Comparative Analysis of the Speeches of Barack Obama and Donald Trump Using NLP

Otari Tchigladze

June 25, 2024

Abstract

This project conducts a comparative analysis of speeches by two public figures, Barack Obama and Donald Trump, using NLP techniques. The objective is to identify and analyze linguistic patterns, emotional tones, and rhetorical styles that distinguish each speaker. The project tries to examine the lexical choices, syntactic structures, and emotional appeals by employing tools such as sentiment analysis, POS tagging, NER, and topic modeling, This analysis highlights commonalities and differences in their rhetoric to help us understand better what makes each speaker stand out and be effective.

1 Introduction

Text Analysis is not an original concept and it has been done many times for different ends. The objective of this project is to conduct a comparative analysis of the speeches of two public figures with peculiar styles and investigate into linguistic and other communication techniques employed by each speaker; Explore the similarities and differences and ask the question: what makes each speaker stand out and be distinctly himself. The goal is to find the right tools and models suited for such investigation. I do not intend to select the best speaker in the end but instead to find indivudual traits alongside some commonalities and differences between the speakers which could potentially allow us to develop a better insight into the aspects of their personas that makes them so charming and captivating; also, to discover a computationally interesting way of doing it. The results can potentially be used for teaching people how to improve their public speaking skills. This project seeks to use NLP techniques, including sentiment analysis and emotion detection models(amongst multiple other instruments) to systematically assess and compare the emotional tones conveyed in the speeches of various political figures. The objective of this analysis is not to rank speakers by effectiveness but to reveal the specific elements such as lexical choices, syntactic structures, and emotional appeals that characterize their speaking styles. This project employs advanced NLP techniques which allows us to automate the analysis of of text data, making it possible to decompose speeches into linguistic features by using tokenization and POS tagging to systematically break down speeches into manageable units for analysis. It analyzes sentiment and emotion. In addition, applying NER to identify key themes and subjects in the speeches gives opportunity to explore the grammatical structures that lend each speaker their distinctive voice.

2 Literature

There are multiple works that inspired this research, one of them being a study by Rameshbhai et al. (2020), titled "Opinion Mining on Newspaper Headlines Using SVM and NLP," which utilized Support Vector Machines and NLP techniques to analyze sentiment in newspaper headlines which highlights the possibility of extracting sentiment from data as small as the article headlines without having to rely on broader context.

A significant study by Manuela Caiani and Jessica Di Cocco (2023), titled "Populism and Emotions: A Comparative Study Using Machine Learning," investigates the use of emotions in the political discourse of populist and non-populist parties. This research highlights the connection between emotions and populism and demonstrates that populists tend to use a broader range of emotional appeals. The study's approach is based on supervised machine learning and provides a comprehensive analysis of the intensity and trends of specific emotions in political discourses.

3 Data

The primary data source for this project is the website *rev.com* that offers various transcripts of famous speeches from different political leaders who are one of the best candidates for this analysis. Otherwise, there are multiple other sources that can be used for more diverse data. The data preprocessing starts with **(1)** *text normalization:* this includes converting all text to lowercase to ensure regularity and removing punctuation and numbers, which are generally irrelevant for our purposes. For such a basic normalization process Python's built-in 're' module can be enough. (2) After the text has been normalized, the next step is the *tokenization process* for which the tool used is spaCy which is a rather flexible NLP library that breaks down text into words/sentences. (3) *Removing the stop words* is also imperative because it allows us to get rid of the noise that can skew the results. This is done with NLTK which includes a pre-defined list of stopwords that can be easily customized and edited according to necessity. (4) Then I use spaCy once again for *part-of-speech tagging* which is crucial for subsequent syntactic analysis and in understanding the grammatical structure of our text. (5) SpaCy also efficiently identifies and classifies *named entities* in a text which I use to extract specific data points that are relevant for deeper contextual analysis. (6) Tf-idf is used as well for identifying and highlighting *the most important terms* in each speech.

4 Model

There are a few noteworthy features in this project. It uses a number of NLP techniques for text processing that provides the possibility to normalize the data and shape it in a way that is appropriate for our objective. One of the NLP libraries used is **SpaCy** for normalization and POS tagging. Firstly, it splits the text into meaningful units for tokenization purposes, in the case of this project these units are words. After tokenization, it does part of speech tagging which is crucial to understand the grammatical structures of sentences. The next step is dependency parsing to detect the relationship between words and sentences between each other, which makes it possible to have better syntactic and semantic understanding of the given speeches. I also use SpaCy for NER, this way I am able to compare the two speeches in more detail and find fine-grained similarities/differences between them in terms of the mentions of certain entities/people/locations etc, which could possibly be contributing to the peculiarity of each speaker. The script also uses SpaCy to identify adjectives in the text, after which they are tagged with the POS label 'ADJ'. Then the **Counter** class from collections is used to count the number of times the top 10 adjectives appear in each text.

I use **NLTK** for processing tasks such as stop words removal as well as the detection of catchphrases of each speaker. Removing stop words is essential for being able to count the relevant words and phrases used by the speakers without much noise, including stop words in the text resulted in useless outcome(logically, the most used words are always going to be the stop words and function words). Using trigrams, I am able to detect the phrases that the speakers are prone to using more often than others this offers insights into repetitive patterns or emphasis areas.

Next, the project employs **NRCLex** library, a useful tool for sentiment analysis that is able to take words and link them to certain emotions. I proceed by passing the string of text to NRCLex as an input. NRCLex has a built-in dictionary that relates words to emotions. In this case my goal is to detect the basic human emotions such as joy, sadness, anger, trust, etc. However, there are a few downsides that need to be addressed regarding the method: one being its lexicon based limitations that is due to the fixed vocabulary that it has. Moreover, the results heavily depend on the quality of the input text. Another significant downside is its lack of ability for context based analysis. I chose to use the method regardless for a few reasons: (a) Based on the scale of analysis I am conducting, I believe NRCLex is a good enough method for the sentiment analysis part of my project; my goal is a general overview of the emotional undertones of each speech, to compare the negative vs positive emotions that each speaker tends to invoke more; I am less interested in the context in which each words are used and more in the frequency of the words typically associated with positive and negative emotions.

Moreover, I calculate **TF-IDF** scores for each speech to see which terms used are more salient for each speaker. This analysis is useful to reveal priorities in the speakers' rhetoric and/or the distinct narration styles. For example, Trump is more likely to use what can be referred rousing and stirring attitude when speaking. whereas Obama tends to be more measured and slow-paced.

My project also utilizes the **paraphrase-MiniLM-L6-v2** model, to measure semantic similarities between the two speeches. This model is able to capture the meanings of sentences based on the surrounding context and help with the comparison at a semantic level rather than just syntactic.

Next part of the project is composed of topic modeling using unsupervised learning. It employs **LDA** to detect prevalent topics in each of the two speaker's speeches and give a representation of it in order to find possible common patterns between the two. Given that I only have two text files to analyze with multiple speeches for each speaker instead of having many files, splitting each file into smaller chunks of text seemed wiser. This allows the LDA model to perform more effectively. The segments are made of 100 words from the speeches. After this, the text data is converted into a document-term matrix with CountVectorizer. For which English stop words are removed. The maximum number of features are set to 10,000.

The LDA parameters are set to:

1. Amount of topics - 10
2. Random state - 42, to ensure reproducibility and consistent results.

To compare the distribution of topics between Obama and Trump speeches, the results are visualized using a bar chart. All findings are represented visually in the form of charts. To compare the distribution of topics between Obama and Trump speeches, the results are visualized using a bar chart. The results are displayed side by side to give a clear insight into the similarities and differences. For this task the script utilizes Matplotlib. Word clouds are also used to represent the most salient features used by each speaker as detected by the TF-IDF module.

The purpose of the project is to perform a comparative analysis of the speeches of Barack Obama and Donald Trump and find similar/different linguistic (and not only) patterns and visualize them in a clear and straightforward manner. As a part of the project, one of the tools used is LDA which is an unsupervised learning method and the purpose of it is to find the distribution of topics. The LDA model is trained with the preprocessed and vectorized text data.
Hyperparameter tuning procedure involved the number of topics (*n_components*). The optimal number of topics depends on the data, in this case, the number was 10. Alpha, which is *doc_topic_prior*, and Beta, *topic_word_prior*, hyperparameters are used to control the sparsity of how the document-topic and topic-word distributions are present in the data. The tuning is done using a grid search to find the values to increase topic consistency. Additionally, *max_iter* is used as the parameter to define the maximum number of iterations the algorithm will perform. Higher values can increase computational time but usually can deliver better results.

5 Results

The analysis provided insights into the linguistic patterns of each speaker, emotional tones, and rhetorical styles. The results are summarized below based on the following figures:

5.1 Topic Distribution in Speeches

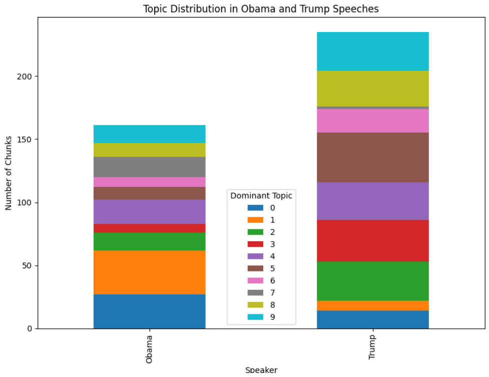

Figure 1: Topic Distribution in Obama and Trump Speeches

The topic modeling analysis, as depicted in Figure 1, reveals the distribution of topics in the speeches of Obama and Trump. The LDA model identified 10 distinct topics within the speech chunks.

- **Obama's speeches:** Display a diverse distribution across several topics with no single topic dominating excessively.
- **Trump's speeches:** Shows a stronger concentration and a narrower focus on specific topics.
- The two speakers seem to prioritize different themes in their speeches and usually they don't seem to align.

5.2 Most Frequently Used Catchphrases

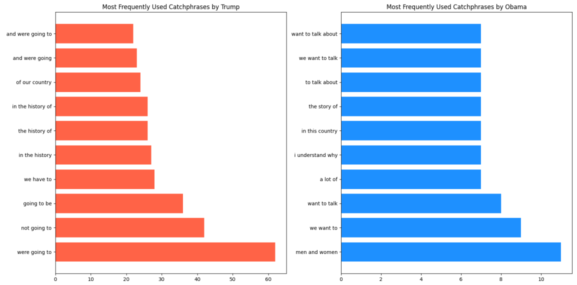

Figure 2: Most Frequently Used Catchphrases by Trump and Obama

The catchphrase analysis shows the most frequently used phrases by each speaker, as seen in Figure 2:

- **Trump:** Phrases like "and were going to", "of our country", and "we have to" are dominant which reflects a forward-looking and directive tone.
- **Obama:** Phrases such as "want to talk about", "we want to", and "i understand why" indicate a conversational, inclusive and empathetic approach.

5.3 Emotion Scores by Speaker

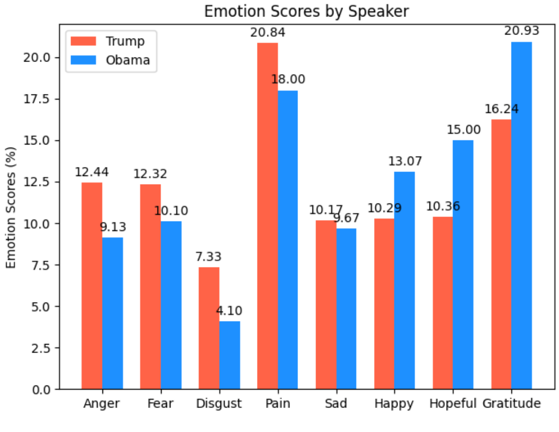

Figure 3: Emotion Scores by Speaker

The emotional analysis, using NRCLex, highlights the proportion of various emotions in their speeches, as shown in Figure 3:

- **Trump:** Higher scores in negative emotions such as Anger (12.44%), Fear (12.32%), and Pain (20.84%). This suggests a more intense and emotionally charged rhetoric.
- **Obama:** Higher scores in positive emotions like Happy (13.07%), Hopeful (15.00%), and Gratitude (20.93%). His speeches tend to convey a more optimistic and reassuring tone.

5.4 Top Adjectives Used

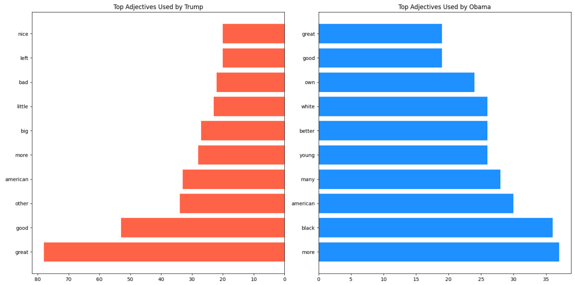

Figure 4: Top Adjectives Used by Trump and Obama

The adjective analysis shows the frequency of descriptive words used by each speaker, as seen in Figure 4:

- **Trump:** Common adjectives include "great", "good", and "american", emphasizing a patriotic and positive image.
- **Obama:** Frequently used adjectives are "great", "good", and "many", reflecting a diverse and inclusive vocabulary.

5.5 Semantic Similarity

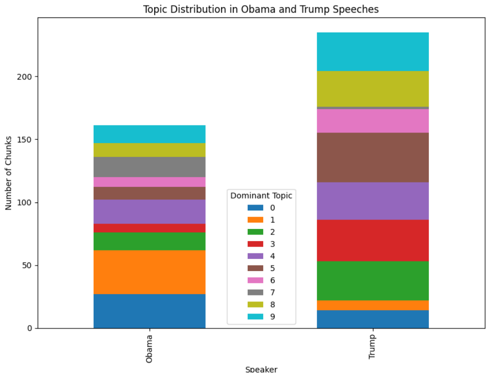

Figure 5: Semantic Similarity Between Obama and Trump Speeches

Using the paraphrase-MiniLM-L6-v2 model, the semantic similarity between the speeches was computed. The highest similarity score was 0.3730 which can be classified as moderate. This suggests that while their rhetorical styles are distinct, there are overlaps in the themes and contexts they address. A lot of it has to do with the public they address, the common talking points they need to address which can be universal across the political spectrum.

6 Limitations

Despite the useful findings, there are several limitations to point out. Firstly, the dataset used was limited to transcripts from a single source, potentially introducing bias in the type of speeches analyzed. Additionally, the normalization process might have led to the loss of certain details in the text which can affect the accuracy of the analyses. The sentiment analysis using NRCLex is limited by its fixed lexicon and lack of context-based understanding, which could result in less accurate emotion detection. Furthermore, the topic modeling and semantic similarity measures depend on the quality of the preprocessing steps and the parameters and sometimes it might not give the as precise results as desired.

7 Discussion

The comparative analysis of Barack Obama and Donald Trump's speeches reveals significant differences in their rhetorical approaches. Obama's speeches tend to be more diverse in topics and emotionally positive, indicating a more inclusive and optimistic rhetorical style. In contrast, Trump's speeches are more focused on specific themes and show higher levels of negative emotions which is reflective of a more intense and a matter-of-fact style. These findings are consistent with their public personas and communication strategies. The use of the NLP techniques in a multidimensional way proves itself reliable for analyzing and comparing the linguistic and emotional aspects of public speeches. However, the limitations identified highlight the need for more comprehensive datasets to improve the accuracy and depth of future analyses.

8 Conclusion

The project demonstrated the use of NLP techniques to conduct a comparative analysis of the speeches of Barack Obama and Donald Trump. The analysis highlights diverse linguistic and emotional patterns for each speaker. This methodology can potentially contribute to a better understanding of their communication strategies and public speaking effectiveness. Future work should aim to include a broader range of speakers and refine the models and techniques used to provide more accurate and comprehensive analyses.

References

- Dunmire, Patricia. (2012). Political Discourse Analysis: Exploring the Language of Politics and the Politics of Language. *Language and Linguistics Compass*, 6. https://doi.org/10.1002/lnc3.365

- Caiani, Manuela, and Di Cocco, Jessica. (2023). Populism and Emotions: A Comparative Study Using Machine Learning. *Faculty of Political and Social Sciences, Scuola Normale Superiore, Florence, Italy, Department of Political and Social Sciences, European University Institute, Florence, Italy, and Department of Political Science, Luiss Guido Carli, Rome, Italy*. Published online on 11 May 2023. https://doi.org/10.1017/S2049847013000057

- Rameshbhai, S., and others. (2020). Opinion Mining on Newspaper Headlines Using SVM and NLP. *International Journal of Data Mining & Knowledge Management Process (IJDKP)*, 10(2), 1-15. https://doi.org/10.5121/ijdkp.2020.10201