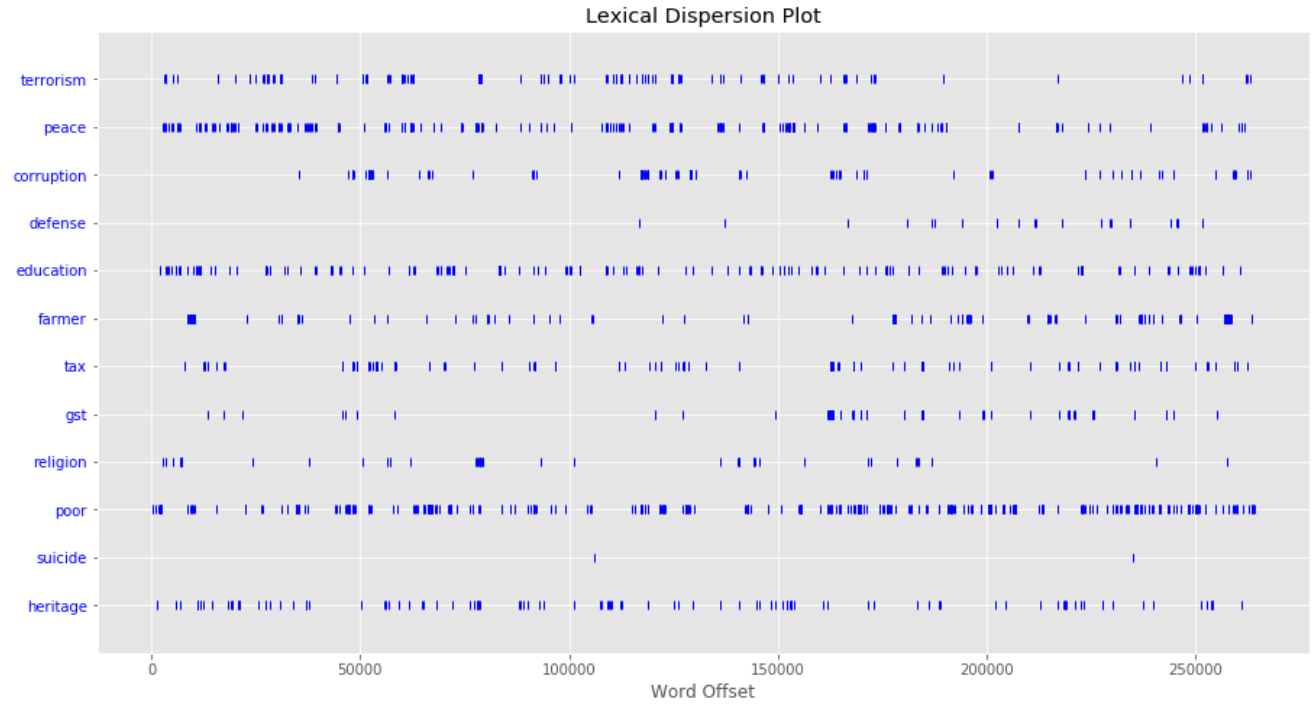

Fig. 3.Lexical Dispersion Plot of 259 Speeches

The lexical dispersion plot shown in Fig. 3 indicates that the issues of 'peace', 'education', 'heritage' etc. appear uniformly through all the speeches. The issue 'poor' i.e. poverty observes a very high word offset and consistent spread through the speeches which tells that issues related to poverty are getting highlighted in the speeches to contextualize the national priority. Whereas, topics of 'GST', 'defense', 'suicide', etc. have a low word offset and spread which indicates that the emphasis given to these issues is dependent on factors like time, audience, event, and overall situation in the country. Thus, a lexical dispersion plot is an effective tool to interpret issues highlighted in the speeches.

In Table- III, the issues from lexical dispersion plot are segregated as per dispersion type.

Table- III: Segregation of Issues by Dispersion Type

Spread & Dispersion Type	Candidate A Topics
Issues with High Dispersion	Peace, Poor (Poverty),
Issues Uniform Dispersion	Heritage, Education, Terrorism
Issues Low Dispersion	Corruption, Gst, Tax, Farmers
Issues having Lack of Coverage	Suicide, Defense

C. Time Series Plot

In predictive analytics, time is a very important factor that needs to be considered. The pattern recognized or predicted needs to be studied and verified with respect to time. Time series data is nothing but a series of data ordered in time. Fig. 5 captures the comparative time-series trend between the issues 'peace' and 'terrorism' as an example.

Holistically, it can be inferred that although the emphasis given to the issue 'peace' fluctuates, it is consistent over the period while the fluctuations for the issue 'terrorism' alleviate and eventually depart. Thus, the comparative study of time-series data with regards to issues raised over time can be used to predict what kind of issues the public can expect from the political candidate.

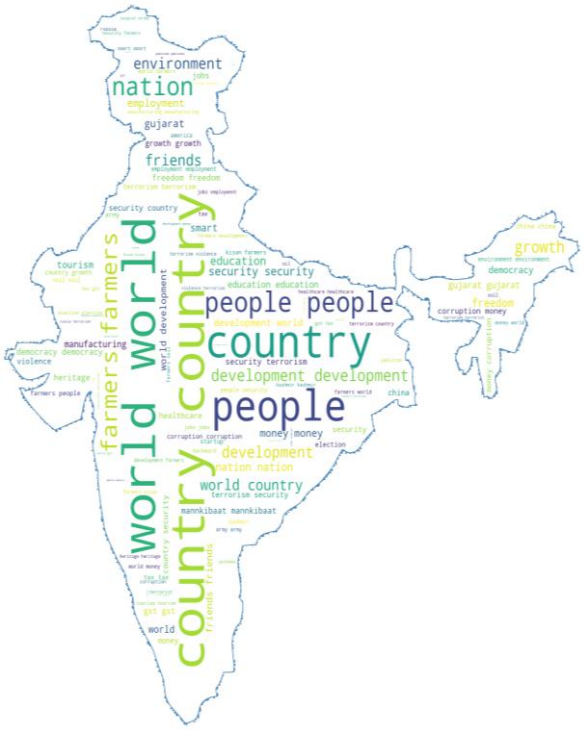

Fig. 4.WordCloud Highlighting the Topics

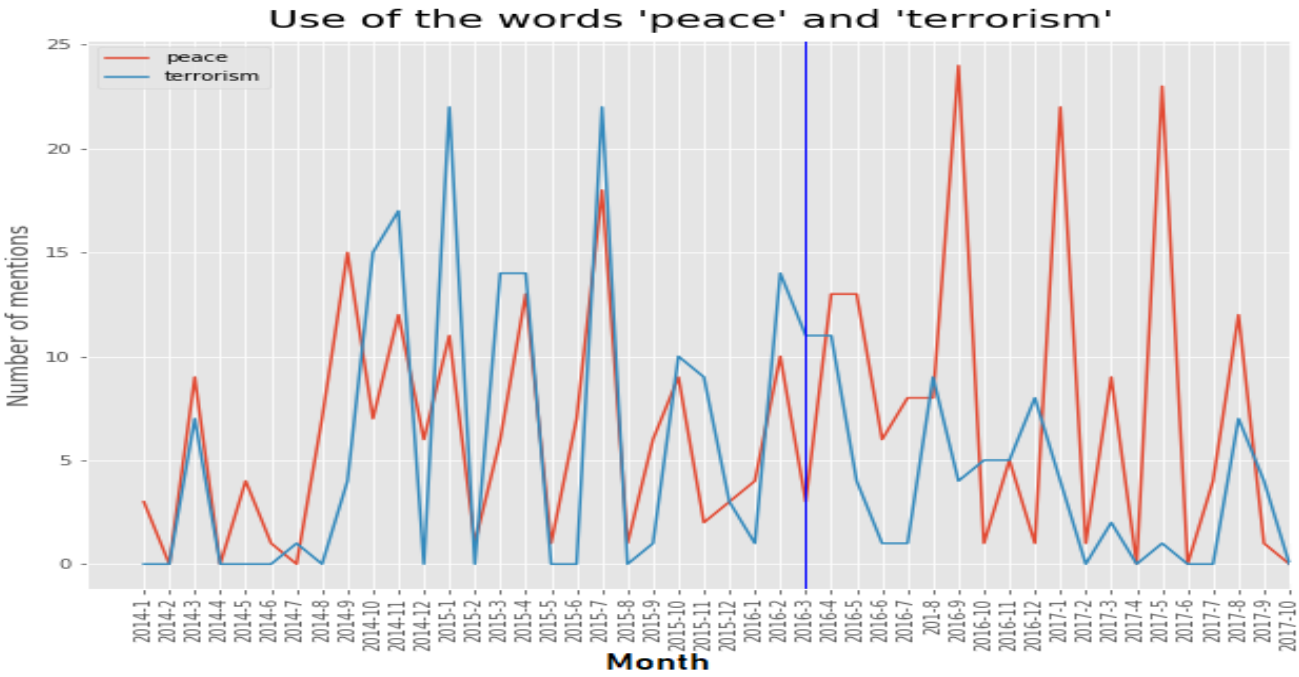

Fig. 5.Time Series Line Plot

D. WordCloud Representation of 259 Speeches

In order to get a quick holistic impression of the speech transcripts under consideration, Word Clouds are created with the help of 'wordcloud' library. Word clouds are a simple and intuitive visualization technique that is often used to provide a first impression of text documents. WordClouds show the most frequent words of the text as a weighted list of words in a specific spatial layout, for example, sequential, circular or random layout [12]. The font sizes of the words change as per their relevance and frequency of occurrence, and other visual properties like colour, position, and orientation often vary for aesthetic reasons or to visually encode additional information [13].

From the sorted Frequency Distribution of tokens, certain tokens are filtered based on the list of words of interest. This text file is exported and referred to as 'token_of_interest.txt' as it contained tokens which are to be added into the word cloud. On a mask of Indian map, the tokens of interest are imprinted and word cloud is generated.

Observations from the word-cloud as shown in Fig. 4:

- Word clouds, although based on rudimentary concepts of Frequency Distribution (FD) and Frequency Count (FC), they provide a holistic understanding of which words are getting emphasis.
- Firstly, 'world', 'country' and 'people' are the tokens having the most frequency count. A hypothesis can be built from that the candidate might be most concerned about certain 'worldly matters' and the 'people'.
- Furthermore, usage of words like 'development', 'security', 'terrorism' and 'growth' are given a secondary level emphasis which hypothesizes that the candidate might be concerned about the development, growth and security aspects of the country or region. Also, terrorism might be a grave issue which the country still needs to deal with.
- Finally, words like 'tax', 'farmers/Kisan', 'education' and countries like 'Russia', 'America', 'China', etc. appear comparatively with less frequency.

E. Scrutinizing Trends/Patterns using Bar-Plot

Bar graphs depict how data is spread over certain potential values. Although a simple looking graph, a bar graph has the capacity to capture the essence of data by judging the spread and answer certain questions.

Fig. 6 shows a representation of 'Topic Name vs. Number of mentions'. For each token name provided, the frequency of the occurrence is calculated and the graph is generated. 'Matplotlib', a data visualization library is used to construct this graph.

Inferences gathered from the graph:

- It is clearly visible that the topic of 'technology' got highlighted the most in the speeches collected.
- On the contrary, the religious issues have got negligible emphasis in the bar-graph.

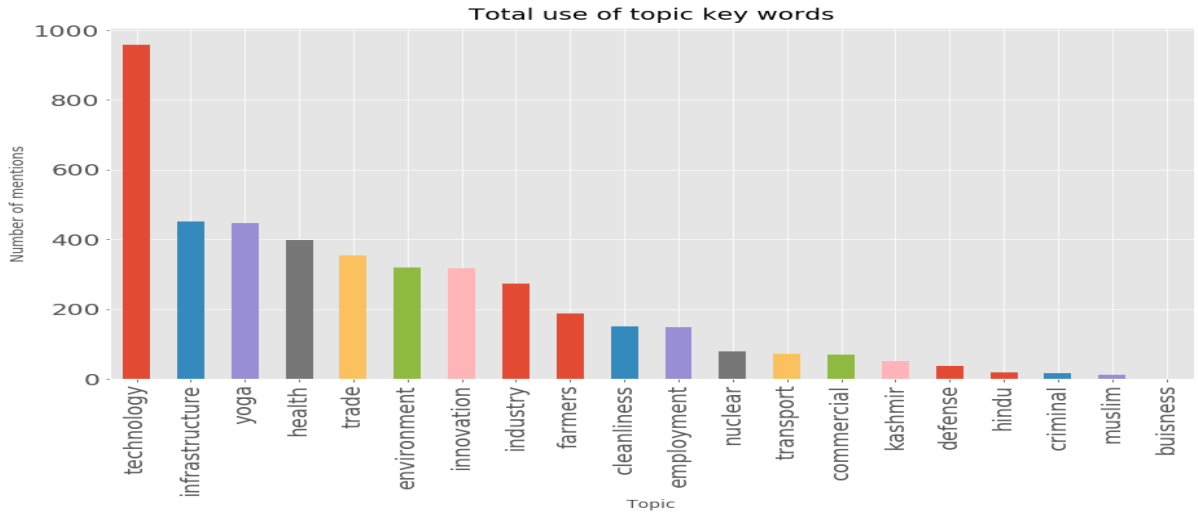

Fig. 6.Topic Bar Plot Highlighting Issues over the 259 transcripts collected

VIII. CONCLUSION

From the above discussion, it can be concluded that Natural Language Processing (NLP) techniques are extremely helpful to understand and interpret the transcripts of political speeches. Text Analytics is useful in tracing issues and topics occurring in the speech transcripts.

Lexical Dispersion Plots effectively depict how multiple topic keywords appear throughout the corpora. The rise and fall of topic keywords in Time Series Plot is a good predictive measure to analyze political discourse. WordClouds and Topic Bar Plots have the capacity to give a holistic understanding of the speech transcripts. Thus, graphical visualizations reflect political ideologies and how they evolve over time.

Thus, Text Analytics and Visualization of political speech transcripts has the potential to assist the political strategists.

IX. FUTURE WORK

In future, it is proposed to extend this work to comparative text analytics on speech transcripts by different candidates from political parties. This will lead to a new direction wherein the results from the speeches by different candidates can be juxtaposed and compared. Topic modelling techniques like Latent Dirichlet Allocation can be implemented to derive greater sense from the corpus. The speech transcripts can also be analyzed along with online news reports and discussions of public on social media. India is a country where languages and dialects changes every significant distance you travel, therefore systems should be developed to analyze the political discourse in multiple languages. India being the largest democracy in the world and being a politically vibrant nation, the challenges to analyze such heterogeneous big-data grow exponentially.

REFERENCES

1. T. A. V. Dijk, “What is political discourse analysis.”, Belgian journal of linguistics, vol. 11, no. 1, pp. 11-52, 1997.
2. V. Wart and S. Jane, “Computer science meets social studies: Embedding cs in the study of locally grounded civic issues.”, Proceedings of the eleventh annual International Conference on International Computing Education Research, pp. 281-282, 2015.
3. O. M. Ayeomoni and O.S. Akinkuolere, "A Pragmatic Analysis of Victory and Inaugural Speeches of President Umaru Musa Yar'Adua.", Theory & Practice in Language Studies, vol. 2, no. 5, 2012.
4. J. Wang, "A critical discourse analysis of Barack Obama’s speeches.", Journal of language teaching and research, vol. 1, no. 3, pp. 254-261, 2010.
5. C. Pu, "Discourse Analysis of President Bush’s Speech at Tsinghua University, China." Intercultural Communication Studies, vol. 16, no. 1, pp. 205-216, 2007.
6. A. Kaur and D. Chopra, “Comparison of Text Mining Tools”, 5th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions), AIIT, Amity University Uttar Pradesh, Noida, India, pp. 186-192, 2016.
7. V. Rajaraman, “Big Data Analytics.”, Resonance, vol. 21, no. 8, pp. 695-716, 2016.
8. M. Schatten, J. Ševa and B. Okreša-Đurić, "An introduction to social semantic web mining & big data analytics for political attitudes and mentalities research." European Quarterly of Political Attitudes and Mentalities, vol. 4, no. 1, 2015.
9. O. Muller, I. Junglas, S. Debortli and J. V. Brocke, "Using text analytics to derive customer service management benefits from unstructured data.", MIS Quarterly Executive, vol. 15, no. 4, pp. 243-258, 2016.
10. D. W. Nickerson and T. Rogers, “Political campaigns and big data." Journal of Economic Perspectives, vol. 28, no. 2, pp. 51-74, 2014.
11. S. Bird, E. Klein and E. Loper, Natural language processing with Python: analyzing text with the natural language toolkit., O'Reilly Media, Inc., 2009.
12. S. Lohmann, J. Ziegler and L. Tetzlaff, "Comparison of tag cloud layouts: Task-related performance and visual exploration." IFIP Conference on Human-Computer Interaction, Springer, Berlin, Heidelberg, pp. 392-404, 2009.
13. S. Lohmann, F. Heimerl, F. Bopp, M. Burch and T. Ertl, "Concentri cloud: Word cloud visualization for multiple text documents." 2015 19th International Conference on Information Visualisation, IEEE, pp. 114-120, 2015.

AUTHORS PROFILE

Paritosh D. Katre is currently a final year student of BE in Computer Engineering at Vishwakarma Institute of Information Technology, India. His areas of interest are Natural Language Processing, Big-Data, Artificial Intelligence, Data Mining and Visual Analytics.