

# NLP Final Presentation

Nina McClure, Lenka Sefcakova, Antonia George

February 22, 2023

# Table of Contents

- 1 Motivation
- 2 Web scraping
- 3 Data
- 4 Word Clouds (TFIDF)
- 5 Sentiment Analysis
- 6 Conclusion

# Motivation

- Using NLP methods, we analyze the language and rhetoric used by Australian politicians during federal election campaigns.
- We explore how politicians used language to appeal to different voter groups, what themes emerged in their speeches, and how their messages changed over time.
- We aim to shed light on the political landscape in Australia, the priorities of different political parties, and the factors that influence voters' decisions.

# Web Scrapping

We obtained the Australian federal election speeches data from the website, compiled by the Museum Of Australian Democracy.

[Read the speeches](#)[Explore the speeches](#)

Australian Federal

ELECTION SPEECHES




[Search](#)

Incumbent Party Leader		Challenger
<b>Scott Morrison</b> Liberal <i>Delivered at Brisbane , May 15th, 2022</i>	} 2022 { <b>ELECTED</b>	<b>Anthony Albanese</b> Labor <i>Delivered at Perth , May 1st, 2022</i>
<b>Scott Morrison</b> Liberal <i>Delivered at Melbourne , May 12th, 2019</i>	<b>ELECTED</b> } 2019 {	<b>Bill Shorten</b> Labor <i>Delivered at Brisbane, May 5th, 2019</i>

# Web Scrapping Methodology

url = "https://electionspeeches.moadoph.gov.au/speeches"

The metadata used in analysis is obtained using BeautifulSoup

Incumbent Party Leader		Challenger
<b>Scott Morrison</b> Liberal Delivered at Brisbane, May 15th, 2022	} 2022 { 	<b>Anthony Albanese</b> Labor Delivered at Perth, May 1st, 2022
<b>Scott Morrison</b>  } 2019 {		<b>Bill Shorten</b> Labor Delivered at Brisbane, May 5th, 2019
<b>Malcolm Turnbull</b> Liberal Party Delivered at Sydney, June 26th, 2016	 } 2016 {	<b>Bill Shorten</b> Australian Labor Party Delivered at Sydney, June 19th, 2016

```
<div class="container">
  <:before
  <section id="shared-header">
  <section id="search">
  <section id="speeches">
    <div class="row">
      <div class="span12">
        <div class="row list-item" id="2022">
          <div class="span12">
            <div class="row list-item" id="2019">
              <div class="span12">
                <div class="row list-item" id="2016">
                  <div class="span12">
                    <div class="row list-item" id="2013">
                      <div class="span12">
                        <div class="row list-item" id="2010">
                          <div class="span12">
                            <div class="row list-item" id="2007">
                              <div class="span12">
                                <div class="row list-item" id="2004">
                                  <div class="span12">
                                    <div class="row list-item" id="2001">
                                      <div class="span12">
                                        <div class="row list-item" id="1998">
                                          <div class="span12">
                                            <div class="row list-item" id="1996">
                                              <div class="span12">
                                                <div class="row list-item" id="1993">
```

Year by year structure as descendants of speeches object, table organization, either class or id carries information, rotated elected sticker - looping needed

The data was obtained from Australian politicians' speeches during federal election campaigns from 1901-2022 and contained:

- full texts of the speeches
- name of the candidate
- political party
- whether they are an incumbent or challenger
- location of speech
- date of speech
- whether the candidate ended up being elected

## Word Cloud- Unprocessed Data

## Unprocessed Data



## Pipeline:

- ① **Lower case**
- ② **Special Character and digit removal** (mentions on budgeting and years of speech as metadata in texts, quoting ””)
- ③ **Stop word removal** including **custom list** of words
- ④ **Lemmatization** (preferred over stemming - enhances interpretability, slightly higher dimension of vocabulary - negligible)



Pass to `TfidfVectorizer((preprocessor=pipeline, max_df = 0.7, min_df = 0.05, ngram_range = (1,1))`

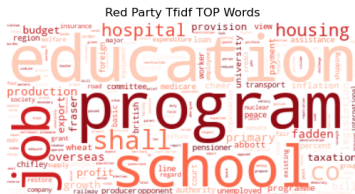
**Uni-grams** (lower vocabulary dimension from 8741 (bi-gram) to 4895 without impact on results, improves computation time)

`min_df = 0.05` each word in vocabulary in at least 5% of documents

`max_df = 0.7` each word in vocabulary in at most 70% of documents



# Word Clouds by Party



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

# Vocabulary by party

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
0.068441	strong	fraser	0.060741
0.067761	worker	bill	0.059214

**Table:** 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 vs. Opposition vocabulary

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:** 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)



# Elected vs. Defeated Vocabulary

Elected Score	Elected	Defeated	Defeated 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
0.068441	strong	fraser	0.060741
0.067761	worker	bill	0.059214

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

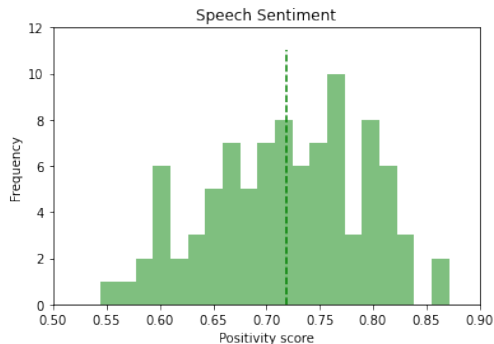


# Sentiment Analysis - Method

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

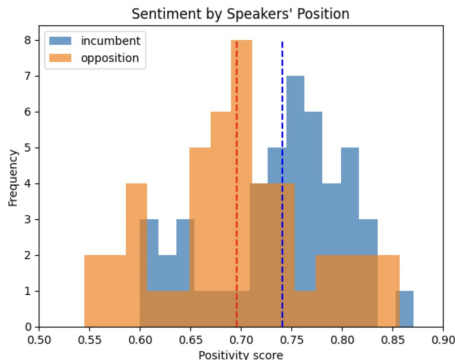
$$\textit{Sentiment Score} = \frac{\textit{number of positive words}}{\textit{number of positive words} + \textit{number of negative words}}$$

# Sentiment Analysis - All Speeches



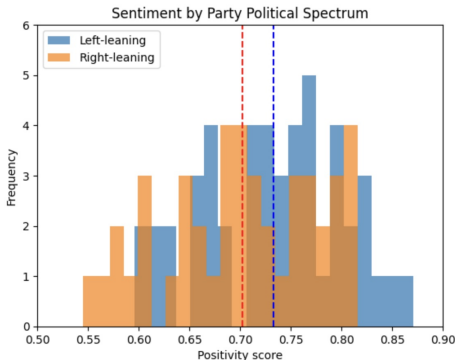
**Figure:** This graph shows the distribution of sentiment scores for all speeches. The mean score is around 0.72, while the minimum and maximum scores are 0.54 and 0.87, respectively.

# Sentiment Analysis - Incumbent vs. Opposition



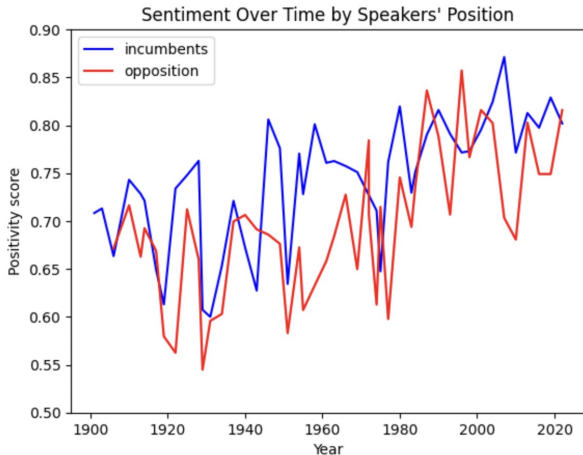
**Figure:** This graph shows the distribution of sentiment scores by position (whether the candidate is an incumbent or opposition). With the hypothesis that incumbents would tend to be more positive, we performed a t-test on the mean scores and obtained a p-value of 0.002. The t-test provides evidence that incumbents give more positive speeches, on average.

# Sentiment Analysis - By Party



**Figure:** This graph shows the distribution of sentiment scores by party. The mean score for left-leaning candidates is higher than for right-leaning candidates. We performed a t-test on the null hypothesis that the mean sentiment scores between parties are equal and obtained a p-value of 0.051, meaning that we cannot conclude (at  $\geq 95\%$  confidence level) that a significant difference between the means exists.

# Sentiment Analysis Over Time



**Figure:** This graph shows sentiment scores over time by position (whether the candidate is an incumbent or opposition).

# Results

- Incumbents tend to make more positive speeches than the opposition.
- There is no conclusive evidence that the mean sentiment score of left vs. right leaning party candidates differs.
- Sentiment appears to have improved over time.