School of Information Studies

From Rhetoric to Reality: Analyzing Presidential Debates

Maliheh Alaeddini, Sarah Elizabeth Jones, Marina Mitiaeva, Rishikesh Thakker
Under Prof. Lu Xiao

Overview

In recent years, there has been a growing fascination with leveraging the capabilities of NLP and AI techniques to forecast election outcomes. This interest spans diverse methodologies including sentiment analysis, fake news identification, fact-checking, and topic modeling. Advanced models intertwine deep learning with foundational statistical approaches. Our study delves into the thematic fabric of primary Presidential campaign debates, employing sentiment analysis to gauge the emotional tone of each speech, organizing sentiments by the respective candidates. Additionally, we explore topic modeling and honesty classification among candidates. Our goal involves thoroughly analyzing these debates using a wide range of ML and NLP techniques. This encompasses various practices like topic modeling, sentiment analysis, and text classification among others.

Literature Review

Before delving into our project's specifics, it is crucial to review the most recent research and advancements in NLP techniques pertinent to this domain.

1 Katalinić et al. [1] examined the 2022 US midterm elections by analyzing tweets from Democratic and Republican candidates (52,688 tweets in total). Using sentiment analysis, topic modeling, and party classification techniques, they uncovered nuanced political dynamics across Senate, Gubernatorial, and House elections. Their methods included Python's Tweepy for data collection, Multinomial Naïve Bayes (MNB) for party-based tweet classification, and sentiment analysis using TextBlob and VADER libraries. Gensim's Latent Dirichlet Allocation (LDA) was employed for topic modeling. The study highlighted heightened subjectivity, polarization, and classification disparities among elections, emphasizing the influence of social media and regional variations in shaping political narratives and outcomes.

2 Gode et al. [2] also delved into US political polarization using language models and a Wikipedia-derived dataset spanning 120 years. Employing Longformer, a Transformer model, they assessed polarization levels and identified divisive words. Their study led to the creation of a website estimating a politician's polarization based on Longformer's analysis of nearest neighbors. Findings highlighted distinct campaign topics for Democrats and Republicans, fostering polarization due to high subjectivity. Issues like *elections*, *cost of living*, and *jobs* were central, forming polarized voting blocs and showcasing the fragmentation within US political discourse.

3 Olabanjo et al. [3] directed their focus on Nigeria's 2023 presidential election, specifically harnessing public opinion expressed through Twitter data. This research, unlike the previous studies, underlines the pivotal role of social media in shaping contemporary elections, employing advanced NLU techniques, notably focusing on the BERT model for sentiment analysis. The methodology involved identifying election-related keywords, hashtags, and Twitter accounts. Data scraping via the Twitter API, followed by meticulous cleaning to refine the dataset, was executed using Python. Three ML models - LSTM, BERT, and LSVC - trained on an IMDB dataset, were utilized for sentiment modeling. Furthermore, the study leveraged NLU techniques like sentence polarity analysis, topic modeling, entity extraction, word frequency analysis, and word clouds for comprehensive insight. Though specific findings were not explicitly detailed, this research presents a thorough exploration at the intersection of political analysis and advanced NLP technologies, shedding light on public sentiment regarding Nigeria's political landscape.

4 Węcel et al. [4] examined the influence of Large Language Models (LLMs), specifically ChatGPT, on fake news operations. Focusing on AI's role in generating and detecting fake news, they explored how LLMs like BERT and GPT-3 influence misinformation identification. The study revealed that while BERT is commonly used in fake news detection models due to its effectiveness in tasks like sentiment analysis and named entity recognition, LLMs such as GPT-3 do not significantly enhance fake news detection rates. Results indicated a marginal performance improvement over random guessing, with various prompts affecting the accuracy and introducing biases in answers without significantly altering accuracy rates. The study also highlighted the issue of *hallucinations*, where the models provided answers that were close but not entirely accurate. Additionally, the research explored the robustness of LLMs over time, finding that the models performed comparably wrong irrespective of data period, suggesting a need for continual adaptation to evolving information. Limitations in model confidence assessment and reliance on past knowledge were identified, prompting future directions like combining LLMs with knowledge graphs for up-to-date facts and investigating the impact of newer language models beyond GPT-3.

5 Mayopu et al. [5] carried out a groundbreaking investigation that centered on analyzing online fake news through the utilization of Latent Semantic Analysis (LSA). Unlike prior computer science-oriented approaches, this research aimed to develop an effective method, blending NLP and LSA techniques, to assist social scientists in dissecting fake news elements. The researchers analyzed the features of real and fake news articles, using the 2016 US presidential campaign to showcase the efficiency of their methodology. Their typologies of fake news included *misinformation*, *disinformation*, *clickbait*, *rumors*, *propaganda*, *satire*, *parody*, and *fabrication*. Employing LSA aided in examining significant sentences from input documents, unveiling latent structures, and exploring word-concept relationships using Singular Value Decomposition. The study identified five concepts through LSA analysis, revealing five topics crucial for identifying fake news during Presidential Elections: *Coalition, Politic, Future, Statement,* and *Issues.*

6 The extensive review authored by Hamed et al. [6] scrutinizes the landscape of fake news detection, highlighting key challenges and innovative approaches utilized in this critical area. The article emphasized key components and challenges in this realm, addressing approaches like data augmentation, feature extraction, and data fusion to bolster accuracy. Recognizing five types of fake news - *rumor*, *disinformation*, *misinformation*, *hoax*, and *clickbait* - the study highlighted the crucial role of datasets, stressing that no benchmark dataset currently encompasses all necessary resources. The dataset's size, diversity, richness, and noise level significantly impact model performance in identifying fake news. The review categorized studies based on their employment of machine learning (ML) and deep learning (DL) methods. The review outlined various ML-based models like Logistic Classifier, Random Forest (RF), and ensemble solutions. Additionally, it highlighted several DL-based models such as LSTM, CNN, CapsNet, among others. Limitations observed in fake news detection models encompassed overfitting, imbalanced datasets, ineffective feature representation, and inadequate data fusion. Prominent techniques used in these models included dataset augmentation techniques like Generative Adversarial Networks (GANs), feature extraction through models like BERT and VGG-19, and multimodal fusion methods including Early Fusion, Joint Fusion, and Late Fusion.

Our Process

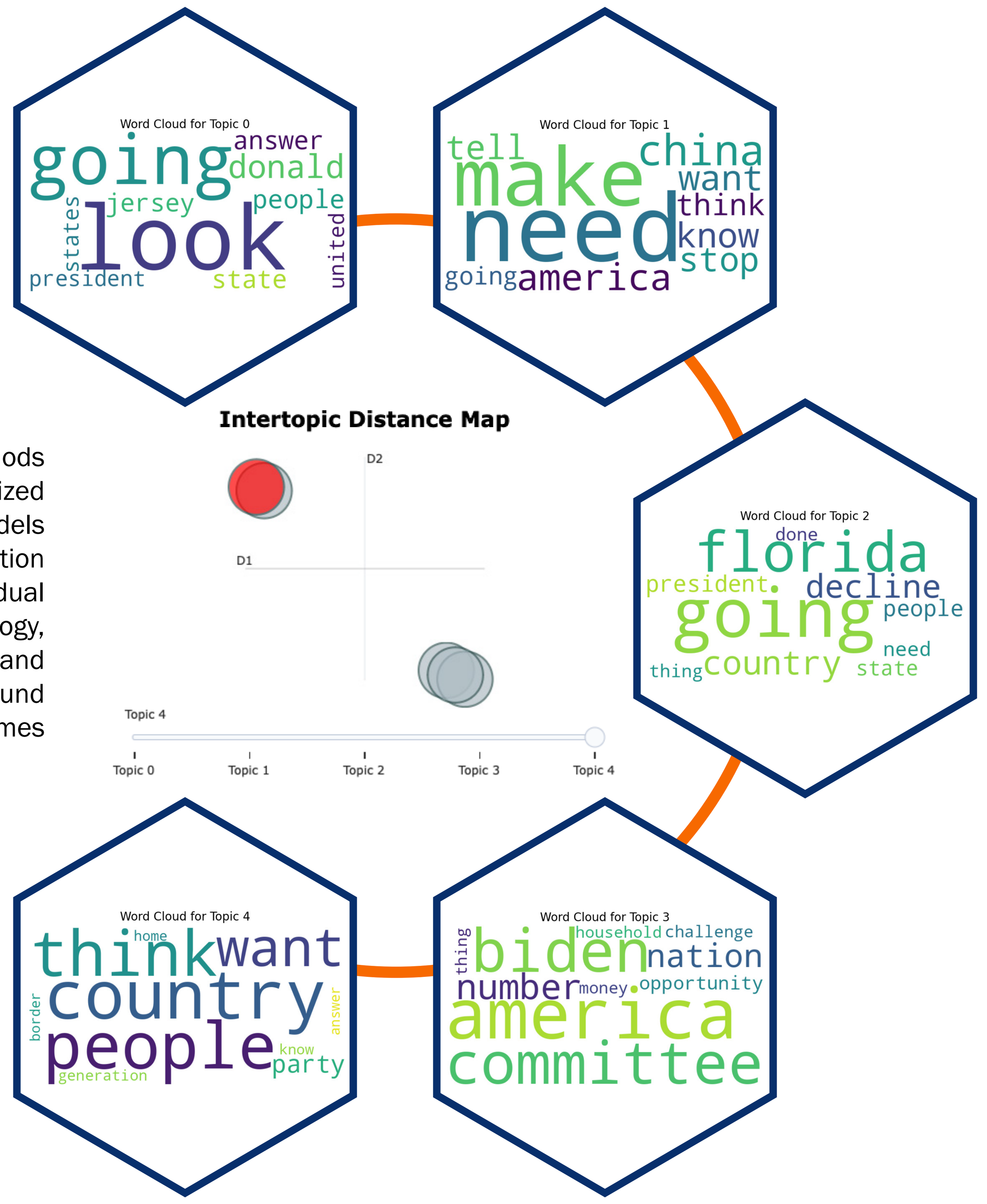
Data Collection & Preprocessing

Data collection was carried out by scraping the transcript of the three Republican Presidential Debates into CSV files and creating a combined dataset with all three debates and an indicator variable for which debate the statement came from. For precision, our analysis excludes the speeches of moderators to ensure a focused dataset aligned with the original topics discussed. We specifically focus on the candidates participating in all three debate rounds.
To preprocess the data and ensure quality analysis we used:

Tokenization | **Lowercasing** | **Punctuation and Numbers Removal** | **Stop Words Removal** | **Lemmatization**

Topic Modelling

Topic modeling uncovers hidden patterns in text, aiding decision -making by revealing abstract topics. Our study used BERTopic, enhanced by SentenceTransformer, to extract and scrutinize topics from textual data. This approach ensured robust data representation through embeddings and employed UMAP for dimensionality reduction, providing a nuanced understanding of content. We evaluated topic coherence using the CoherenceModel, finding that BERTopic surpassed traditional LDA methods by approximately %10. Our personalized approach assigned unique BERTopic models to each speaker, enabling a detailed exploration of topic distributions and highlighting individual differences in the dataset. This methodology, integrating advanced NLP techniques and personalized topic modeling, offered profound insights into underlying textual themes and patterns.



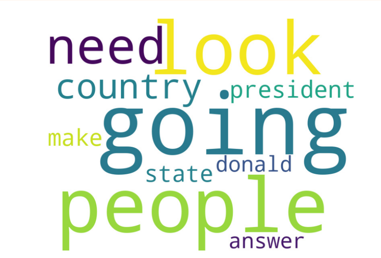
Distance analysis reveals that Topics 1 and 5, *future* and *hope*, are highly related. However, there were three topics (Topics 2–4) that focused on negative aspects of where the country is going: *demands*, *fears*, and *constraints*. Under these topics, the candidates were discussing what people *need* and the *challenges* that people face. Topic 4, *constraints*, offered intriguing insights into discussions revolving around *challenges*, *money*, and *opportunities*, often linked to current President Biden, possibly reflecting attempts by Republican candidates to differentiate themselves from his policies. Additionally, of particular interest is Topic 1, *future*. It reveals that as candidates are using words like *look*, *going*, and *answer*, the candidates are also discussing former President Donald Trump. Though he was not present at any of the debates, he continued to run for the Republican nomination to the Presidency. This is indicative of the party's focus on whether or not former President Trump will continue to be the leader of the party in the future.

References

1 Katalinić, J., Dunđer, I., & Seljan, S. (2023). Polarizing Topics on Twitter in the 2022 United States Elections. Information, 609 ,(11)14. https://doi.org/10.3390/info14110609
2 Gode, S., Bare, S., Raj, B., & Yoo, H.K. (2023). Understanding Political Polarisation using Language Models: A dataset and method. https://doi.org/10.48550/arXiv.2301.00891
3 Olabanjo, O., Wusu, A., Afisi, O., Asokere, M., Padonu, R., Olabanjo, O., Ojo, O., Folorunso, O., Aribisala, B., & Mazzara, M. (2023). From Twitter to Aso-Rock: A sentiment analysis framework for understanding Nigeria 2023 presidential election. Heliyon, 5)9), e16085. https://doi.org/10.1016/j.heliyon.2023.e16085
4 Hamed, S. K. H., Ab Aziz, M. J., & Yaakub, M. R. (2023). A review of fake news detection approaches: A critical analysis of relevant studies and highlighting key challenges associated with the dataset, feature representation, and data fusion. Heliyon, 10)9), e20382. https://doi.org/10.1016/j.heliyon.2023.e20382
5 Mayopu, R. G., Wang, Y.-Y., & Chen, L.-S. (2023). Analyzing Online Fake News Using Latent Semantic Analysis: Case of USA Election Campaign. Big Data and Cognitive Computing, 81 ,(2)7. https://doi.org/10.3390/bdcc7020081
6 Węcel,K.,Sawiński,M.,Stróżyna,M.,Lewoniewski,W.,Księżniak,E.,Stolarski,P. & Abramowicz,W.(2023).Artificial intelligence—friend or foe in fake news campaigns. Economics and Business Review,70-41 (2)9. https://doi.org/10.18559/ebr.2023.2.736

Topic Modelling Per Candidate

Candidate-specific modeling suggests that Chris Christie was highly focused on where the country is going. He frequently referenced former President Donald Trump. Similar to Topic 1, *future*, Christie's interest appears to be on looking toward the future and assessing whether or not former President Trump is the most qualified candidate for the Republican Presidential nomination.

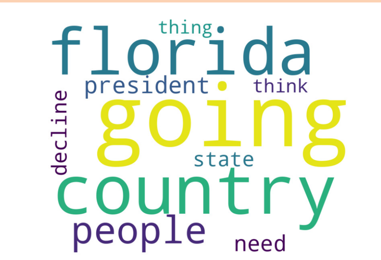


A. Chris Christie

The most negative candidate across all three debates, Nikki Haley's specific language suggests a focus on what people *need* and *want*. Her frequent references to *China* and *America* suggest a focus on international affairs that is consistent with her history as the UN Ambassador for the United States. Additionally, her frequent use of the word *make* implies a level of force not present in other candidates' language.

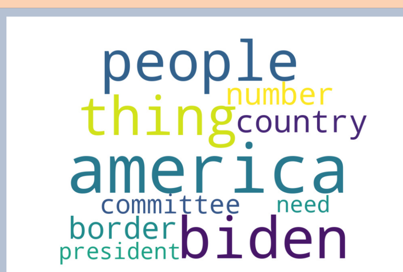
B. Nikki Haley

Ron DeSantis' language reveals that he relied heavily on his experience as a Florida governor to talk about the country as a whole. In fact, *Florida* and *country* appear to be used almost the same amount, suggesting that DeSantis may have leveraged the debates as an opportunity to market how his tenure in Florida can be exported to a national level. He also relies on future-looking language with *going*, but his use of *decline* and *need* suggests negativity in his outlook.



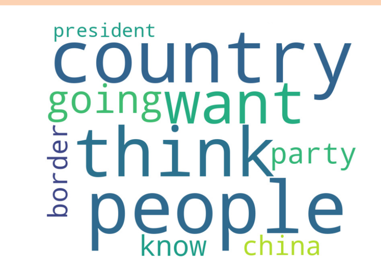
C. Ron DeSantis

While immigration is a high-profile issue in the US, only Tim Scott and Vivek Ramaswamy have an explicitly immigration-related word on their most frequently used list. Scott frequently used both *border*, *America*, and *country*, suggesting that he is particularly focused on migration into the US. Additionally, his high use of *Biden* suggests that he may have been raising immigration as a way to lodge a critique at current President Biden. Tim Scott has not stepped out of the primary race. However, his participation in the debates is still useful for demonstrating that a high focus on immigration throughout all three debates was not a winning strategy for this particular candidate, despite its salience to the Republican Party.



D. Tim Scott

Vivek Ramaswamy, like Tim Scott, frequently used the word *border*, however, he also frequently used *China* rather than *America*. Like Nikki Haley, he may be particularly interested in foreign affairs. However, unlike Nikki Haley, who focused entirely on countries in her most frequently used words, Ramaswamy's high use of the word *border* suggests that he is particularly interested in immigration. Additionally, his use of *think*, *want*, and *people* far and above other words suggests that Ramaswamy may be relying on populist rhetoric for his campaign.



E. Vivek Ramaswamy

Sentiment Analysis

Chris Christie	NEGATIVE	0.96
Nikki Haley	NEGATIVE	0.97
Ron DeSantis	NEGATIVE	0.95
Tim Scott	NEGATIVE	0.93
Vivek Ramaswamy	NEGATIVE	0.93

We executed thorough sentiment analysis and misinformation detection on a text dataset, harnessing advanced NLP techniques. Employing a sentiment analysis pipeline, we segmented the text data into manageable chunks, evaluating sentiment within each segment. This approach allowed us to aggregate sentiment scores by speaker, unveiling nuanced sentiments expressed by individuals in the dataset. The overall balance of the topics skewing toward negative subjects is aligned with our sentiment analysis, which found that all of the candidates were highly negative throughout all three of the debates. Nikki Haley's comments were classified as the most negative, with Vivek Ramaswamy and Tim Scott having the least negative, however, there is very little variation across all the candidates. Therefore, we can conclude that the Republican Primaries presented a highly negative view of American affairs.

Honesty Classification

Chris Christie	1.0
Nikki Haley	1.0
Ron DeSantis	1.0
Tim Scott	1.0
Vivek Ramaswamy	1.0

In parallel, we addressed misinformation detection by leveraging a labeled dataset distinguishing real and fake news. Preprocessing involved handling missing values and splitting data for training and testing. Our machine learning pipeline, combining a TF-IDF vectorizer with an SVM classifier—an established method for text classification—was trained and evaluated with a classification report. Applying this trained model to our dataset, we gauged the ‹honesty› of each speaker based on their text classifications. Interestingly, all candidates' statements were classified as honest. This outcome likely stems from their focus on future plans and aspirations for the country, making it challenging to label them as untrue before implementation. This dual approach, integrating sentiment analysis and misinformation detection, provided a comprehensive view of the textual data. It shed light on emotional tone and factual reliability, offering valuable insights.