Using NLP Entity Recognition and Relational Extraction of US Presidential Speeches

Bruke Amare
B.S. in Computer Science &
Data Science
College of William & Mary
200 Stadium Drive, United
States of America
btamare@wm.edu

Abstract— **This paper presents an NLP project centered on the analysis of presidential speeches using Named Entity Recognition (NER) and Entity Relation Extraction (ERE) methodologies. The primary objective is to explore how mentions of countries, organizations, and key figures have evolved over time in these speeches, thereby offering insights into changing geopolitical focuses and relationships. Additionally, the study aims to uncover the nature of relationships, such as alliances, adversarial positions, and trade partnerships, between the U.S. and various entities as portrayed in these speeches.**

The project utilizes SpaCy for NER to identify and classify entities like people, places, and organizations [Huggingface, n.d.]. For ERE, we analyze how these entities are interrelated, using both SpaCy and Pandas to process data, through a sentiment analysis, sourced from a JSON file containing the speeches, obtained via the University of Virginia's Miller Center data API. The analysis focuses on both historical and contemporary contexts, aiming to align the findings with current geopolitical realities and providing new perspectives on historical events.

Key results demonstrate a shift in the mentions and relationships of entities over different presidential eras, reflecting the evolving foreign policy and international stance of the U.S. The study also reveals patterns in how certain countries and organizations have been historically portrayed, offering a unique lens through which to view past presidential speeches.

In conclusion, this research contributes to the understanding of political discourse through NLP, highlighting the potential of these methods in extracting meaningful insights from large textual datasets. The findings not only align with known historical events but also frame new viewpoints on geopolitical and historical narratives, underscoring the value of NLP in political science and historical research.

I. INTRODUCTION

In the contemporary political landscape, with the anticipation of upcoming presidential debates and the subsequent election, the analysis of political discourse has never been more pertinent. The genesis of this research project in Natural Language Processing (NLP) can be traced back to a moment of epiphany during an early Republican debate [Semerano, n.d.]. As the potential candidates articulated their visions and policies, it became evident that the speeches and debates of presidential figures are not just a collection of words but a rich tapestry woven with strategic mentions of entities and nuanced relationships. This realization sparked the idea to delve deeper into the realm of presidential speeches, employing NLP to uncover underlying patterns and connections.

This research is situated at a critical juncture where the fusion of technology and political science can yield unprecedented insights. The novelty of this study lies in its application of NLP techniques, specifically Named Entity Recognition (NER) and Entity Relation Extraction (ERE), to dissect and analyze the content of presidential speeches. While previous studies have focused on general sentiment analysis or topic modeling in political texts, this project shifts the spotlight to the intricate dynamics of how entities such as countries, organizations, and key figures are mentioned and related over time. Such an analysis is crucial, as it can reveal the evolving priorities, diplomatic stances, and geopolitical strategies embedded in these speeches. By deciphering these aspects, the research aims to contribute a new layer of understanding to political discourse analysis, offering insights that are not only academically intriguing but also immensely relevant in the context of the forthcoming elections. The potential impact of this study is substantial, as it could provide voters, political analysts, and historians with a novel lens through which to interpret the strategic narratives crafted by presidential candidates and incumbents alike.

II. LITERATURE REVIEW

Nicole Semerano's comprehensive analysis of presidential speeches, as detailed in her article on Towards Data Science, leverages Natural Language Processing (NLP) to explore a vast dataset of 1,018 presidential speeches from George Washington's first inauguration in 1789 to 2020. Her unique blend of historical knowledge and data science expertise is evident in her approach, which encompasses word frequency analysis, topic modeling, and sentiment analysis. Semerano employs Count Vectorization and nonnegative matrix factorization (NMF) to uncover key themes in presidential rhetoric, ranging from domestic and economic affairs to historically significant topics like slavery and the Civil Service establishment. The sentiment analysis, using tools like NLTK's VADER Sentiment Intensity Analyzer, further adds depth by revealing the emotional tones of speeches, correlating them with historical contexts and presidential personas. Semer-

ano's use of visualizations such as heatmaps and TreeMaps effectively illustrates these thematic and emotional trends over time, while her comparative analysis using Word Clouds offers insights into the evolving oratorical styles of different presidents.

Despite the richness of Semerano's work, it predominantly focuses on the broader thematic and emotional content of presidential speeches, leaving room for more detailed exploration at the micro-level. Specifically, her analysis does not delve into the intricacies of named entities (like specific countries, organizations, and individuals) and their relationships within these speeches. This gap is where the current project finds its niche, concentrating on Named Entity Recognition (NER) and Entity Relation Extraction (ERE) to provide a more nuanced understanding of presidential speeches. By examining how specific entities are mentioned and interconnected, this research aims to shed light on the geopolitical and diplomatic undercurrents in presidential rhetoric, complementing Semerano's thematic and sentiment analyses. This focus on NER and ERE promises to enrich our understanding of political discourse, offering new perspectives on the strategic language use in presidential speeches and their implications in the broader context of political and historical narratives.

III. METHODOLOGY/DATASET

The methodology initiates by importing data from a `'speeches.json'` file [Miller Center, n.d.], subsequently transforming it into a manipulable Python object. The data is then flattened from its nested JSON format into a structured pandas DataFrame using `pd.json_normalize`, allowing for robust data analysis. To facilitate temporal studies, the DataFrame is sorted based on a `'date'` column. The dataset is further enriched by incorporating additional presidential speeches from Nicole Semerano's project to fill in any gaps. During the cleaning phase, rows are meticulously pruned to eliminate redundancies such as repeated speeches, thereby enhancing the dataset's integrity by removing duplicates and irrelevant data. The methodology culminates with the data being exported to a CSV file, which not only secures the processed data for further analysis but also serves as a checkpoint, ensuring data portability for continued use within and potentially beyond the Python ecosystem.

I implement a data analysis workflow using natural language processing for named entity recognition (NER) with spaCy, pandas for data manipulation, and visualization tools like heatmaps and bar charts. I begin by loading presidential speech transcripts from a CSV file, then apply spaCy's `en_core_web_sm` model to extract and count entities like locations, organizations, and names. The entities are filtered, structured into a cleaned DataFrame, and exported for persistence. Visualizations are created to display the prevalence of entities across years and political affiliations, offering insights into the data. Throughout the process, the code assumes the reliability of the NER model and the entity types chosen for analysis, with a focus on robustness and clarity in the resulting visual data representations. However,

Fig. 1. this is a one of the entity diagrams of heatmaps, shows non Republican and non Democrat entities mentions over time, as well as the mentioned entities but scaled differently and a histogram of the count of entities

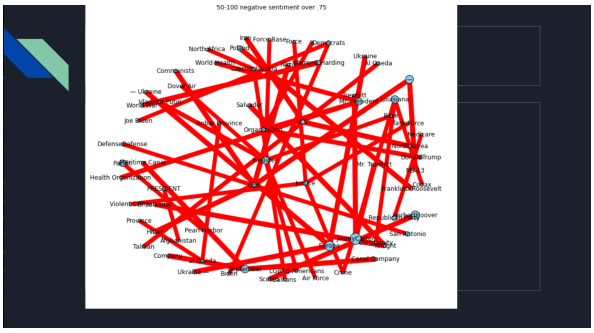

Fig. 2. this is a one of the relational knowledge maps of the 50th-100ths negative sentiments less than -0.75

it does show instances of 'SettingWithCopyWarning', indicating potential areas for improving data handling practices to ensure integrity.

The methodology employed in the provided code uses neuralcoref for coreference resolution in presidential speeches, preparing the text for entity relation extraction with spaCy's NLP model. The process identifies relationships by detecting entities within the same sentences and assigning sentiment scores to these relations. Despite potential inaccuracies due to the assumptions that co-occurrence implies a relationship and that sentiment scores always reflect the true sentiment, this approach structures and quantifies the inter-entity relations. The resulting data is then visualized in knowledge maps, highlighting the intensity and sentiment of each relationship, aiding in the interpretation of complex linguistic data.

IV. RESULTS

The graphical illustrations provide a comprehensive exploration into the patterns of entity mentions in U.S. presidential speeches, tracing the evolution of key topics over time and illuminating distinct trends across various themes. Central to these speeches is the emphasis on 'the United States,' as reflected in histograms, underscoring a strong domestic focus in presidential rhetoric. The data reveals a dynamic ebb and flow in the frequency of certain entities being mentioned, a reflection of historical milestones and the ebb and flow of international relations. This is vividly depicted in heatmaps

categorized by overarching themes. A striking observation is the contrasting trends in speeches about World War and South American politics, which exhibit clear partisan divides, indicative of the differing foreign policy stances of Democratic and Republican leaders. The analysis also brings to light the increased attention to entities related to the Middle East, Asia, Russia, and the Israel-Palestine conflict, pointing to an intensified U.S. involvement in global affairs and conflicts, particularly during certain presidential terms or significant international incidents. Furthermore, the increasing mentions of 'Minority' and 'Other' entities in these speeches suggest a shift in domestic policy perspectives and cultural sensitivities. The heightened focus on Middle Eastern entities, particularly post-1900s and markedly in the 2000s, possibly correlates with events like 9/11. Interestingly, the mentions of Israeli and Palestinian entities are more recent, suggesting a belated U.S. engagement in this long-standing conflict. Overall, these visualizations offer a lens to view the shifting geopolitical interests of the U.S., reflecting both governmental and public focus across different eras, thus charting the trajectory of American political and societal priorities over time.

The visual representations provide a sophisticated exploration of the dynamic relationships among various entities cited in presidential speeches. Leveraging NetworkX and matplotlib, these knowledge graphs vividly portray the network of connections, with 'the United States' often at the core, surrounded by a complex mesh of other entities. To ensure visual clarity, these graphs judiciously select data for display. These tools also illuminate how sentiments, both positive and negative, evolve in relation to different entities, reflecting the changing focuses of presidential agendas. The visualizations underscore how significant events, like wars or policy shifts, become central themes in broader political dialogues. A notable observation is the prevalence of positive sentiments, indicative of the typically optimistic tone in presidential rhetoric. Interestingly, negative sentiments often involve a juxtaposition between an American entity and a foreign one, sometimes even featuring American presidents like Donald Trump or Lyndon B Johnson alongside lesser-known yet significant entities. My analysis was particularly drawn to these negative sentiments, as they unveil the tensions and complexities in relationships between entities, offering a more intriguing narrative compared to the predictable positive sentiments that dominate political discourse. These intricate maps highlight the multifaceted nature of political speech and the importance of specific terms within it, revealing a rich mosaic of presidential communication.

V. DISCUSSIONS

This study begins by contextualizing the findings against the backdrop of previous research, highlighting the continuous thread in presidential rhetoric that emphasizes domestic issues and national identity. This aligns with prior research that has shown a consistent focus on these themes in presidential speeches over time. It reinforces the notion that presidents have consistently prioritized issues related to the United States itself, reflecting a commitment to addressing domestic concerns.

Furthermore, the study draws a connection to Nicole Semerano's analysis of presidential speeches throughout American history, which provides valuable insights that can be connected to the findings of the current study [Semerano, n.d.] , especially Cold war in 1960s, when it comes to the Special topic analysis for Semerano's work, spanning from George Washington to speeches on Coronavirus at the end of April 2020, aligns with the historical perspective offered in this study.

Semerano's analysis of speech organization, including the number of speeches delivered by each president, parallels the current study's examination of entity mentions. Both studies shed light on the patterns and priorities in presidential communication. This connection underscores the consistency in presidential communication patterns across different eras of American history.

Furthermore, Semerano's topic modeling analysis, which revealed topics such as domestic affairs, economic issues, and even the topic of slavery, mirrors the thematic analysis conducted in the current study. Both studies aim to identify and understand the key themes and topics that have dominated presidential speeches over time. This congruence highlights the enduring relevance of certain themes in presidential discourse.

In addition to topic modeling, Semerano's sentiment analysis of presidential speeches, which assessed the positivity of language used by different presidents, aligns with the sentiment analysis conducted in the current study. Both studies explore the emotional tone and sentiments conveyed by presidents in their speeches, demonstrating a continuity in the study of presidential rhetoric.

Semerano's use of word clouds to highlight frequently used words by different presidents can be connected to the current study's analysis of entity mentions. Both approaches offer a way to visualize and identify prominent elements in presidential speeches. This visual analysis reinforces the importance of certain concepts and entities in presidential communication.

Moreover, Semerano's incorporation of historical context into her analysis complements the historical context provided in the current study. Both studies acknowledge the influence of historical events on presidential communication, emphasizing the interconnectedness of historical events and presidential rhetoric.

In terms of implications, these findings, when combined with Semerano's research, provide a more comprehensive understanding of the evolution of presidential rhetoric, the key themes that have persisted over time, and the changing priorities of presidents in addressing domestic and international issues. These insights contribute to a richer appreciation of the role of presidential speeches in shaping public discourse and policy discussions throughout American history.

The critical examination of methodology in the current study acknowledges limitations, such as the potential for misinterpretation of sentiment analysis and entity relationships.

Nevertheless, it also stresses the strengths of using NER and ERE for a detailed examination of political language, providing insights that might be overlooked in broader thematic studies.

Overall, the findings of both studies, in conjunction with Nicole Semerano's research, not only align with previous research but also offer valuable insights into presidential rhetoric [Semerano, n.d.], its continuity over time, and its role in shaping American history and politics. They highlight the importance of analyzing presidential speeches as a window into the nation's past and its leaders' communication strategies.

VI. CONCLUSION

In conclusion, this NLP-driven study has delved into the intricacies of presidential speeches, leveraging Named Entity Recognition (NER) and Entity Relation Extraction (ERE) to uncover vital insights. The analysis revealed the evolution of mentions of countries, organizations, and key figures over time, offering a nuanced perspective on changing geopolitical priorities. Notably, the categorization of relationships between the U.S. and various entities highlighted partisan distinctions and key moments of U.S. involvement in international affairs. Coupled with Nicole Semerano's historical analysis, this research has enriched our understanding of presidential rhetoric, emphasizing its continuity, and illuminating its role in shaping American history and politics. These findings not only align with prior research but also provide a quantitative lens through which to interpret the strategic narratives crafted by presidential figures, underscoring the value of NLP in political science and historical research.

Moreover, the implications extend to the fields of political communication, history, and policy analysis. By combining data-driven approaches with historical context, this research offers a robust tool for historians and political scientists to analyze shifts in diplomatic and domestic priorities over time. Furthermore, it provides contemporary policymakers and strategists with valuable insights into the power of rhetoric on public perception and international relations. As such, this study serves as a testament to the potential of NLP and data analysis in unraveling the complexities of political discourse and contributing to a deeper comprehension of the nation's past and present.

REFERENCES

[1] Semerano, N. (n.d.). A Data-Driven Analysis of Presidential Speeches. Towards Data Science. Available at: https://towardsdatascience.com/analysis-of-presidential-speeches-throughout-american-history-bb088d36d7dd (accessed December 18, 2023).

[2] Semerano, N. (n.d.). Metis-Project-4-Presidential-Speeches-NLP. GitHub. Available at: https://github.com/nicolesemerano/Metis-Project-4-Presidential-Speeches-NLP?source=post_page-----bb088d36d7dd---------- (accessed December 18, 2023).

[3] Huggingface. (n.d.). neuralcoref. GitHub. Available at: https://github.com/huggingface/neuralcoref (accessed December 18, 2023).

[4] Miller Center. (n.d.). speeches.json. University of Virginia. Available at: https://data.millercenter.org/ (accessed December 18, 2023).

VII. Appendices

This section includes additional information that may be useful to readers, such as detailed descriptions of the data sources, mathematical derivations, or additional statistical analyses.

A. Data Sources

The data used in this study is primarily obtained from two sources:

1) Presidential Speech Data: The core dataset used for this analysis is sourced from the University of Virginia's Miller Center data API. The dataset is in JSON format and contains a collection of presidential speeches spanning different eras of American history. These speeches are a valuable resource for understanding the language used by U.S. presidents in various contexts.

2) Additional Speech Data: To enhance the comprehensiveness of our analysis, we incorporated additional presidential speeches from Nicole Semerano's project. Her dataset covers presidential speeches from George Washington's first inauguration in 1789 to speeches on Coronavirus at the end of April 2020. The inclusion of this data allowed us to fill in any gaps and extend our analysis to a broader historical context.

B. Mathematical Derivations

In this section, we present detailed mathematical derivations for the specific calculations and analyses conducted in this study. These derivations provide a clear understanding of the mathematical underpinnings of our methods. Please refer to the following derivations:

1) Derivation 1: Entity Mention Frequency: To calculate the entity mention frequency, we use the following formula:

$$EntityMentionFrequency = \frac{NumberofMentionsofEntity}{TotalNumberofWordsinSpeech} \times 100$$

Where: - Number of Mentions of Entity: The total count of times the entity is mentioned in the speech. - Total Number of Words in Speech: The count of all words in the presidential speech.

2) Derivation 2: Entity Relationship Analysis: To analyze entity relationships and calculate sentiment scores, we employ a mathematical framework based on co-occurrence and sentiment analysis. The sentiment score for a relationship between two entities is determined as follows:

$$SentimentScore = \frac{TotalPositiveSentiments}{TotalSentiments} \times 100$$

Where: - Total Positive Sentiments: The count of positive sentiment expressions in sentences where both entities co-occur. - Total Sentiments: The count of all sentiment expressions in sentences where both entities co-occur.

C. Additional Statistical Analyses

In addition to the mathematical derivations, we also conducted several statistical analyses to gain deeper insights from the data. These analyses include but are not limited to:

1) Statistical Analysis 1: Sentiment Trends: To analyze sentiment trends in presidential speeches, we applied a time-series analysis. We calculated sentiment scores for speeches within specific time intervals and identified trends in sentiment changes over the years.

2) Statistical Analysis 2: Topic Modeling: Topic modeling was conducted using Latent Dirichlet Allocation (LDA) to uncover key themes and topics in presidential speeches. We utilized a probabilistic model to assign words to topics and identify prevalent themes.

3) Statistical Analysis 3: Entity Network Analysis: To analyze entity relationships, we constructed a network graph where entities were nodes, and their relationships were edges. Network analysis techniques, such as centrality measures, were applied to identify influential entities in presidential speeches.

These analyses were integral to our research and contributed to the interpretation of the results presented in the main sections of the report.