Analyzing the Voters' View: Emotional Sentiment vs Long Run Development, of Casting Votes in Elections using NLP

Rajini A.
Department of Mathematics and Statistics
Bhavan's Vivekananda College of Science, Humanities and Commerce
(OU)
Secunderabad, India
rajinigupta.peddi@gmail.com

Raju Kommarajula
Sr.Data Scientist
Verizon
India
rajstats.2010@gmail.com

Chakravadhanula Naga Pranav
Bhavan's Vivekananda College of Science, Humanities and Commerce
(OU)
Secunderabad, India
chakravadhanula.pranav@gmail.com

Mummadi Sai Prasanna
Bhavan's Vivekananda College of Science, Humanities and Commerce
(OU)
Secunderabad, India
spmummadi2301@gmail.com

2023 International Conference on Computational Intelligence, Networks and Security (ICCINS) | 979-8-3503-1379-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICCINS58907.2023.10450030

***Abstract*—In the realm of electoral politics, campaigning serves as a crucial strategy for politicians to market themselves and persuade voters to support them. Voters must critically analyze election-related speeches of candidates and discern the underlying goals of these speeches to make informed decisions when selecting leaders. This research aims to investigate the basis upon which voters cast their votes, specifically examining whether emotional sentiment or considerations of long-term development play a significant role. The study utilizes Natural Language Processing (NLP) techniques and relies on a dataset comprising videos of campaigns converted into text featuring speeches made by candidates from the parties under consideration. NLP techniques are used to gain insight into voters' decision-making processes and determine their inclination towards emotional sentiment versus long-term development. The analysis involves cleaning the speech text, followed by sentiment analysis to assess the positivity and negativity expressed by the candidates. Furthermore, topic modeling is performed to identify the themes discussed in the speeches. This research found that voters tend to cast their votes by considering the candidates' long-term development plans and positive attitudes. Emotional sentiments do not appear to manipulate voters, and negative statements made by candidates are generally not appreciated. The findings of this research provide valuable insights into the factors that shape electoral choices.**

Keywords*—Campaigning, Emotional sentiment, Long-term development, NLP, Topic modeling.

I. INTRODUCTION

One of the best strategies politicians use to market themselves during elections is campaigning. It mostly aids in persuading others to vote for them. It is highly important for voters to analyze their candidates' election-related speeches and determine the key goals of the speech while choosing their leader. To find on what basis the voters are casting their votes to a particular candidate is a very interesting area to study. To find whether the voters are casting their vote based on emotions or long-term development is the key objective of our research. The process of analyzing these speeches includes speech transcription, translation into English, text mining, text analysis using NLP (Natural Language Processing) techniques, and present visualizations utilizing tools available. The figure presented as Fig.1 illustrates the step-by-step flow chart that outlines the analytical process employed in the article. This visual representation visually depicts the sequential progression of activities and procedures involved in the analysis, providing a clear and concise overview of the methodology employed.

Computational thinking and statistics are combined in data science. We can analyze several aspects of a text with the aid of tools from data science. Since NLP (Natural Language Processing) techniques are currently only available for the English language, it is difficult to analyze speeches delivered by politicians in their regional language. Therefore, transcription and translation are the main steps that must be followed to analyze speeches delivered by politicians during the pre-election campaign.

The technique of sentiment analysis, employed within Natural Language Processing (NLP), allows for the determination of the sentiment expressed in a text. NLP utilizes sentiment analysis to analyze various aspects such as positive and negative sentiment, as well as emotion recognition. On the other hand, topic modeling serves as a machine learning technique that extracts abstract topics from text by grouping relevant words associated with each topic.

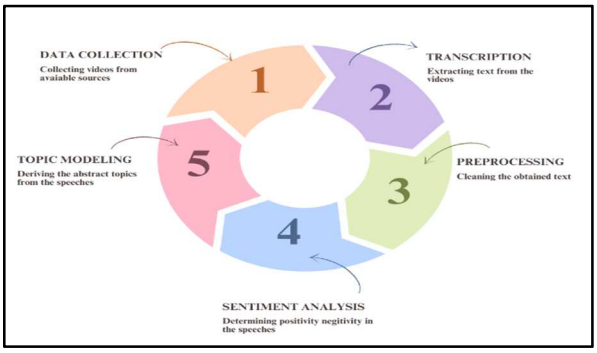

Fig. 1 Flow chart of the process of analysis

II. LITERATURE REVIEW

Dr. Manpreet Kaur, Dr. Rajesh Verma and Dr. Sandeep Ranjan (2022)[1] conducted a location-based sentiment analysis in order to determine the emotions that were being expressed in the tweets. Martin Haselmayer (2021)[2] delved into broader issues surrounding the impact of positive and negative political communication on elections and democracies. Additionally, authors explored the role of mood in electoral competition, shedding light on its function and significance. Yogev Matalon, Ofir Magdaci, Adam Almozlino and Dan Yamin (2021)[3] by utilizing a Random Forest model that takes into account user attributes and employs Natural Language Processing techniques on the Source text, the researchers successfully predicted whether a retweet would result in Online Incivility (O.I.). This case study highlights the significant role O.I. plays in shaping online political discourse, emphasizing its importance in understanding and analyzing such interactions. Gavin Abercrombie and Riza Batista-Navarro (2020)[4] found limited indications suggesting that parliamentary language exhibits greater stability compared to ordinary English. They also observed that semantic changes in parliamentary language largely align with those documented in other corpora, suggesting consistency across different contexts. Budi Haryanto, YovaRuldeviyani, FathurRohman, Julius Dimas T. N., Ruth Magdalena and Muhamad Yasil F (2019)[5] collected the data by engaging in comment sections of significant Indonesian news media's Facebook posts. The study employed the Naive Bayes Classifier method in data mining to effectively categorize opinions. Zulfadzli Drus and Haliyana Khalid (2019)[6] showcased that a majority of articles utilized the opinion-lexicon method to analyze the sentiment of text in social media. The research involved extracting data from microblogging platforms, with a particular focus on Twitter, and employing sentiment analysis techniques across various domains such as business, politics, healthcare, and more. Veny Amilia Fitri, Rachmadita Andreswari and Muhammad Azani Hasibuan (2019)[7] reported the results of testing data using different algorithms in RapidMiner tools. Among the algorithms tested, the Naive Bayes method achieved the highest accuracy rate of 86.43%, outperforming the Decision Tree and Random Forest algorithms, which achieved accuracies of 82.91%. Prima Widyaningrum, Yova Ruldeviyani and Ramanti Dharayani (2019)[8] presented their findings regarding sentiment analysis. The results indicated that there were 547 instances of "anticipation" sentiment and 728 instances of "trust" sentiment. These findings served as valuable insights for the company, leading them to develop chatbot technology. However, it's worth noting that despite the positive assessment of the analysis, there still exists negative sentiment within the community. Paritosh D. Katre (2019)[9] the use of text analytics based on Natural Language Processing (NLP) was experimented with, revealing its high effectiveness in political discourse analysis (PDA). Claus Boye Asmussen and Charles Møller (2019)[10] introduced a comprehensive framework that offers a step-by-step approach for researchers interested in topic modeling. This framework includes a code template that enables researchers to effectively conduct topic modeling on their respective cases. Widodo Budiharto and Meiliana (2018)[11] devised an algorithm and technique to identify crucial information, determine the most frequently used words, train a model, and predict sentiment polarity. By utilizing the R language, experimental results revealed a prediction that one party is likely to emerge as the winner in the election. Justin Grimmer (2007)[12] introduced a statistical model designed to assess the priorities of political actors based on their statements, incorporating the rhetorical structure. This model was implemented on a dataset consisting of 24,000 press releases from senators in 2007. The model serves as a valuable tool to test hypotheses regarding the interactions between members of Congress and their constituents. Olessia Koltsova, Sergei Koltcov (2013)[13] made an intriguing discovery regarding the behavior of highly engaged Live Journal users. The research revealed that these users consistently allocated an equal amount of time to both "social/political" and "private/recreational" topics in their online activities. Doris Baum (2012)[14] conducted a study focusing on the utilization of topic data in speech analysis. The research involved examining a speech segment obtained from an automatic speech recognition transcript. The findings revealed that each segment could be effectively represented as a combination of themes using a domain-specific topic model. Kevin M. Quinn, Burt L. Monroe, Michael Colaresi, Michael H. Crespin, Dragomir R. Radev (2009)[15] conducted a comprehensive analysis of the Senate's agenda spanning from 1997 to 2004, employing a topic model. The implementation of this model allowed for a more comprehensive understanding of the dynamics within the democratic agenda, surpassing previous limitations in distinguishing and meaningfully associating speech subject types.

III. DATA AND METHODOLOGY

A. Data

News outlets recorded and uploaded speeches made by the participants in these gatherings to YouTube. Approximately 60 hours of video recordings classified into 30 speeches made by the members of the parties under consideration served as the only data source for this investigation.

The collected videos belong to three different parties. For better illustration of the results, they are named as party 1, party 2 and party 3.

The videos must be converted into text as raw data to conduct additional text mining and analysis. The Google API and Python's speech recognition module are used to transcribe the videos. Once we had the regional language version of the video's transcript, we translated it into English. The text was personally reviewed after it had been translated, and it was then put into a CSV-format spreadsheet and used for further procedures.

B. Methodology

The analysis encompasses a series of sequential steps. The initial step involves text cleaning to ensure data accuracy, followed by sentiment analysis to assess the positive and negative aspects of the speeches. The subsequent step entails topic modeling to identify the thematic content discussed by the candidates in their speeches. Each of these steps will be elaborated on in detail below.

C. Preprocessing

Data preparation is an essential step that must be completed before any text analysis because the text may

contain undesirable information such as stop words, numerals, and special characters like @, #, etc. These factors will reduce the accuracy of the analysis. By deleting these, the results are more accurate. Making all the words lowercase is the first step in data preparation. String functions are utilized for this purpose. The analysis is unaffected by punctuation. Since it is more accurate to analyze the text without them, removing them is the next stage in preprocessing. Additional symbols were also dropped along with the punctuation. Commonly used words like "is", "are", "the" etc. are considered stop words. Utilizing a stop word library from the NLTK corpus, they are eliminated. Utilizing the textblob library, the spellings are fixed. Next in the preprocessing procedure, the data is tokenized and stemmed using the Lancaster Stemmer from the NLTK package. The text is then lemmatized utilizing the textblob library.

The flow chart depicted as Fig. 2 illustrates the series of preprocessing steps undertaken in the analysis. This visual representation outlines the specific actions and procedures carried out prior to the main analysis, showcasing the necessary preparations and transformations applied to the data. By presenting these steps in a graphical format, the figure provides a comprehensive overview of the preprocessing methodology utilized in the study.

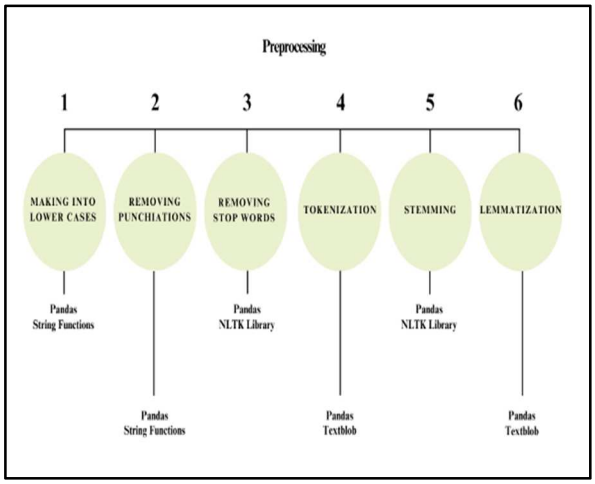

Fig. 2 Flow chart of preprocessing steps

D. Sentiment analysis

To conduct sentiment analysis on the data, the VADER (Valence Aware Dictionary and Sentiment Reasoner) library is utilized. This library is employed to evaluate the sentiment of the text and determine its valence and sentiment polarity. A list of words and their corresponding sentiment ratings make up the sentiment lexicon that VADER uses. Based on how frequently a term appears in the lexicon, every word in the text is given a sentiment score. The results are then added together to get the text's overall sentiment polarity, which ranges from negative to positive. By interpreting the sentiment polarity score, the text's sentiment is subsequently determined as favorable, negative, or neutral. This allows for a comprehensive understanding of the sentiment expressed within the text.

E. Topic Modeling

For topic modeling, the Latent Dirichlet Allocation (LDA) algorithm, which is a probabilistic statistical model based on unsupervised learning, is employed. This algorithm utilizes two hyperparameters, α and β, which correspond to the per-topic distribution and per-topic word distribution, respectively, and requires manual initialization. In the context of a document denoted as M, Z represents the topic assigned to the Nth word, while W refers to the specific word. As W (specific words) is the only observable variable and the others are latent, their identification is limited to the papers. θ represents a matrix in which the rows represent documents, and the columns represent topics. θ(i, j) denotes the likelihood that document i^{th} contains topic j^{th}. Like this, ϕ is a matrix where the words are the columns, and the topics are the rows. ϕ(i, j) denotes the likelihood that the word j will appear in a topic i. K unique words are generated for the themes based on the ϕ distribution. We calculated the number of words associated with each topic using this LDA, and we identified the topics in documents as a result. The Genism library and plotting tool pyLDAvis are used in the procedure.

The image displayed as Fig. 3 presents the working mechanism of the Latent Dirichlet Allocation (LDA) model, sourced from Wikipedia. This visual depiction illustrates the underlying processes and principles of the LDA model, highlighting how it operates and generates topics from a given corpus of documents. By incorporating this image, the figure serves as a valuable visual aid in understanding the functioning of the LDA model in the context of the article or study.

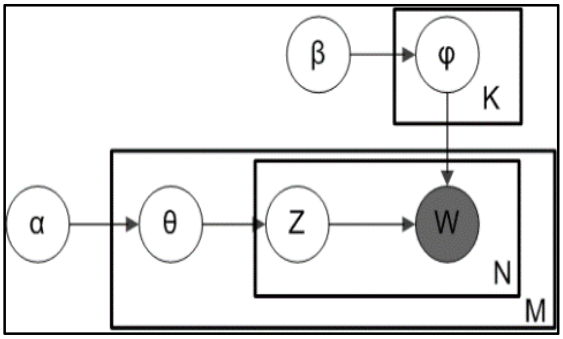

Fig. 3 Working mechanism of LDA model

IV. RESULTS

During our analysis, sentiment analysis was conducted on the speeches, yielding measurements for positivity, negativity, neutrality, and a compound score. These metrics provided insights into the overall sentiment expressed within the speeches. Compound score is basically “Normalized, Weighted Composite Score”. Based on compound score the sentiment of the speech is identified. We found that most of the speeches are positive, and few speeches are negative. Very few speeches fell under the neutral category.

Table I displays the compound scores obtained from the sentiment analysis of the speeches included in the study.

TABLE I. SENTIMENT ANALYSIS COMPOUND SCORES OF THE CONSIDERED SPEECHES

Speech number	Compound Score	Speech number	Compound Score
1	0.9843	16	0.7003

2	-0.9862	17	0.6398
3	-0.9973	18	0.8316
4	0.9973	19	0.9862
5	0.9987	20	0.193
6	-0.9393	21	0.9858
7	0.8561	22	0.97
8	-0.8779	23	0.9865
9	0.999	24	0.9371
10	-0.9869	25	0.9001
11	0.9877	26	0.9607
12	-0.9648	27	0.7845
13	0.9991	28	0.918
14	0.9988	29	0.9818
15	0.9949	30	0.1779

The speeches with compound score less than -0.5 are negative, if the compound score is in between -0.5 and 0.5 it is termed as neutral and if the compound score is greater than 0.5 it is termed as positive.

Party-1 has 6 negative speeches and 6 positive speeches. Party-2 has 2 neutral and 9 positive speeches. Party-3 has 1 neutral and 8 positive speeches.

The figure represented as Fig. 4 displays a multiple bar chart showcasing the distribution of positive, negative, and neutral terms within each speech. This visual representation presents a comparative analysis of the sentiment composition in the speeches, allowing for a clear understanding of the proportions of positive, negative, and neutral terms used in each speech. By utilizing this bar chart, the figure effectively communicates the sentiment distribution across the speeches in a concise and visually accessible manner.

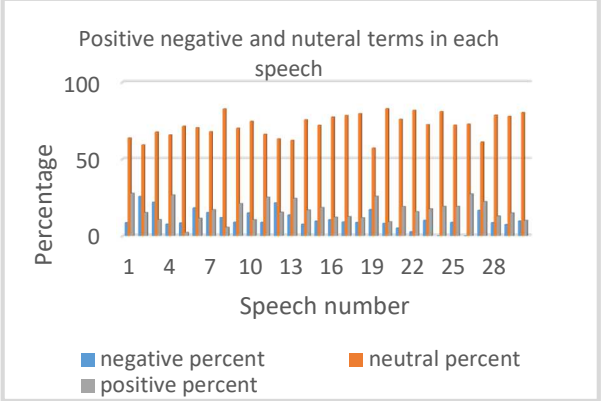

Fig. 4 Multiple bar chart representing positive, negative, and neutral terms in each speech.

Topic modelling is the next step of our analysis. We used the LDA method for topic modelling which gave us 5 topics for each party.

Party-1 nominees made few false promises like giving gas cylinder for cheaper price which is “Promissory estoppel”, they trolled other parties for their words while campaigning which comes under “Trolling”, they criticized the schemes introduced by previous leaders and advertising their schemes which is “Criticism and political advertisement”. They also spoke about employment creation and building multi-speciality hospitals in that respective constituency which is “Employment and development of medical facilities” and how other parties are deceiving the voters by taking bribe which can be understood as “Deception of other parties”. Here employment and medical facilities comes under development category and rest all comes under emotional category.

Party-2 nominees spoke about their party’s contribution towards the state which comes under “Political advertisement”, they also spoke about developing the roads and infrastructure which is “Infrastructural development”, then they spoke about farmer problems like limited electricity, water, and agricultural loans etc., and their schemes towards farmers welfare which is “Agricultural development and Schemes for Farmers”. They spoke about the corruption that was done by other candidates which comes under “Corruption”, and they promised to fulfil all the promises and asked for votes which are “Seeking for votes”. Here Infrastructural development, Agricultural development and Schemes for Farmers and Corruption falls under the development category and political advertisement and seeking for votes falls under the emotional category.

Party-3 nominees spoke about voting for girl candidate and supporting females in all fields which falls under “Supporting feminism”, they also promoted themselves which can be termed as” Political advertisement”, they made some fake promises which is “Promissory estoppel”, they spoke about farmer problems that falls under “Farmer support” and commented on the other party candidates which is” Trolling”. Supporting feminism, Political advertisement, Promissory estoppel, and Trolling falls under the emotional category and farmer support is in development category.

Table II presents the classification of topics, categorizing them as either emotional or development-oriented, for each political party.

TABLE II. CLASSIFICATION OF TOPICS: EMOTIONAL VS DEVELOPMENT OF EACH PARTY

Party	Emotional topics	Development topics
Party-1	4	1
Party-2	2	3
Party-3	1	4

The word cloud depicted as Fig. 5 visually represents the frequently used words in the speeches obtained. This visual presentation showcases a dynamic arrangement of words, where the size and prominence of each word reflect its frequency or importance within the speech dataset. By employing a word cloud, the figure effectively highlights and visually emphasizes the most used words, providing a quick and intuitive overview of the key themes or topics present in the speeches.

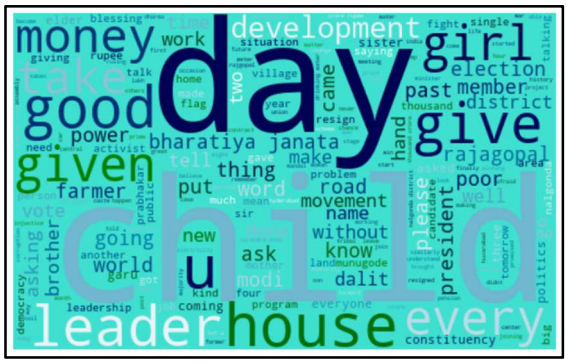

Fig. 5Word Cloud

V. CONCLUSION

The analysis incorporates sentiment analysis and topic modeling to determine the factors influencing voters' decisions during elections. By analyzing the results and comparing them with the election outcomes, it is concluded that the voter is casting the vote by considering the long-term development and positive attitude of the candidates. The voters are not manipulated by emotional sentiments, and they did not appreciate the negative statements made by the candidates.

VI. SCOPE OF FUTURE

The study is constrained to a sample of 30 speeches, and its scope can be expanded by incorporating a larger number of speeches from diverse constituencies. By including a broader range of speeches, the study can capture a more comprehensive understanding of the subject matter. Furthermore, while the current analysis does not involve predictive modeling or consider additional variables, future research could explore the potential for prediction by incorporating relevant factors beyond the speeches themselves. By incorporating these elements, the study can potentially enhance its predictive capabilities and provide deeper insights into the topic under investigation.

REFERENCES

[1] M. Kaur, R. Verma, and S. Ranjan, "Location-based Sentiment Analysis of Indian Political Campaign on Twitter," International Journal of Electronics, Communication and Soft Computing Engineering, vol. 14, no. S3, pp.10, August 2022. DOI: 10.9756/INT-JECSE/V14S3.3.

[2] M. Haselmayer, "Candidates rather than context shape campaign sentiment in French Presidential Elections (1965–2017)," French Politics, vol. 19, pp. 394-420, June 2021. DOI: 10.1057/s41253-021-00159-5

[3] Y. Matalon, O. Magdaci, A. Almozlino, and D. Yamin, "Using sentiment analysis to predict opinion inversion in Tweets of political communication," Scientific Reports, vol. 11, no. 1, pp. 7250, Mar. 2021. DOI: 10.1038/s41598-021-86510-w.

[4] G. Abercrombie, and R. Batista-Navarro, "ParlVote: A Corpus for Sentiment Analysis of Political Debates," in Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), Marseille, France, pp. 5073-5078, May 2020. https://aclanthology.org/2020.lrec-1.624.

[5] B. Haryanto, Y. Rule Deviyani, F. Rohman, J. D. T. N., R. Magdalena, and M. Y. F., "Facebook Analysis of Community Sentiment on 2019 Indonesian Presidential Candidates from Facebook Opinion Data," The Fifth Information Systems International Conference 2019, Procedia Computer Science, vol. 161, pp. 715-722, January 2019. DOI: 10.1016/j.procs.2019.11.175.

[6] Z. Drus and H. Khalid, "Sentiment Analysis in Social Media and Its Application: Systematic Literature Review," in Proceedings of The Fifth Information Systems International Conference 2019, Procedia Computer Science, vol. 161, pp. 707-714, July 2019. DOI: 10.1016/j.procs.2019.11.174.

[7] V. A. Fitri, R. Andreswari, and M. A. Hasibuan, "Sentiment Analysis of Social Media Twitter with Case of AntiLGBT Campaign in Indonesia using Naïve Bayes, Decision Tree, and Random Forest Algorithm," in Proceedings of The Fifth Information Systems International Conference 2019, Procedia Computer Science, vol. 161, pp. 765-772, July 2019. DOI: 10.1016/j.procs.2019.11.181.

[8] P. Widyaningrum, Y. Ruldeviyani, and R. Dharayani, "Sentiment Analysis to Assess the Community's Enthusiasm Towards the Development Chatbot Using an Appraisal Theory," in Proceedings of The Fifth Information Systems International Conference 2019, Procedia Computer Science, vol. 161, pp. 723-730, July 2019. DOI: 10.1016/j.procs.2019.11.176.

[9] P. D. Katre, "NLP Based Text Analytics and Visualization of Political Speeches," International Journal of Recent Technology and Engineering (IJRTE), vol. 8, no. 3, pp. 8574, September 2019. DOI: 10.35940/ijrte.C6503.098319.

[10] C. B. Asmussen and C. Møller, "Smart literature review: a practical topic modelling approach to exploratory literature review," Big Data, vol. 6, pp. 93, October 2019. DOI: 10.1186/s40537-019-0255-7.

[11] W. Budiharto and M. Meiliana, "Prediction and analysis of Indonesia Presidential election from Twitter using sentiment analysis," Big Data, vol. 5, pp. 51, December 2018. DOI: 10.1186/s40537-018-0164-1.

[12] J. Grimmer, "A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases," Political Analysis, vol. 18, no. 1, pp. 1-35, Winter 2010. DOI: https://doi.org/10.1093/pan/mpp034.

[13] O. Koltsova and S. Koltcov, "Mapping the public agenda with topic modeling: The case of the Russian LiveJournal," Policy & Internet, vol. 5, no. 2, pp. 207-227, June 2013. https://doi.org/10.1002/1944-2866.POI331.

[14] D. Baum, "Recognising speakers from the topics they talk about," Speech Communication, vol. 54, no. 10, pp. 1132-1142, December2012. https://doi.org/10.1016/j.specom.2012.06.003.

[15] K. M. Quinn, B. L. Monroe, M. Colaresi, M. H. Crespin, and D. R. Radev, "How to analyze political attention with minimal assumptions and costs," American Journal of Political Science, vol. 54, no. 1, pp. 209-228, January 2010. https://doi.org/10.1111/j.1540-5907.2009.00427.x