NLP Final Project

Nina McClure, Lenka Sefcakova, Antonia George

February 2023

1 Introduction

The use of natural language processing (NLP) techniques to analyze political speeches has gained significant attention in recent years. In this paper, we examine Australian federal election data, including full text of speeches from election candidates, combined with metadata about the candidates and election outcomes.
The motivation behind this project is to use NLP methods to gain a deeper understanding of the language and rhetoric used by Australian politicians during federal election campaigns. Specifically, we aim to explore how politicians use language to appeal to different voter groups and what themes emerge in their speeches. This analysis aims to shed light on the political landscape in Australia and the priorities of different political parties. We do this by using techniques including sentiment analysis and TF-IDF.

2 Data

We extracted data containing election campaign launch speeches for Australian federal elections for the years 1901-2022 from an archive by the Museum of Australian Democracy [1]. We used the scraping techniques learned in class to extract the full texts of the speeches, as well as accompanying metadata, including name of the candidate, political party, whether they are an incumbent or challenger, location of speech, date of speech, and whether the candidate ended up being elected. The resulting data set consists of 92 speeches (two for each election year minus two speeches that were missing).

2.1 Web Scraping

The data is extracted from `"https://electionspeeches.moadoph.gov.au/speeches"`. The metadata used in analysis is obtained using `BeautifulSoup` python package. The website is organized in a table like structure, with either `class` or `id` carrying information (Figure 1).

3 Methods

Using methods we learned in class, we examined some descriptive statistics and visualizations of the data. We used the TF-IDF measure in order to obtain and visualize which words were most "important" across different speeches.

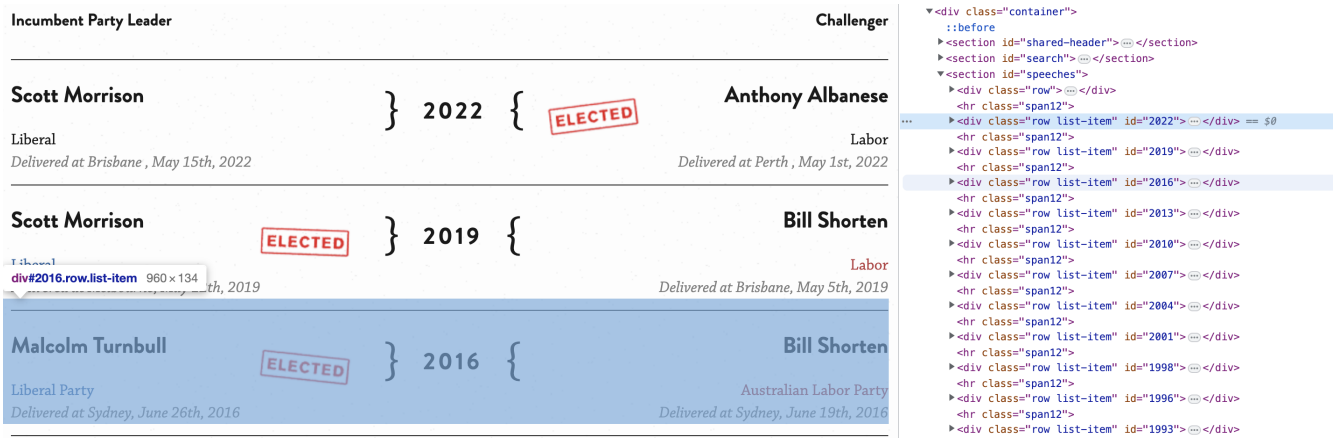

Figure 1: *Museum of Australian Democracy website with speeches and metadata structure.*

We then performed sentiment analysis techniques, using the python package 'nltk' to obtain opinion lexicon. By incorporating the opinion lexicon, we were able to analyze whether incumbent speeches tend to be more positive than challenger speeches. Additionally, we analyzed whether there are sentiment differences between parties. We also analyzed whether candidates who ended up being elected made more positive speeches than those that were not elected. We used t-tests to test for significant differences in average speech sentiment between the different groups.

3.1 Pre-processing for TF-IDF analysis

As we can see in Figure 2, the raw text in a word cloud form has little informative value and one can hardly tell what the context is. In this section we will analyse and argue for specific pre-processing choices we made for the TF-IDF analysis and provide concrete examples to emphasise the impact these choices have on the results of the analysis. The choices will be illustrated on the whole corpus unless otherwise specified. We will work with already lowercased text.

Stop words: As shown in Figure 2, arbitrary stop word removal has helped as now we can understand that the data is clearly about Australia, we however already possess this information and therefore the word cloud in the middle of the figure is still not useful. We therefore opted to also include a custom stop word list including terms such as 'Australia', 'government', 'election' etc.. The full list of stop words used can be found in Figure 14 in the appendix. Additionally, we remove all non-letter characters, such as special characters and digits as the quoting and budget planning presented in the speeches has little value to our analysis.

Stemming vs. Lemmatizing: As can be seen in Figure 3 the stemmed results are much easier to interpret. Moreover the size of our vocabulary is 10,455 for stemming and 14,913 for lemmatization. Albeit the vocabularies are relatively large, with a small dataset the price for higher interpretability that is available with using lemmatized words is relatively low, especially as we will decrease this dimension in future steps.

The final pre-processing pipeline is then:

1. Lower case
2. Special character and digit removal and stop word removal, including a custom list of words
3. Lemmatization of words

Figure 2: *The most common words used in the Australian federal election speeches from 1901-2022 before any pre-processing steps with no stop word removal (left), and after common stop words are removed (middle) and after custom stop words are removed (right)*

Figure 3: *The word clouds generated using stemming (left), and lemmatizing (right) after stop words and other characters defined in previous step are removed*

3.2 Parameters used in TF-IDF analysis

We used the `TfidfVectorizer((preprocessor=pipeline,...)` to compute the TF-IDF scores for each group of interest. As we have passed through our pre-processing steps we now work with a vocabulary of lemmatized words present in our corpus. The lemmatized result in Figure 3 (right) shows the word cloud generated using the frequencies (TF-IDF scores) without any limitations on word document frequency within the corpus.

Bounding the document frequency of a word can help us reduce the vocabulary in two ways. Firstly, setting an upper bound on this further strips the corpus of corpus-specific "stop words" that we have not added to the custom list. By setting the value to 70% we restrict the words in our vocabulary to be in at most 70% of all documents. The difference can be spotted in Figure 4 as now we have more policy specific terms such as education and job present.

The size of the vocabulary used by the model is now 14,652. This is where the second bound comes in, as the TFIDF score measures relative frequencies we can turn our focus on words that reappear, specifically only allowing for words that are in at least n% of all the documents. In

Figure 4: *The word clouds generated bounding the word document frequency above by 70% (left), and pre-processing without applying the upper bound (right)*

Figure 5 we see that in the most frequent words there has been no change by adding a lower bound of 20%. The vocabulary size has been pushed down to 1,653 words.

Figure 5: *The word clouds generated bounding the word document frequency above by 70% (left), and adding a lower bound of 20% (right) after pre-processing*

As we would like to pick up on trends in a possibly smaller group, we argue that the bound might be too tight and lower it to 5% of all documents: i.e., a word has to be present in at least 4 documents in our corpus which translates to at least 2 elections. As we have seen, setting the lower bound has no effect on the figure representation of the corpus and so we will not include an illustration. The resulting vocabulary size is 4,897. It is important to say that we are setting this lower bound while knowing that all of our future analysis will be on evenly split data, i.e., approximately 1:1 ratio in document counts. The resulting representations by different groups can be seen in Figure 6 for blue (right-leaning) and red (left-leaning) parties and in Figure 7 for visual comparison between incumbents and opposition. Additionally in Tables 5, we can see unique vocabularies used by each group.

Figure 6: *These word clouds show the most common words used in the Australian federal election speeches by party.*

Figure 7: *These word clouds show the most common words used in the Australian federal election speeches split up by incumbents vs. opposition candidates.*

The blue party appears to focus more socialistic views where the red party has words like budget, housing, profit and taxation present in their speeches. In the incumbent and opposition separation, however, this is not as clear, and unfortunately no concrete results can be drawn. We therefore continue with sentiment analysis in the hope of shedding some more light onto the Australian political landscape.

4 Sentiment Analysis Results

We analyzed the speeches by calculating a "positivity" score. To do this, we calculated how many positive and negative words there were in each speech and calculated the "positivity" score as follows:

$$Sentiment\ Score = \frac{number\ of\ positive\ words}{number\ of\ positive\ words\ +\ number\ of\ negative\ words}$$

Using the sentiment "positivity" scores, we tested and analyzed 3 different questions. We tested whether "positivity" was significantly different between different parties, between incumbents vs. opposition candidates, and between candidates that were elected vs. not elected.

4.1 Sentiment Analysis - By Party

We first tested whether sentiment is significantly different between parties. We had two categories for parties: left-leaning and right-leaning. The mean "positivity" score for left-leaning candidates was 0.733 and the mean "positivity" score for right-leaning candidates was 0.703 (Figure 8). We performed a t-test on the null hypothesis that the mean scores are equal and obtained a p-value of 0.051, meaning that we cannot conclude (at at least a 95% confidence level) that a significant difference between the means exists.

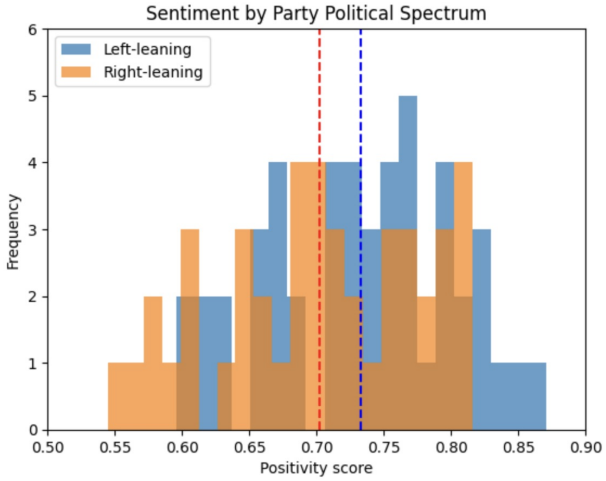

Figure 8: *This graph shows the distribution of sentiment scores by party. The mean score for left-leaning candidates is higher than for right-leaning candidates, but the t-test performed could not conclude that there is a stastically significant higher mean for left-leaning candidates.*

4.2 Sentiment Analysis - Incumbent vs. Opposition

We also tested whether sentiment is significantly different between candidates who are currently incumbents versus those who are not incumbents. We hypothesized that incumbents may tend to be more positive than their opposition. The mean "positivity" score for incumbent candidates was 0.741 and the mean "positivity" score for non-incumbent candidates was 0.696 (Figure 9). We performed a t-test on the null hypothesis that the mean scores are equal and obtained a p-value of 0.002, meaning that a significant difference between the means exists. The t-test results provide evidence that incumbents give more positive speeches, on average.

4.3 Sentiment Analysis - Elected vs. Not Elected

Lastly, we analyzed sentiment between candidates who ended up being elected versus those were not elected. The mean "positivity" score for elected candidates was 0.735 and the mean "positivity" score for non-elected candidates was 0.701 (Figure 10). We performed a t-test on the null hypothesis that the mean scores are equal and obtained a p-value of 0.028, meaning that a significant difference

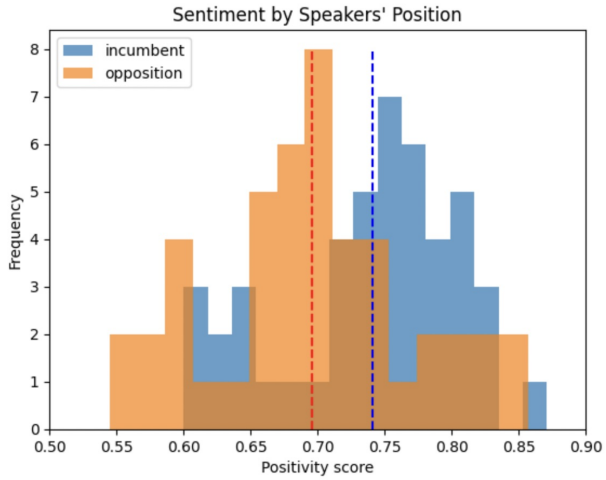

Figure 9: *This graph shows the distribution of sentiment scores by position (whether the candidate is an incumbent or opposition). The mean sentiment score for incumbents is higher than for opposition candidates, meaning that incumbents are more positive, on average.*

between the means exists. The t-test results provide evidence that candidates who ended up being elected make more positive speeches, on average.

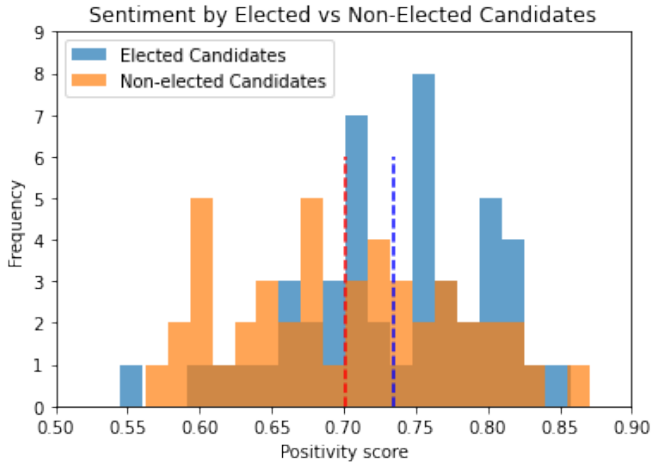

Figure 10: *This graph shows the distribution of sentiment scores by whether candidate was elected or not. The mean sentiment score for those who were elected was higher than for those who were not elected, on average.*

4.4 Sentiment Analysis - Further Analysis

These results bring up an interesting question. Since incumbents tend to be more positive in their speeches and candidates who were elected were more positive - is there an inherent problem when trying to disentangle whether this correlation is at play when looking at the difference in sentiment between the groups? We run into another issue here because our sample set is relatively small, so it's difficult to run many statistical models that provide a clear answer.
We find that incumbents tend to be re-elected more often than not (roughly 70 per cent of the

time), and so it may be the case that the results we found in the sentiment analysis on elected versus non-elected candidates in section 5.3 may just reflect that incumbents are over-represented in the elected group and also tend to have a more positive tone. To attempt to answer this question, we chose to look at sentiment across elected versus non-elected candidates stratified by incumbents and opposition. We observe that elected incumbents tend to have a more positive tone on average, although a t-test yields an insignificant result (Figure 11). On the other hand, elected and non-elected opposition candidates have very similar tones (Figure 12). So, it does appear that the result we saw in section 5.3 was to some extent driven by the over-representation of incumbents within the elected group.

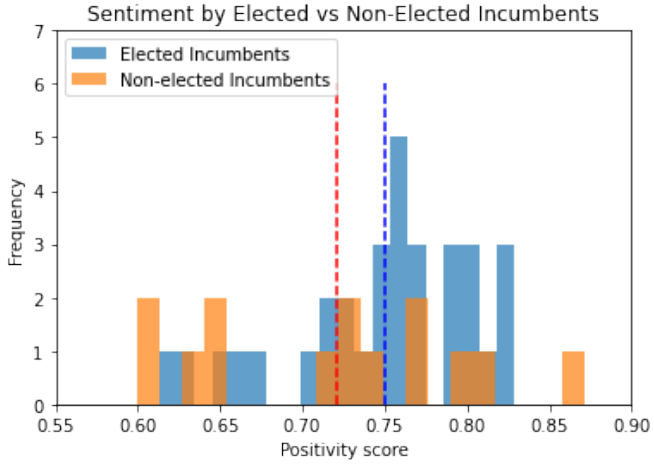

Figure 11: *This graph shows the distribution of sentiment scores by elected vs. non-elected incumbents.*

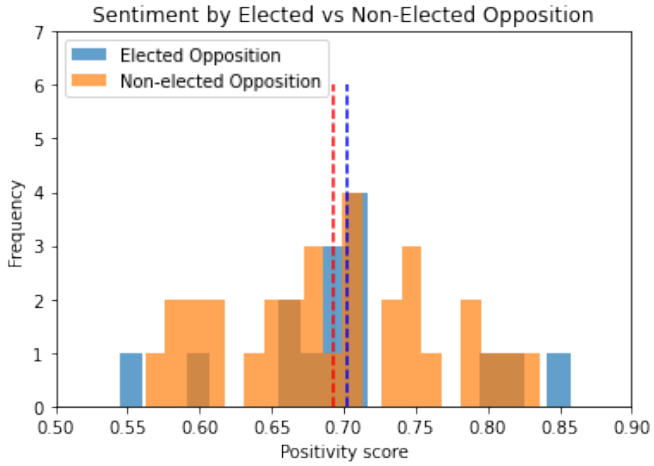

Figure 12: *This graph shows the distribution of sentiment scores by elected vs. non-elected opposition candidates.*

Finally, we looked at how sentiment changed over time. We thought this would be an interesting exercise given that we have over 100 years' worth of data. However, it is difficult to infer meaning from the results, as we find that sentiment was quite volatile over the period (Figure 13). Nonetheless, we observe an upward trend in sentiment over time.

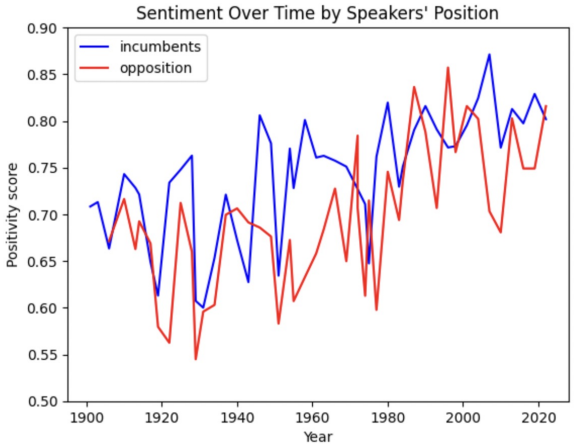

Figure 13: *Sentiment over time.*

5 Conclusion

This paper has demonstrated that using natural language processing techniques is a unique and effective way to gain insights into the political landscape of federal elections. The pre-processing choices we have made for the analysis have shown that by lemmatizing, bounding the document frequency of a word, and removing stop words, we are able to focus our analysis on policy-specific terms and obtain more meaningful results from our data. Through the use of sentiment analysis and TF-IDF measures, our analysis has shed light on the themes that emerge in candidate speeches, the differences between top words/topics among different types of candidates, and the differences between mean sentiment among groups. Our sentiment analysis has yielded several important insights. The results of t-tests showed us that:

- There is no conclusive evidence that the mean sentiment score of left vs. right leaning party candidates differs.
- Incumbents tend to have more positive speeches than opposition.
- Candidates who were elected had more positive speeches, on average, compared to candidates who were defeated.
- After examining the sentiment results stratified by incumbents and opposition, it appears as though the results from the sentiment analysis that saw elected candidates being more positive was in fact, to some extent, driven by over-representation of incumbents in the elected group, who tend to be more positive.

Overall, our findings highlight the potential usefulness of NLP techniques in analyzing political language. Our results also have implications for political campaigning and messaging, suggesting that voters respond positively to more positive messaging.

References

[1] "All speeches." [Online]. Available: https://electionspeeches.moadoph.gov.au/speeches

Appendix

australia government prime minister labor party commonwealth australian election vote politics minister howard menzies hear hear billion 000 000 000 hear labor labour parties party

Figure 14: *List of custom stop words used in TF-IDF analysis after lower casing the corpus*

Blue Score	Blue	Red	Red Score
0.128245	empire	program	0.141436
0.114761	cheer	hospital	0.078541
0.105096	li	housing	0.076302
0.104700	programme	overseas	0.072461
0.079508	tariff	fadden	0.072060
0.076982	british	profit	0.062833
0.072670	communist	budget	0.062556
0.072533	socialist	taxation	0.062493
0.072178	coalition	growth	0.061867
0.071835	applause	legislation	0.061087

Table 1: *Unique vocabulary for subset Blue from their top 20 scoring words cross referenced with subset Red and vice-versa with corresponding TFIDF scores (top 10)*

Incumbent Score	Incumbent	Opposition	Opposition Score
0.108992	programme	housing	0.081086
0.086360	li	fadden	0.073701
0.083563	export	inflation	0.071071
0.076927	british	profit	0.070161
0.069625	tariff	hospital	0.069720
0.066623	assistance	taxation	0.069401
0.064291	provision	wheat	0.068212
0.063822	peace	worker	0.067082
0.063496	bill	overseas	0.064361
0.063216	communist	socialist	0.061195

Table 2: *Unique vocabulary for subset Incumbent from their top 20 scoring words cross referenced with subset Opposition and vice-versa with corresponding TFIDF scores (top 10)*