Topic Modeling and Electoral Impact of Canadian Parliamentary Speeches (1980–2020)

Sebnem Sera Uysal

Matriculation No: 12629559

ABSTRACT

This project analyzes Canadian parliamentary speeches from 1980 to 2020 with three main objectives:

- To identify the dominant political topics discussed in Parliament over time,
- To examine the thematic focus of speeches by Prime Ministers and opposition party leaders,
- To investigate the relationship between Members of Parliament's topic preferences and their subsequent electoral performance.

TABLE OF CONTENTS

INTRODUCTION

In democratic societies, speeches delivered in parliament play a key role in shaping policies and guiding public sentiment. From 1980 to 2020, the Parliament of Canada has been a platform for a wide range of themes and priorities, which have shifted over time to reflect the country's socio-political changes.

Why is this research important? Recognizing themes in parliamentary speeches helps us understand a nation's changing priorities and challenges. For those in policy-making and political analysis, understanding these trends is essential for planning and predicting future political directions.

Purpose of the Paper: This paper presents a detailed analysis of the content and themes of Canadian parliamentary discussions from 1980 to 2020. This research delves into these speeches with three main objectives:

1. Identifying the main topics discussed over the four decades.

2. Analyzing the focus placed by prime ministers and opposition leaders on these subjects.

3. Exploring potential correlation between the most frequently discussed topics by each parliamentarian and their electoral outcomes.

Considering the vast amount of speeches over this period, traditional methods fall short. This is where Natural Language Processing (NLP) and Machine Learning (ML) come into play. These advanced techniques offer solutions to process large text data, identify patterns and extract conclusions.

The data was sourced directly from the Parliament of Canada's website using web scraping. The speeches were then combined with detailed profiles of the parliamentarians from the same period.

My analysis comprised several steps:

Data Acquisition and Cleaning: Web scraping was coupled with dynamic webpage interactions. Potential inconsistencies and missing values in the datasets were addressed and the data was then transformed and merged.

Natural Language Processing: NLP libraries were used to process the speech data, emphasizing relevant terms from the speeches.

Machine Learning: To classify the speeches into specific categories, I utilized K-means clustering. This revealed topics such as "People and Governance", "Procedure", "Trade", "Health and Childcare", "Broad Governance Topics", "Gender and Social Issues", "Finance and Taxes", "Employment and Economy", "Crime and Law" and "Agriculture".

Early results indicate some interesting patterns. For instance, Prime Ministers from 1980 to 2020 frequently addressed topics such as "People and Governance", "Trade" and "Employment and

Economy". On the other hand, Opposition Party Leaders have mostly addressed topics such as "People and Governance", "Trade", "Broad Governance Topics", "Finance and Taxes", "Employment and Economy", "Crime and Law" and "Health and Childcare". In addition, opposition party leaders expanded their focus over the years, with a significant emphasis on "Gender and Social Issues" after 2000.

Correlation Analysis: My core interest lies in understanding the link between the most frequently talked about topics and electoral outcomes. As my outcome variable, I used the percentage of votes each parliamentarian received in each election. I believe this approach, focusing on vote percentage (vote share in a particular election) over binary outcomes like "Elected" or "Defeated", would provide a clearer insight into the correlation between discussed topics and electoral performance. To control for potential variables, I included factors like "Age_at_Election", "duration_of_service" and Party affiliation in my regression analysis. I employed fixed effects regression because it enables us to control for variables that remain constant over time, offering a more unbiased coefficient estimation for the variables of interest.

My analysis suggests that there might be an overall positive correlation between the most frequently talked about topics of each parliamentarian where I consider both the speeches just before the election (election year and one year before it) and all four-year periods (election year and three years before it). Particularly, considering just before the election, discussing topics like "employment and economy," "agriculture," and "health and childcare" are highly correlated with parliamentarians' electoral results. Over a broader four-year period, emphasis on "health and childcare" consistently results in the most significant increase in electoral performance, followed closely by "employment and economy" and "agriculture," highlighting the continuous importance of these subjects to voters.

DATA DESCRIPTION

The data for this study are collected from the online profiles of members of the Parliament of Canada[1]. These profiles provide a comprehensive overview of each parliamentarian's career, including their electoral history, federal experience, committee memberships, provincial experience, family ties and municipal experience. These data are the foundation for the subsequent analyses and for understanding the topics they discuss and the analysis to determine the correlation with their electoral performance.

My additional input sources are

1) speech data made in the Parliament of Canada, including the date of the speech, speaker's name, party, position and riding. It also provides a categorization of the speech into the main topic, subtopic and sub-subtopic, alongside the complete text of the speech and additional identifiers.

[1] https://lop.parl.ca/sites/ParlInfo/default/en_CA/People/parliamentarians

2) FED_Results, which provides details about the elections in Canada, information about candidates and their political affiliations, votes received and results of each election. It also includes data about provinces, regions, start and end dates of terms and unique identifiers for federal electoral districts.

DATA PROCESSING

The following figure summarizes my data processing workflow based on Python. It contains 5 main scripts, their inputs and outputs that will be explained in detail in the next sections.

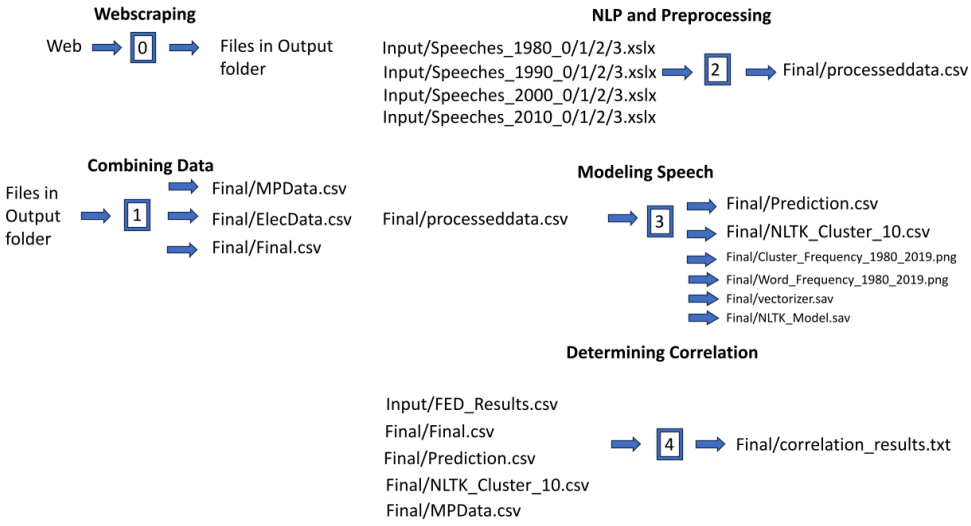


Figure 1: Data Processing Workflow

The data are first gathered using a web scraping tool developed in Python[2]. We use the "selenium" library in conjunction with the Chrome browser to automate the process of accessing and extracting data from the web pages of interest. This automated tool navigates to the profile page of each parliamentarian and extracts the required information from the HTML content using the "BeautifulSoup" library. Extracted data are then saved in different CSV and Excel files for further analysis in the "Output" folder of the project root.

The following data are collected for each parliamentarian if it exists on the web page of the Parliament of Canada:

Name: Full name of the parliamentarian.

Electoral History: Detailed records of the parliamentarian's electoral performances, including the elections they contested, votes they received and the outcome of each election.

[2] See "0_Download_Files.py"

Federal Experience: Information on the parliamentarian's experience at the federal level, including the positions they have held and the duration of their service.

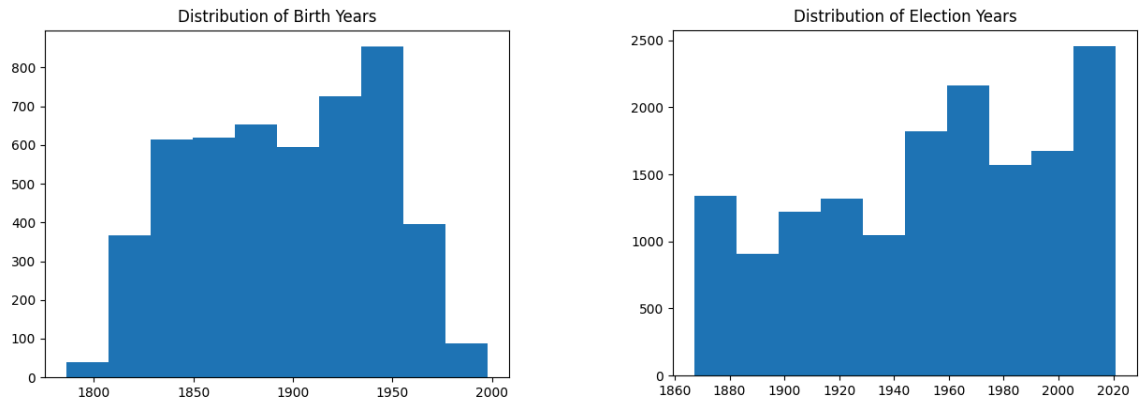
Committee Membership: Details of the committees the parliamentarian has been a part of, along with their roles and responsibilities within those committees.

Provincial Experience: Insights into the parliamentarian's career at the provincial level.

Family Ties in Parliament: Information on any family connections the parliamentarian may have within the Parliament of Canada.

Municipal Experience: Details of the parliamentarian's experience in municipal governance.

I first pull together individual MP profiles by merging data of parliamentarians and saving them to a single file, i.e. Final/MPData.csv. During this step, I pay special attention to cleaning and standardizing their names and extracting their birth years, which I then visualize to understand the distribution across the dataset[3]. Similar to the MP profiles, I aggregate the electoral history data of these MPs from the downloaded Excel files and save them into "Final/ElecData.csv". Plots 1 and 2 below show the distribution of birth years and election years of all parliamentarians for all periods, based on the extracted information from these files, respectively.



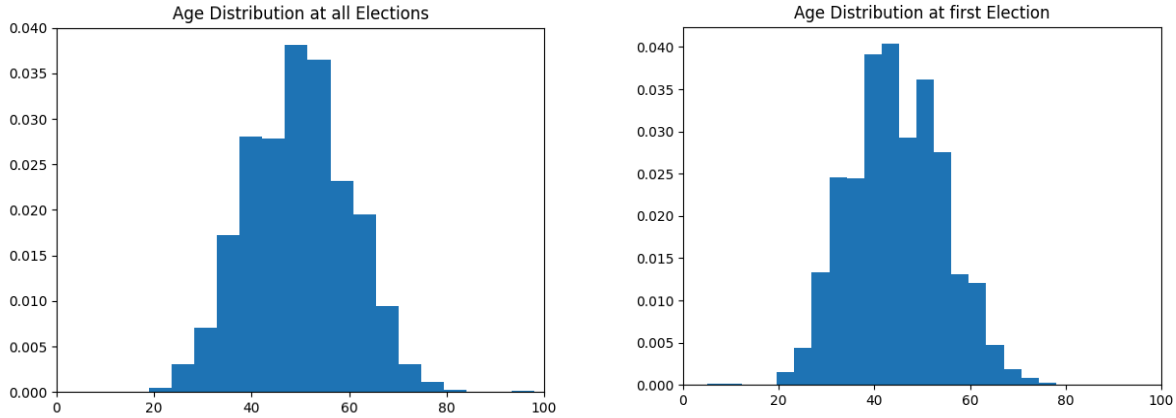
Plot 1 and 2: Distribution of Birth Years and Election Years of Parliamentarians

Following this, I process data on Federal Districts (FEDs). I clean the FED data for each MP and merge them into a single data frame. Then I merge this data frame with "Final/ElecData.csv" on the common Constituency and Name column. Finally, I combine this merged dataset with "Final/MPData.csv" file on the common MP_ID column to create a comprehensive dataset[4]. With

[3] See "1_Combine_MPInformation.py".

[4] MP_ID appears to be a unique identifier for each Member of Parliament (MP). MP_ID is associated with one specific MP and all data related to that MP (such as speeches, topics discussed, years of service, etc.) would reference this ID to ensure accuracy and consistency.

this in place, I calculate the age of MPs at the time of their elections by subtracting birth year from the election year. Plot 3 below visualizes the age distribution of MPs at the time of all elections, whereas Plot 4 below visualizes the age distribution of MPs at the time of their first election.



Plots 3 and 4: Age Distribution at all Elections and Age Distribution at First Election

I continue with preprocessing the speech data from Parliamentarians over the years 1980 and 2020[5]. My goal is to obtain clean and tokenized speeches free of stopwords, certain banned words and other potential noises. For this purpose, I

- convert speech text to lowercase,
- tokenize, i.e. split the text into individual words,
- remove stopwords for English (from NLTK library) and specific banned words, i.e. 'i', 'thank', 'you', 'canada', 'minister', 'speaker', 'mr.', 'mrs.', 'ms.', 'madame', 'hon', 'dear',
- find stems of the words,
- apply NLTK's post-tagging algorithm to keep only certain types of words, i.e. verbs, nouns and adverbs, to focus on the main content of the speeches.

To speed up the cleaning and processing of speeches, especially given the large volume of data, I use a "ThreadPoolExecutor". This approach allows us to process multiple batches of data concurrently, significantly speeding up the overall process.

While preprocessing the speech data, I determined the average length of raw speeches as:

- 1009 characters for speeches between 1980 and 1990
- 1377 characters for speeches between 1990 and 2000
- 1380 characters for speeches between 2000 and 2010
- 1274 characters for speeches between 2000 and 2010

[5] See "2_Prepare_Speech.py".

Based on these findings, I filtered out the raw speeches whose lengths are less than 1000 characters to remove short speeches that might not contain enough meaningful content for my analysis.

Next, I transform raw speech data into a clean, structured format suitable for analysis. The processed data have each speech tokenized and cleaned, ready for further steps such as topic modeling. The output of this process is "Final/processeddata.csv".

After getting the data ready, I start on speech modeling[6]. My main goal is to cluster the speeches based on their main topic and figure out the most frequent words in each group. I am interested in clustering the speeches based on their content because clustering provides us with a structured way to understand the main themes or topics discussed in the speeches over the years.

To proceed with speech modeling, I load the data from "Final/processeddata.csv" and tokenize the words using the TfidfVectorizer to ensure they are in a numerical format optimal for clustering. This step is essential to obtain the term frequency and inverse document frequency (TF-IDF) representation, which captures the importance of terms within the speech relative to the entire dataset. This transformation gives more weight to words that are frequent in a single speech but not across all speeches, emphasizing the uniqueness of each speech. I then employ the K-Means clustering model on this matrix, aiming to group the speeches into a specified number of clusters (cluster_groups). K-Means clustering works by selecting random centroids and iteratively optimizing these points to minimize the variance within each cluster while maximizing the variance between different clusters. Each resulting cluster ideally represents a collection of speeches with similar themes or content. During training and testing the data, I do not use sampling.

After clustering, I identify the most prevalent words in each cluster. These key terms offer insights into the primary themes of each cluster. The table below presents the clusters, where each cluster is represented as a column. The words are ranked from most to least frequent in each cluster:

Cluster/ Words	0	1	2	3	4	5	6	7	8	9
	peopl	membe r	trade	health	member	women	tax	program	crime	farme r
	member	commit tee	agreeme nt	care	peopl	violenc	incom	budget	crimin	farm
	nation	motion	trade agreeme	health care	govern	men	budget	job	victim	grain

[6] See "3_Model_Speech.py".

			nt							
	govern	order	unit state	provinc	year	work	pay	year	sentenc	wheat
	act	questio n	state	child	question	program	year	cent	offend	agricu ltur
	quebec	debat	unit	child care	time	children	cent	govern	offenc	produ c
	legisl	stand	export	feder	say	woman	govern	peopl	code	wheat board
	countri	rule	countri	drug	work	gender	busi	tax	law	board
	time	report	industri	govern	countri	famili	credit	money	crimin code	marke t
	say	chair	market	need	want	member	tax credit	work	court	produ ct

Table 1: Frequent Words in Clusters

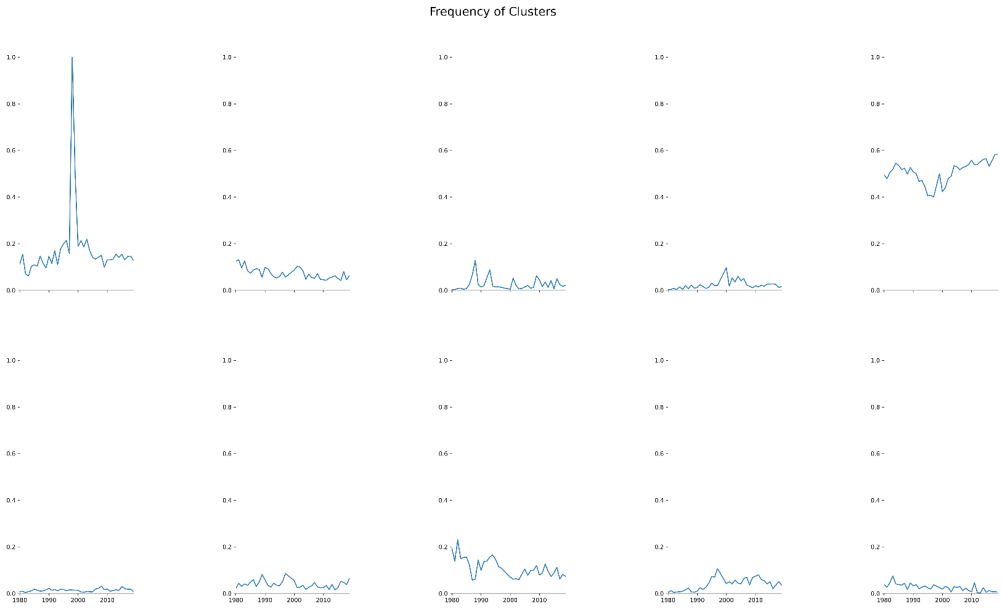
Based on these frequent words and clusters, I came up with the following terms that define each cluster:

Cluster index	**Cluster Label**
0	People and Governance
1	Procedure
2	Trade
3	Health and Childcare
4	Broad Governance Topics
5	Gender and Social Issues
6	Finance and Taxes
7	Employment and Economy
8	Crime and Law
9	Agriculture

Table 2: Theme of Each Cluster

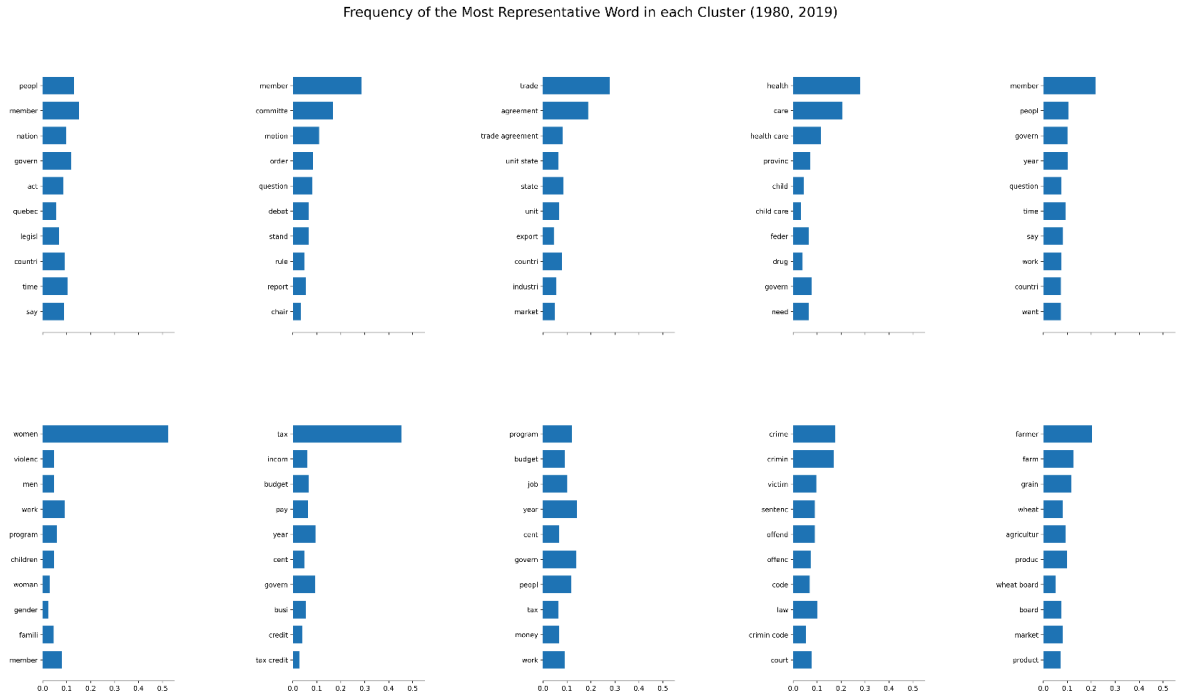
I use the trained clustering model to assign each speech in the dataset to a cluster. This assigns a cluster number to every speech, reflecting the cluster it most closely aligns with based on its content. I also visualize how these clusters are distributed over various years. This clustering approach provides a structured way to understand the main themes or topics discussed in the speeches over the years.

The following plots are generated by the Python script[7]. The first plot visualizes the frequencies of cluster assignments over time, whereas the second plot visualizes the frequency of the most representative words. By looking at both charts, I can understand the main discussions in parliament and the words that were often used for those topics. For example "people and governance", "broad governance" and "employment and economy" base these three topics of speeches more than other topics between 1980-2020.



Plot 5: Cluster History Plot[8]

[7] See "3_Model_Speech.py".
[8] This shows the frequency of each cluster over time, helping to visualize trends.



Plot 6: Word Frequencies in each Cluster[9]

What are the topics that people were talking about 1980-2020?

From 1980 to 2020, parliamentarians discussed a variety of topics, as identified by the clusters of speeches. These topics give a comprehensive overview of the discussions. Dominant themes include the following

People and Governance: This theme includes discussions about the general society and how governance affects them or is influenced by them.

Procedure: This covers the procedural aspects of governance or institutional operations, perhaps focusing on the mechanics of how things are done.

Trade: This theme includes discussions related to trade agreements, exports, imports and international commerce that fall under this theme.

Health and Child Care: This theme includes discussions about health policies, childcare provisions and related subjects.

[9] This visualizes the frequency of the most representative word in each cluster, providing insights into what each cluster represents.

Broad Governance Topics: A more general theme, this likely includes talks about the structure and function of the government.

Gender and Social Issues: Topics related to gender equality, women's rights and other societal issues are clustered here.

Finance and Taxes: Discussions about the financial structure, budgeting, taxation and similar fiscal matters belong to this theme.

Employment and Economy: Topics related to jobs, employment policies and broader economic trends and strategies are clustered here.

Crime and Law: Speeches that discuss the legal system, criminal activities, laws and regulations form this cluster.

Agriculture: This theme focuses on farming, agricultural policies and related subjects.

"What is the prime minister talking about? What is the opposition?"

To answer this question, I extend the existing Stata code. First, I find the parliamentarians and their MP_IDs[10], who were Prime Ministers (PM) or Opposition Party Leaders between 1980-2020. The table below represents these parliamentarians:

Prime Ministers	
Name	**Time Period**
Justin Trudeau	2015 - Present
Stephen Harper	2006-2015
Paul Edgar Philippe Martin	2003-2006
Joseph Jacques Jean Chretien	1993-2003
Martin Brian Mulroney	1984-1993
Charles Joseph Clark	1979-1980
Pierre Elliott Trudeau	1968-1984
Opposition Party Leaders	

[10] We gather this information from https://lop.parl.ca/sites/ParlInfo/default/en_CA/People/primeMinisters and https://lop.parl.ca/sites/ParlInfo/default/en_CA/People/LeadersOfficialOpposition.

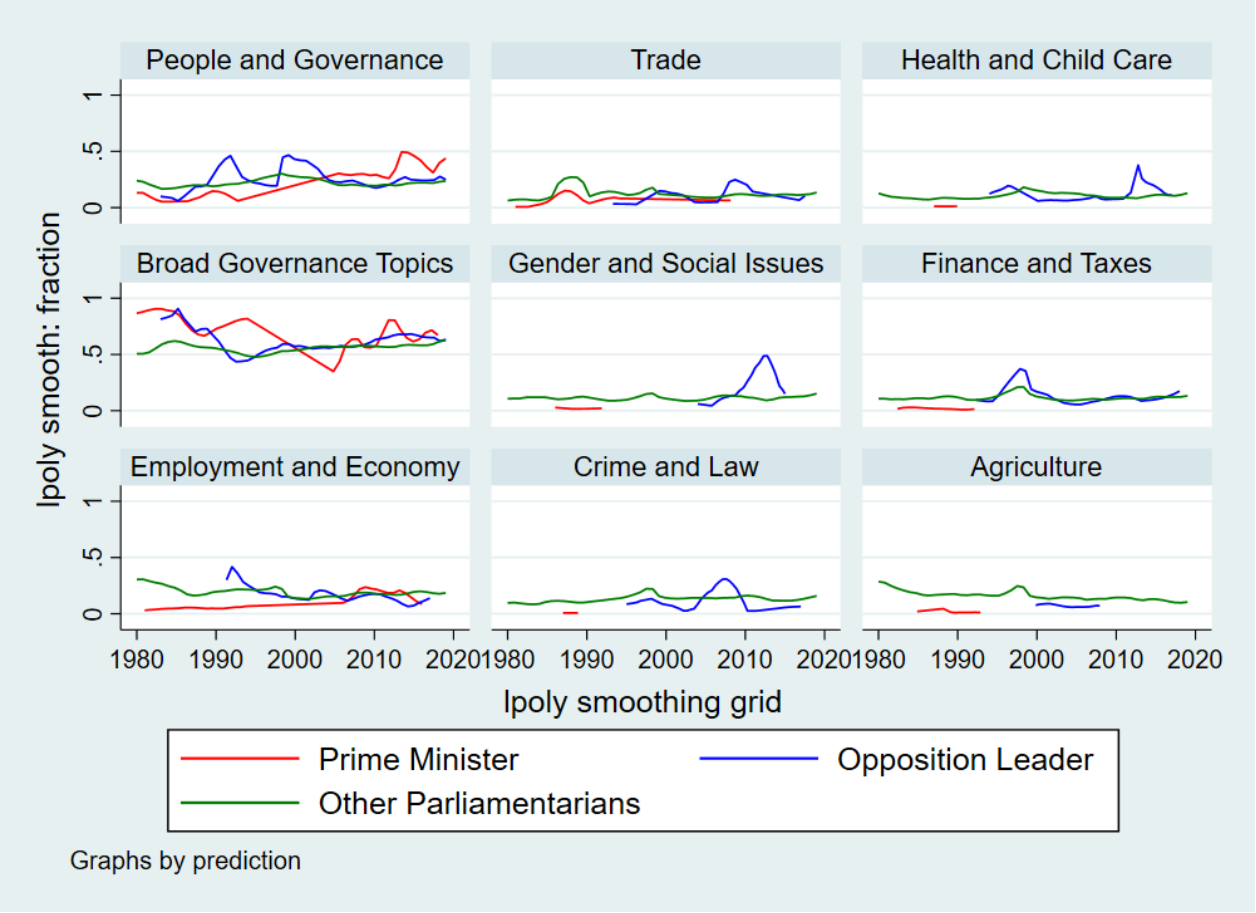
Name	Time Period
Andrew Scheer	2017-2020
Rona Ambrose	2015-2017
Thomas J. Mulcair	2012-2015
Jack Layton	2011
Michael Ignatieff	2008-2011
Stephane Dion	2006-2008
Stephen Harper	2002-2006
Stockwell Burt Day	2000-2001
Ernest Preston Manning	1997-2000
Michel Gauthier	1996-1997
Gilles Duceppe	1996-1997
Lucien Bouchard	1993-1996
Joseph Jacques Jean Chretien	1990-1993
John Napier Turner	1984-1990
Martin Brian Mulroney	1983-1984
Charles Joseph Clark	1980-1983

Table 3: Prime Ministers and Opposition Party Leaders Between 1980-2020

Using the above-mentioned MPs and their MP_IDs, I further process the dataset to create variables indicating if an MP was a Prime Minister (**is_pm**) or an Opposition Leader (**is_opposition**) during a certain year. I discard prime ministers and opposition leaders who were in charge for less than a year between 1980 and 2020. Parliamentarians who were not the prime minister or opposition party leader are marked as "Other Parliamentarians".

One of the critical stages in the data processing is matching names in the dataset. Matching names in large datasets can be tricky due to inconsistencies or slight variations in the way names are recorded. To address this challenge, I apply the Jaro-Winkler distance metric (Winkler, 1990), a specialized measure that calculates the similarity between two strings. This technique is particularly useful for matching names as it takes into account the order of characters and gives

more weight to the beginning of the strings. After using this distance metric, I am able to identify and merge records with similar names. From this matching process, 1357 out of 1363 perfect matches are identified. These matches allow us to merge datasets (Final/ElecData.csv, Final/MPData.csv, Input/FED_Results.csv, Input/Speeches/1980_0.xlsx etc.) seamlessly based on MP_ID, ensuring that each speech was correctly attributed to the right individual. With this cleaned and structured dataset in hand, I then proceed to analyze the content of the speeches, particularly focusing on what was discussed by prime ministers and opposition leaders over the years. I visualize these findings in the following plot (for the sake of simplicity, procedure topic with index number 1 is removed from the plot), revealing patterns in the topics that dominated political discussions across 40 years.



Plot 7: Topic Frequencies across Years in Stata

From 1980 to 2020, **Prime Ministers** have mostly talked about **"People and Governance", "Employment and Economy"**, **"Broad Governance Topics"** and **"Trade".** In Particular, there has been an increase in the fraction of speeches based on **"People and Governance"** after 1995. **"Trade"** has been one of the most frequent topics until 2010. Prime Ministers talked about **"Employment and Economy"** consistently over the years, with a slight peak around 2015. As

for **"Broad Governance Topics"**, these were discussed often. However, between 1980 and 2005, they didn't talk about it as much; but after 2005, it became a popular topic again.

On the other hand, Opposition Party Leaders have mostly addressed topics such as **"People and Governance"**, **"Trade"**, **"Broad Governance Topics"**, **"Finance and Taxes"**, **"Employment and Economy"**, **"Crime and Law"** and **"Health and Childcare"**. Notably, discussions on **"Gender and Social Issues"** increased after the year 2000 with a peak around 2015. Topics like **"Trade"**, **"People and Governance"**, **"Employment and Economy"** and **"Broad Governance Topics"** have been consistently discussed, indicating their continued relevance in parliamentary discussions. **"Finance and Taxes"**, **"Crime and Law"** and **"Health and Childcare"** relatively remained constant with slight peaks around 1995, 2005 and 2015 respectively.

EMPIRICAL STRATEGY AND RESULTS

1. Empirical Strategy

In the last analysis, I try to answer whether there is a correlation between the electoral performance of parliamentarians (measured as **perc_votes**) and the topics they frequently discuss (represented by **cluster_freq_X**). For this analysis, I employ fixed effects regression, which is particularly suitable for panel data structure, to control for unobserved time-invariant heterogeneity across MPs (MP_ID).

The main advantage of using fixed effects is its ability to control for unobserved time-invariant heterogeneity. In this respect, by using fixed effects I can account for consistent, unchanging characteristics of a parliamentarian, such as charisma, communication style, or inherent popularity. By considering these characteristics, which could influence both the independent and dependent variables, I gain a clearer understanding of the relationship between the variables I am interested in.

By focusing on within-parliamentarians changes, fixed effects regression provides a clearer picture of how changes in the independent variables (e.g., the topics spoken about by a parliamentarian) correlate with changes in the outcome variable (e.g., electoral performance), free from the influence of time-invariant characteristics of that parliamentarian. Including time-fixed effects in the regression along with MP-fixed effects provides even more granularity and control in the analysis.

To sum up, MP fixed effects ensure that the correlation is not driven by unobserved, consistent characteristics of parliamentarians. Time-fixed effects ensure that the correlation is not driven by events or trends that impact an MP's topic choice and electoral performance in any given year. My identification assumption is that, after controlling for fixed effects, changes within an MP's choice of topics are exogenous with respect to unobserved factors affecting their electoral performance.

Therefore, while fixed effects regression is often associated with causal inference, it is a robust method for establishing correlations in panel data, especially when there's potential for omitted variable bias or when unobserved time-invariant characteristics might be influencing the relationship. By using both parliamentarian and time-fixed effects, I am asking: "Given a parliamentarian's characteristics and the events of a particular year, how does a change in the topics they talk about correlate with a change in their electoral performance?"

Before examining the correlation, federal general elections between 1980 and 2020 can be seen below:

Parliament	Date of Election
32nd	February 18, 1980
33rd	September 4, 1984
34th	November 21, 1988
35th	October 25, 1993
36th	June 2, 1997
37th	November 27, 2020
38th	June 28, 2004
39th	January 23, 2006
40th	October 14, 2008
41st	May 2, 2011
42nd	October 19, 2015
43rd	October 21, 2019
44th	September 20, 2021

Table 4: Federal General Elections in Canada between 1980-2020[11]

2. Data Preparation

For this purpose, I implement a Python script[12] to work on and look into datasets about Members of Parliament (MPs), their speeches and election results[13] between 1980 and 2020. I start by cleaning speaker names in the Final/Prediction.csv file[14] by removing certain characters and prefixes.

By using Levenshtein distance metric (Levenshtein, 1966), I measure how close the speaker names in the Final/Prediction.csv file are to those in the "Final/MPData.csv" file. I then pick rows with a similarity score of 1 or less and add an MP_ID column to the Final/Prediction.csv file, as I did in the Stata code in the previous section. This file is then merged with the "Final/Final.csv" file. Next, I add the top topics from the "Final/NLTK_Cluster_10.csv" file to this dataset. I then make a "**uid**" column by joining together the district name, FED_ID and election year to make a

[11] Source: https://en.wikipedia.org/wiki/List_of_Canadian_federal_general_elections
[12] See "4_Determine_Correlation.py".
[13] See "Input/FED_Results.csv".
[14] It is the output of "3_Model_Speech.py".

unique identifier. I also create the same identifier in "FED_Results.csv" file for every row so that I can merge it based on this identifier with the previously-mentioned merged dataset. Therefore, for each parliamentarian, I obtain their speeches in the parliament between 1980 and 2020. In order to filter out the unrelated speeches, I focus on the ones made:

(i) in an election year and the year before,

(ii) in an election year and three years before[15].

Using the Votes1…20 columns in FED_Results.csv, I calculate "total_votes" of each election and each parliamentarian (MP) over the years. With the actual and the total votes count, I obtain the "perc_votes" column to reflect the percentage of votes of an MP for every election.

My independent variables are "cluster_freq_0…9", which indicates the frequency of a specific topic that can be seen below within the specified time frame:

cluster_freq_0 → people and governance

cluster_freq_1 → procedure

cluster_freq_2 → trade

cluster_freq_3 → health and childcare

cluster_freq_4 → broad governance

cluster_freq_5 → gender and social issues

cluster_freq_6 → finance and taxes

cluster_freq_7 → employment and economy

cluster_freq_8 → crime and law

cluster_freq_9 → agriculture

Since I am not interested in procedural topics, I assign zero to the "cluster_freq_1" column.

3. Regression

I aim to understand the impact of various topics (**represented by cluster frequencies**) on the electoral performance (**perc_votes**) of parliamentarians. I employ a fixed effects regression model to control for unobserved time-invariant heterogeneity across MPs (MP_ID) and years (year). Apart from the main independent variables (topic frequencies), I also control for the age of the parliamentarian at the time of the election, the duration of their service and their party affiliation.

[15] Federal General Elections in Canada generally take place in four year intervals.

The fixed effects regression is as follows:

$$\begin{aligned}\text{perc\_votes}_{it} = \beta_0 &+ \beta_1 \times \text{cluster\_freq\_0}_{it} + \beta_2 \times \text{cluster\_freq\_2}_{it} + \ldots \\ &+ \beta_9 \times \text{cluster\_freq\_9}_{it} \\ &+ \beta_{10} \times \text{Age\_at\_Election}_{it} + \beta_{11} \times \text{duration\_of\_service}_{it} \\ &+ \gamma \times \text{Party}_{it} + \mu_i + \lambda_t + \epsilon_{it}\end{aligned}$$

Where:

- i denotes each MP and t denotes each year.
- μ_i represents MP fixed effects.
- λ_t represents year fixed effects.
- ϵ_{it} is the error term.
- γ represents a set of coefficients for each political party.
- βs are the coefficients of interest, representing the effect of each predictor on electoral performance.

My outcome variable is the percentage of votes each parliamentarian receives in each election. Instead of using a binary outcome like "Defeated" or "Elected", or the absolute number of votes, my focus is on understanding the relationship between discussing different topics and electoral performance. Specifically, I am interested in seeing how a parliamentarian's choice of topics correlates with their vote percentage compared to other politicians in the same election.

In addition to MP and year-fixed effects, I add the following variables to my regression as controls:

Age_at_Election: Age can influence voting patterns. Older or younger candidates might appeal to different demographics, or their age might be associated with their experience, charisma or policy preferences.

duration_of_service: This could serve as a proxy for experience or familiarity to the electorate. A longer duration might suggest more experience, which could influence electoral performance.

Party: Different parties have different platforms, resources and levels of popularity. By controlling for party affiliation, the regression accounts for these inherent differences between parties.

I apply fixed effects regressions in two scenarios:

(i) for speeches made in the election year and the year before it,

(ii) for speeches made in the four years leading up to an election, including the election year.

They both control for the age at an election, duration of service and political party and include fixed effects for MP and year.

4. Results

My analysis with fixed effects regression, provides insights into the relationship between the topics parliamentarians frequently discussed and their electoral performance between 1980 and 2020. The regression results are as follows:

	FE (1-year interval)	FE (4-year interval)
cluster_freq_0 *(people and governance)*	0.141 ***	0.192 ***
	(0.008)	(0.007)
cluster_freq_2 *(trade)*	0.163 ***	0.158 ***
	(0.009)	(0.008)
cluster_freq_3 *(health and childcare)*	0.168 ***	0.241 ***
	(0.009)	(0.008)
cluster_freq_4 *(broad governance)*	0.091 ***	0.150 ***
	(0.007)	(0.006)
cluster_freq_5 *(gender and social issues)*		
cluster_freq_6 *(finance and taxes)*	0.158 ***	0.171 ***
	(0.009)	(0.008)
cluster_freq_7 *(employment and economy)*	0.180 ***	0.236 ***
	(0.008)	(0.007)
cluster_freq_8 *(crime and law)*	0.097 ***	0.149 ***
	(0.009)	(0.008)
cluster_freq_9 *(agriculture)*	0.169 ***	0.212 ***
	(0.009)	(0.008)
Age_at_Election		-0.002 ***
		0
duration_of_service		

factor(Party)Bloc Québécois	0.225 ***	0.241 ***
	(0.018)	(0.013)
factor(Party)Canadian Reform Conservative Alliance	0.125 ***	0.152 ***
	(0.018)	(0.013)
factor(Party)Conservative Party of Canada	0.138 ***	0.163 ***
	(0.018)	(0.013)
factor(Party)Green Party of Canada	-0.114 ***	-0.102 ***
	(0.018)	(0.014)
factor(Party)Independent	-0.235 ***	-0.223 ***
	(0.018)	(0.013)
factor(Party)Liberal Party of Canada	0.113 ***	0.121 ***
	(0.017)	(0.013)
factor(Party)Marijuana Party	-0.047 *	-0.035 *
	(0.020)	(0.014)
factor(Party)Natural Law Party of Canada	0.009	0.009
	(0.022)	(0.017)
factor(Party)New Democratic Party	0.084 ***	0.059 ***
	(0.017)	(0.013)
factor(Party)No affiliation to a recognised party	-0.115 ***	-0.139 ***
	(0.021)	(0.015)
factor(Party)People's Party of Canada	-0.167 ***	-0.166 ***
	(0.039)	(0.023)
factor(Party)Progressive Conservative Party	0.091 ***	0.098 ***
	(0.017)	(0.013)
factor(Party)Reform Party of Canada	0.103 ***	0.115 ***
	(0.018)	(0.013)
factor(Party)Strength in Democracy	-0.191 ***	-0.207 ***
	(0.025)	(0.0251)

N	98983	195124
R2	0.772	0.782
*** p < 0.001; ** p < 0.01; * p < 0.05.		

Table 5: Fixed Effects Regression Results

At 1-year interval:

In the short term, the topics parliamentarians talk about can be linked to how well they perform in elections. For instance, talking often about **"employment and economy" (cluster_freq_7)** can improve a parliamentarian's election results by 0.180 percentage points. This is the most influential topic. It is closely followed by **"agriculture" (cluster_freq_9)** with an increase of 0.169 percentage points and **"health and childcare" (cluster_freq_3**) with an increase of 0.168 percentage points. Speaking about **"trade" (cluster_freq_2)** and **"finance and taxes" (cluster_freq_6)** can also help, increasing scores by 0.163 and 0.158 percentage points, respectively. The topic **"people and governance" (cluster_freq_0)** gives a smaller increase of 0.141 percentage points and **"crime and law" (cluster_freq_8)** 0.091 percentage point increase. Overall, these findings suggest that in the short term, topics related to the economy, agriculture and health and child care are particularly correlated with a parliamentarian's electoral performance.

At 4-year interval:

When I expand the time interval I am looking at, the influence of topic discussions on electoral performance exhibits distinct patterns. **"Health and childcare" (cluster_freq_3)** consistently emerges as the most influential topic in terms of electoral performance. Parliamentarians who frequently address this topic see their electoral performance increased by a significant 0.241 percentage points. Not far behind, **"employment and economy" (cluster_freq_7)** discussions are correlated with an increase of 0.236 percentage points in vote share, emphasizing the importance of economic well-being that voters care about. Meanwhile, the topic of **"agriculture" (cluster_freq_9)** preserves its importance correlating with a 0.212 percentage point increase in electoral outcomes. **"People and governance" (cluster_freq_0)**, has a slightly milder effect over this longer period, increasing performance by 0.192 percentage points. In addition, **"finance and taxes" (cluster_freq_6)** and **"trade" (cluster_freq_2)** discussions lead to gains of 0.171 and 0.158 percentage points, respectively. I can complete with the broader themes of **"broad governance" (cluster_freq_4)** and **"crime and law" (cluster_freq_8)**, each correlating with an increase of around 0.150 percentage points in electoral performance.

Hence, the analysis suggests there might be an overall positive correlation between the most frequently talked about topics of each parliamentarian where I consider both the speeches just before the election (election year and one year before it) and all four-year periods (election year and three years before it).

Party affiliations also display varied associations with electoral outcomes. For example, affiliations with old and major parties are generally positively correlated with electoral success, whereas certain smaller and relatively young parties and independents show a negative relationship. For instance, MPs affiliated with the "Conservative Party of Canada" or "Liberal Party of Canada" are consistently better in terms of electoral performance. In contrast, affiliations like the "Strength in Democracy Party" or being an "Independent" often lead to lesser performance.

POLICY IMPLICATIONS AND LIMITATIONS

Natural Language Processing has made the analysis of large volumes of speech data feasible, which has helped us to reveal significant insights into electoral performance in my analysis. In addition, through Machine Learning techniques—such as K-means clustering—I have managed to infer the key topics addressed in these speeches by parliamentarians, including Prime Ministers and opposition party leaders. Moreover, with that processed data, I was able to examine whether there is any correlation between most frequent speech topics and electoral performance.

A critical takeaway from this study is that there might be a correlation between the topics parliamentarians choose to discuss and their electoral success. I have come to that conclusion by using fixed effects regression analysis. Politicians who align their speeches with issues the public feels strongly about—such as employment and economy, health and childcare, agriculture, or trade—stand a higher chance of gaining more votes.

This study does have some limitations. For instance, I only consider the difference between election years and speech years, not their specific dates, a detail that could matter if an election was mid-year. Second, a potential for omitted variable bias could still exist, even though I have controlled for various confounders. Beyond the current scope, elements like media influence and shifts in public opinion on various topics might profoundly impact election outcomes. These factors set the stage for future research, promising an even deeper understanding of the complex interplay between speeches and electoral dynamics.

On the other hand, this study has some limitations. For instance, I only consider the difference between election years and speech years, not their specific dates, a detail that could matter if an election was mid-year and the speech of a parliamentarian just comes after that. Second, a potential for omitted variable bias could still exist, even though I have controlled for various confounders. It would also be worth considering the effect of media attention or public opinion on specific topics, which might also play an important role in the election results. Future research could delve into these aspects, offering a more comprehensive understanding of the electoral dynamics.

REFERENCES

Levenshtein, Vladimir I. "Binary codes capable of correcting deletions, insertions and reversals." Soviet physics doklady. Vol. 10. No. 8. 1966.

Winkler, William E. "String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage." (1990).