Analyzing Sentiment in Political Speeches

Dhrithi Guntaka
MehtA+

Mael Camilo Lefevre Castaneda
MehtA+

Ayazhan Zhumaken
MehtA+

Lorezno Rivera-Alvarez
MehtA+

July 25, 2024

Abstract

Political discussions play a heavy role in today's society, it often influences opinions and shapes communities. Due to its prevalence, understanding the tone and sentiment within political speeches is crucial. Our project analyzes the sentiment within political campaign speeches throughout the years using natural language processing (NLP) techniques. We originally implemented two models; the first uses one-hot encoding with linear regression and the second uses TF- IDF vectorization and linear regression. By calculating the Mean Square Error (MSE), we then determined the TF-IDF model was better fit to providing the text that had not already been scored with a sentiment score. The one-hot encoding model was best used for visualizing the sentiments trends averaged per year. The TF- IDF model has an MSE score of 0.21, while the One- Hot Encoding model had an MSE score of 0.24. We were successfully able to determine the evolution of sentiment throughout a speech for various political figures.

1 Introduction

Speeches are a fundamental medium of communication between political leaders and the public. They serve as a crucial tool for conveying messages and shaping how the public responds to things like proposed policies and legislations. Today's politicians are practically required to be skilled orators, their choice of works and ability to communicate effectively have a significant impact on their success in office. For these reasons, our project aims to analyze sentiment within speeches to help better understand the strategy behind how politicians speak with the public and how it has evolved.

Sentiment is the underlying tone and emotions within any text. It can be positive, negative, or neutral. Sentiment analysis used Natural Language Processing (NLP) to determine how positive or negative any given text is. Sentiment analysis can reveal persuasion techniques, calculate emotional engagement, and shifts in political strategy. We defined the sentiment score as an overall rating of the mood and enthusiasm of the speech, ranging from -1 to +1, lower scores indicate negative sentiment.

2 Related Work

Politics has remained a hot topic over the years with numerous researches analysing the different aspects of the US elections. One of the work that contributed to this field is a research paper "Sentiment is all you need to win US Presidential elections"[1], where they demonstrated that sentiment analysis could effectively capture public opinion and forecast election results based on the sentiment of speeches and social media posts. However, their study primarily focused on sentiment analysis during election periods, not covering the sentiment evolution over longer periods.

Our model will fills in those gaps, allowing us to visualise the sentiment trends and accurately score new, unscored texts, contributing new insights to the field of political sentiment analysis.

3 Methodology

3.1 Dataset

Our primary dataset was the Campaign Documents from "The American Presidency Project" at UC Santa Barbara. This dataset includes 23,469 documents starting from 1860 and is up to date. We began by selecting one campaign speech, at random, from each election and using Google's LLM (Large Language Model), Gemini, to synthetically generate a sentient score. In order to do this we split each speech into sections of four sentences and generated a sentiment score for each of those sections. This approach allowed us to see exactly how the speakers sentiment fluctuated within a single speech. The sentiment scores we generated ranged from -1 to +1. We considered scores from -1 to -0.1 as negative, scores from -0.1 to 0.1 as neutral, and scores from 0.1 to 1 as positive.

Preprocessing Preprocessing is key step in preparing data to make the model more accurate and efficient. To preprocess we used text splitting, one-hot encoding, and TF-IDF vectorization. In order to make the speeches smaller and more manageable, we broke it down into groups of four sentences. This made sure the data was suitable for analysis later on. We utilized one-hot encoding to convert the text data into numerical values which could be in the linear regression model.

3.2 Model

In our model, we utilized one-hot encoding to convert the text data into numerical values to be used in the linear regression model. Using TF-IDF vectorization, we converted the text into lowercase, removed any stop words, and convert it to vectors. We then trained a linear regression model using the training data that has been vectorized. This model was then used for predicted values on the test data. We used Mean Square Error (MSE) to assess the models performance.

Mean Square Error (formula shown below), was used as the loss function to measure the regression models performance. MSE is calculated by averaging the squared differences between the predicted and actual values from our model. The lower the MSE value, the better fit the model is to the data.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{1}$$

For the One-Hot Encoding model, our MSE error rate was 0.24. For the TF-IDF model, it was slightly lower at 0.21. This shows that both models are accurate but the TF-IDF model is more effective. TF-IDF stands for Term Frequency-Inverse Document Frequency, a statistic measure that can evaluate how important any word in a document is relative to the rest of the words in it.

4 Results

Our model can also work with individual speeches to analyze the sentiment fluctuation thought the speech as see in the graph below.

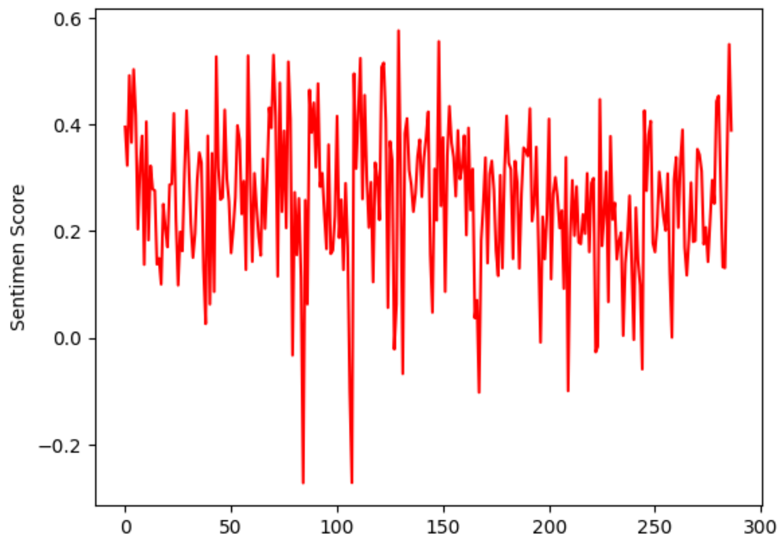

Figure 1: Donald Trump speech accepting the candidacy for the 2024 elections

We are able to compare speaking styles of different candidates as they give speeches regarding the same topic at the same time.

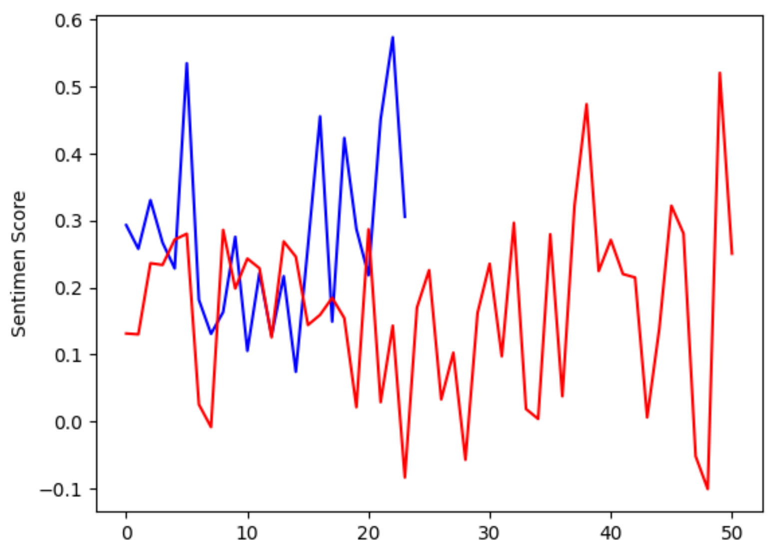

Figure 2: Joe Biden and Donald Trump speeches at the end of the 2020 elections

Below is a graph of three speeches by Woodrow Wilson where the green speech happened before WW1, the red speech during WW1, and the blue speech after. we can see how during the war the speech slowly went down in tone while the others remained somewhat constant.

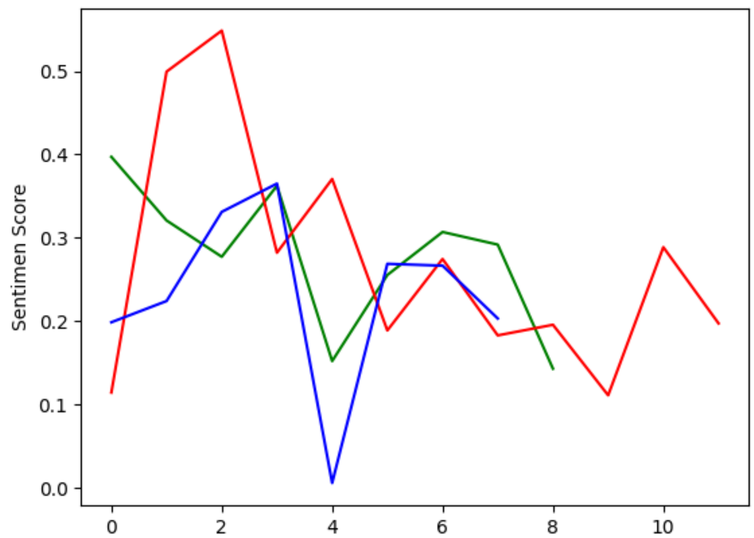

Figure 3: 3 speeches of Woodrow Wilson

5 Conclusion and Future Work

In this project we were able to successfully track sentiment in speeches, however there are many ways to expand on this research. First, we could expand the dataset used to include all of the campaign speeches available rather than using speeches chosen at random. Due to the sheer amount of documents available and our time constraints, we could not. Next, is to include debates and other speeches within the dataset, this would provide an even larger view of how sentiment changes.

6 Division of Labor

We divided the work as follows:

- Data Generation: Ayazhan Zhumaken, Dhrithi Guntaka, Mael Camilo Lefevre Castaneda
- Paper: Dhrithi Guntaka, Mael Camilo Lefevre Castaneda, Ayazhan Zhumaken
- Model: Mael Camilo Lefevre Castaneda, Dhrithi Guntaka
- Poster: Dhrithi Guntaka, Mael Camilo Lefevre Castaneda, Ayazhan Zhumaken, Lorenzo Rivera-Alvarez

7 Acknowledgements

We would like to acknowledge and thank Ms. Haripriya and Mr. Bhagirath for their advice and support throughout this the course of this project.

References

[1] Sovesh Mohapatra and Somesh Mohapatra. Sentiment is all you need to win US presidential elections. In Mika Hämäläinen, Khalid Alnajjar, Niko Partanen, and Jack Rueter, editors, *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*, pages 15–20, Taipei, Taiwan, November 2022. Association for Computational Linguistics.