Measuring Political Polarization in Canada's Parliament

by

Connor Glossop

Supervisor: Rohan Alexander

April 2023

Abstract

This thesis analyzes recent trends in political polarization in Canada through the use of topic modelling algorithms. Due to recent trends of increasing political polarization in many parts of the world, especially the United States of America, analysis on the change in political polarization in Canada could help to identify possible causes and mitigate them before societal conflict occurs. Speeches taken from the Canada's House of Commons between 2004 and 2019 are processed and modelled using Latent Dirichlet Allocation and BERTopic, two popular topic modelling algorithms. The data is then aggregated, with topic frequency distributions measured for each political party. These are compared using Jensen-Shannon Divergence to establish a measure of political polarization between parties over time, as well as Generalized Jensen-Shannon Divergence to find an measure how political polarization in Canada has changed between 2004 and 2019. The level of political polarization in Canada's House of Commons increased between 2004 and 2019, with notable increases in polarization observed prior to each election.

ACKNOWLEDGEMENTS

I would like to thank my thesis supervisor, Rohan Alexander, for his guidance and support on this thesis project. Without his help and passion for Canada's Hansard and the LiPaD dataset, this thesis would never have even gotten off the ground. I would also like to thank my father, Neil Glossop (NΨ 8T3) for his help. Without his help in editing, his encouragement, and the constant supply of milk he brought to keep me awake while I worked, I surely could not have completed this thesis.

Contents

List of Figures

List of Tables

1 Introduction

Political polarization has been shown to reduce the effectiveness of governments and increase the occurrences of bad policies [1]. To combat this, research has been conducted to try to measure political polarization in various countries, often leveraging the transcripts from government proceedings and outcomes of votes. The Canadian Hansard [2] contains the transcripts of all debates in the Canadian Parliament (both the House of Commons and the Senate). The full transcript of every speech, question and response by Members of Parliament and Senators are recorded every day that the parliament is in session. The results of parliamentary votes are recorded both within the Hansard document for a given day and a separate Votes database [3]. This database contains records of all votes from October 2004 onwards.

In 2021, Alsinet et al. [4] developed metrics for evaluating the political polarization of Reddit users as a substitute for the general population. Hanretty et al. [5] used statistical regression techniques to analyze the left-right split of UK Members of Parliament based on their voting results, while Goet [6] analyzed the UK's Hansard dataset to estimate dyadic representation[1] in various historical periods. Research has been done into analyzing dyadic representation in Canada [8], however little analysis has been performed on the political polarization present within Canada's parliament and its change over time.

This thesis analyzes recent trends in political polarization in Canada through the use of topic modelling algorithms. Specifically, speeches taken from the Hansard as published via the Linked Parliamentary Data Project [9] are processed using natural language processing techniques, and modelled using Latent Dirichlet Allocation [10] and BERTopic [11] to establish a measure of polarization. Sentiment analysis using VADER [12] is also used for generating summary statistics.

1.1 Canadian Political Landscape

Canada's electoral system is based on the Westminster System of the United Kingdom, with the Monarch of the United Kingdom serving as Canada's Head of State and the Prime Minister serving as the Head of Government. Canada's parliament has two houses: the Senate (the upper house with

[1] Dyadic Representation refers to "how well the sitting legislator acts as an agent for the constituency on legislative decisions" [7]

105 seats) consisting of appointed members, and the House of Commons (the lower house with 338 seats) whose members are elected by Canadian citizens. Following an election, the political party with the largest number of seats (regardless of whether that number is an absolute majority of seats) forms the government, with the leader of that party becoming the Prime Minister. The party with the second largest number of seats is known as the Official Opposition [13].

According to The Canadian Encyclopedia, there are currently five major political parties in Canada. For most of its history, Canadian politics had been dominated by the Liberal and the Conservative parties [14]. Only the Liberal and Conservative parties (and their predecessors) have ever formed government [15].

The Conservative Party (the Conservatives) is a centre-right party traditionally connected to conservative ideologies [16]. Founded in 1864, the Conservatives (then called the Liberal-Conservative Party) led by Sir John A. Macdonald formed the first post-confederation government of Canada [14]. Renamed to the Progressive Conservative Party in 1942, the Conservatives remained an important political party throughout much of the 19th and 20th centuries. In the 1990s the Conservatives experienced a sharp decline in popularity, leading to them merging with the Canadian Alliance (a much smaller, more right-wing party) to form the Conservative Party of Canada [17]. Today, the Conservatives are the official opposition to Justin Trudeau's Liberal government and are led by Pierre Poilievre. They hold 115 seats (34%) in the House of Commons [18].

The Liberal Party (the Liberals) is a centre-left party historically connected to the ideas of liberalism, neoliberalism, globalism and biculturalism [2] [16]. The Liberal Party was founded as a coalition of reformist opposition groups in both Quebec and Ontario shortly before Canadian Confederation in 1867 [20]. Since then the Liberals have become Canada's most successful political party, having won 26 of 44 Canadian elections [15]. The current leader of the Liberal Party is Justin Trudeau, who has served as Canada's Prime Minister since 2015. The Liberals have ruled Canada in a minority government since the 2019 election, and control 156 seats (46%) in the House of Commons [18].

After being founded in 1961, the **New Democratic Party** (the NDP) became a progressively

[2] Biculturism is used in Canadian contexts to refer to the cultural pluralism of the two most historically significant cultures in Canada: those of the English and French Canadians [19]

more important political party in Canadian politics. A staunchly left-wing party, the NDP rose to national prominence in the 2011 election where they placed second. Since then, they have become a highly influential party (often holding the balance of power during Liberal minority governments) despite lacking the support needed to win an election [14]. The NDP are currently led by Jagmeet Singh and control 25 seats (7.4%) in the House of Commons [18].

The Bloc Québécois (the Bloc) is a Quebecois regionalist party which first appeared in the 1993 federal election. The Bloc only runs candidates in Quebec and promotes policies primarily related to Quebec's interests [14]. The Bloc has historically been closely tied to the Quebec Sovereignty Movement, advocating that Quebec succeed from Canada and form its own independent nation [21]. Today, the Bloc Québécois is led by Yves-François Blanchet and controls 32 seats (9.5%) in the House of Commons, all of which are in Quebec [18].

The Green Party is a small, left-wing environmentalist party founded in 1983. After winning its first seat in the House of Commons in 2011 under the leadership of Elizabeth May, the Green Party has becoming an increasingly important player in Canadian politics. After a tumultuous 2021 election, Elizabeth May returned as party leader [14]. The Green Party currently controls 2 seats (0.6%) in the House of Commons [18].

Despite more parties gaining in political influence in recent years, Canada currently operates under a "two party-plus" system according to journalist J.J. McCullough, with the Liberals and Conservatives still acting as dominant forces [22].

1.2 Reproducibility

In order to aid in reproducibility and demonstrate exactly what steps were taken during this thesis, all code will be made public in a dedicated GitHub repository found at `https://github.com/Maelstrum127/hansard-canada-thesis`. Furthermore, all data used in this thesis project is publicly available at `https://www.lipad.ca/` [23].

2 Literature Review

2.1 Political Polarization

According to Cook et al. [24], polarization is the irrational occurrence of when "two people respond to the same evidence by updating their beliefs in opposite directions". Similarly, DiMaggio et al. [25] more directly defines political polarization as the "measure" of how opposed opinions on a given issue are, and how that opposition increases over time. Specifically, they state that polarization refers to the "extremity of and distance between responses" to an issue rather than the content of those responses.

Political polarization is commonly cited as either the root cause or a key compounding factor in many of modern society's problems [26], [27]. Despite this, the exact causes and effects of political polarization are often debated. In *Linking Conflict to Inequality and Polarization*, Esteban and Debraj [28] analyze the causes of societal conflict, specifically income inequality and polarization, with the goal of creating a behavioural model for conflict. Their final model was constructed using a linear combination of three metrics (income inequality, social fragmentation and polarization), and was able to correctly predict the prevalence and scale of societal conflict (defined as a broad range of conflicts up to and including civil wars). As the model weights were all positive, this shows that an increase in polarization is correlated to an increased level of societal conflict.

In *Is Polarization Bad?* Testa [1] analyzes the effects of elections in polarized societies, and how they can affect the quality of the resulting government. She states "more [polarized] societies... tend to be ruled by bad governments that choose poor policies." Despite this, some amounts of political polarization can be seen to increase accountability of elected officials, especially on policies unrelated to the government's ideology. This is further confirmed by Maoz et al. [29] who found that in many European democracies, the level of political polarization is positively correlated to the duration of sessions of cabinet. Due to the ubiquity of political polarization within modern democracies (despite it being seen as a flaw of those democracies due to it undermining the political power of the majority of the population), understanding its effects are vital.

2.1.1 Measures of Polarization

As polarization is a sociological concept rather than a mathematical one, many attempts have been made to formulate accurate measures for it. *Have American's Social Attitudes Become More Polarized?* [25], written by DiMaggio et al., proposes four measures which can be used to evaluate the polarization of responses to a given issue.

Dispersion of opinion measures the "far apart-ness" of opinions, and can be directly measured using the variance of opinions. Greater variance indicates a more polarized population. In the equation below s^2, N and $\bar{x}$ are the variance, total number and mean value of opinions, with x_i being the numerical value of a given opinion (as found e.g. via a poll where respondents rank items on numerical scales).

$$s^2 = \sum_{i=1}^{N} \frac{(x_i - \bar{x})^2}{N-1}$$

Bimodality measures the extent to which opinions are divided into two separate "camps," with most opinions normally distributed around one of the two centroids and very few occupying the space between them. While it is theoretically possible for more than two centroids to exist, they are rarely seen in practice. DiMaggio et al. propose kurtosis to measure bimodality, a metric often used to differentiate between "peaked" (i.e. few outliers), "flat" (i.e. many outliers), and bimodal distributions [30]. Kurtosis can be defined as

$$k = \frac{\sum(x - \bar{x})^4}{s^4} - 3$$

where s, x, $\bar{x}$ and N are as described above. The factor of -3 ensures that an unaltered normal distribution has a kurtosis value of 0. Large values (approaching 2) indicate a highly non-polarized population, while lower values (approaching -2) indicate a highly bimodal, polarized population.

Constraint refers to how correlated a given person's views on one issue are to another issue. DiMaggio et al. note that a society that is more polarized is more likely to follow different ideologies, where each side has a "narrative" that connects various issues. They propose Cronbach's alpha (α), a metric which measures association between items. As α increases (corresponding to a strengthening in constraint as defined above), the polarization in the society increases. Cronbach's alpha is defined as

$$\alpha = \frac{k}{k-1}\left[1 - \frac{\sum \sigma_i^2}{\sigma_{yi}^2}\right]$$

where k is the number of items in the scale, σ_i^2 is the diagonal covariance for the i^{th} item, and σ_{yi} is the sum of the diagonal and off covariances for all items.

Finally, **Consolidation** is the amount of "in-population" bimodality. This measures the amount of polarization within a given group, with the idea that less-polarized groups likely lead to increased overall polarization than highly polarized groups. This can also be measured with kurtosis.

In their paper *Measuring Polarization in Online Debates*, Alsinet et al. explore various methods of measuring polarization via debates on online social networks [4]. While some social networks, such as Facebook, have actively implemented policies and algorithms to attempt to mitigate polarization on their platforms, these have largely been unsuccessful. Alsinet et al. attempt to create a model to measure political polarization in debates on Reddit, a social media website focused on social news, user discussions and the sharing of web-based content. Alsinet et al. use a graph-based method to identify connections between user comments as they exist on the website. Through constructing a graph, they partition the graph into two sets: those who agree with the root comment (R) and those who don't (L). Using this partition method, they were able to create a measure of group partition by finding the number of interactions where members of one group are in agreement with another member of that group and contrasting that to the number of interactions where members of the two groups interact and are in disagreement. The partition is generated by randomly assigning users to one of the two groups, and iterating until the polarization measured is maximized. This method was able to accurately estimate the polarization of Reddit threads, with threads about Halloween tending to have significantly lower polarization measures than those about World Politics.

In *Linking Conflict to Inequality and Polarization* [28], metrics were used to quantify the levels of income inequality, social fragmentation and polarization in a society. The **Gini Index** is a measure of distance between two groups, and is commonly used to measure income inequality. The **Hirschman-Herfindahl Fractionalization Index** estimates the probability that two individuals belong to different societal groups, making it an effective measure of societal fragmentation. It has been demonstrated to relate ethnolinguistic diversity to conflict and growth; however, no clear link between conflict and either income inequality or fragmentation has been shown. The third metric

used is Polarization, expressed with the following formula:

$$\tilde{P} = \sum_{i=1}^{m} n_i^2 (1 - n_i)$$

where m is the total number of groups and n_i is the share of the total that is a member of that group. Groups are assumed to be uniformly "distant" from one another.

In their 2017 paper *Cross-national measurement of polarization in political discourse: Analyzing floor debate in the U.S. the Japanese legislatures*, Sakamoto and Takikawa [31] use topic modelling to analyze the relative levels of polarization during debates in the Japanese Diet (i.e. the Japanese parliament) compared to the United States Congress. Records from January 1994 to December 2016 were taken from both the US Congress and Japanese Diet. After constructing a bag-of-words model and using multinomial regression to identify topics, a stochastic estimate for each group was calculated by aggregating the probability distributions of each topic over a specified group (e.g. party, gender) and over a specified timeframe. Sakamoto and Takikawa propose using Jensen-Shannon divergence (JS divergence) between the distributions' estimates of the aforementioned group distributions to measure the level of political polarization between the two groups. The equation for JS divergence can be seen below, where p and q (Θ_1 and Θ_2 in the original paper) refer to the distributions for two groups with the same sample space χ:

$$D_{JS}(p||q) = \frac{1}{2} D_{KL}\left(p||\frac{p+q}{2}\right) + \frac{1}{2} D_{KL}\left(q||\frac{p+q}{2}\right)$$

where D_{KL} is the Kullback–Leibler divergence, defined as follows:

$$D_{KL}(p||q) = \sum_{x \in \chi} p(x) \log\left(\frac{p(x)}{q(x)}\right)$$

These measures were then used to compare the major political parties in both the Japanese Diet and the United States Congress over the specified period. Due to limitations in the methods, each party could only be compared to a single party at a time. Sakamoto and Takikawa found that Japan was significantly more polarized than the United States in the period studied.

Although rarely used, **Generalized Jensen-Shannon Divergence** is a form of Jensen-Shannon Divergence which allows for the weighted divergence of multiple probability distributions to be found [32]. It can be defined as follows:

$$JS_{\pi}(p_1, p_2, ..., p_n) = H\left(\sum_{i=1}^{n} \pi_i p_i\right) - \sum_{i=1}^{n} \pi_i H(p_i)$$

where $p_1, p_2, ..., p_n$ are n probability distributions with the same sample space and $\pi_1, \pi_2, ..., \pi_n$ are weightings (or prior probabilities) assigned to each of the n probability distributions. H is Shannon Entropy, defined as follows:

$$H(p) = -K\sum_{i=1}^{N} p_i \ln(p_i)$$

where p is a series of N events (e.g. a discrete probability distribution) and K is a positively-valued constant, usually set equal to 1 [33].

2.1.2 Polarization in the United States of America

A significant amount of research has been performed on analyzing political polarization in the United States of America.

DiMaggio et al. applied the metrics they developed to the General Social Survey and National Election Study, two large scale surveys of American political opinion, between 1972 and 1994 [25] They found that on the overall population, Americans had become less politically polarized over the selected timeframe. Variance decreased on General Social Survey responses despite staying approximately the same on National Election Study responses, while kurtosis increased towards a value of 0. A decrease in polarization was also found in most issues when comparing both inter-group and intra-group polarization.

Despite the decrease in polarization in general population, Jacobson [34] noticed that a sharp increase in political polarization can be seen in the American Congress between the 93rd Congress (1973 - 1975) and the 105th Congress (1997 - 1999, which voted to impeach Bill Clinton following his high-profile affair with staffer Monica Lewinsky). He goes on to argue that this increasing polarization in Congress could cause an increase in political polarization among the general public. Jacobson cites several notable issues in American political history (such as the debate over the legality of abortion) where polarization among Members of Congress predated polarization of the American public on that issue. Jacobson concludes that "In reality [...] the relationship between mass and elite partisan consistency is inherently interactive" [34], where partisan consistency (i.e.

the likelihood a voter agrees with the policies of their chosen party) is a key indicator of polarization [35]. It can therefore be seen that the positions of the political elite (such as Members of Congress) strongly influence the responses to and resulting polarization from political issues among the general public. With an increasingly polarized political elite, this may result in an increasingly polarized public.

In 2008, Fiorina and Abrams considered the evidence for and effects of political polarization on the American public [36]. After dismissing multiple commonly-cited claims of evidence of polarization in American society as being unrelated to political polarization, they analyzed the claim was that Americans held increasingly opposing positions on many beliefs related to politics. To investigate this claim, Fiorina and Abrams expanded upon DiMaggio et al.'s study [25] on the National Election Study to include results from 1972 to 2004. Through analyzing the results of the seven-point scale measure of ideology taken National Election Study in 1972 and 2004, Fiorina and Abrams found that there was almost no visible difference in the ideological landscape of 1972 compared to 2004 (i.e. the share of respondents identified as liberals, moderates, conservatives, etc. stayed the same). Similar results were found when analyzing the General Social Survey, also analyzed by DiMaggio et al. From this analysis, Fiorina and Abrams concluded that "there is little indication of increasing polarization," but rather that the public, which predominantly held centrist views, was drifting slightly to the left or right depending on the issue considered [36]. Fiorina and Abrams consider the effects of an increasingly polarized societal elite on ordinary citizens. Upon consulting existing literature, they conclude that there is "little evidence that the hypothesized dire consequences of polarized politics [...] are showing up in the American public" [36].

Contrary to Fiorina and Abrams' predictions, Heltzel and Laurin concluded in 2020 that political polarization in America had reached an all-time high [35]. Compared to previous decades, Americans had become more hostile towards other Americans who held opposing political views, were more likely to agree with their chosen political party's stance across all issues, and increasingly identified with extremist ideologies such as Neo-Nazism. With this context, Heltzel and Laurin propose two possible pathways for polarization in America. The idea that polarization will continue to increase is considered first, with Heltzel and Laurin stating that this is unlikely as most Americans currently overestimate their degree of polarization compared to members of rival political parties. However, this overestimation in polarization could cause some people to distrust those

with opposing views, leading to a further increase in polarization [37]. Second, Heltzel and Laurin consider that polarization can be thought of as a pendulum that is currently at the apex of its arc. Heltzel and Laurin state that Americans have a "growing resentment for polarization and its consequences," possibly leading to a decrease in the long term. Heltzel and Laurin do not state which outcome is more likely. They do however note that despite indications that the level of political polarization has not increased between 2014 and 2017, the negative effects of polarization have significantly worsened [35]. This paints an ominous picture for the future of politics in America: even if polarization decreases, its negative effects could be felt for decades to come.

Currently, social media is often seen as a catalyst for increasing the political polarization of groups through the creation of "echo chambers" where members of those groups are pushed further to the extremes via the lack of contact with contradictory viewpoints. Bail et al. investigated this concept by exposing a large group of politically diverse American social media users to a Twitter bot representing views opposite to their own (i.e. a Republican was exposed to a Democrat Twitter bot while a Democrat was exposed to a Republican Twitter bot) [38]. This Twitter bot used tweets from various sources, including elected officials, opinion leaders, media organizations, and nonprofit groups. Study participants had their political beliefs measured via a short survey about various policies, and were ranked on a seven-point scale based on their responses. Despite many studies reporting that regularly exposing people to beliefs opposing their own reduces political polarization [39], Bail et al. observed that users became significantly more polarized following the experiment. Liberal study participants outside of the control group became slightly more liberal after exposure to the Twitter bot. In contrast, Bail et al. found that "Republicans exhibited substantially more conservative views posttreatment that increase in size with level of compliance, and these effects are highly significant." On average, conservatives saw an increase of 0.6 points on the seven-point scale in the direction of conservatism, compared to just 0.14 points for liberals in the direction of liberalism (although this was not statistically significant) [38].

2.1.3 Polarization in Canada

Significantly less research has been performed on the level of political polarization in Canada than in the United States.

One of the most commonly cited papers on political polarization in Canada analyzed the causes

of political polarization between Canadians living in cities and those living in suburbs. This polarization, with suburban voters tending to vote more conservatively than urban voters, as been demonstrated both in the United States and Canada. Alan Walks analyzes seven possible causes. In the United States, "much of the [political polarization] between cities and the suburbs [can] be explained in reference to the segregation of individuals by race and class." Alan Walks contends that although there are significant differences in Canadian and American demographics, "sociodemographic characteristics are still strong explainers of voting behaviour in both countries" [40]. Housing tenure (the length a given person acts as a homeowner for their home) is then analyzed as a possible cause of polarization. Housing tenure has been shown to cause an increase in individualism [41], and has been shown to have moderate effects on political choices in Canada. Alan Walks states that despite evidence showing that housing tenure has political impacts in Canada, no research has been performed on its effects on political polarization. Other causes analyzed include the higher abundance of citizens reliant on welfare programs in cities than in suburbs and the idea that citizens "self-sort" into regions with more like-minded people (which can further reinforce political beliefs via social interaction). Through analyzing survey data collected from the Beaches-East York constituency in Toronto, Alan Walks found that the most important factor in explaining political polarization between cities and suburbs was the self-selection of residents into their environments — "supporters of political parties on the left [moved] into the inner city based on a conscious decision to link their political convictions to their lifestyle choices." Neither segregation nor housing tenure were found to have significant effects [40]. While no firm insights can be extrapolated from this research to explain political polarization on the national level, self-sorting of citizens combined with reinforcement via social interaction could indicate a general trend of an increase in political polarization over time.

In 2022, Pennycook et al. [42] used beliefs about COVID-19 as a method for measuring political polarization in the United Kingdom, the United States and Canada. In particular, Canada is directly compared to the United States due to them sharing a similar culture and for the tendency of American cultural issues (such as the "culture wars") to "spill over into Canadian politics." Pennycook et al. primarily use "cognitive sophistication" (i.e. the general understanding of basic science, maths and probabilities, media skepticism, ... of the society), the level of which has been shown to be negatively correlated with political polarization, as a measure of political polarization. Penny-

cook et al.'s analysis was based on two surveys conducted in March and December of 2020 on the public opinion of COVID-19 media coverage and COVID-19 vaccination efforts. They found that political ideology was a stronger predictor of a person's beliefs and attitudes regarding COVID-19 in the United States and Canada than in the United Kingdom. Furthermore, Pennycook et al. noted that both Canada and the United States experienced an increase in political polarization between March and December 2020, though this was far more marked in the United States.

Polarization Eh? Ideological Divergence and Partisan Sorting in the Canadian Mass Public is a 2022 paper by Merkley [43] which uses the Canadian Election Study to analyze trends in ideological divergence, ideological consistency, and partisan-ideological sorting (i.e. the level to which "policy beliefs and ideology become increasingly intertwined with partisanship"). The Canadian Election Study is a annual survey designed to analyze the Canadian public's attitudes towards various political issues and aspects of the election itself such as the parties and leaders present [44]. Using this data, Merkley found that between 1993 and 2019 there is little evidence of a direct increase in political polarization in Canada. However, Merkley does find evidence that "partisanship, ideological identification, and policy beliefs are increasingly interconnected" [43], something shown in the United States to increase political polarization over time [35].

2.2 Natural Language Processing

Natural Language Processing (NLP) is a branch of Computer Science and Artificial Intelligence focused on using computers to understand textual data (i.e. "natural language") and try to extract meaning and information from it. NLP forms one of the most promising and exciting fields in Artificial Intelligence research today.

Historically, most NLP was performed using hand-written rules based on Noam Chomsky's seminal 1956 paper *Three Models for the Description of Language* [45]. In the 1990s, these rule-based methods were replaced by statistical and data driven models such as Hidden Markov Models [46]. The advent of Deep Learning caused Artificial Neural Networks such as Recurrent Neural Networks to become the industry standard [47]. Since the 2018 release of BERT, the first Large Language Model (LLM) [48], the power of NLP has greatly increased. LLMs have brought NLP to the forefront of Artificial Intelligence research and the public conscience, as demonstrated by the incredible success of OpenAI's ChatGPT [49].

Described below are some of the most important metrics and techniques used in NLP.

2.2.1 Metrics

Pointwise Mutual Information (PMI) computes how the measured probability of the co-occurrence of two events (in NLP, events are usually words or tokens, defined below) p(x, y) differs from the expected probability of the co-occurrence of these two events assuming independence of the events p(x)p(y). For example, given a text and the words "apple" and "pie", PMI computes the difference between the probability of the two words appearing side-by-side (i.e. "apple pie" or "pie apple") compared to the chance that this occurs completely randomly in the text [50]. PMI is calculated as shown below:

$$\mathrm{PMI}(w_i, w_j) = \log\left(\frac{P(w_i, w_j) + \varepsilon}{P(w_i) \cdot P(w_j)}\right)$$

Normalized Pointwise Mutual Information (NPMI) is a normalized version of PMI proposed by Bouma in 2009 [51]. As NPMI is normalized, its value can vary between a value of 1 (words always co-occur) and -1 (words never co-occur). The ability to find patterns in textual data makes NPMI a commonly used tool in text mining [50] and topic modelling [52]. NPMI is calculated as follows:

$$\mathrm{NPMI}(w_i, w_j) = \frac{\log\left(\frac{P(w_i, w_j)+\varepsilon}{P(w_i)\cdot P(w_j)}\right)}{-\log(P(w_i, w_j) + \varepsilon)}$$

Where P(x) is the probability of event x occurring (word x appearing in the text). The numerator of NPMI is PMI.

Term Frequency Inverse Document Frequency (TF-IDF) is one of the most commonly used NLP methods [53]. Given a number of documents each containing a number of words $\vec{w} = (w_0, ..., w_M)$, Term Frequency (TF) refers to how many times a given word w_i appears in a given document d. Similarly, Document Frequency (DF) refers to the number of documents in which the word w_i appears at least once. Inverse Document Frequency (IDF) is calculated by taking the log of the normalized inverse of DF:

$$\mathrm{IDF}(w_i) = \log\left(\frac{|D|}{\mathrm{DF}(w_i)}\right)$$

where $|D|$ is the total number of documents considered. TF-IDF is calculated as follows:

$$\text{TF-IDF}(w_i, d) = \text{TF}(w_i, d) \cdot \text{IDF}(w_i)$$

TF-IDF is widely used in the fields of information retrieval [54], [55] and text summarization [56].

2.2.2 Common Data Processing Techniques

Tokenization is the act of breaking a text corpus into a set of discrete units ("tokens") to allow for ease in future analysis. While basic tokenization is often performed by considering each word as a separate token, this has the potential to lose significant amounts of meaning if the order of tokens is not taken into account (e.g. via the breaking up of idioms) [57]. Tokenization has become an important tool in textual data processing for NLP analysis [58].

As an example, if a tokenization algorithm is run on the text "I like analyzing data using NLP," with each word being considered as a separate token, the output would be ["I", "like", "analyzing", "data", "using", "NLP"].

Stemming is the act of converting different grammatical forms of words to their "base form" via computing their word stem. This allows words with the same meaning to be equated for NLP purposes (e.g. "flying", "flew" and "fly", all of which are based on the verb "to fly"). There are multiple different styles of Stemming algorithms including truncation algorithms (such as Porter stemming), statistical algorithms (such as Hidden Markov Models) and mixed models [59]. While stemming algorithms produce useful results, they often output "words" which are not valid on their own (e.g. "studies" $\rightarrow$ "studi") [60].

If one applies the NLTK's implementation of the Porter stemming algorithm [61] to each word in the list ["I", "like", "analyzing", "data", "using", "NLP"], the output would be ["i", "like", "analyz", "data", "use", "nlp"].

Lemmatization is a closely related concept to stemming. Rather than converting words to their word stem form, lemmatization converts words to their lemma: the most basic, meaningful form of a word (e.g. corpora $\rightarrow$ corpus, better $\rightarrow$ good). Dictionaries (such as the Oxford English Dictionary [62]) can be seen as an ordered list of word lemmas. Lemmatized documents tend to be much easier to read than stemmed documents [60].

If one applies the NLTK's implementation of the WordNet Lemmatizing algorithm [61] to each word in the list ["I", "like", "analyzing", "data", "using", "NLP"], the output would be ["I", "like", "analyzing", "data", "using", "NLP"].

Stopword Removal is a commonly performed technique in natural language processing. Simply stated, stopwords are defined as "words with low discrimination power." Stopwords usually include determiners (such as "a", "the", "another"), coordinating conjunctions (such as "for", "but", "so") and prepositions (such as "in", "over", "before"), but may also be augmented with domain-specific stopwords (e.g. "doctor" or "patient" for medical literature) [63]. Stopword removal is a common practice in topic modelling, though the supposed positive effects of their removal have been called into question [64].

2.2.3 Topic Modelling

Topic modelling is a statistical natural-language based approach to discovering latent patterns (topics) in collections of textual data, and forms one of the largest fields within NLP. In particular, topic modelling algorithms are a set of "techniques for revealing the underlying semantic structure in large collections of documents" [65]. These underlying structures emerge from the statistical properties of the documents, making manual human annotation unnecessary prior to training. In addition, these statistical properties can be used to automatically categorize or summarize similar documents, making topic modelling a powerful tool for document classification [52]. Topic models have been used in various fields, including semantic analysis [66], word-sense analysis [67], bioinformatics [68], information retrieval [69] and the measurement and analysis of political polarization [31].

Most topic modelling algorithms learn via a statistical analysis of the Bag of Words (BoW) representation of an unstructured set of documents [65], and are therefore considered unsupervised models. There are, however, some models that are trained non-probabilistically [70] and some that use a Sequence of Words (e.g. n-gram) approach [71] rather than a BoW representation when calculating topics. Some research has been performed on creating supervised topic models, specifically for Latent Dirichlet Allocation models [72]. With the advent of large language models, new BERT-based topic modelling algorithms have been proposed [11].

2.2.3.1 Latent Semantic Analysis

Latent Semantic Analysis (LSA) is one of the foundational methods for Topic Modelling [65]. Proposed by Landauer and Dumais in 1997, LSA is a non-probabilistic algorithm based on the Singular Value Decomposition (SVD) of the term-document matrix of an input text. This leads to a reduced, higher-dimensional representation of each word and their contexts [70]. LSA was first used in a topic modelling context by Bellegarda in 2005 [73]. While versions using non-SVD methods of dimensionality reduction have been proposed [74], LSA has been supplanted by Latent Dirichlet Allocation in most topic modelling research [75].

2.2.3.2 Latent Dirichlet Allocation

One of the most popular topic modelling algorithms is Latent Dirichlet Allocation (LDA) [65]. LDA models are a class of generative probabilistic models proposed by Blei et al. in 2003 that focus primarily on large text-like datasets [10]. LDA's process can be summarised as follows, as detailed by Syed and Spruit [52]:

1. For each topic $k \in \{1, ..., K\}$
    - draw a distribution over the vocabulary $V, \beta_k \sim \text{Dir}(\eta)$
2. For every document d
    - draw a distribution over topics, $\theta_d \sim \text{Dir}(\alpha)$
    - for each word w in document d
        - draw a topic assignment $z_{d,n} \sim \text{Mult}(\theta_d)$ where $z_{d,n} \in \{1, ..., K\}$
        - draw a word $w_{d,n} \sim \text{Mult}(\beta_{z_{d,n}})$ where $w_{d,n} \in \{1, ..., V\}$

Topics β_k are generated via a Dirichlet distribution and act as multinomial distributions over the vocabulary V. Likewise, each document is generated via a Dirichlet distribution and represents a distribution over the K topics. α and η are values used for smoothing the topics and words within documents and topics respectively.

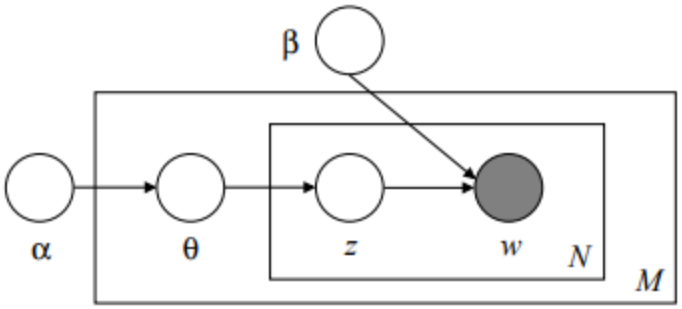


Figure 1: Representation of the LDA method in "plate notation." Image taken from [10]

The central goal of LDA is to "determine the posterior distribution of latent variables given the document", as shown in the equation below [65]:

$$P(\theta, z|w, \alpha, \beta) = \frac{P(\theta, z, w|\alpha, \beta)}{P(w|\alpha, \beta)}$$

While most implementations of LDA use non-supervised learning methods [65], some research has been performed to train supervised LDA models [72].

2.2.3.3 BERTopic

BERTopic is a recently proposed method for topic modelling using BERT (see 2.2.4.2 for more information on BERT). BERTopic leverages a class-based TF-IDF algorithm (c-TF-IDF) to perform clustering on input documents. From each cluster of documents, topic-word distributions can be generated. These clusters are then iteratively merged until the desired level of topics is reached. This allows for a far more dynamic modelling of classes compared to prior topic modelling algorithms such as LSA and LDA [11].

As BERTopic was only released in 2021 (being published in 2022), relatively little research has been performed on its efficacy beyond the original paper. It has however been shown to work effectively in languages other than English [76] and on social media data [77], making it an extremely promising candidate for future advances in topic modelling.

2.2.3.4 Model Comparison and Evaluation

To compare different topic models, the Coherence and Perplexity scores of each model is calculated. This is done through the retrieving of the top-N words for a given topic.

Coherence measures the "supportiveness" of the top-N words for each topic. For example, the words "game, ball, racket" would be considered to have a high coherence (as they all concern sports) while the words "rock, computer, truck" would have a low level of coherence (as they have no clear connection). As such, coherence can be thought of as "how much the documents collected into a topic makes sense as being put into the same topic" [52].

There are a number of different algorithms used to calculate coherence. The most prevalent among these are C_V, C_P, C_{UCI}, C_{NPMI}, and C_{UMass} [78] In all cases, the coherence score for each individual topic is calculated from its top-N words. The arithmetic mean of the coherence score for each topic is then taken to serve as the overall coherence score for the model [52]. Average coherence often serves as the primary method of model selection [78].

For all coherence measures discussed below, W is the corpus of top-N words for a given topic. NPMI and PMI are Normalized Pointwise Mutual Information and Pointwise Mutual Information, as defined above. ε is a smoothing factor "which guarantees that score returns real numbers" by avoiding logarithms of 0 [79]. Historically, ε was chosen to be equal to 1, however, Stevens et al. found that an epsilon closer to 0 produces better results [79]. γ is an additional feature proposed by Aletras and Stevenson which gives words with a larger PMI or NPMI a higher weight in subsequent calculations. Aletras and Stevenson recommend a value of 2 for γ for best results [80]. ϕ_{S_i} is the coherence for a single topic. The overall model coherence is taken as the arithmetic mean of all ϕ_{S_i} [78].

UCI Coherence (C_{UCI}) is a method proposed by Newman et al. in 2010 which uses PMI to measure topic coherence [81]. Röder et al. [78] defined C_{UCI} as follows:

$$\phi(W) = \frac{1}{N \cdot (N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \text{PMI}(w_i, w_j)$$

NPMI Coherence (C_{NPMI}) is a measure closely related to C_{UCI}, but using NPMI rather than PMI. C_{NPMI} was demonstrated by Aletras and Stevenson to perform better than C_{UCI} in most cases [80]

$$\phi(W) = \frac{1}{N \cdot (N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \text{NPMI}(w_i, w_j)$$

UMass Coherence (C_{UMass}) was proposed by Mimno et al. in 2008 [82]. C_{UMass}) relies on document frequency D(w)—the number of documents where a given word w occurs. Co-document

frequency D(w_i,w_j) is the number of documents where both words w_i and w_j occur.

$$\phi(W) = \frac{1}{N \cdot (N-1)} \sum_{i=2}^{N} \sum_{j=1}^{n-1} \log \left(\frac{D(w_i, w_j) + \varepsilon}{D(w_j)} \right)$$

C_V is a coherence score proposed by Röder et al. in 2015 as a combination of the indirect cosine measure with normalized pointwise mutual information (NPMI) and a Boolean sliding window [78]. The method for calculating C_V was detailed by Syed and Spruit in 2017 [52]:

$$\vec{v} = \left\{ \sum_{w_i \in W\prime} \text{NPMI}(w_i, w_j)^{\gamma} \right\}_{j=1,..,|W|}$$

$$\phi_{S_i}(\vec{u}, \vec{w}) = \frac{\sum_{i=1}^{|W|} u_i \cdot w_i}{||\vec{u}||_2 \cdot ||\vec{w}||_2}$$

C_P is a coherence score proposed by Röder et al. in 2015 in the same paper they proposed C_V coherence. They defined C_P as "a combination of Fitelson's confirmation measure with the boolean sliding window" [78]. The exact algorithm for computing C_P was detailed by Alkhodair et al. in 2018 [83]:

$$\phi(W) = \frac{1}{N \cdot (N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left(\frac{\frac{p(w_j|w_i) - p(w_j|\neg w_i)}{p(w_j|w_i) + p(w_j|\neg w_i)} + \frac{p(w_i|w_j) - p(w_i|\neg w_j)}{p(w_i|w_j) + p(w_i|\neg w_j)}}{2} \right)$$

2.2.3.5 Coherence Score Comparison

In the 2015 paper *Exploring the Space of Topic Coherence Measures* Röder et al. explore the results of seven common coherence measures, including C_V, C_P, C_{UCI}, C_{NPMI}, and C_{UMass}. Röder et al. conclude that C_V is the best coherence measure in terms of performance, closely followed by C_P, C_{UCI} and C_{NPMI}, with C_{UMass} performing significantly worse than those listed before. C_{UMass} is however over 20 times faster to compute than C_V, making it a viable alternative in cases where compute time is important [78]. Despite these conclusions, Röder expressed in a 2017 GitHub comment that C_V was found to have several notable issues not noticed during the development of the paper two years prior. As such, he recommended the use of C_P, C_{NPMI} and C_{UCI} when evaluating topics [84].

In the paper *Exploring Topic Coherence over Many Models and Many Topics*, Stevens et al. [79] compared C_{UCI} and C_{UMass} on three different topic models (LDA, SVD and Non-negative

Matrix Factorization). They found that the C_{UMass} coherence score, while generally producing similar results to C_{UCI}, produced less divergent topics in the three models. The two coherence scores returned similar accuracies overall indicating that the choice of coherence is not extremely important in deciding a model's success. Further, they found that a smoothing coefficient $\varepsilon = 10^{-}12$ reduces overall coherence but leads to a higher average coherence in the top 10% of topics, indicating that better topics were identified.

2.2.4 Sentiment Analysis

Sentiment analysis is the study of opinions, feelings, emotions and attitudes from textual data [85]. The basic sentiment analysis process is to first find opinions in a text, then identify the sentiments they express, and finally classifying the polarity of these sentiments in a human-readable and machine-usable way. Sentiment analysis can be performed on the document level, the sentence level, or the aspect level [86].

Sentiment analysis is most commonly performed on user reviews for online products as well as social media data (e.g. collections of tweets about a certain topic from Twitter) [85]. It is also widely used in financial, news and political polarization analysis [87].

2.2.4.1 VADER Scores

Short text posts on social media websites such as Twitter have caused challenges for text-based sentiment analysis—how can sentiment be detected with so few words? With often limited amounts of text available, more efficient methods of sentiment analysis have been developed. VADER (Valence Aware Dictionary for sEntiment Reasoning) is a rule based method for sentiment analysis focusing specifically on the analysis of text from social media websites such as Twitter [12].

When a text is analyzed using VADER, the word choice, word order and word degree are compared to output four scores. The first three scores ('pos','neg' and 'neu') represent the level of positivity, neutrality and negativity for the input text, with each being on a scale from 0 (not at all) to 1 (completely). The final score, 'compound', represents an aggregation of the three main scores. Compound scores range from -1 (completely negative statement) to 1 (completely positive statement) [88]. For example, if the phrase "VADER is super cool" is evaluated using VADER, the resulting scores are {'neg': 0.0, 'neu': 0.244, 'pos': 0.756, 'compound': 0.7351} showing that this

phrase is more positive than negative. Alternatively, if the phrase "I think VADER is useless" is evaluated, the resulting scores are {'neg': 0.483, 'neu': 0.517, 'pos': 0.0, 'compound': -0.4215}.

VADER shows large advances over previous lexical baselines when performing sentiment analysis on text retrieved from social media websites, with an overall F1 score of 0.96 on a corpus of 4200 tweets (a value higher than humans). While VADER is primarily used for analyzing social media text, it also outperforms seven other lexical baselines on Amazon.com product reviews and NY Times Editorials, and performs only slightly worse than best on a corpus of Movie Reviews. In these cases however, it performs significantly worse than on social media data (F1 scores of 0.63, 0.55 and 0.61 respectively) and does not outperform humans [12].

VADER Scores have been used to analyze sentiment on Twitter [88], healthcare platforms [89] and educational settings [90].

2.2.4.2 Neural Network Based-Methods

While research as early as 2013 indicated that neural networks could offer advances in sentiment analysis [86], neural network-based sentiment analysis was first performed at scale in 2015. These analyses primarily used Convolutional Neural Networks and Long Short-Term Memory methods to evaluate sentiment across all three levels of sentiment analysis (document; sentence; and aspect) [85].

Following its release in 2018, BERT became one of the most popular statistical machine learning methods for sentiment analysis [91]. BERT is designed for "language inference without substantial task-specific architecture modifications" [48], making it usable for a large variety of use cases. With additional pre-training providing domain knowledge, BERT outperforms many machine learning and statistical methods for sentiment analysis [91].

2.3 Measuring Polarization using NLP

With levels of political polarization increasing in many parts of the world, efforts have been made to analyze it using NLP techniques. Two of the most popular methods for doing this are with topic modelling and sentiment analysis.

2.3.1 Topic Modelling

Topic modelling is a commonly used method to analyze polarization in text data. In 2020, Xu and Xiong [92] used topic modelling on social media data to analyze polarization in debates on a Gilette advertisement focused on toxic masculinity. While they did not directly measure polarization, they did find that topic modelling was able to identify key positions and talking points as separate topics. This shows that topic modelling can be used to identify polarization in text data.

As noted in 2.1.1, Sakamoto and Takikawa used LDA-based topic modelling to identify key topics discussed in the Japanese and American parliaments between 1994 and 2016. After aggregating the occurrences of each topic on a party level, the "popularity distribution" of topics for each party was be compared via JS divergence to establish a measure of political polarization [31]. Similar work (though not directly focused on polarization) was performed by da Silva et al. in 2021 [93].

2.3.2 Sentiment Analysis

Sentiment analysis can be applied to parliamentary debates to perform various tasks, one of which is measuring polarization [94]. According to Abercrombie and Batista-Navarro, there are five primary methods for analyzing the sentiment of parliamentary debates. These include statistical machine learning and rule-based systems, both of which will be used in this thesis. In 2008, Monroe et al. [95] proposed that an increase in negativity as detected using sentiment analysis algorithms can be seen as an increase in political polarization. Haselmayer and Jenny used sentiment analysis algorithms on communications released by political parties to attempt to measure political polarization (among many other goals) using Monroe et al.'s results [87].

It can therefore be seen that sentiment analysis is likely a valid method of measuring political polarization, especially when inherently political statements (such as parliamentary debates and official party communications) are considered.

3 Materials and Methods

This section details the materials and methods used to measure Canadian political polarization. Results from each method are discussed below in 4. All data preprocessing and analysis was performed on Google Colab using Python version 3.9.16.

3.1 The Hansard

The Hansard is a collection of parliamentary debates taken from the Canadian House of Commons. This data is published online at `https://www.lipad.ca/` by The Linked Parliamentary Data Project (LiPaD) [23], a project spearheaded by group of researchers based primarily at the University of Toronto [9]. LiPaD maintains the digitized Hansard for all sessions of parliament between 1901 and 2019 (the 9th through 42nd Parliaments), though only data from 2004-2019 will be considered in this analysis (the 38th through 42nd Parliaments). All speeches contained in the LiPaD dataset are recorded in English, with speeches being translated if necessary. The French version of the LiPaD dataset is currently unavailable [23].

The data is structured as follows: each year in the LiPaD dataset has a corresponding folder (e.g, 2015), within which are folders for each month in which a speech occurred (e.g. 3, where 3 corresponds to the month of March). These subfolders contain one .csv file for each day on which speeches occurred, with all speeches on that day being contained within that .csv file as a separate row. Each .csv file is named in the "yyyy-mm-dd" format of the date for which the file corresponds (e.g. 2015-3-9.csv). The .csv files contain 15 fields for each speech, some or all of which may be populated for a given speech. These fields include but are not limited to the entire text of the speech, the speaker's name, party and position, and the primary topic (e.g. Statements By Members, Oral Questions, Routine Proceedings) and subtopic (e.g. Questions on the Order Paper, Foreign Affairs, National Defence) of the speech. The most important fields for this analysis are:

- speechtext
- speakerparty
- speakername

The full list of fields along with an example a .csv file can be seen in appendices A.1 and A.2 respectively.

During the period considered for this analysis (2004-2019), there were seven general elections resulting in two changes of power (the 2006 and 2015 elections). Of these seven elections, only two resulted in a majority government (2011 and 2015). This period will therefore have examples of both majority and minority governments for each of the Liberal and Conservative parties. This showcases a variety of situations where political polarization can be analyzed.

3.2 Preprocessing

This preprocessing procedure was adapted from Sakamoto and Takikawa [31].

Files were aggregated on a year-by-year basis. First, the .csv files for each day were read in as DataFrames using the Pandas library [96] These DataFrames were the concatenated together to form one large DataFrame containing each speech given during the selected year. All columns except for *speakername*, *speakerparty* and *speechtext* were removed, as these are not considered during analysis.

When manually analyzing the data, it became apparent that some of the most common words were place names (e.g. used as '...the hon. member for Gatineau', spoken by Mr. Gérard Deltell on April 16 2018) and personal names (e.g. '...that old shtick by Johnny Carson', spoken by Rodger Cuzner on February 9 2012). As these words do not carry any important information and may introduce noise to algorithms used on the processed data, an attempt was made to remove them. A list of the 100 largest population centres in Canada was scraped from Wikipedia [97], with certain other important Canadian cities which are not in the 100 largest cities added (such as Iqaluit). To remove names, a list of 18,239 first names from various cultures compiled by the United States Naval Academy [98] was scraped, and was further augmented through the addition of all names of members of parliament who spoke in the year considered.

Each speech was then converted entirely to lowercase and tokenized on a word-by-word basis using NLTK's RegExpTokenizer [61]. Stopwords from NLTK's Stopwords set were then combined with the list of cities and names described above and removed from each speech. In addition the words "mr" and "speaker", two of the most common words in the corpus, were removed as they address the Speaker of the House, something which contains no information about the contents of the speech. Each word was then stemmed using NLTK's PorterStemmer.

Following Sakamoto and Takikawa's recommendations, all words which appear fewer than 50

times in the remaining corpus for a given year were removed for being too rare. Next, words which appeared in more than half of all speeches were removed for being too common, and therefore not providing enough useful information. All numeric tokens (tokens whose first character is numeric, e.g. “8” or “21st”) were then removed. Finally, all speeches with fewer than four remaining words were removed as they do not contain enough information for algorithmic purposes, such as topic modelling.

The resulting DataFrame of speeches (the “processed data”) was used in all subsequent analysis.

3.3 Topic Modelling with LDA

Using the *gensim* package [99], multiple parallelized LDA models (gensim.models.ldamulticore) were trained on the combined dataset with the number of topics used ranging between 30 and 70 (with one model trained per number of topics specified, as suggested by Sakamoto and Takikawa [31]). 40 models were trained in total. Each model was trained with 3 passes and 80 interactions (DEFINE) to help improve model performance using a “combined dataset” formed from the processed data from every year being concatenated together acting as the dataset. All models were then compared using C_{NPMI} coherence using a boolean sliding window size of 10 (as recommended by Röder et al. [78]), with the model achieving the best average coherence across all topics being selected.

Processed data from each year was then evaluated by the selected model so that each speech could be assigned to a topic. These topic/speech pairs were then aggregated on a party level. This resulted in a count of the number of speeches by each party that was classified into each topic, which were then normalized to values between 0 and 1 to create a “composite probability distribution” for each party over the selected year, as performed in Sakamoto and Takikawa’s analysis [31]. Finally, the composite probability distributions were compared pairwise using JS divergence to find a polarization score for each pair of parties in a given year.

3.4 Topic Modelling with BERTopic

Unlike with LDA models, where the number of topics must be predefined, BERTopic's algorithm determines the optimal number of topics based on the training data. As such, only one model was trained, and used BERTopic's default hyperparameters. Due to computational limitations in Google Colab, this model was trained using BERTopic's *model.fit()* function in two year batches (i.e the model was first trained on data from 2004 and 2005, then on data from 2006 and 2007, and so forth until data from each year had been used).

After being trained, processed data from each year was predicted upon by model so that each speech could be assigned to a topic. All speeches assigned to topic -1 (the "outlier" topic) were discarded while the remaining speeches were aggregated on a party level. As above, the number of speeches per party was normalized to values between 0 and 1, and the resulting composite probability distribution was compared to each other party in the selected year using JS divergence.

3.5 Generalized Jensen-Shannon Divergence as a Measure of Polarization

As Generalized Jensen-Shannon Divergence (Generalized JS divergence) is rarely used, this thesis will evaluate its ability to generalize trends seen in JS divergence to establish a "overall polarization score." The composite probability distributions of each party will be used to calculate Generalized JS divergence, with the weights assigned to each distribution being the percent of speeches given that year by the party. Generalized JS divergence will be calculated for each year and for distributions generated from both the selected LDA and BERTopic models. As Generalized JS divergence has not been implemented in any commonly-used python package, such as SciPy or NumPy, it was implemented manually using SciPy's *scipy.stats.entropy* to calculate Shannon Entropy [100].

4 Results and Discussion

4.1 LDA

Of the 40 models generated using LDA, the best model used 54 topics with a C_{NPMI} coherence score of -0.320. This indicates that although the model had the largest degree of topic coherence among those trained, each of the topics used by the model were themselves not very coherent. Intertopic distance (a two-dimensional visualization of the "closeness of topics") of all 54 topics was plotted using the pyLDAvis plotting interface for gensim models.

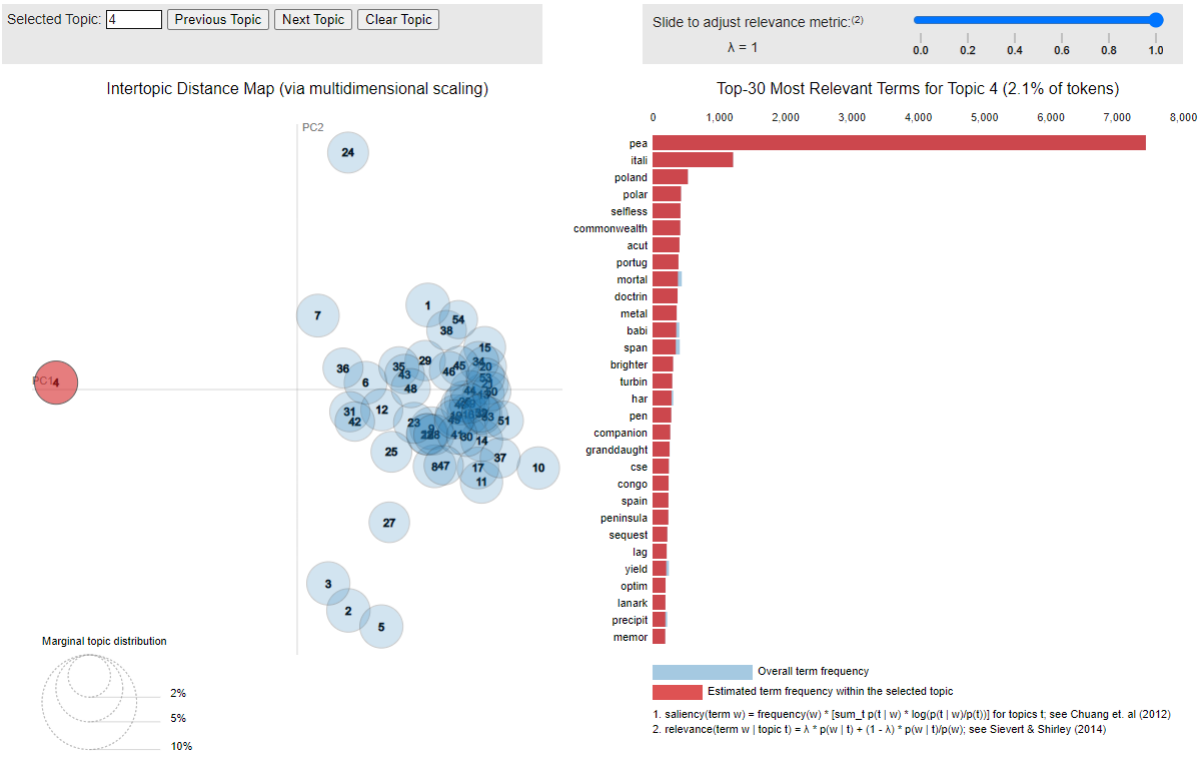


Figure 2: Intertopic Distance Map of Topics Identified by LDA

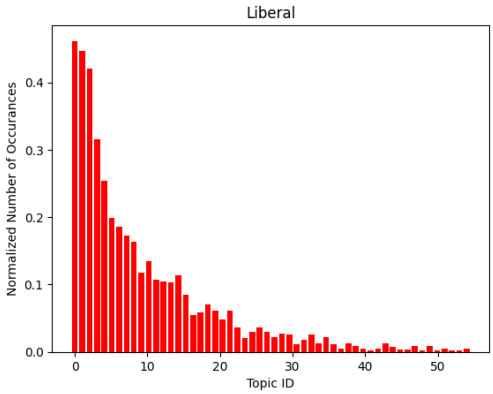
A large cluster of topics can be observed in the left half of the plot, while very few topics are present on the left half. This indicates that the majority of topics are closely connected, sharing many of the same words.

Proposed Topic Name	Topic ID	Top Words
Global Politics	4	Peace, Italy, Poland, Polar, Selfless
Food Production and Selling	24	Lentil, Ninety, Merchant, Dominion, Freshwater
???	36	Feminine, Prostitution, Hull, Tout, Weird
Indigenous Groups	44	Solidarity, Haida, Shipment, Salish, Ramadan

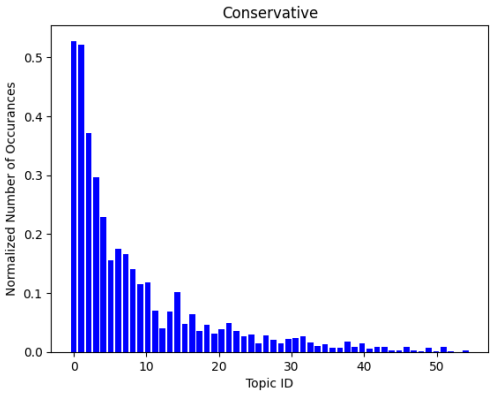
Table 1: Top Words from Various Topics Identified by LDA, with Proposed Topic Names

As can be seen in Table 1, the topics proposed by LDA have little obvious connection between the top words. In topic 4, many of the top words (not all shown above) are the names of countries or regions such as "Congo", "Italy", "Portugal", "Poland" and "Spain". Despite this, other top words are "Selfless", "Brighter" and "Pen", which are seemingly completely unrelated to the aforementioned countries.

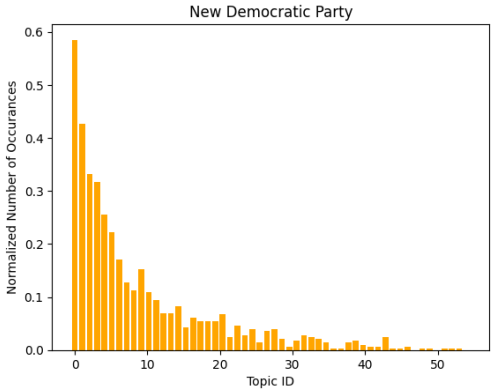
The topic frequencies for each party were identified for each year. Graphs representing these frequencies for the year 2017 can be seen below in Figure 3. Graphs for each other year between 2004 and 2019 can be seen in Appendix B.1.



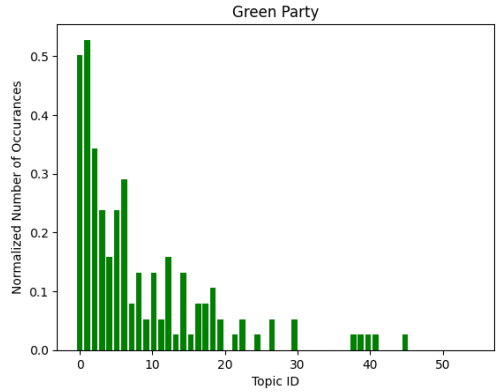
(a) Liberal Party 2017



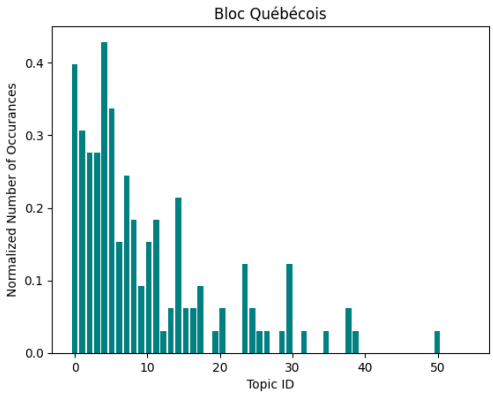
(b) Conservative Party 2017



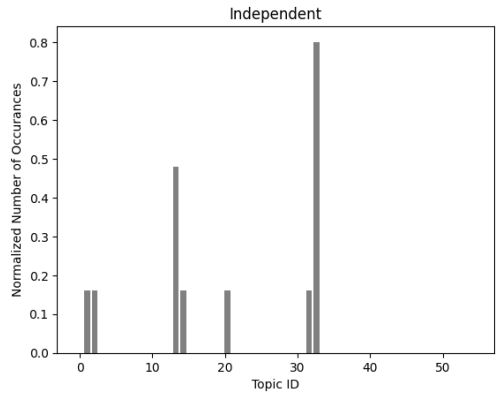
(c) New Democratic Party 2017



(d) Green Party 2017



(e) Bloc Québécois 2017



(f) Independents 2017

Figure 3: LDA Topic Frequencies for Each Party in 2017

Significant differences in topic occurrences across parties can clearly be observed. For example, topic 5 (top 5 words: “Calendar”, “Cigarette”, “Curtail”, “Parcel”, “Melt”) is frequently discussed by members of the Bloc Québécois, while it was discussed much more rarely by mem-

bers of the Green Party. The differences in these distributions were evaluated by computing the JS Divergence of each pair of parties, the results of which can be seen below in Table 2.

	Liberal	**Conservative**	**NDP**	**Bloc Québécois**	**Green Party**	**Independent**
Liberal	0.000000	0.008378	0.011408	0.059385	0.051533	0.457769
Conservative	0.008378	0.000000	0.009555	0.062068	0.048799	0.462300
NDP	0.011408	0.009555	0.000000	0.063487	0.053424	0.467013
Bloc Québécois	0.059385	0.062068	0.063487	0.000000	0.092989	0.489642
Green Party	0.051533	0.048799	0.053424	0.092989	0.000000	0.513615
Independent	0.457769	0.462300	0.467013	0.489642	0.513615	0.000000

Table 2: Pairwise Jensen-Shannon Divergences of Party Topic Frequencies on LDA Topics in 2017

These pairwise JS divergences were calculated for each year between 2004 and 2019 (shown in Appendix C.1) and then plotted. In Figure 4, the pairwise JS divergences for each of the Liberals and Conservatives, the Liberals and the NDP, and the Conservatives and NDP are plotted.

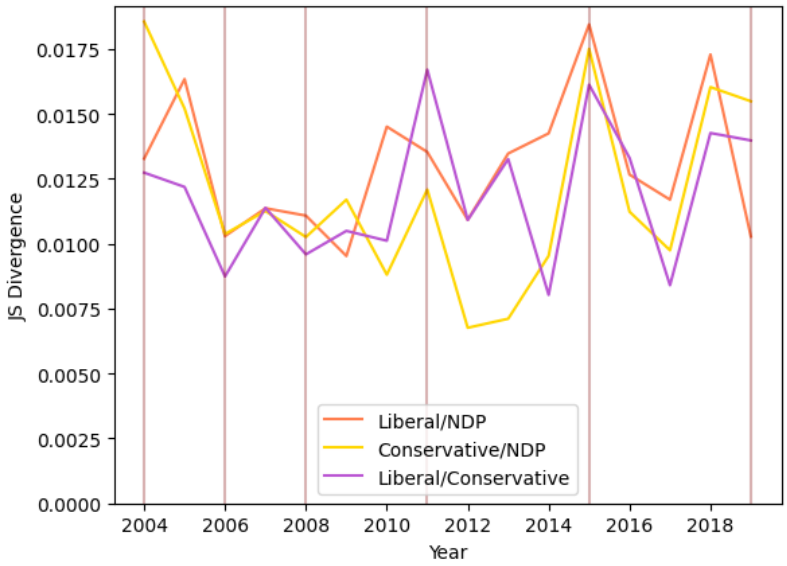


Figure 4: LDA Pairwise Jensen-Shannon Divergences for the Liberal Party, Conservative Party, and New Democratic Party Between 2004 and 2019

Clear increases in political polarization (party divergences) can be seen in election years, noted on the graph by maroon lines. These large increases seem to disappear in the year following the

election. It is interesting to note that between 2010 and 2013, as well as in 2016, the JS divergence of the Conservatives (a right-wing party) and the NDP (a left-wing party) is lower than either of their divergences with the Liberals (a centrist party). In 2016, this may have been caused by the newly elected Liberal majority government but there is no obvious explanation for the reduced level of polarization between 2010 and 2013 (as this was a Conservative majority government) [15]. Due to the level of noise present, it is difficult to discern if polarization has increased or decreased during this time period.

4.2 BERTopic

In total, BERTopic identified 170 topics (excluding topic -1, composed entirely from outliers). The number of documents assigned to each topic ranged from 204 for topic 0 (top 5 words: "Misconduct", "Restitution", "Martial", "Tlicho" and "Sahtú", where Tlicho and Sahtú are both First Nations groups) to 10 for topics 167 through 170. The mean number of speeches assigned to each topic was 26.57. 5006 speeches were assigned to topic -1 and were therefore discarded, accounting for 52.4% of the combined dataset (a total of 9549 speeches). A graph showing the number of speeches assigned to each topic (excluding topic -1) is shown in Figure 5. This model achieved a C_{NPMI} coherence score of 0.0268.

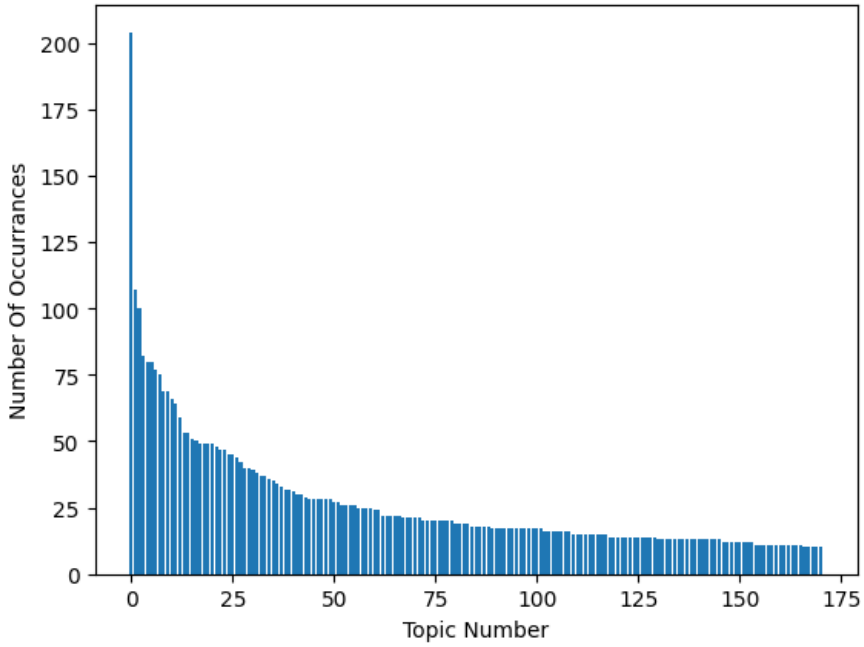


Figure 5: Number of Speeches Assigned to Each Topic by BERTopic

The intertopic distance of all topics is plotted below in Figure 6 using BERTopic's inbuilt *model.visualize_topics* function.

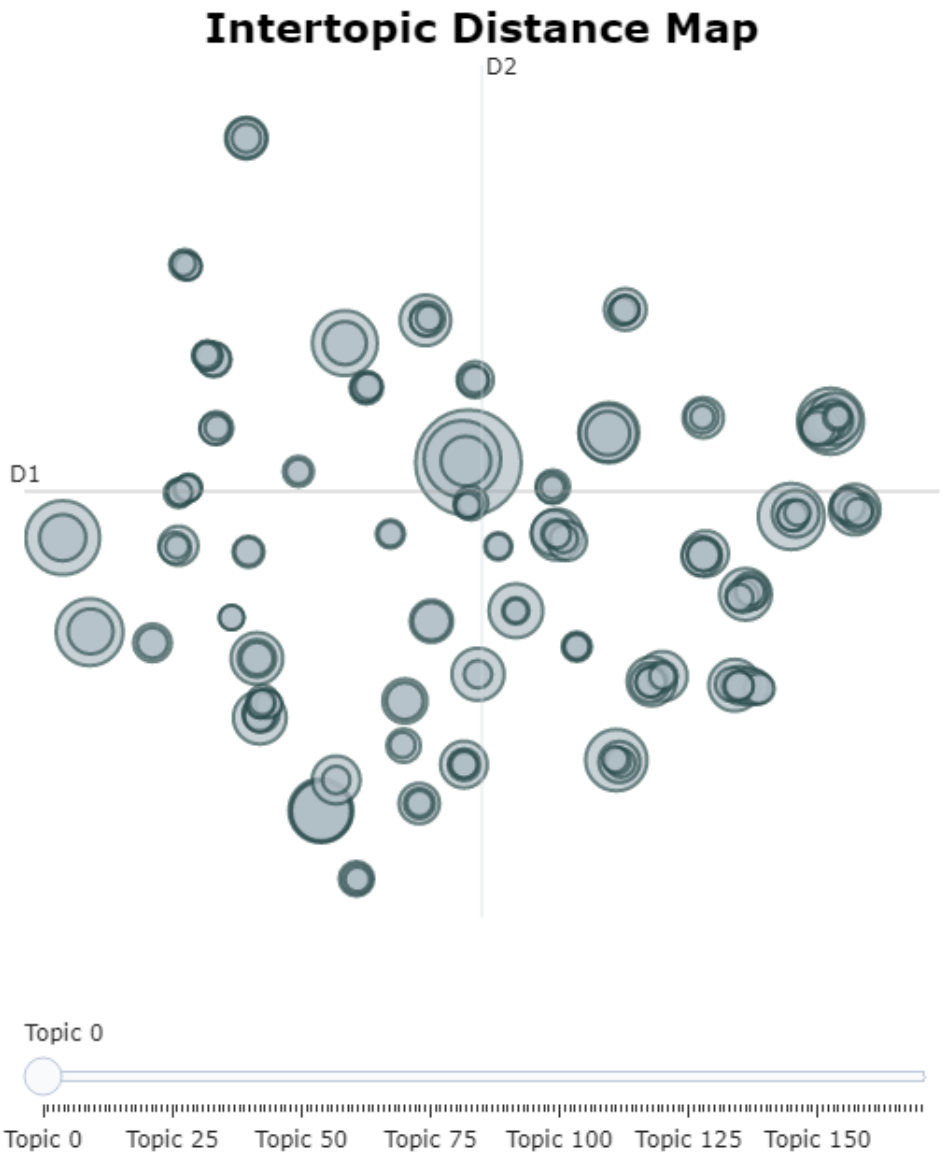


Figure 6: Intertopic Distance Map of Topics Identified by BERTopic

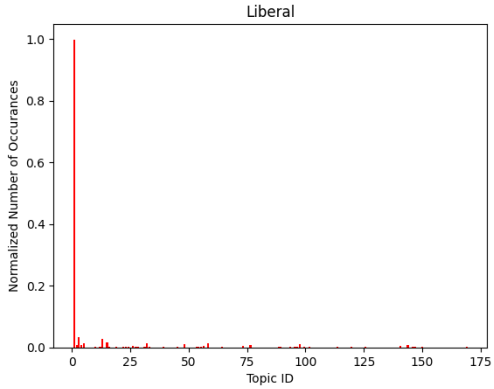
Topics are quite evenly distributed on the intertopic distance plot, with no clear location holding the majority of topics. Some topics identified by BERTopic are displayed below in Table 3, with five of the top ten words displayed. A proposed name for each topic is also given.

Proposed Topic Name	Topic ID	Top Words
Military	18	Crewman, Regiment, Sergeant, Colonel, Lieutenant
Healthcare	26	Cardiac, Transplant, Liver, Tumor, Kidney
Middle Eastern Politics	27	Hezbollah, Kurdish, IRGC, Extremist, Revolutionary
Cars	39	Detroit, Automobile, Motor, Toyota, Chrysler
Wine	74	Wine, Winery, Vine, Vineyard, Distillery
Climate Change	78	Diesel, IPCC, Hydrogen, Greener, Smog

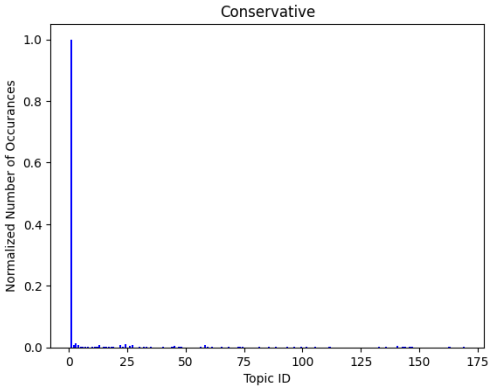
Table 3: Top Words from Various Topics Identified by BERTopic, with Proposed Topic Names

All words shown in this table were converted from their stemmed form to a more human-readable form. Due to bigrams being allowed when training the model, some top words were duplicated (e.g. both "ipcc ipcc" and "ipcc" were included in the top 10 words for topic 78). Some acronyms also appear in topics, such as IRGC (the Iranian Revolutionary Guard Corps) and IPCC (Intergovernmental Panel on Climate Change). As can be seen, BERTopic identifies highly human-understandable topics, with clear connections being observable in many of the top words.

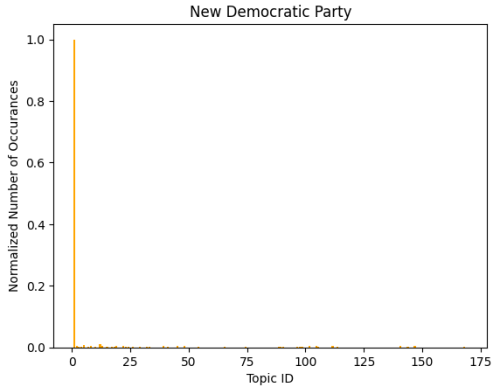
As shown for LDA models, the topic distributions for each party for 2017 is shown below in Figure 7. The full set of topic distributions by party for each year between 2004 and 2019 can be seen in Appendix B.2.



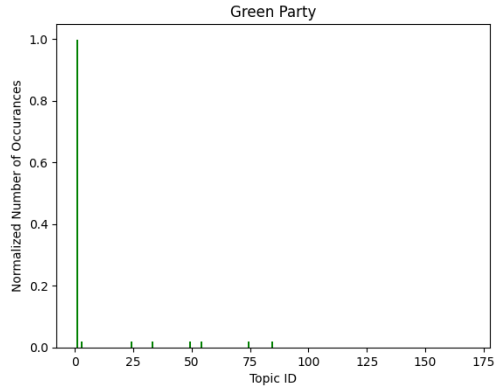
(a) Liberal Party 2017



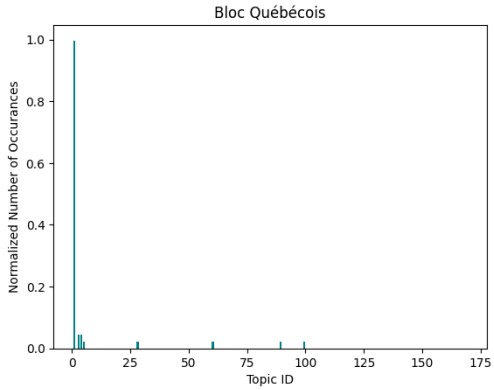
(b) Conservative Party 2017



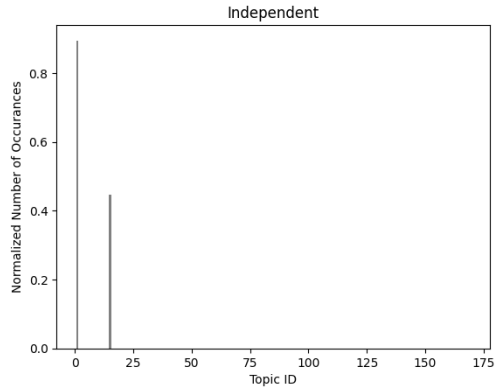
(c) New Democratic Party 2017



(d) Green Party 2017



(e) Bloc Québécois 2017



(f) Independents 2017

Figure 7: BERTopic Topic Frequencies for Each Party in 2017

This result looks strikingly different from the LDA results, with the vast majority of each party's discussions appearing to be in reference to a single topic. BERTopic seems to have grouped an extremely large number of speeches into topic 1 (top 6 words: "Tsilhqot", "Libyan", "Dilute",

"Gadhafi", "Newsroom" and "Anabaptist", where the Tsilhqot are a First Nations group located in British Colombia). The model seems to have selected a set of highly diverse ethnic groups but it is unclear why. The differences in these distributions was then evaluated using JS Divergence, the results of which can be seen below in Table 4. The full set of pairwise JS Divergence tables can be seen in Appendix C.2.

	Liberal	**Conservative**	**NDP**	**Bloc Québécois**	**Green Party**	**Independent**
Liberal	0.000000	0.039384	0.048900	0.078934	0.089792	0.161260
Conservative	0.039384	0.000000	0.031774	0.076625	0.065761	0.163063
NDP	0.048900	0.031774	0.000000	0.078604	0.065683	0.157112
Bloc Québécois	0.078934	0.076625	0.078604	0.000000	0.081458	0.177037
Green Party	0.089792	0.065761	0.065683	0.081458	0.000000	0.164111
Independent	0.161260	0.163063	0.157112	0.177037	0.164111	0.000000

Table 4: Pairwise Party JS Divergence of BERTopic Topics in 2017

As with LDA, the pairwise JS divergences found using BERTopic for each year between 2004 and 2019 were plotted, as shown in Figure 8. Once again, only the Liberals and Conservatives, the Liberals and the NDP, and the Conservatives and NDP are plotted to ensure legibility.

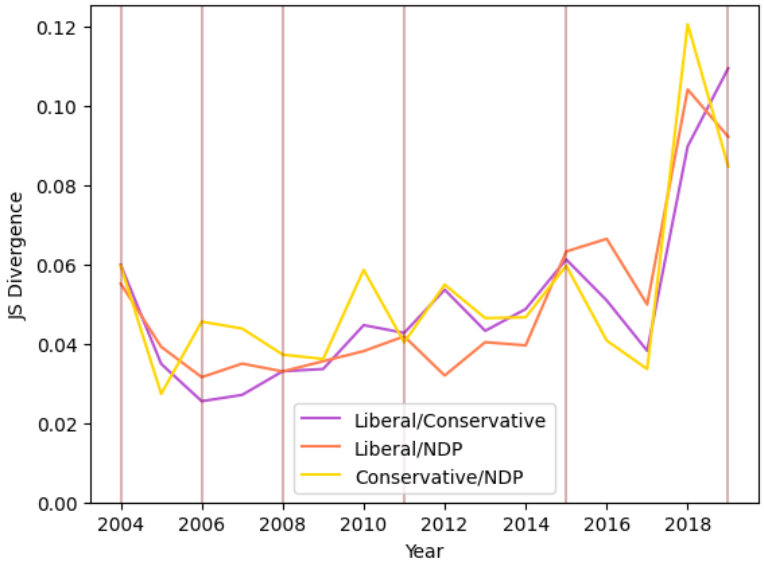


Figure 8: BERTopic Pairwise JS Divergences Between 2004 and 2019

Unlike with LDA (Figure 4), inter-party polarization as found using BERTopic shows a clear and marked increase in JS divergence over time, with a large spike in polarization occurring in 2018. In addition, increases in polarization between parties are visible in most election years. It is interesting to note that, as in Figure 4, polarization between the Conservatives and NDP decreases significantly following the 2015 election.

4.3 Generalized Jensen-Shannon Divergence Over Time

Using Generalized JS divergence, the overall divergence of each years' parliament was calculated based on data obtained from both the selected LDA and BERTopic models. This data can be found in Appendix D in Tables 35 and 36. A graph of Generalized JS Divergence calculated on LDA data can be seen below in Figure 9.

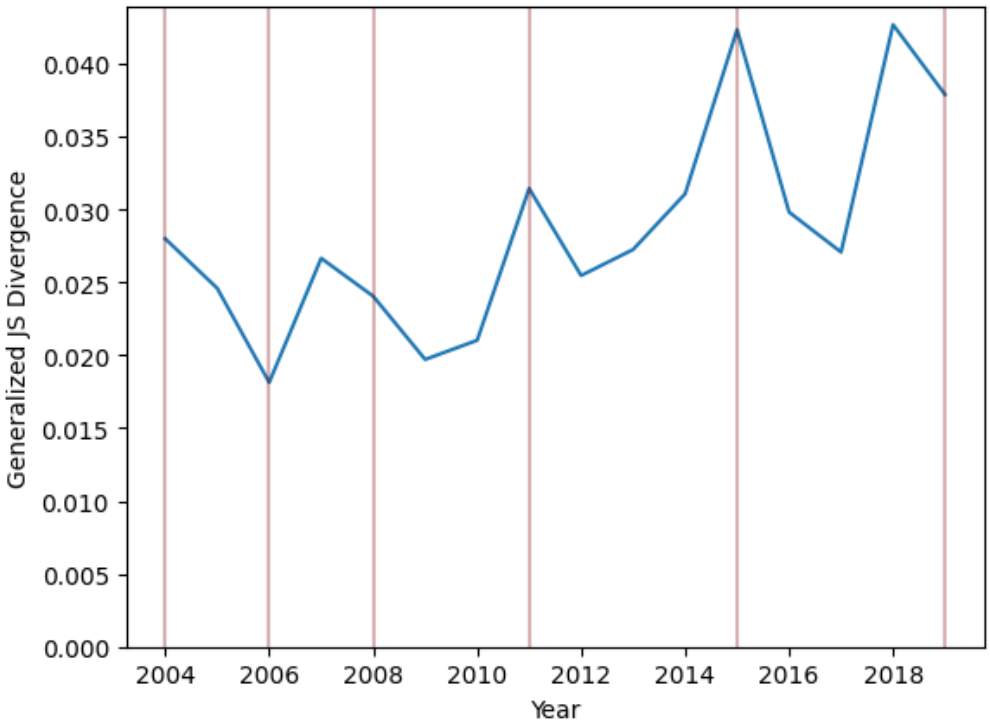


Figure 9: Generalized Jensen-Shannon Divergence Applied to LDA Model Results in Canadian Parliament Between 2004 and 2019.

As above, election years are noted by a horizontal maroon line. As can be seen, Generalized Jensen-Shannon Divergence found using LDA generally increases as time goes on, with significant

spikes observed in or just prior to election years and significant reductions in years immediately following elections. It is interesting to note a decrease in divergence in 2006. As this election occurred on January 26th [15], no speeches occurred in that year prior to the election. As such, the drop in divergence may be due to 2006 being better considered as a "year after an election" than a "year before/during an election."

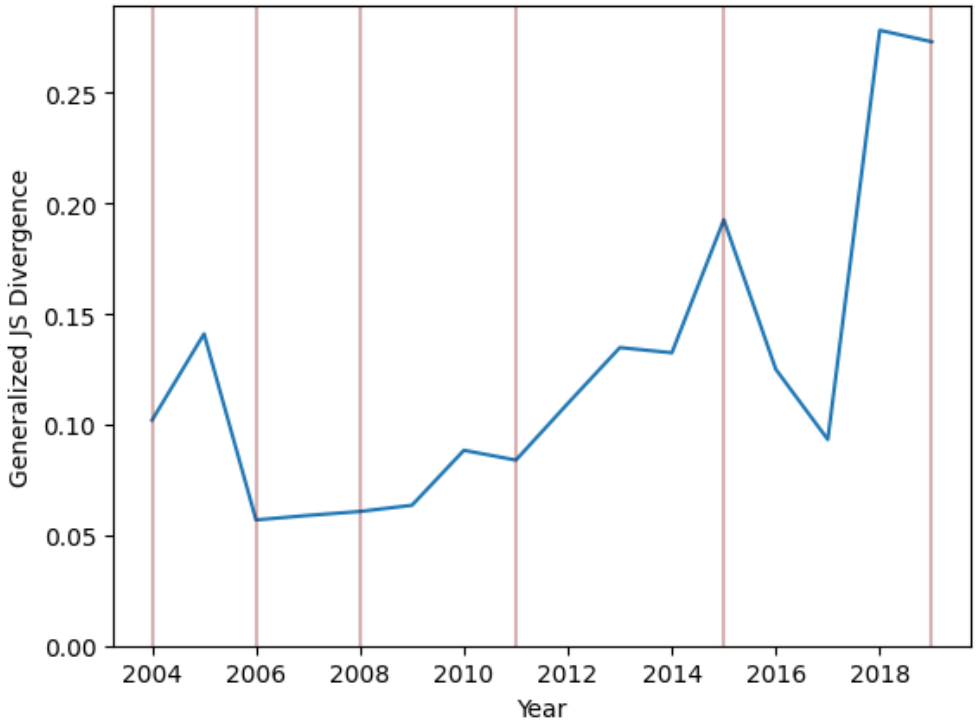


Figure 10: Generalized Jensen-Shannon Divergence Applied to BERTopic Model Results in Canadian Parliament Between 2004 and 2019.

Generalized Jensen-Shannon Divergence found using BERTopic tends towards an increase over time, showing notable increases in 2005, 2015 and 2018. In general, these increases in divergence disappear in the year following elections (except for 2006, where the election took place in January, and 2011, where divergence was lower than both 2010 and 2012). Interestingly, two of the three largest observed spikes in polarization appear in the years where a change of government occurred. In 2006, Stephen Harper's Conservative Party defeated Paul Martin's Liberal minority government to win the election while in 2015, Justin Trudeau's Liberal Party defeated Stephen Harper's Conservative government. This trend does not continue to 2019, where despite a large

increase in polarization, the Liberal Party won the election (though they did lose the popular vote) [15].

The generalized JS divergence as calculated using results from BERTopic is roughly one order of magnitude larger than the values calculated using LDA results. This may be due to the abundance of topics used in BERTopic's modelling process (3.2 times as many topics as LDA), causing far more pronounced levels of divergence to be observed.

4.4 LDA vs BERTopic

There are multiple ways in which LDA and BERTopic can be compared for political polarization analysis.

The coherence score using C_{NPMI} of the selected LDA model was -0.320, compared to a value of 0.0268 for the BERTopic model. This indicates that BERTopic is able to create "better" topics than LDA. Furthermore, many of the topics found by BERTopic had clear, humanly understandable meanings to them (as shown in Table 3), whereas those identified by the LDA model (Table 1) had no easily discernible meaning. It is important to note that BERTopic identified significantly more topics (170 in addition to the outlier topic) compared to the selected LDA model (which used only 53). This difference in the number of topics, when combined with BERTopic's ability to classify data as outliers, may have allowed BERTopic to find more meaningful topics than LDA, which was forced to overgeneralize.

As both models produced similar results, neither model is clearly "better" than the other. While BERTopic identified topics which were very understandable to humans, it did so by ignoring over half of the data and appears to have over-grouped many of the speeches into topic 1. LDA, while identifying less understandable topics, generates a far more "believable" distribution of topics. BERTopic has demonstrated that like LDA, it is able to measure political polarization via topic modelling.

4.5 Limitations and Possible Solutions

One of the largest limitations of using topic modelling as performed in this thesis to measure political polarization is that it assumes that if the speeches of two parties are frequently assigned

to a certain topic, then those two parties are considered "not very polarized" (at least with respect to that topic). This assumption is clearly erroneous—consider bill C-14: "An Act to amend the Criminal Code and to make related amendments to other Acts (medical assistance in dying)" [101]. Due to extensive debates on bill C-14 in May 2016, the Conservative and Liberal parties both discussed the topic of assisted dying frequently. This would, using the methods described in 3, indicate that the Conservatives and Liberals are not highly polarized. If the texts of the speeches themselves are analyzed however, it becomes clear that there is a large degree of polarization between the Conservatives and the Liberals on this topic. The Liberals, being in favour of the bill, tend to use the term "medical assistance in dying" when referring to the contents of the bill and frequently speak of the "right to die". Conversely, members of the Conservative Party use the terms "euthanasia" and "assisted suicide," with one member stating "Palliative care provides death with true dignity and not a forced death, which is what physician-assisted suicide is" (Alice Wong, May 5th 2016 [23]).

To prevent this mistaken reduction in polarization, a second level of topic modelling could be performed on significant topics identified by the original topic model. Hypothetically, this second model would be able to identify key terms used by each side of the debate. As each side would use different terminology (e.g. 'medical assistance in dying" (Liberal Party terminology) versus "assisted suicide" (Conservative Party terminology)), these positions would be classified as separate subtopics. This method could also serve as a measure of in-group cohesion—consolidation—mentioned by DiMaggio et al. as another measure of political polarization [25].

5 Conclusion

Both topic modelling algorithms display the same overall trend: political polarization in Canada's parliament has markedly increased since 2004. Furthermore, political polarization increases significantly in the years leading up to an election but then falls following the completion of the election.

This general increase in political polarization in speeches delivered to the House of Commons is worrying. As noted by Jacobson [34] when considering the United States of America, an increase in polarization among members of a country's parliament (which for Canada is the House

of Commons) directly leads to an increase in polarization among that country's general population. As political polarization causes an increase in societal conflict [28], the results from this thesis imply that Canada can expect to see a rising level of conflict between its citizens over the coming years.

This analysis could be extended in multiple ways to better understand the change over time and possible effects of political polarization in Canada. One interesting future avenue of research would be to compare the methods explored in this thesis to similar work using sentiment analysis (e.g. via VADER) rather than topic modelling as the basis for Jensen-Shannon divergence-based analysis of political polarization. As VADER outputs four sentiment-based scores for each input, these could be aggregated and compared in a similar manner to what was performed in this thesis (where, for example, the number of documents classified as positive, neutral, and negative is taken as the composite probability distributions).

As noted above in Section 4.3, the Generalized Jensen-Shannon Divergence calculated using BERTopic results *appears* to be somewhat able to predict when governmental change occurs. This is demonstrated by a large spike in polarization being observed in the data whenever an election which causes change of government happens. This large spike is also seen in 2018 and 2019 which, while featuring an election, did not feature change of government (or even a particularly close election). More research could be done to expand this analysis to cover past election years to see if the predictive quality of the model holds.

Finally, political polarization is something which has existed for all of human civilization. As many societies have survived periods of intense political polarization without complete collapse (such as with the Civil Rights Movement in the United States [34]), Canada's recent increase in political polarization may not indicate that the country is at the brink of destruction. Future work could expand the analysis presented in this thesis into the past and to other societies. This could help to determine if the current increase in political polarization is an expected part of living in a democracy or if it seems to suggest a coming spiral out of control.

References

[1] Cecilia Testa. "Is polarization bad?" In: *European Economic Review* 56.6 (2012), pp. 1104–1118.

[2] *Welcome to the House of Commons*. URL: https://www.ourcommons.ca/en.

[3] *Votes 44th parliament, 1st Session(November 22, 2021 to present)*. URL: https://www.ourcommons.ca/members/en/votes.

[4] Teresa Alsinet et al. "Measuring polarization in online debates". In: *Applied Sciences* 11.24 (2021), p. 11879.

[5] Chris Hanretty, Benjamin E Lauderdale, and Nick Vivyan. "Dyadic representation in a Westminster system". In: *Legislative Studies Quarterly* 42.2 (2017), pp. 235–267.

[6] Niels D Goet. "Measuring polarization with text analysis: Evidence from the UK House of Commons, 1811–2015". In: *Political Analysis* 27.4 (2019), pp. 518–539.

[7] Stephen Ansolabehere and Philip Edward Jones. "Dyadic representation". In: (2011).

[8] Callie Moore and Rohan Alexander. "Dyadic Representation in Canadian Parliament". In: (Apr. 2022).

[9] Kaspar Beelen et al. "Digitization of the Canadian parliamentary debates". In: *Canadian Journal of Political Science/Revue canadienne de science politique* 50.3 (2017), pp. 849–864.

[10] David M Blei, Andrew Y Ng, and Michael I Jordan. "Latent dirichlet allocation". In: *Journal of Machine Learning Research* 3.Jan (2003), pp. 993–1022.

[11] Maarten Grootendorst. "BERTopic: Neural topic modeling with a class-based TF-IDF procedure". In: *arXiv preprint arXiv:2203.05794* (2022).

[12] Clayton Hutto and Eric Gilbert. "Vader: A parsimonious rule-based model for sentiment analysis of social media text". In: *Proceedings of the international AAAI conference on web and social media*. Vol. 8. 1. 2014, pp. 216–225.

[13] *The Electoral System of Canada*. Mar. 2023. URL: https://www.elections.ca/content.aspx?section=res&dir=ces&document=part1&lang=e.

[14] William Christian and Harold Jansen. *Canadian Party system*. Dec. 2020. URL: https://www.thecanadianencyclopedia.ca/en/article/party-system.

[15] *Appendices - General Election Results Since 1867*. URL: https://www.ourcommons.ca/MarleauMontpetit/DocumentViewer.aspxDocId=1001&Sec=Ch25&Seq=11.

[16] Alain Gagnon and Brian Tanguay. "Canadian parties in transition: recent trends and new paths for research". In: *University Of Toronto Press* (2017).

[17] *Progressive Conservative Party of Canada*. URL: https://www.britannica.com/topic/Progressive-Conservative-Party-of-Canada.

[18] *Current members of Parliament*. URL: https://www.ourcommons.ca/members/en/search.

[19] Claude Couture and A. Davidson Dunton. *Biculturalism*. Dec. 2013. URL: https://www.thecanadianencyclopedia.ca/en/article/biculturalism.

[20] *Liberal Party of Canada*. Mar. 2023. URL: https://www.britannica.com/topic/Liberal-Party-of-Canada.

[21] *Bloc Québécois*. Feb. 2023. URL: https://www.britannica.com/topic/Bloc-Quebecois.

[22] J.J. McCullough. *Political parties*. Feb. 2023. URL: https://thecanadaguide.com/government/political-parties/.

[23] *Data*. URL: https://www.lipad.ca/data/.

[24] John Cook and Stephan Lewandowsky. "Rational irrationality: Modeling climate change belief polarization using Bayesian networks". In: *Topics in cognitive science* 8.1 (2016), pp. 160–179.

[25] Paul DiMaggio, John Evans, and Bethany Bryson. "Have American's social attitudes become more polarized?" In: *American journal of Sociology* 102.3 (1996), pp. 690–755.

[26] Kevin Arceneaux, Martin Johnson, and Chad Murphy. "Polarized political communication, oppositional media hostility, and selective exposure". In: *The Journal of Politics* 74.1 (2012), pp. 174–186.

[27] William Park. *How the views of a few can determine a country's fate*. Aug. 2019. URL: https://www.bbc.com/future/article/20190809-how-the-views-of-a-few-can-determine-the-fate-of-a-country.

[28] Joan Esteban and Debraj Ray. "Linking conflict to inequality and polarization". In: *American Economic Review* 101.4 (2011), pp. 1345–1374.

[29] Zeev Maoz and Zeynep Somer-Topcu. "Political polarization and cabinet stability in multiparty systems: A social networks analysis of European parliaments, 1945–98". In: *British Journal of Political Science* 40.4 (2010), pp. 805–833.

[30] David Ruppert. "What is kurtosis? An influence function approach". In: *The American Statistician* 41.1 (1987), pp. 1–5.

[31] Takuto Sakamoto and Hiroki Takikawa. "Cross-national measurement of polarization in political discourse: Analyzing floor debate in the US the Japanese legislatures". In: *2017 IEEE international conference on big data (Big Data)*. IEEE. 2017, pp. 3104–3110.

[32] Jianhua Lin. "Divergence measures based on the Shannon entropy". In: *IEEE Transactions on Information theory* 37.1 (1991), pp. 145–151.

[33] PA Bromiley, NA Thacker, and E Bouhova-Thacker. "Shannon entropy, Renyi entropy, and information". In: *Statistics and Inf. Series (2004-004)* 9 (2004), pp. 2–8.

[34] Gary C Jacobson. "Party polarization in national politics: The electoral connection". In: *Polarized politics: Congress and the president in a partisan era*. Vol. 5. 2000, pp. 17–18.

[35] Gordon Heltzel and Kristin Laurin. "Polarization in America: Two possible futures". In: *Current opinion in behavioral sciences* 34 (2020), pp. 179–184.

[36] Morris P Fiorina and Samuel J Abrams. "Political polarization in the American public". In: *Annu. Rev. Polit. Sci.* 11 (2008), pp. 563–588.

[37] Adam M Enders and Miles T Armaly. "The differential effects of actual and perceived polarization". In: *Political Behavior* 41 (2019), pp. 815–839.

[38] Christopher A Bail et al. "Exposure to opposing views on social media can increase political polarization". In: *Proceedings of the National Academy of Sciences* 115.37 (2018), pp. 9216–9221.

[39] Stefano Balietti et al. “Reducing opinion polarization: Effects of exposure to similar people with differing political views”. In: *Proceedings of the National Academy of Sciences* 118.52 (2021). DOI: 10.1073/pnas.2112552118.

[40] R Alan Walks. “The causes of city-suburban political polarization? A Canadian case study”. In: *Annals of the Association of American Geographers* 96.2 (2006), pp. 390–414.

[41] John Agnew. “Home ownership and identity in capitalist societies”. In: *Housing and identity: Cross-cultural perspectives* (1981), pp. 60–97.

[42] Gordon Pennycook et al. “Beliefs about COVID-19 in Canada, the United Kingdom, and the United States: A novel test of political polarization and motivated reasoning”. In: *Personality and Social Psychology Bulletin* 48.5 (2022), pp. 750–765.

[43] Eric Merkley. “Polarization Eh? Ideological Divergence and Partisan Sorting in the Canadian Mass Public”. In: *Public Opinion Quarterly* 86.4 (2022), pp. 932–943.

[44] *Canadian Election Study*. May 2020. URL: https://www.elections.ca/content.aspx?section=res&dir=rec%5C%2Feval%5C%2Fces&document=index&lang=e.

[45] Noam Chomsky. “Three models for the description of language”. In: *IRE Transactions on information theory* 2.3 (1956), pp. 113–124.

[46] Prakash M Nadkarni, Lucila Ohno-Machado, and Wendy W Chapman. “Natural language processing: an introduction”. In: *Journal of the American Medical Informatics Association* 18.5 (2011), pp. 544–551.

[47] Daniel W Otter, Julian R Medina, and Jugal K Kalita. “A survey of the usages of deep learning for natural language processing”. In: *IEEE transactions on neural networks and learning systems* 32.2 (2020), pp. 604–624.

[48] Jacob Devlin et al. “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *arXiv preprint arXiv:1810.04805* (2018).

[49] OpenAI. Nov. 2022. URL: https://openai.com/blog/chatgpt.

[50] Sean M Watford et al. “Novel application of normalized pointwise mutual information (NPMI) to mine biomedical literature for gene sets associated with disease: Use case in breast carcinogenesis”. In: *Computational Toxicology* 7 (2018), pp. 46–57.

[51] Gerlof Bouma. “Normalized (pointwise) mutual information in collocation extraction”. In: *Proceedings of GSCL* 30 (2009), pp. 31–40.

[52] Shaheen Syed and Marco Spruit. “Full-Text or Abstract? Examining Topic Coherence Scores Using Latent Dirichlet Allocation”. In: *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. 2017, pp. 165–174. DOI: 10.1109/DSAA.2017.61.

[53] Thorsten Joachims. *A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization.* Tech. rep. Carnegie Mellon University Department of Computer Science, 1996.

[54] Amit Singhal et al. “Modern information retrieval: A brief overview”. In: *IEEE Data Eng. Bull.* 24.4 (2001), pp. 35–43.

[55] Hongli Yuan et al. “A detection method for android application security based on TF-IDF and machine learning”. In: *Plos one* 15.9 (2020), e0238694.

[56] Hans Christian, Mikhael Pramodana Agus, and Derwin Suhartono. “Single document automatic text summarization using term frequency-inverse document frequency (TF-IDF)”. In: *ComTech: Computer, Mathematics and Engineering Applications* 7.4 (2016), pp. 285–294.

[57] Jonathan J Webster and Chunyu Kit. “Tokenization as the initial phase in NLP”. In: *COLING 1992 volume 4: The 14th international conference on computational linguistics*. 1992.

[58] S Vijayarani, R Janani, et al. “Text mining: open source tokenization tools-an analysis”. In: *Advanced Computational Intelligence: An International Journal (ACII)* 3.1 (2016), pp. 37–47.

[59] Anjali Ganesh Jivani et al. “A comparative study of stemming algorithms”. In: *Int. J. Comp. Tech. Appl* 2.6 (2011), pp. 1930–1938.

[60] Divya Khyani et al. “An interpretation of lemmatization and stemming in natural language processing”. In: *Journal of University of Shanghai for Science and Technology* 22.10 (2021), pp. 350–357.

[61] Edward Loper and Steven Bird. *NLTK: The Natural Language Toolkit*. 2002. DOI: 10.48550/ARXIV.CS/0205028. URL: https://arxiv.org/abs/cs/0205028.

[62] *Oxford English Dictionary*. URL: https://www.oed.com/.

[63] Jashanjot Kaur and P Kaur Buttar. “A systematic review on stopword removal algorithms”. In: *International Journal on Future Revolution in Computer Science & Communication Engineering* 4.4 (2018), pp. 207–210.

[64] Alexandra Schofield, Måns Magnusson, and David Mimno. “Pulling out the stops: Rethinking stopword removal for topic models”. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, short papers*. 2017, pp. 432–436.

[65] Kherwa Pooja and P Bansal. “Topic Modeling: A Comprehensive Review”. In: *EAI Endorsed Trans. Scalable Inf. Syst* 7 (2019), e2.

[66] David Jurgens and Keith Stevens. “The S-Space package: an open source package for word space models”. In: *Proceedings of the ACL 2010 system demonstrations*. 2010, pp. 30–35.

[67] Tim Van de Cruys and Marianna Apidianaki. “Latent semantic word sense induction and disambiguation”. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. 2011, pp. 1476–1485.

[68] Lin Liu et al. “An overview of topic modeling and its current applications in bioinformatics”. In: *SpringerPlus* 5.1 (2016), pp. 1–22.

[69] Xing Yi and James Allan. “A Comparative Study of Utilizing Topic Models for Information Retrieval”. In: *Advances in Information Retrieval*. Ed. by Mohand Boughanem et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 29–41. ISBN: 978-3-642-00958-7.

[70] Thomas K Landauer and Susan T Dumais. “A solution to Plato’s problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge.” In: *Psychological review* 104.2 (1997), p. 211.

[71] Hanna M Wallach. “Topic modeling: beyond bag-of-words”. In: *Proceedings of the 23rd international conference on Machine learning*. 2006, pp. 977–984.

[72] Jon Mcauliffe and David Blei. “Supervised topic models”. In: *Advances in neural information processing systems* 20 (2007).

[73] J Bellagarda. “Latent Semantic Mapping: A Data driven Framework for Modeling Global Relationships Implicit in Large Volumes of Data”. In: *IEEE signal processing magazine* 22 (2005), pp. 70–80.

[74] V. Paul Pauca et al. “Text Mining using Non-Negative Matrix Factorizations”. In: *Proceedings of the 2004 SIAM International Conference on Data Mining (SDM)*, pp. 452–456. DOI: 10.1137/1.9781611972740.45. eprint: https://epubs.siam.org/doi/pdf/10.1137/1.9781611972740.45. URL: https://epubs.siam.org/doi/abs/10.1137/1.9781611972740.45.

[75] Jordan Boyd-Graber, Yuening Hu, David Mimno, et al. “Applications of topic models”. In: *Foundations and Trends® in Information Retrieval* 11.2-3 (2017), pp. 143–296.

[76] Abeer Abuzayed and Hend Al-Khalifa. “BERT for Arabic topic modeling: An experimental study on BERTopic technique”. In: *Procedia computer science* 189 (2021), pp. 191–194.

[77] Roman Egger and Joanne Yu. “A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts”. In: *Frontiers in sociology* 7 (2022).

[78] Michael Röder, Andreas Both, and Alexander Hinneburg. “Exploring the space of topic coherence measures”. In: *Proceedings of the eighth ACM international conference on Web search and data mining*. 2015, pp. 399–408.

[79] Keith Stevens et al. “Exploring topic coherence over many models and many topics”. In: *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*. 2012, pp. 952–961.

[80] Nikolaos Aletras and Mark Stevenson. “Evaluating topic coherence using distributional semantics”. In: *Proceedings of the 10th international conference on computational semantics (IWCS 2013)–Long Papers*. 2013, pp. 13–22.

[81] David Newman et al. “Automatic evaluation of topic coherence”. In: *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*. 2010, pp. 100–108.

[82] David Mimno et al. “Optimizing semantic coherence in topic models”. In: *Proceedings of the 2011 conference on empirical methods in natural language processing*. 2011, pp. 262–272.

[83] Sarah A Alkhodair et al. “Improving interpretations of topic modeling in microblogs”. In: *Journal of the Association for Information Science and Technology* 69.4 (2018), pp. 528–540.

[84] Michael Röder. *Not being able to replicate coherence scores from paper*. Mar. 2018. URL: https://github.com/dice-group/Palmetto/issues/13#issuecomment-371553052.

[85] Lei Zhang, Shuai Wang, and Bing Liu. “Deep learning for sentiment analysis: A survey”. In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 8.4 (2018), e1253.

[86] Walaa Medhat, Ahmed Hassan, and Hoda Korashy. “Sentiment analysis algorithms and applications: A survey”. In: *Ain Shams engineering journal* 5.4 (2014), pp. 1093–1113.

[87] Martin Haselmayer and Marcelo Jenny. “Sentiment analysis of political communication: Combining a dictionary approach with crowdcoding”. In: *Quality & quantity* 51 (2017), pp. 2623–2646.

[88] Shihab Elbagir and Jing Yang. “Twitter sentiment analysis using natural language toolkit and VADER sentiment”. In: *Proceedings of the international multiconference of engineers and computer scientists*. Vol. 122. 2019, p. 16.

[89] Maria Chiara Martinis, Chiara Zucco, and Mario Cannataro. “An Italian lexicon-based sentiment analysis approach for medical applications”. In: *Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*. 2022, pp. 1–4.

[90] M Umair et al. “Sentiment analysis of students’ feedback before and after covid-19 pandemic”. In: *International Journal on Emerging Technologies* 12.2 (2021), pp. 177–182.

[91] Hu Xu et al. “BERT post-training for review reading comprehension and aspect-based sentiment analysis”. In: *arXiv preprint arXiv:1904.02232* (2019).

[92] Sifan Xu and Ying Xiong. “Setting socially mediated engagement parameters: A topic modeling and text analytic approach to examining polarized discourses on Gillette’s campaign”. In: *Public Relations Review* 46.5 (2020), p. 101959.

[93] Nádia FF da Silva et al. “Evaluating topic models in Portuguese political comments about bills from brazil’s chamber of deputies”. In: *Intelligent Systems: 10th Brazilian Conference, BRACIS 2021, Virtual Event, November 29–December 3, 2021, Proceedings, Part II 10*. Springer. 2021, pp. 104–120.

[94] Gavin Abercrombie and Riza Batista-Navarro. “Sentiment and position-taking analysis of parliamentary debates: a systematic literature review”. In: *Journal of Computational Social Science* 3.1 (2020), pp. 245–270.

[95] Burt L Monroe, Michael P Colaresi, and Kevin M Quinn. “Fightin’words: Lexical feature selection and evaluation for identifying the content of political conflict”. In: *Political Analysis* 16.4 (2008), pp. 372–403.

[96] The pandas development team. *pandas-dev/pandas: Pandas*. Version 1.4.4. Feb. 2020. DOI: 10.5281/zenodo.3509134. URL: https://doi.org/10.5281/zenodo.3509134.

[97] *List of the largest population centres in Canada*. Apr. 2023. URL: https://en.wikipedia.org/wiki/List_of_the_largest_population_centres_in_Canada.

[98] URL: https://www.usna.edu/Users/cs/roche/courses/s15si335/proj1/files.php%5C%3Ff=names.txt&downloadcode=yes.

[99] Radim Rehurek and Petr Sojka. “Gensim–Python Framework for Vector Space Modelling”. In: *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic* 3.2 (2011).

[100] Pauli Virtanen et al. “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python”. In: *Nature Methods* 17 (2020), pp. 261–272. DOI: 10.1038/s41592-019-0686-2.

[101] *C-14: An Act to amend the Criminal Code and to make related amendments to other Acts (medical assistance in dying)*. URL: https://www.parl.ca/LegisInfo/en/bill/42-1/C-14.

A Appendix A: Additional Data and Notes

A.1 List of Fields Included in LiPaD Dataset

- basepk
- hid
- speechdate
- pid
- opid
- speakeroldname
- speakerposition
- maintopic
- subtopic
- subsubtopic
- speechtext
- speakerparty
- speakerriding
- speakername
- speakerurl

A.2 Example of Speech Contained in LiPaD Dataset

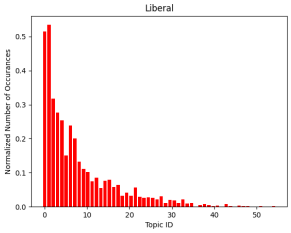
basepk	4647937
hid	ca.proc.d.2015-12-10.17039.0
speechdate	2015-12-10
pid	3176068f-62ee-4b2f-aacd-1131201c6db1
opid	242.0
speakeroldname	The Speaker
speakerposition	NaN
maintopic	Routine Proceedings
subtopic	Office of the Privacy Commissioner
subsubtopic	NaN
speechtext	Pursuant to section 38 of the Privacy Act, I have the honour to lay upon the table the annual report of the Privacy Commissioner for the fiscal year ending March 31, 2015. Pursuant to Standing Order 108(3)(h), this report is deemed permanently referred to the Standing Committee on Access to Information, Privacy and Ethics.
speakerparty	Liberal
speakerriding	Halifax West
speakername	Geoff Regan
speakerurl	http://www.parl.gc.ca/parlinfo/Files/Parliamentarian.aspx?Item=3176068f-62ee-4b2f-aacd-1131201c6db1&Language=E&Section=ALL

Figure 11: Example of Speech Data Contained in LiPaD Dataset

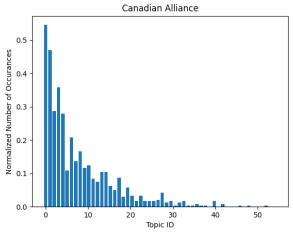
This speech was given on December 10, 2015 by Geoff Regan, the Speaker of the House. As can be seen, not all fields have an associated value—neither the speakerposition nor subsubtopic columns have been assigned a value, and are therefore displayed as NaN (not a number) by Pandas.

B Appendix B: Graphs of Topic Occurrences by Party and Topic Model

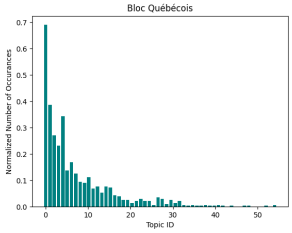
B.1 LDA Topic Frequency Graphs



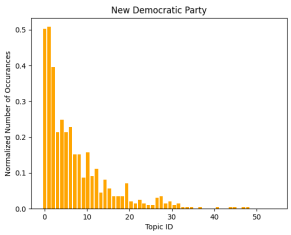
(a) Liberal Party 2004



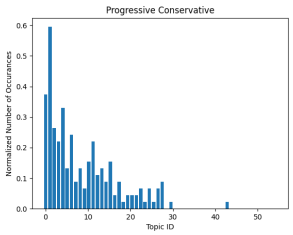
(b) Canadian Alliance 2004



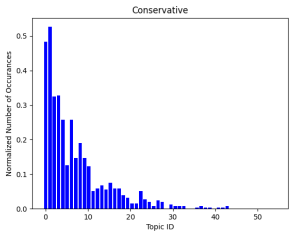
(c) Bloc Québécois 2004



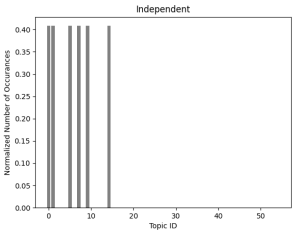
(d) New Democratic Party 2004



(e) Progressive Conservative 2004

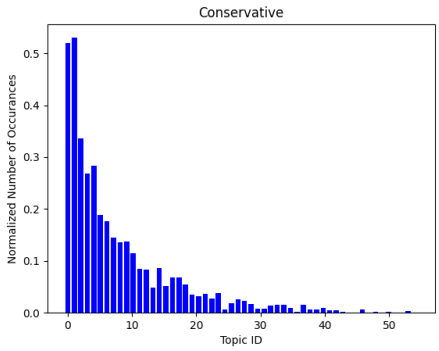


(f) Conservative Party 2004

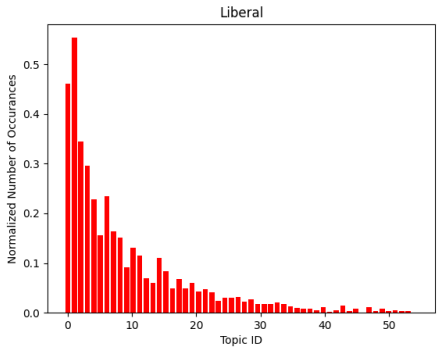


(g) Independent 2004

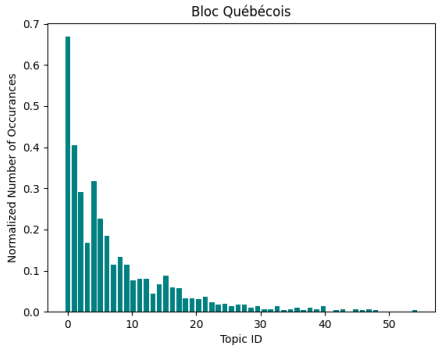
Figure 12: LDA Topic Frequencies for Each Party in 2004



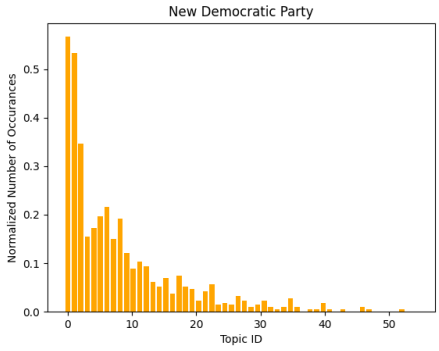
(a) Conservative Party 2005



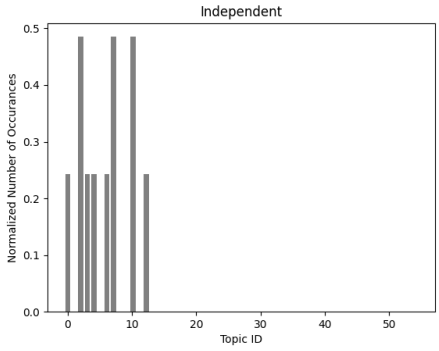
(b) Liberal Party 2005



(c) Bloc Québécois 2005

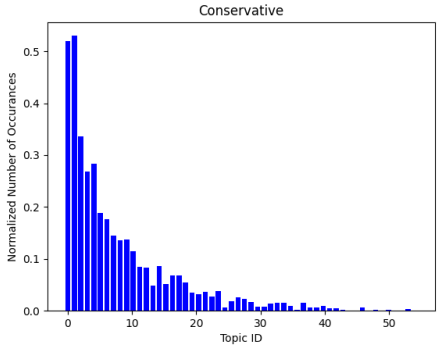


(d) New Democratic Party 2005

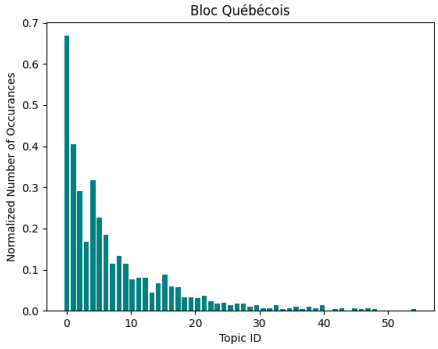


(e) Independent 2005

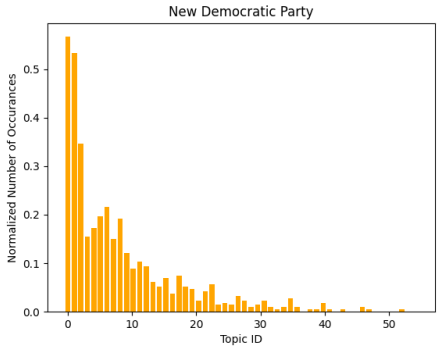
Figure 13: LDA Topic Frequencies for Each Party in 2005



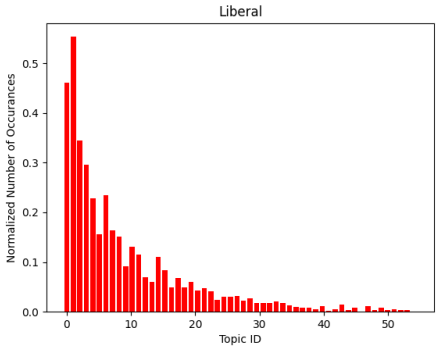
(a) Conservative Party 2006



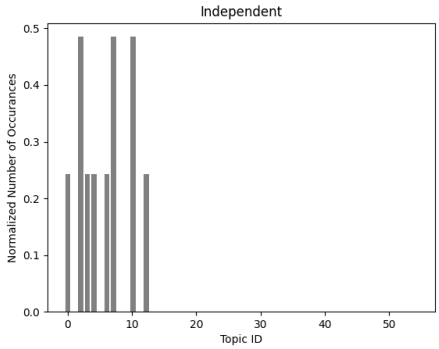
(b) Bloc Québécois 2006



(c) New Democratic Party 2006

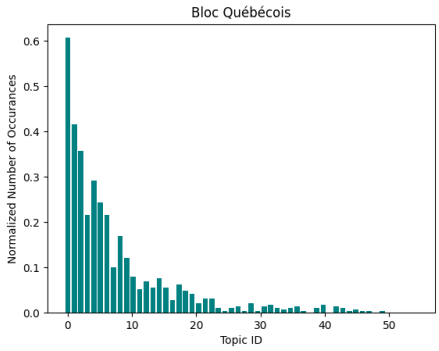


(d) Liberal Party 2006

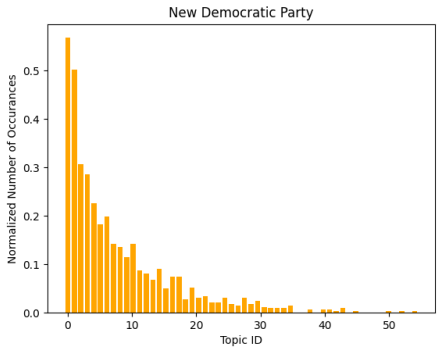


(e) Independent 2006

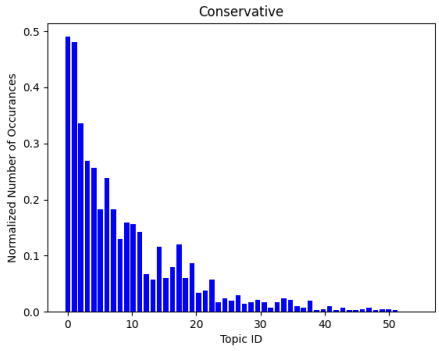
Figure 14: LDA Topic Frequencies for Each Party in 2006



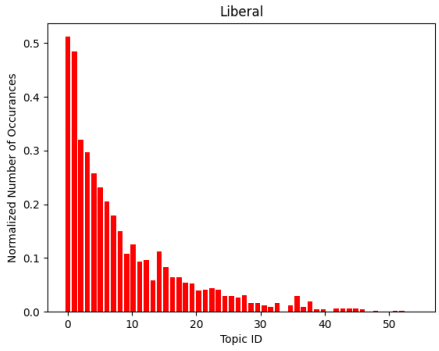
(a) Bloc Québécois 2007



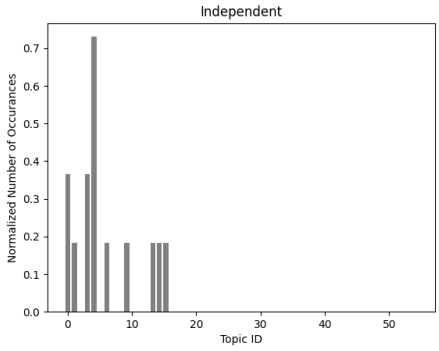
(b) New Democratic Party 2007



(c) Conservative Party 2007

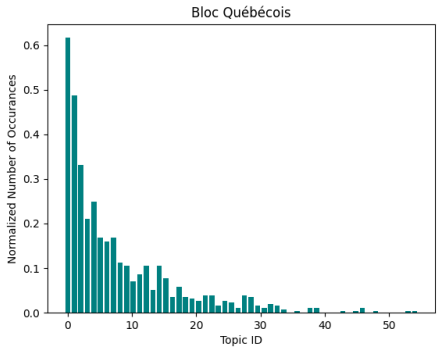


(d) Liberal Party 2007

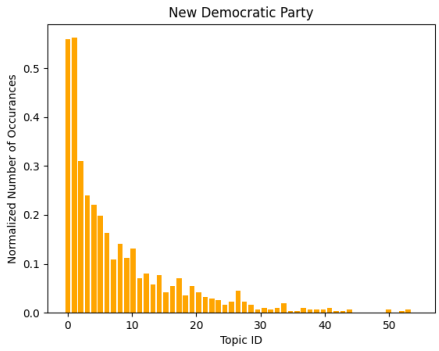


(e) Independent 2007

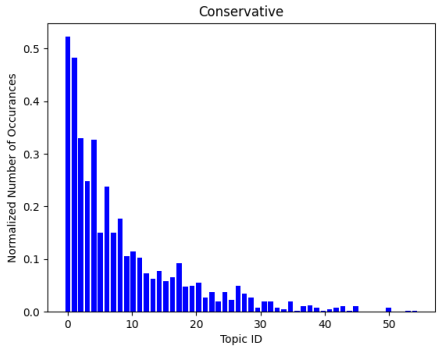
Figure 15: LDA Topic Frequencies for Each Party in 2007



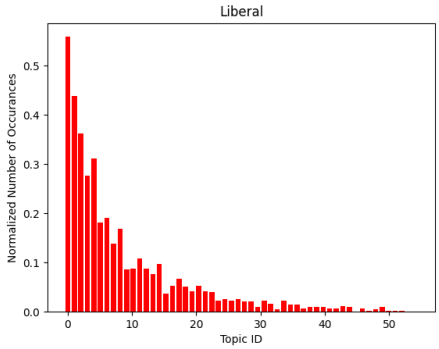
(a) Bloc Québécois 2008



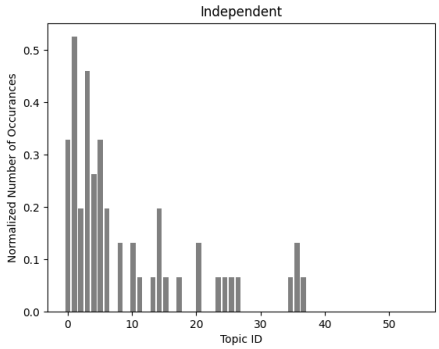
(b) New Democratic Party 2008



(c) Conservative Party 2008

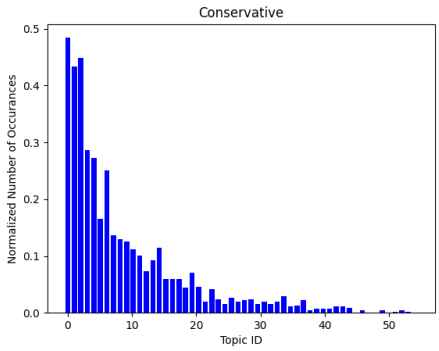


(d) Liberal Party 2008

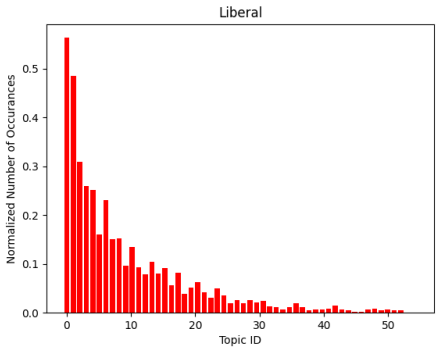


(e) Independent 2008

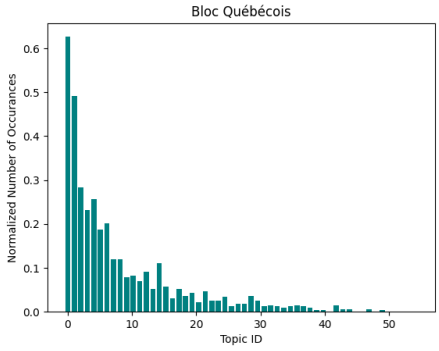
Figure 16: LDA Topic Frequencies for Each Party in 2008



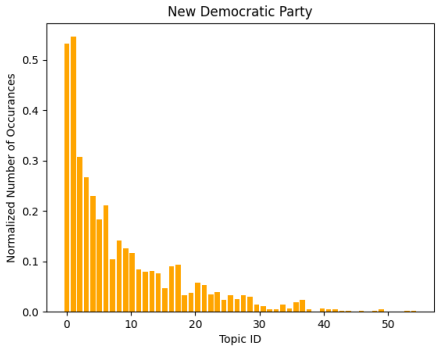
(a) Conservative Party 2009



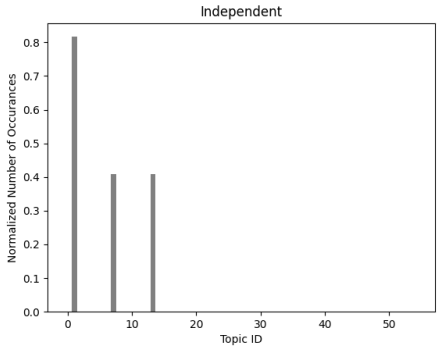
(b) Liberal Party 2009



(c) Bloc Québécois 2009

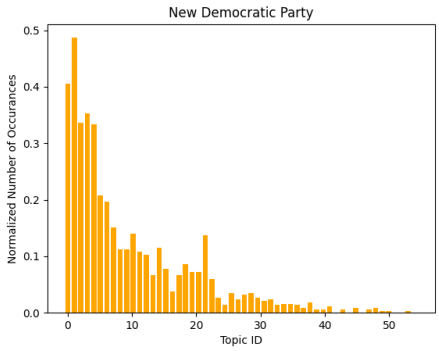


(d) New Democratic Party 2009

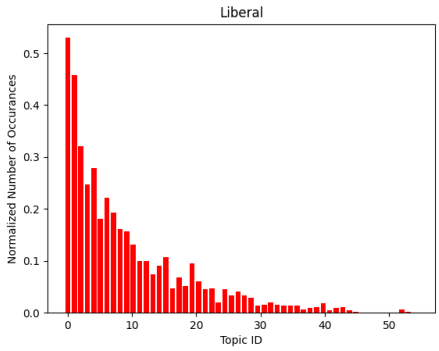


(e) Independent 2009

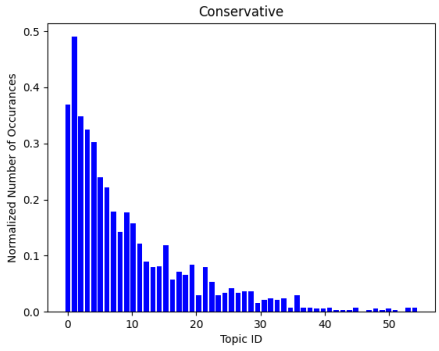
Figure 17: LDA Topic Frequencies for Each Party in 2009



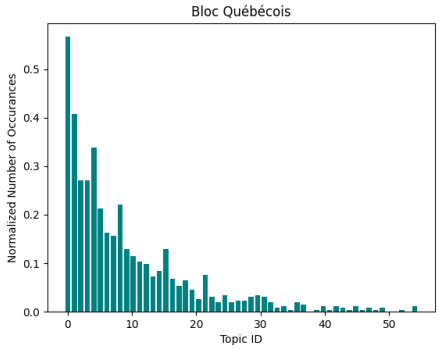
(a) New Democratic Party 2010



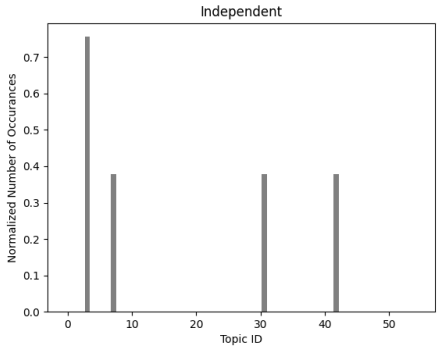
(b) Liberal Party 2010



(c) Conservative Party 2010

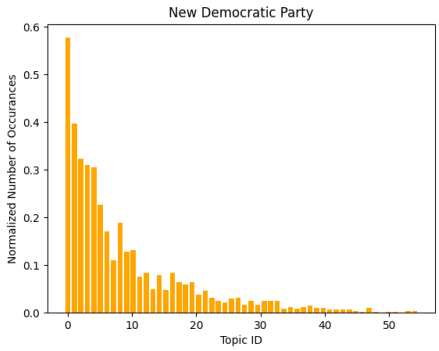


(d) Bloc Québécois 2010

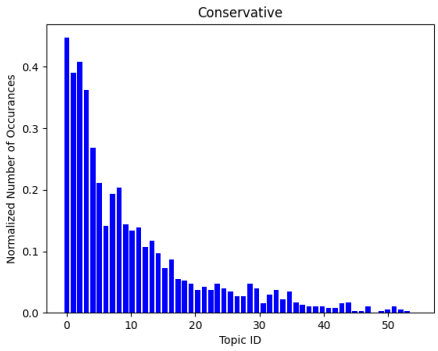


(e) Independent 2010

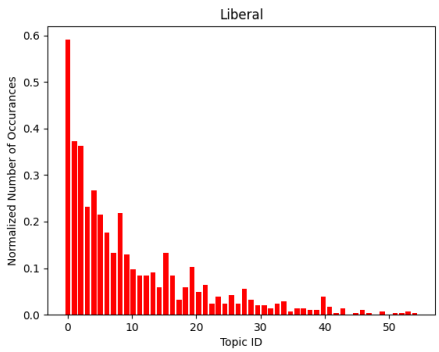
Figure 18: LDA Topic Frequencies for Each Party in 2010



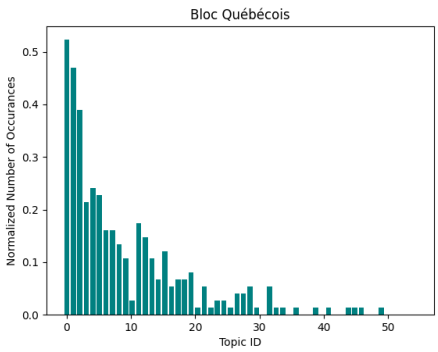
(a) New Democratic Party 2011



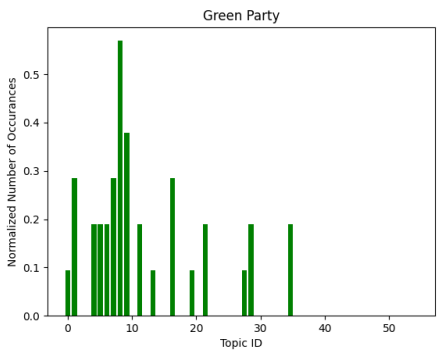
(b) Conservative Party 2011



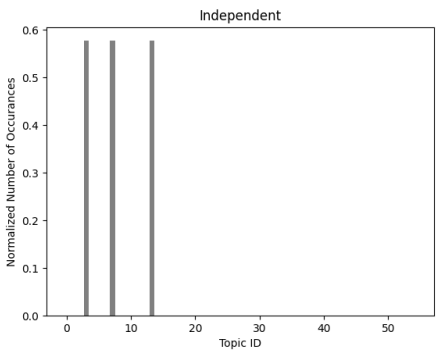
(c) Liberal Party 2011



(d) Bloc Québécois 2011

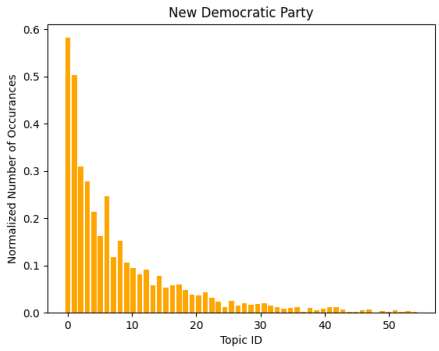


(e) Green Party 2011

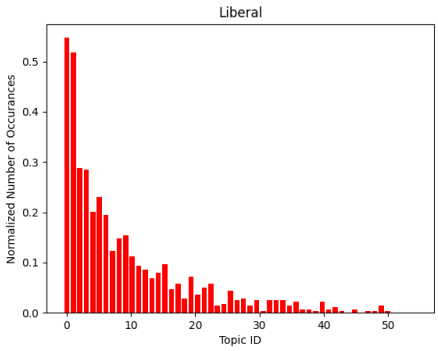


(f) Independent 2011

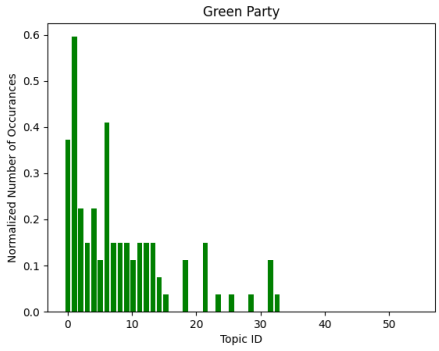
Figure 19: LDA Topic Frequencies for Each Party in 2011



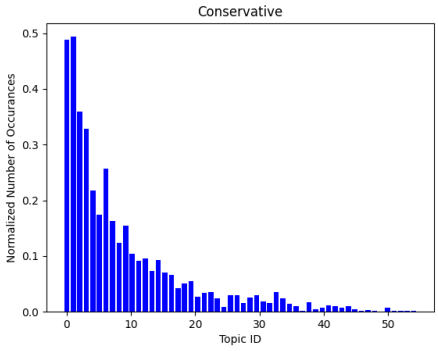
(a) New Democratic Party 2012



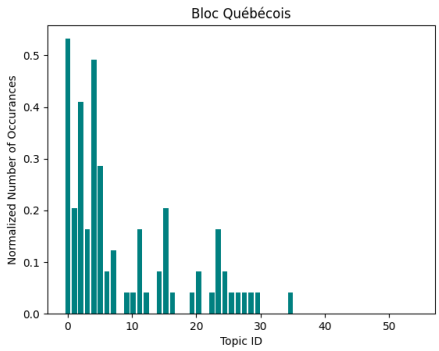
(b) Liberal Party 2012



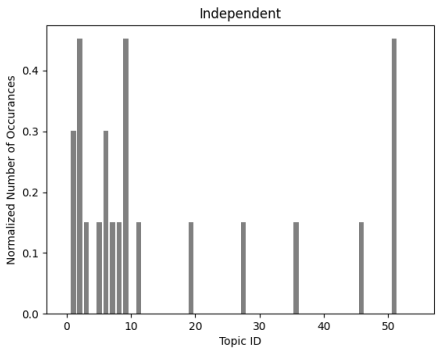
(c) Green Party 2012



(d) Conservative Party 2012

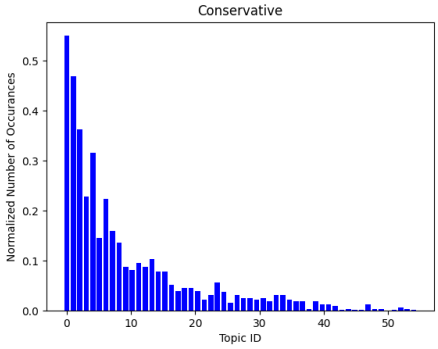


(e) Bloc Québécois 2012

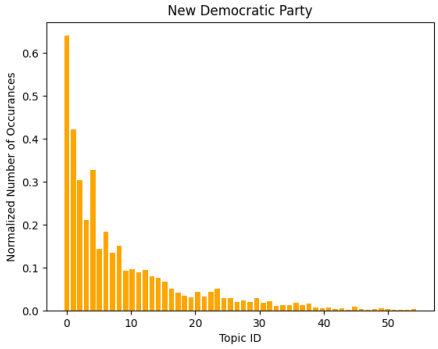


(f) Independent 2012

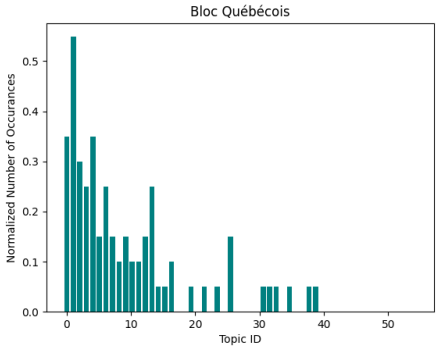
Figure 20: LDA Topic Frequencies for Each Party in 2012



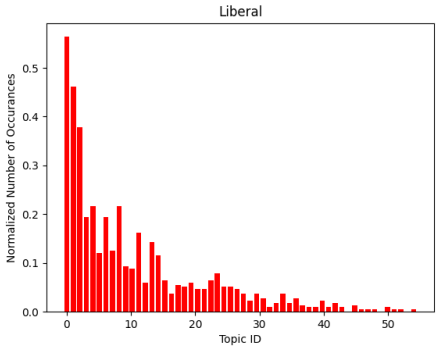
(a) Conservative Party 2013



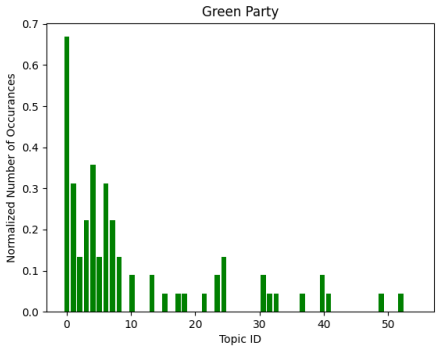
(b) New Democratic Party 2013



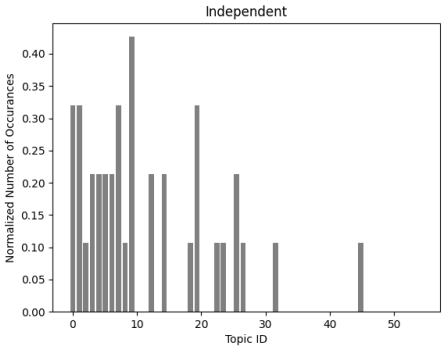
(c) Bloc Québécois 2013



(d) Liberal Party 2013

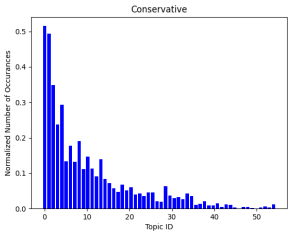


(e) Green Party 2013

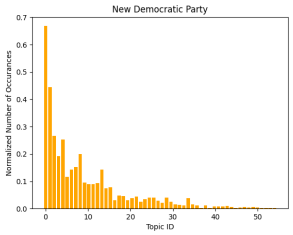


(f) Independent 2013

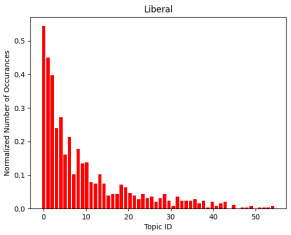
Figure 21: LDA Topic Frequencies for Each Party in 2013



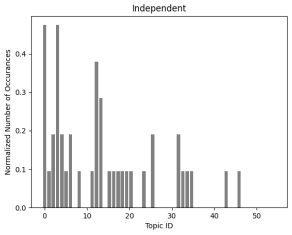
(a) Conservative Party 2014



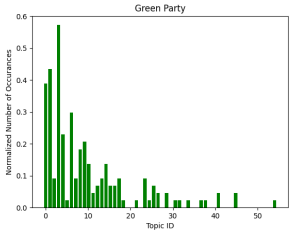
(b) New Democratic Party 2014



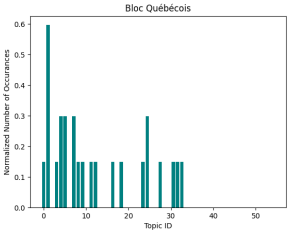
(c) Liberal Party 2014



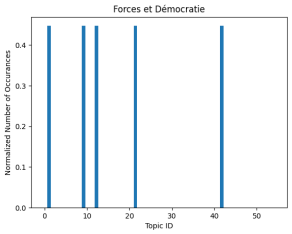
(d) Independent 2014



(e) Green Party 2014

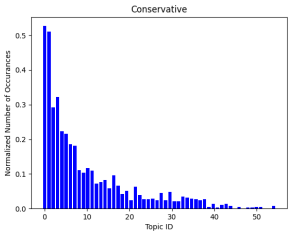


(f) Bloc Québécois 2014

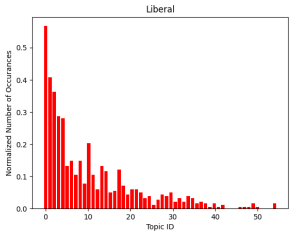


(g) Forces et Démocratie 2014

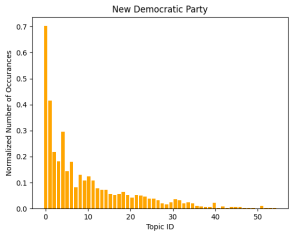
Figure 22: LDA Topic Frequencies for Each Party in 2014



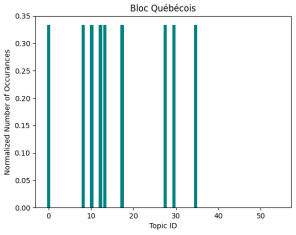
(a) Conservative Party 2015



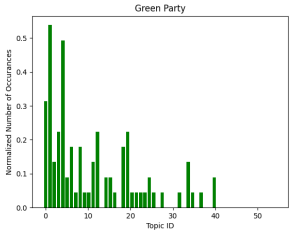
(b) Liberal Party 2015



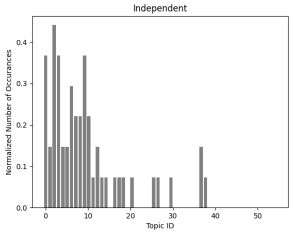
(c) New Democratic Party 2015



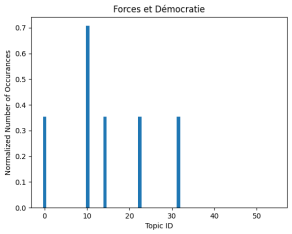
(d) Bloc Québécois 2015



(e) Green Party 2015

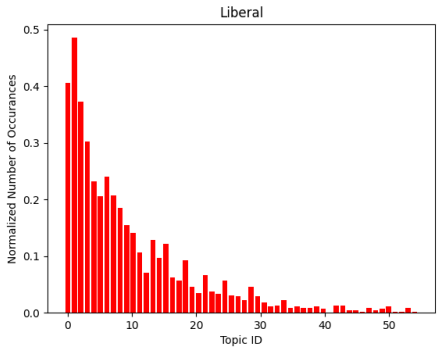


(f) Independent 2015

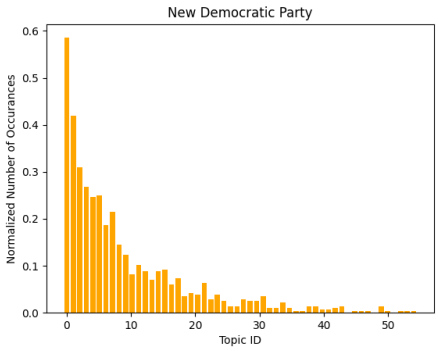


(g) Forces et Démocratie 2015

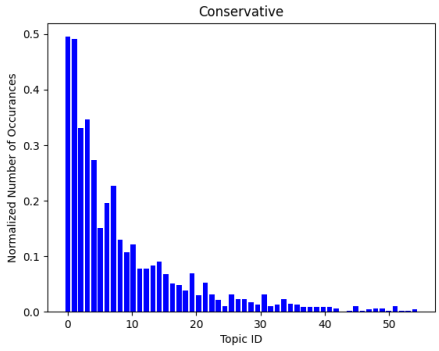
Figure 23: LDA Topic Frequencies for Each Party in 2015



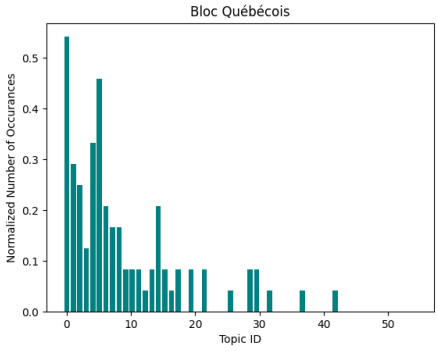
(a) Liberal Party 2016



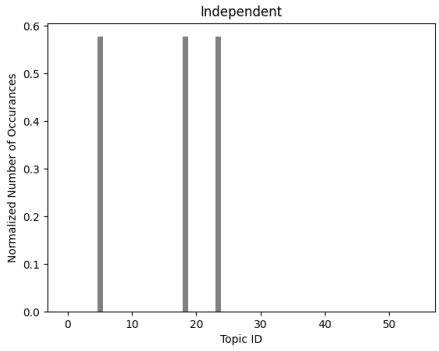
(b) New Democratic Party 2016



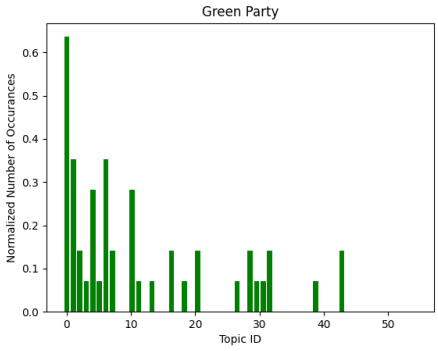
(c) Conservative Party 2016



(d) Bloc Québécois 2016

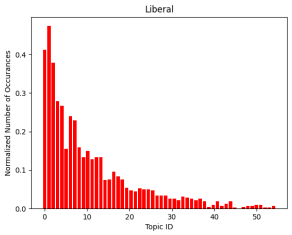


(e) Independent 2016

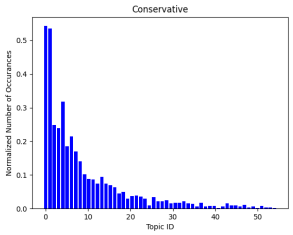


(f) Green Party 2016

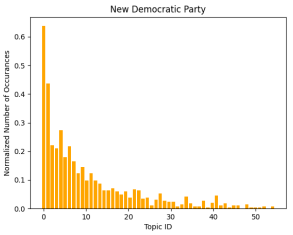
Figure 24: LDA Topic Frequencies for Each Party in 2016



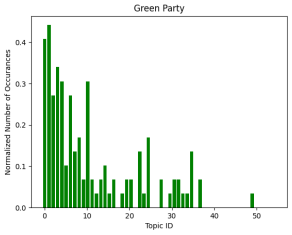
(a) Liberal Party 2018



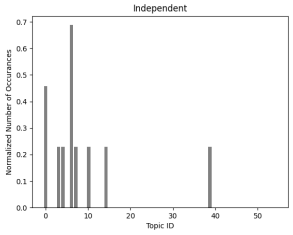
(b) Conservative Party 2018



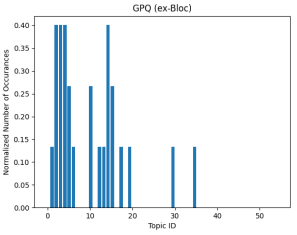
(c) New Democratic Party 2018



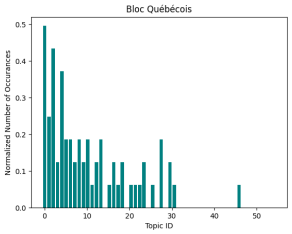
(d) Green Party 2018



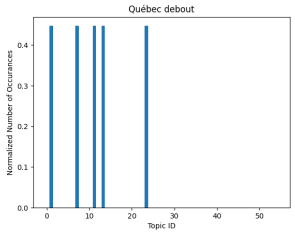
(e) Independent 2018



(f) GPQ (ex-Bloc) 2018

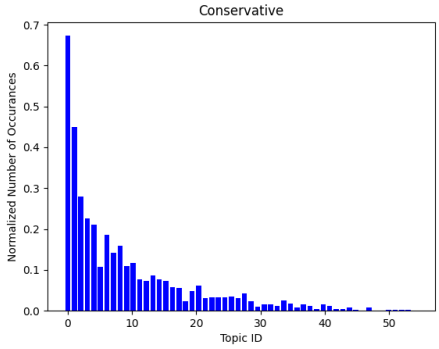


(g) Bloc Québécois 2018

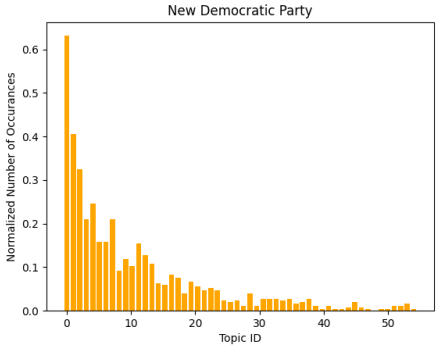


(h) Québec debout 2018

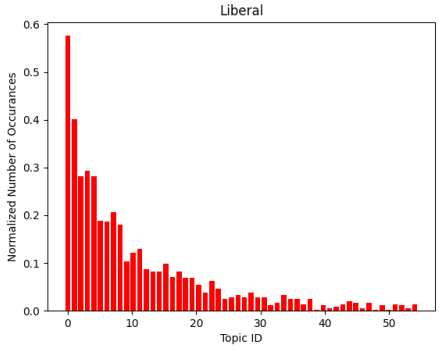
Figure 25: LDA Topic Frequencies for Each Party in 2018



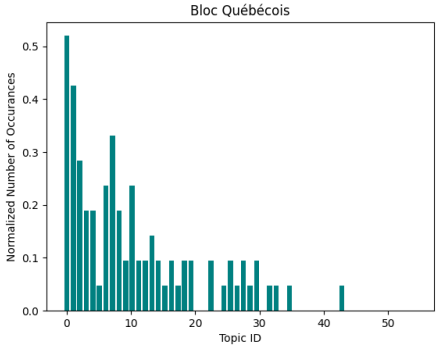
(a) Conservative Party 2019



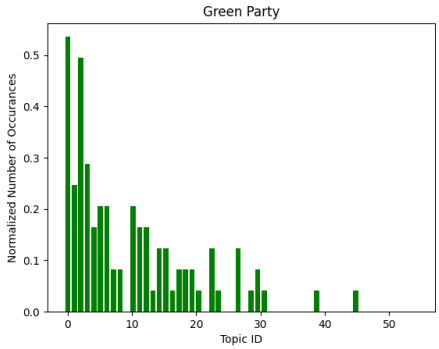
(b) New Democratic Party 2019



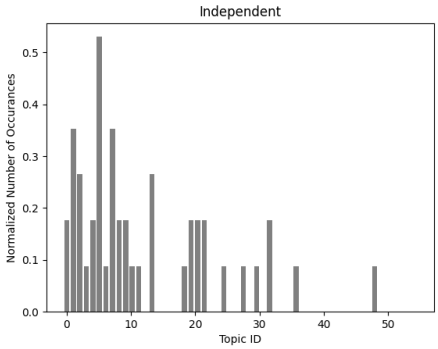
(c) Liberal Party 2019



(d) Bloc Québécois 2019



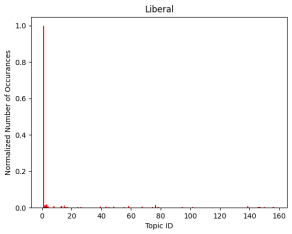
(e) Green Party 2019



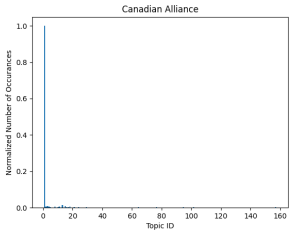
(f) Independent 2019

Figure 26: LDA Topic Frequencies for Each Party in 2019

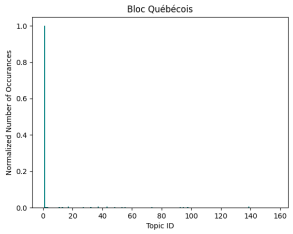
B.2 BERTopic Topic Frequency Graphs



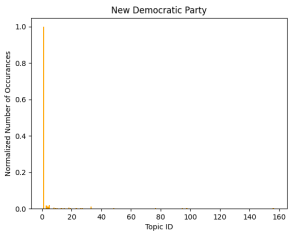
(a) Liberal Party 2004



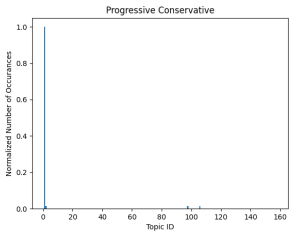
(b) Canadian Alliance 2004



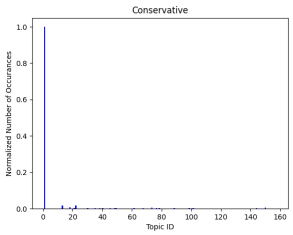
(c) Bloc Québécois 2004



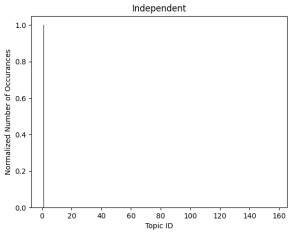
(d) New Democratic Party 2004



(e) Progressive Conservative 2004

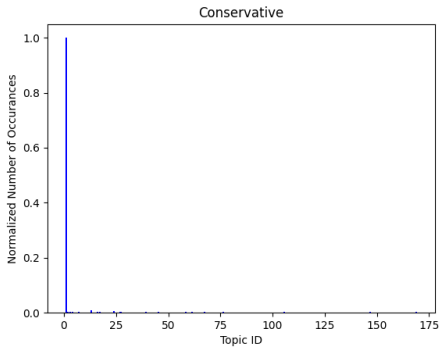


(f) Conservative Party 2004

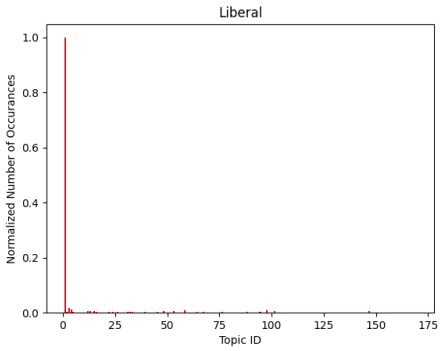


(g) Independent 2004

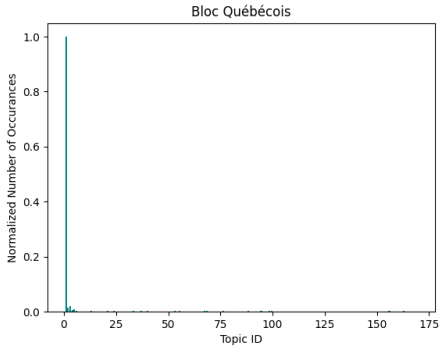
Figure 27: BERTopic Topic Frequencies for Each Party in 2004



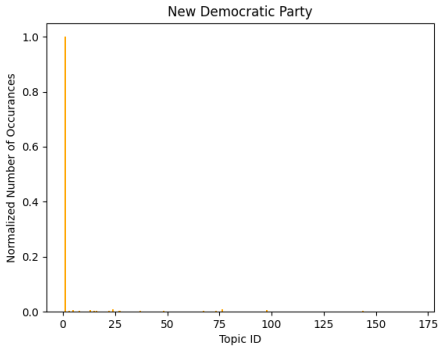
(a) Conservative Party 2005



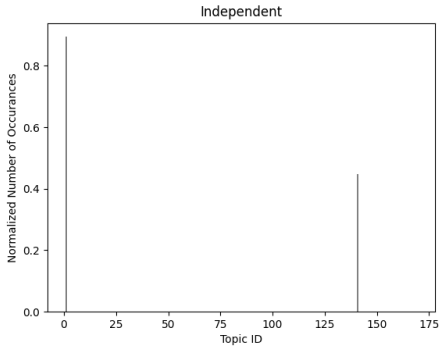
(b) Liberal Party 2005



(c) Bloc Québécois 2005

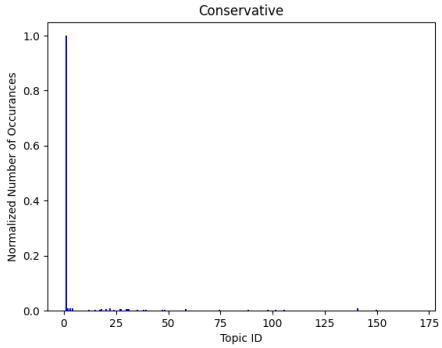


(d) New Democratic Party 2005

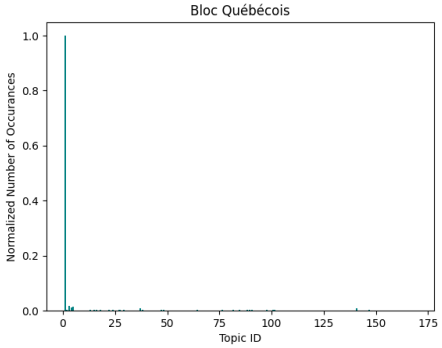


(e) Independent 2005

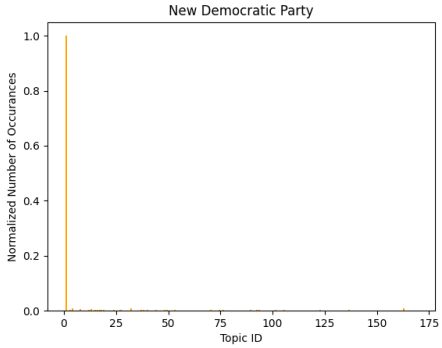
Figure 28: BERTopic Topic Frequencies for Each Party in 2005



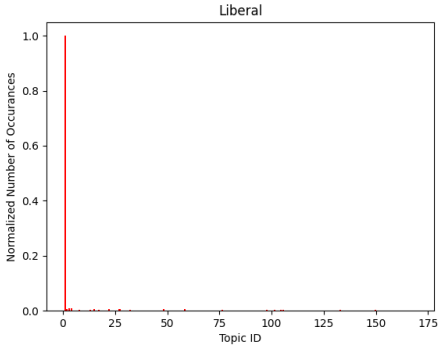
(a) Conservative Party 2006



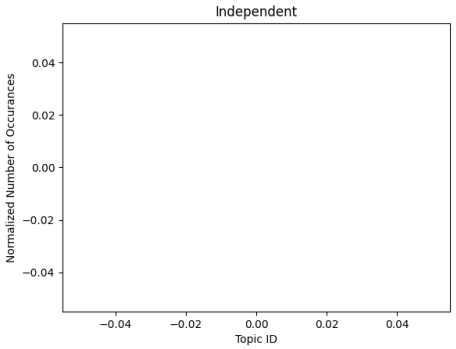
(b) Bloc Québécois 2006



(c) New Democratic Party 2006

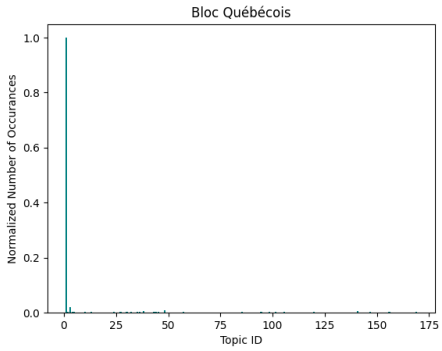


(d) Liberal Party 2006

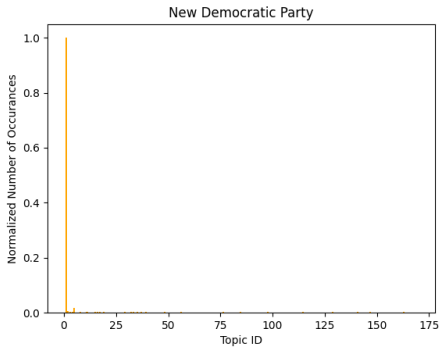


(e) Independent 2006

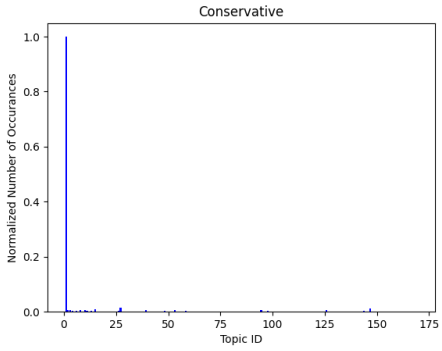
Figure 29: BERTopic Topic Frequencies for Each Party in 2006



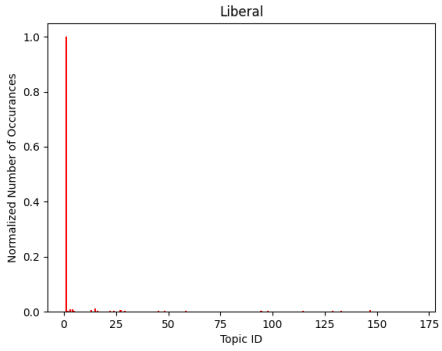
(a) Bloc Québécois 2007



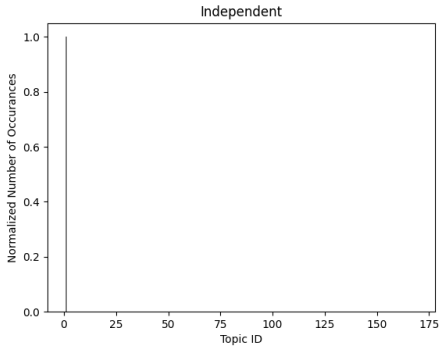
(b) New Democratic Party 2007



(c) Conservative Party 2007

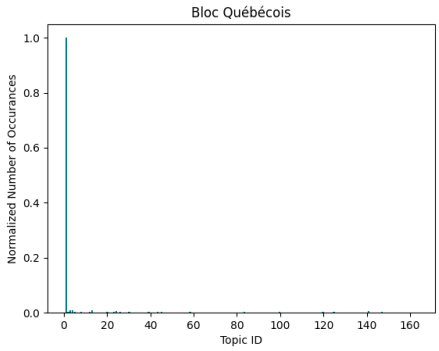


(d) Liberal Party 2007

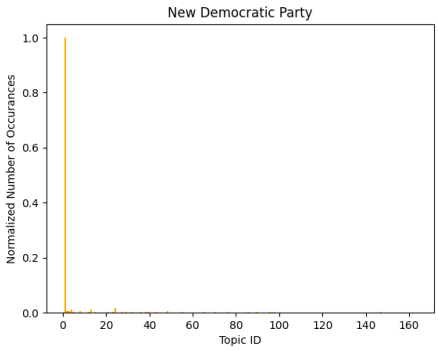


(e) Independent 2007

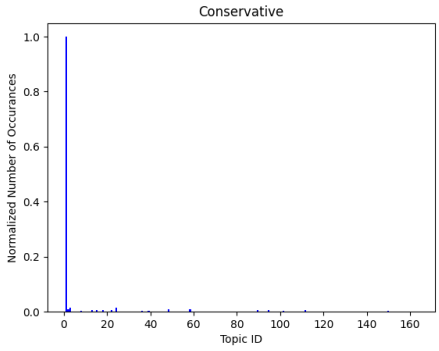
Figure 30: BERTopic Topic Frequencies for Each Party in 2007



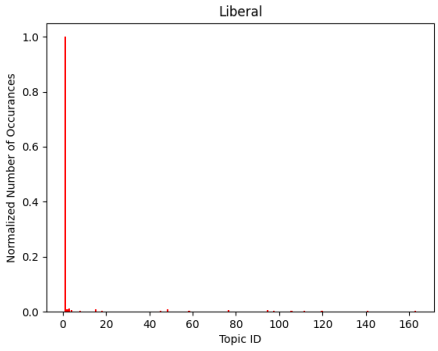
(a) Bloc Québécois 2008



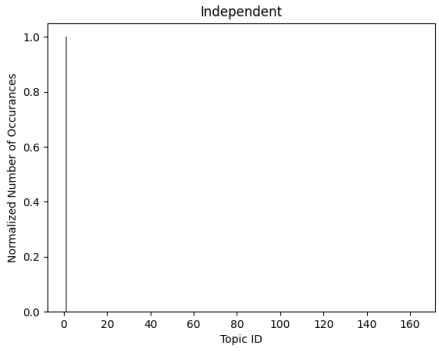
(b) New Democratic Party 2008



(c) Conservative Party 2008

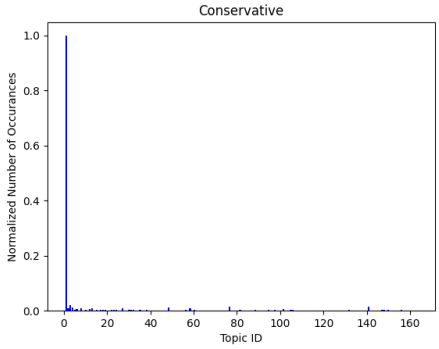


(d) Liberal Party 2008

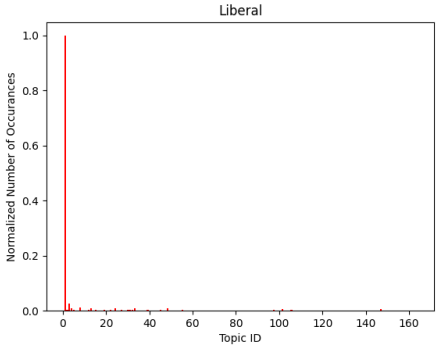


(e) Independent 2008

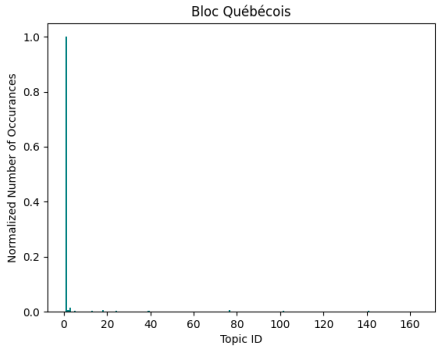
Figure 31: BERTopic Topic Frequencies for Each Party in 2008



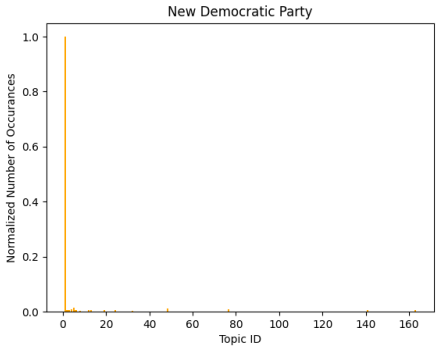
(a) Conservative Party 2009



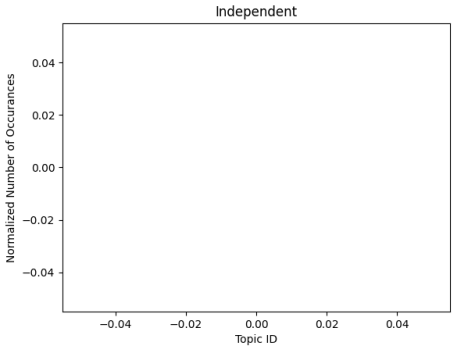
(b) Liberal Party 2009



(c) Bloc Québécois 2009

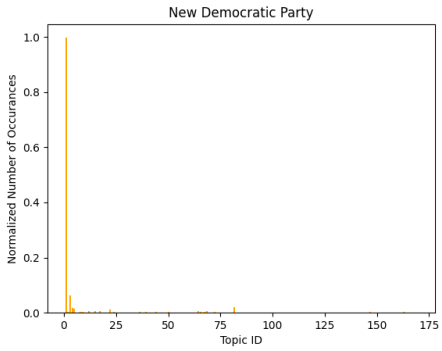


(d) New Democratic Party 2009

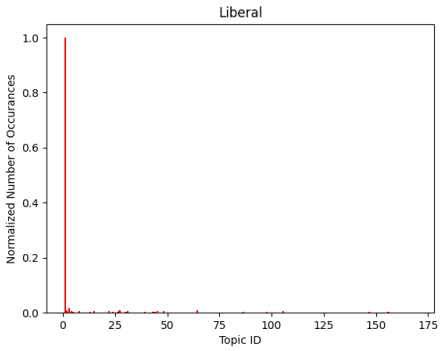


(e) Independent 2009

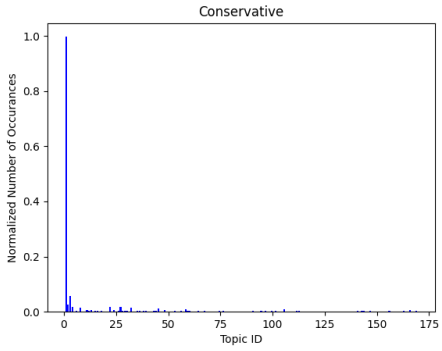
Figure 32: BERTopic Topic Frequencies for Each Party in 2009



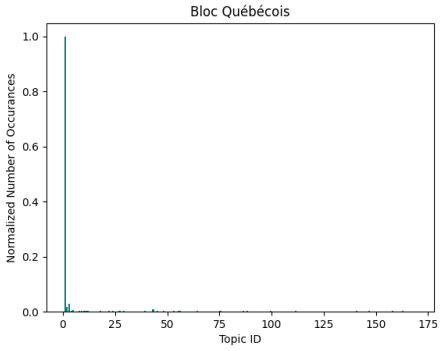
(a) New Democratic Party 2010



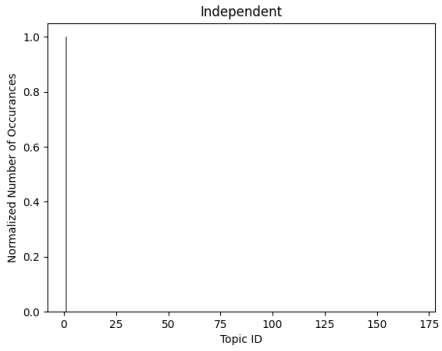
(b) Liberal Party 2010



(c) Conservative Party 2010

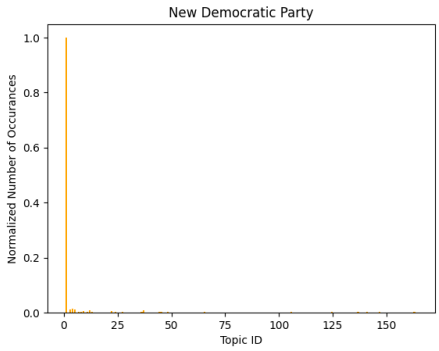


(d) Bloc Québécois 2010

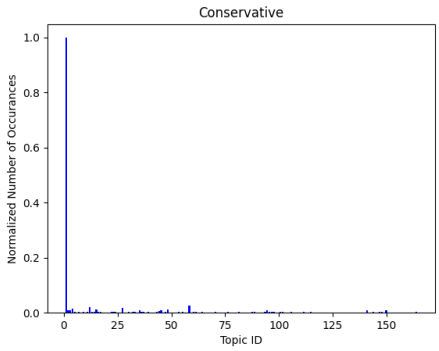


(e) Independent 2010

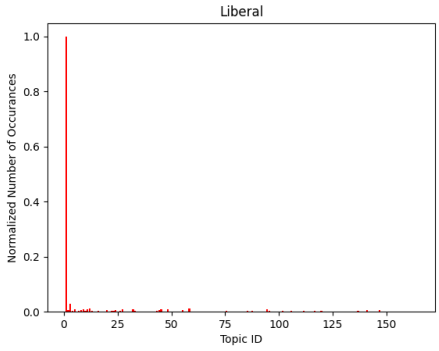
Figure 33: BERTopic Topic Frequencies for Each Party in 2010



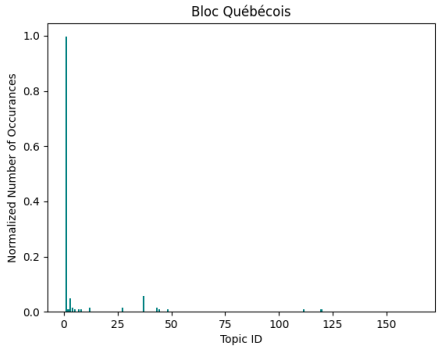
(a) New Democratic Party 2011



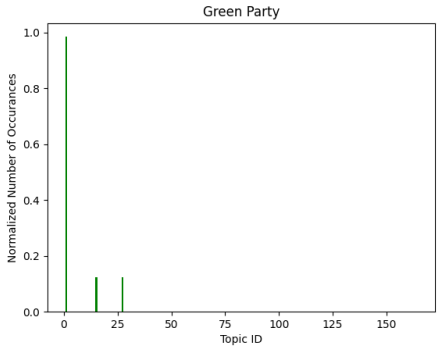
(b) Conservative Party 2011



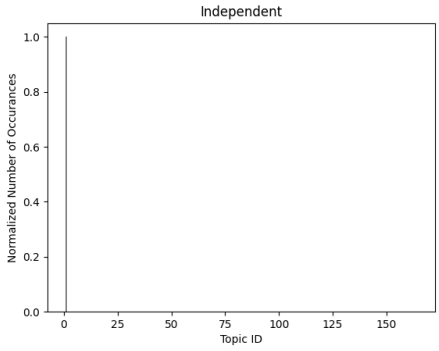
(c) Liberal Party 2011



(d) Bloc Québécois 2011

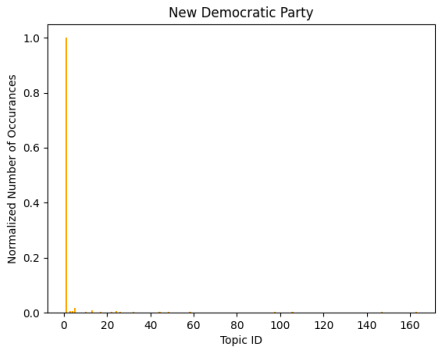


(e) Green Party 2011

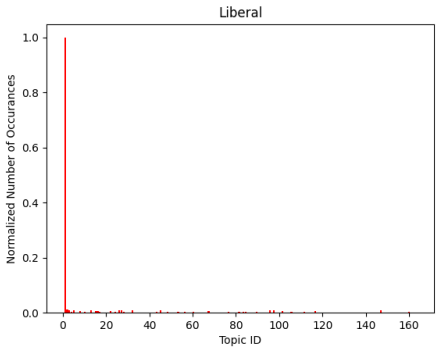


(f) Independent 2011

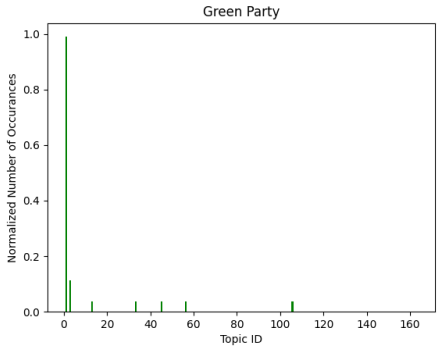
Figure 34: BERTopic Topic Frequencies for Each Party in 2011



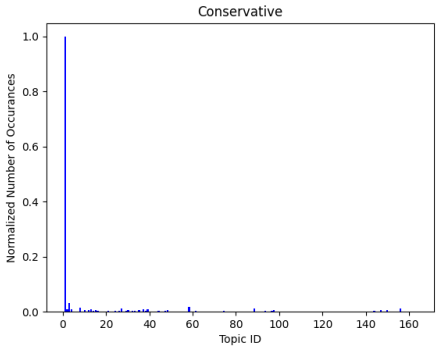
(a) New Democratic Party 2012



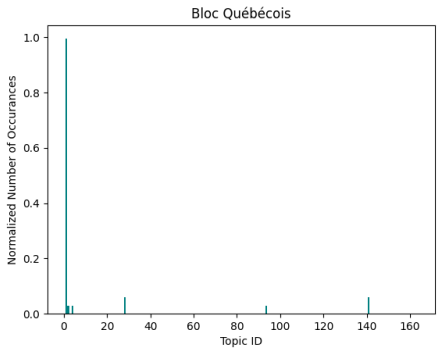
(b) Liberal Party 2012



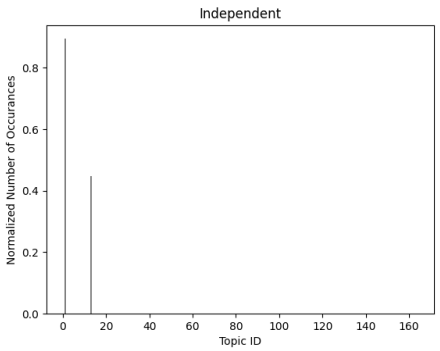
(c) Green Party 2012



(d) Conservative Party 2012

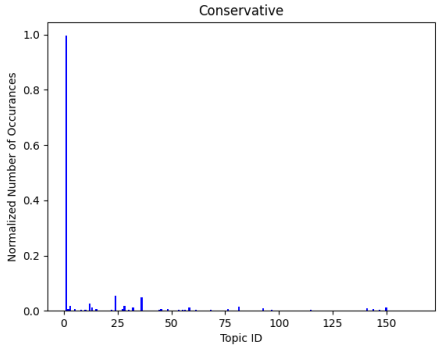


(e) Bloc Québécois 2012

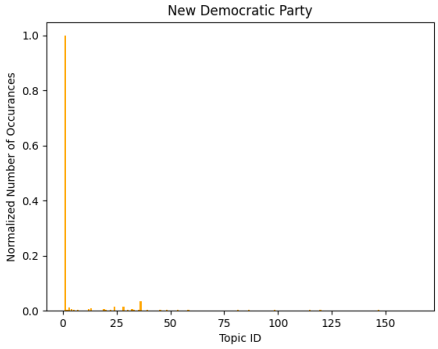


(f) Independent 2012

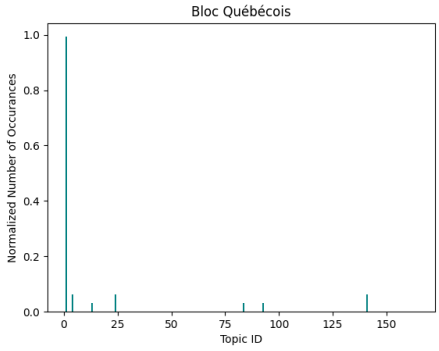
Figure 35: BERTopic Topic Frequencies for Each Party in 2012



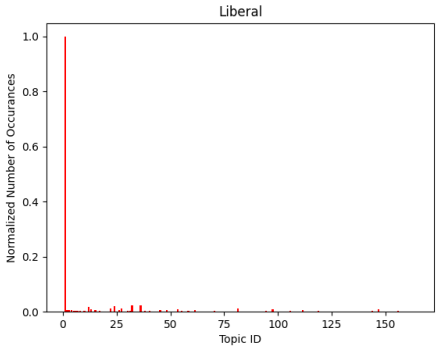
(a) Conservative Party 2013



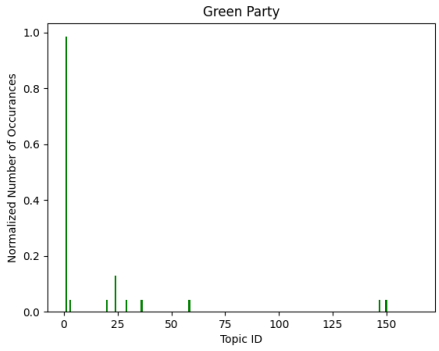
(b) New Democratic Party 2013



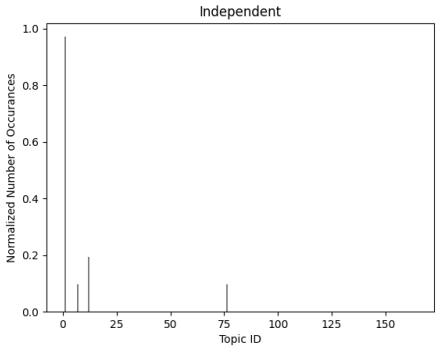
(c) Bloc Québécois 2013



(d) Liberal Party 2013

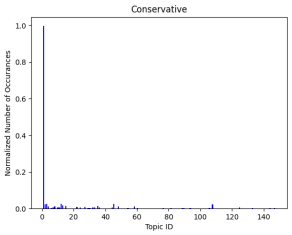


(e) Green Party 2013

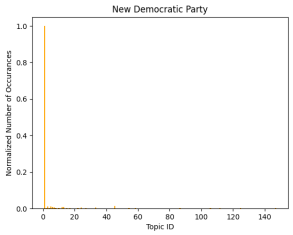


(f) Independent 2013

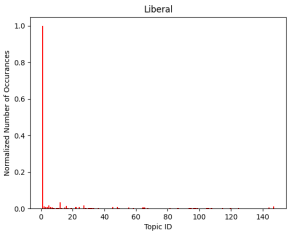
Figure 36: BERTopic Topic Frequencies for Each Party in 2013



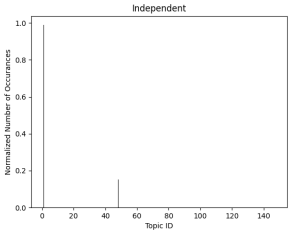
(a) Conservative Party 2014



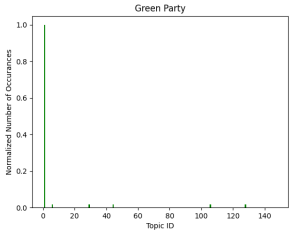
(b) New Democratic Party 2014



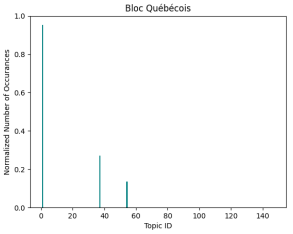
(c) Liberal Party 2014



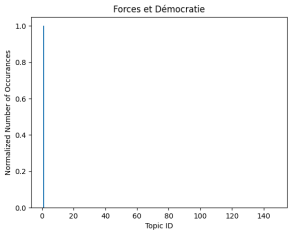
(d) Independent 2014



(e) Green Party 2014

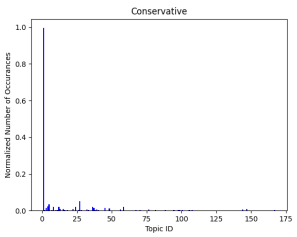


(f) Bloc Québécois 2014

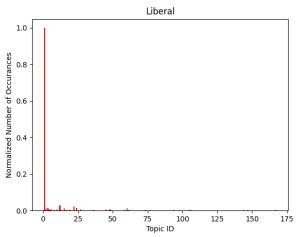


(g) Forces et Démocratie 2014

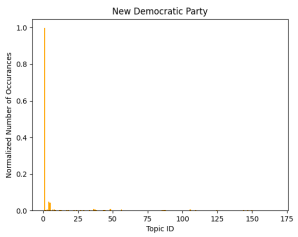
Figure 37: BERTopic Topic Frequencies for Each Party in 2014



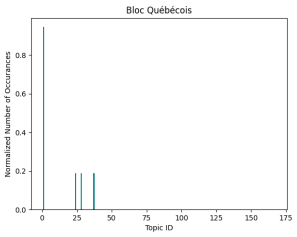
(a) Conservative Party 2015



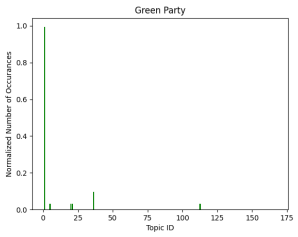
(b) Liberal Party 2015



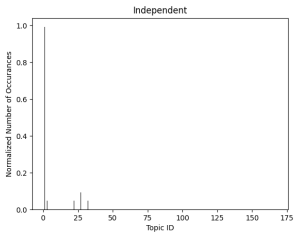
(c) New Democratic Party 2015



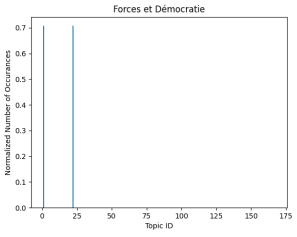
(d) Bloc Québécois 2015



(e) Green Party 2015

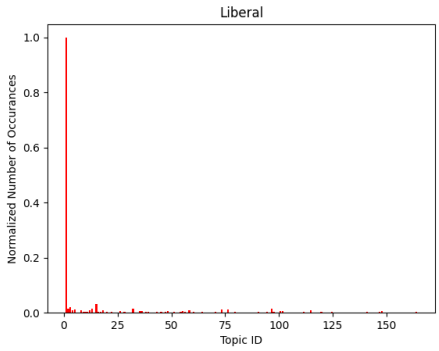


(f) Independent 2015

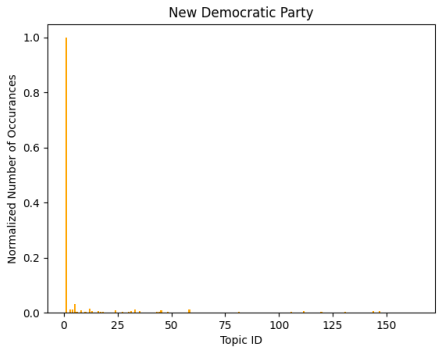


(g) Forces et Démocratie 2015

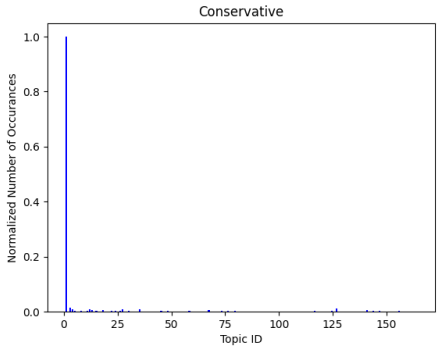
Figure 38: BERTopic Topic Frequencies for Each Party in 2015



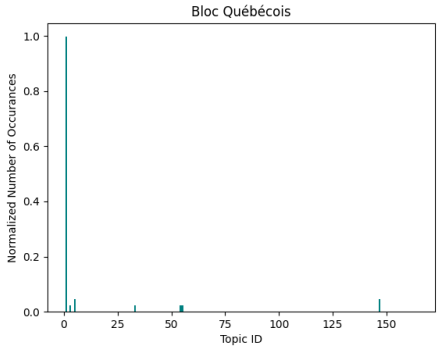
(a) Liberal Party 2016



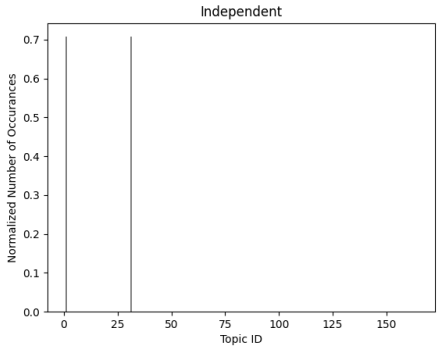
(b) New Democratic Party 2016



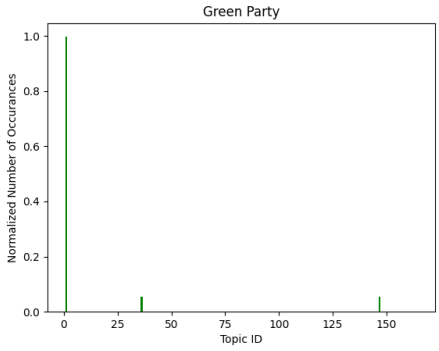
(c) Conservative Party 2016



(d) Bloc Québécois 2016

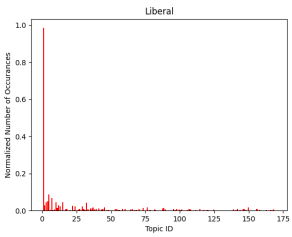


(e) Independent 2016

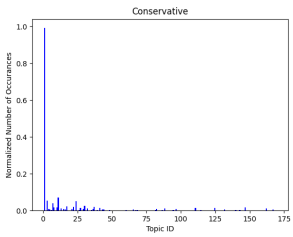


(f) Green Party 2016

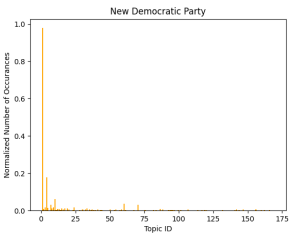
Figure 39: BERTopic Topic Frequencies for Each Party in 2016



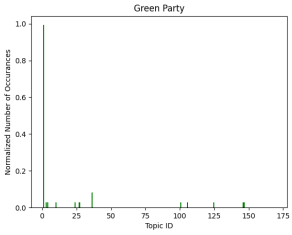
(a) Liberal Party 2018



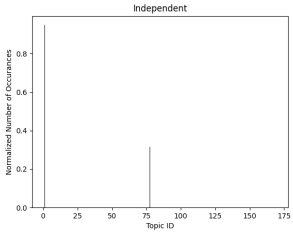
(b) Conservative Party 2018



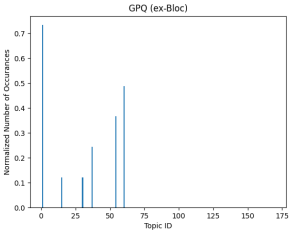
(c) New Democratic Party 2018



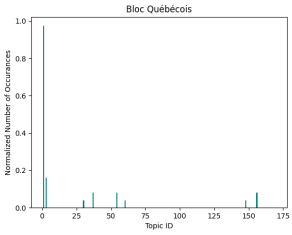
(d) Green Party 2018



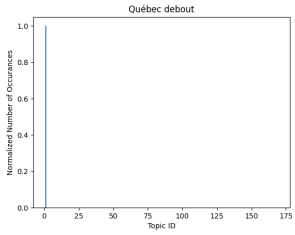
(e) Independent 2018



(f) GPQ (ex-Bloc) 2018

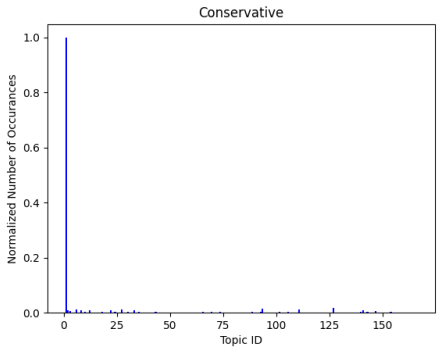


(g) Bloc Québécois 2018

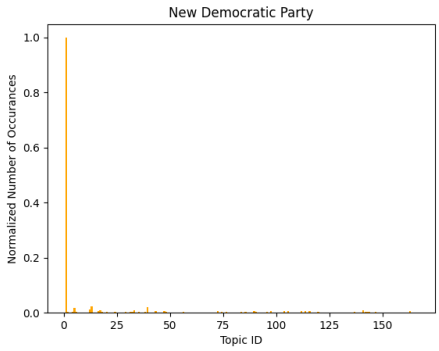


(h) Québec debout 2018

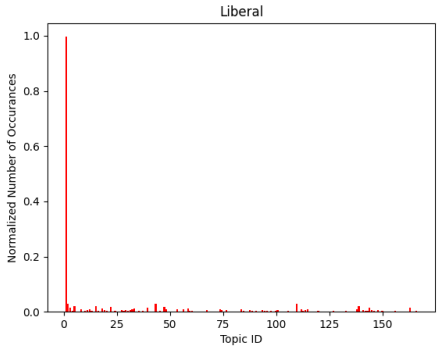
Figure 40: BERTopic Topic Frequencies for Each Party in 2018



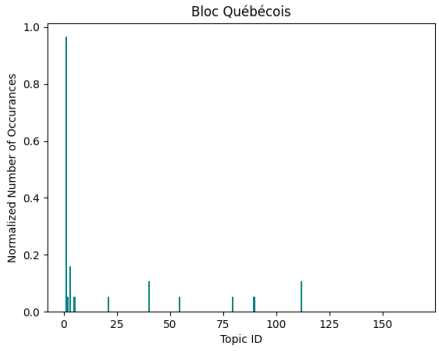
(a) Conservative Party 2019



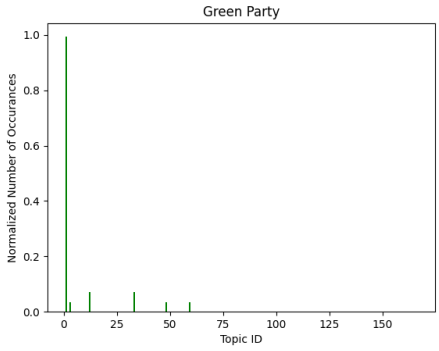
(b) New Democratic Party 2019



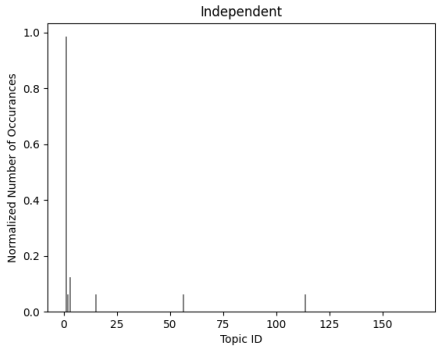
(c) Liberal Party 2019



(d) Bloc Québécois 2019



(e) Green Party 2019



(f) Independent 2019

Figure 41: BERTopic Topic Frequencies for Each Party in 2019

C Appendix C: Party Jensen-Shannon Divergences by Year and Topic Model

C.1 LDA Jensen-Shannon Divergence Tables

Table 5: Pairwise Party JS Divergence of LDA Topics in 2004

	Liberal	Canadian Alliance	Bloc Québécois	NDP	Progressive Conservative	Conservative	Independent
Liberal	0.000000	0.013183	0.013859	0.013276	0.033049	0.012738	0.298366
Canadian Alliance	0.013183	0.000000	0.015426	0.018565	0.033414	0.013294	0.315373
Bloc Québécois	0.013859	0.015426	0.000000	0.016269	0.042070	0.019541	0.302905
NDP	0.013276	0.018565	0.016269	0.000000	0.036355	0.018557	0.299001
Progressive Conservative	0.033049	0.033414	0.042070	0.036355	0.000000	0.033850	0.356785
Conservative	0.012738	0.013294	0.019541	0.018557	0.033850	0.000000	0.309745
Independent	0.298366	0.315373	0.302905	0.299001	0.356785	0.309745	0.000000

Table 6: Pairwise Party JS Divergence of LDA Topics in 2005

	Conservative	Liberal	Bloc Québécois	NDP	Independent
Conservative	0.000000	0.012188	0.012798	0.015207	0.250760
Liberal	0.012188	0.000000	0.019142	0.016343	0.258562
Bloc Québécois	0.012798	0.019142	0.000000	0.015361	0.261856
NDP	0.015207	0.016343	0.015361	0.000000	0.271167
Independent	0.250760	0.258562	0.261856	0.271167	0.000000

Table 7: Pairwise Party JS Divergence of LDA Topics in 2006

	Conservative	Bloc Québécois	NDP	Liberal	Independent
Conservative	0.000000	0.013640	0.010377	0.008740	0.589691
Bloc Québécois	0.013640	0.000000	0.013533	0.012289	0.602358
NDP	0.010377	0.013533	0.000000	0.010294	0.603808
Liberal	0.008740	0.012289	0.010294	0.000000	0.594899
Independent	0.589691	0.602358	0.603808	0.594899	0.000000

Table 8: Pairwise Party JS Divergence of LDA Topics in 2007

	Bloc Québécois	NDP	Conservative	Liberal	Independent
Bloc Québécois	0.000000	0.020015	0.022171	0.017993	0.222653
NDP	0.020015	0.000000	0.011261	0.011367	0.233269
Conservative	0.022171	0.011261	0.000000	0.011384	0.245794
Liberal	0.017993	0.011367	0.011384	0.000000	0.238806
Independent	0.222653	0.233269	0.245794	0.238806	0.000000

Table 9: Pairwise Party JS Divergence of LDA Topics in 2008

	Bloc Québécois	NDP	Conservative	Liberal	Independent
Bloc Québécois	0.000000	0.016833	0.016269	0.016088	0.129629
NDP	0.016833	0.000000	0.010259	0.011083	0.106347
Conservative	0.016269	0.010259	0.000000	0.009590	0.108902
Liberal	0.016088	0.011083	0.009590	0.000000	0.108351
Independent	0.129629	0.106347	0.108902	0.108351	0.000000

Table 10: Pairwise Party JS Divergence of LDA Topics in 2009

	Conservative	Liberal	Bloc Québécois	NDP	Independent
Conservative	0.000000	0.010495	0.013981	0.011698	0.457956
Liberal	0.010495	0.000000	0.010762	0.009527	0.440557
Bloc Québécois	0.013981	0.010762	0.000000	0.012694	0.449611
NDP	0.011698	0.009527	0.012694	0.000000	0.442237
Independent	0.457956	0.440557	0.449611	0.442237	0.000000

Table 11: Pairwise Party JS Divergence of LDA Topics in 2010

	NDP	Liberal	Conservative	Bloc Québécois	Independent
NDP	0.000000	0.014511	0.008808	0.018380	0.520476
Liberal	0.014511	0.000000	0.010116	0.013008	0.529850
Conservative	0.008808	0.010116	0.000000	0.013949	0.520594
Bloc Québécois	0.018380	0.013008	0.013949	0.000000	0.522055
Independent	0.520476	0.529850	0.520594	0.522055	0.000000

Table 12: Pairwise Party JS Divergence of LDA Topics in 2011

	NDP	Conservative	Liberal	Bloc Québécois	Green Party	Independent
NDP	0.000000	0.012085	0.013534	0.032418	0.240859	0.516934
Conservative	0.012085	0.000000	0.016704	0.029713	0.229633	0.472319
Liberal	0.013534	0.016704	0.000000	0.028164	0.227979	0.517668
Bloc Québécois	0.032418	0.029713	0.028164	0.000000	0.231250	0.504053
Green Party	0.240859	0.229633	0.227979	0.231250	0.000000	0.542724
Independent	0.516934	0.472319	0.517668	0.504053	0.542724	0.000000

Table 13: Pairwise Party JS Divergence of LDA Topics in 2012

	NDP	Liberal	Green Party	Conservative	Bloc Québécois	Independent
NDP	0.000000	0.010920	0.070496	0.006763	0.108909	0.270578
Liberal	0.010920	0.000000	0.079385	0.010917	0.104140	0.268855
Green Party	0.070496	0.079385	0.000000	0.070598	0.184837	0.290719
Conservative	0.006763	0.010917	0.070598	0.000000	0.107182	0.263754
Bloc Québécois	0.108909	0.104140	0.184837	0.107182	0.000000	0.347628
Independent	0.270578	0.268855	0.290719	0.263754	0.347628	0.000000

Table 14: Pairwise Party JS Divergence of LDA Topics in 2013

	Conservative	NDP	Bloc Québécois	Liberal	Green Party	Independent
Conservative	0.000000	0.007111	0.065572	0.013260	0.104436	0.165627
NDP	0.007111	0.000000	0.067176	0.013475	0.108752	0.159771
Bloc Québécois	0.065572	0.067176	0.000000	0.082306	0.151672	0.163975
Liberal	0.013260	0.013475	0.082306	0.000000	0.128595	0.168172
Green Party	0.104436	0.108752	0.151672	0.128595	0.000000	0.252221
Independent	0.165627	0.159771	0.163975	0.168172	0.252221	0.000000

Table 15: Pairwise Party JS Divergence of LDA Topics in 2014

	Conservative	NDP	Liberal	Independent	Green Party	Bloc Québécois	Forces et Démocratie
Conservative	0.000000	0.009535	0.008030	0.155654	0.073903	0.212865	0.490380
NDP	0.009535	0.000000	0.014251	0.159417	0.076398	0.218219	0.481450
Liberal	0.008030	0.014251	0.000000	0.159901	0.077345	0.221736	0.482095
Independent	0.155654	0.159417	0.159901	0.000000	0.195735	0.304917	0.564287
Green Party	0.073903	0.076398	0.077345	0.195735	0.000000	0.248792	0.478615
Bloc Québécois	0.212865	0.218219	0.221736	0.304917	0.248792	0.000000	0.461359
Forces et Démocratie	0.490380	0.481450	0.482095	0.564287	0.478615	0.461359	0.000000

Table 16: Pairwise Party JS Divergence of LDA Topics in 2015

	Conservative	Liberal	NDP	Bloc Québécois	Green Party	Independent	Forces et Démocratie
Conservative	0.000000	0.016127	0.017505	0.412450	0.095565	0.099003	0.479582
Liberal	0.016127	0.000000	0.018448	0.370816	0.099392	0.103675	0.438759
NDP	0.017505	0.018448	0.000000	0.405444	0.080859	0.115016	0.444912
Bloc Québécois	0.412450	0.370816	0.405444	0.000000	0.472249	0.398725	0.474710
Green Party	0.095565	0.099392	0.080859	0.472249	0.000000	0.195748	0.515059
Independent	0.099003	0.103675	0.115016	0.398725	0.195748	0.000000	0.501063
Forces et Démocratie	0.479582	0.438759	0.444912	0.474710	0.515059	0.501063	0.000000

Table 17: Pairwise Party JS Divergence of LDA Topics in 2016

	Liberal	NDP	Conservative	Bloc Québécois	Independent	Green Party
Liberal	0.000000	0.012663	0.013306	0.071641	0.562309	0.166019
NDP	0.012663	0.000000	0.011235	0.061377	0.563842	0.161214
Conservative	0.013306	0.011235	0.000000	0.072275	0.595663	0.165442
Bloc Québécois	0.071641	0.061377	0.072275	0.000000	0.562441	0.188761
Independent	0.562309	0.563842	0.595663	0.562441	0.000000	0.616291
Green Party	0.166019	0.161214	0.165442	0.188761	0.616291	0.000000

Table 18: Pairwise Party JS Divergence of LDA Topics in 2018

	Liberal	Conservative	NDP	Green Party	Independent	GPQ (ex-Bloc)	Bloc Québécois	Québec debout
Liberal	0.000000	0.014265	0.017291	0.065025	0.321859	0.226569	0.077214	0.421636
Conservative	0.014265	0.000000	0.016033	0.071533	0.295000	0.236486	0.077462	0.429749
NDP	0.017291	0.016033	0.000000	0.076727	0.308353	0.255887	0.079867	0.441566
Green Party	0.065025	0.071533	0.076727	0.000000	0.286161	0.227641	0.119374	0.474971
Independent	0.321859	0.295000	0.308353	0.286161	0.000000	0.364771	0.357238	0.602807
GPQ (ex-Bloc)	0.226569	0.236486	0.255887	0.227641	0.364771	0.000000	0.260073	0.587794
Bloc Québécois	0.077214	0.077462	0.079867	0.119374	0.357238	0.260073	0.000000	0.451452
Québec debout	0.421636	0.429749	0.441566	0.474971	0.602807	0.587794	0.451452	0.000000

Table 19: Pairwise Party JS Divergence of LDA Topics in 2019

	Conservative	NDP	Liberal	Bloc Québécois	Green Party	Independent
Conservative	0.000000	0.015488	0.013978	0.054402	0.081211	0.166697
NDP	0.015488	0.000000	0.010276	0.066268	0.073987	0.161555
Liberal	0.013978	0.010276	0.000000	0.060790	0.070830	0.159549
Bloc Québécois	0.054402	0.066268	0.060790	0.000000	0.100448	0.172142
Green Party	0.081211	0.073987	0.070830	0.100448	0.000000	0.234444
Independent	0.166697	0.161555	0.159549	0.172142	0.234444	0.000000

C.2 BERTopic Jensen-Shannon Divergence Tables

Table 20: Pairwise Party JS Divergence of BERTopic Topics in 2004

	Liberal	Canadian Alliance	Bloc Québécois	NDP	Progressive Conservative	Conservative	Independent
Liberal	0.000000	0.044062	0.052828	0.055103	0.069745	0.059837	0.063779
Canadian Alliance	0.044062	0.000000	0.030413	0.029889	0.046824	0.041910	0.027160
Bloc Québécois	0.052828	0.030413	0.000000	0.044914	0.038006	0.044220	0.020743
NDP	0.055103	0.029889	0.044914	0.000000	0.048437	0.059593	0.036091
Progressive Conservative	0.069745	0.046824	0.038006	0.048437	0.000000	0.051584	0.026587
Conservative	0.059837	0.041910	0.044220	0.059593	0.051584	0.000000	0.034152
Independent	0.063779	0.027160	0.020743	0.036091	0.026587	0.034152	0.000000

Table 21: Pairwise Party JS Divergence of BERTopic Topics in 2005

	Conservative	Liberal	Bloc Québécois	NDP	Independent
Conservative	0.000000	0.034886	0.035455	0.027326	0.330076
Liberal	0.034886	0.000000	0.041342	0.039170	0.321059
Bloc Québécois	0.035455	0.041342	0.000000	0.029997	0.330804
NDP	0.027326	0.039170	0.029997	0.000000	0.319823
Independent	0.330076	0.321059	0.330804	0.319823	0.000000

Table 22: Pairwise Party JS Divergence of BERTopic Topics in 2006

	Conservative	Bloc Québécois	NDP	Liberal	Independent
Conservative	0.000000	0.032007	0.045527	0.025473	nan
Bloc Québécois	0.032007	0.000000	0.042958	0.034543	nan
NDP	0.045527	0.042958	0.000000	0.031533	nan
Liberal	0.025473	0.034543	0.031533	0.000000	nan
Independent	nan	nan	nan	nan	nan

Table 23: Pairwise Party JS Divergence of BERTopic Topics in 2007

	Bloc Québécois	NDP	Conservative	Liberal	Independent
Bloc Québécois	0.000000	0.031348	0.041475	0.033999	0.035597
NDP	0.031348	0.000000	0.043777	0.034943	0.028559
Conservative	0.041475	0.043777	0.000000	0.027067	0.048838
Liberal	0.033999	0.034943	0.027067	0.000000	0.040707
Independent	0.035597	0.028559	0.048838	0.040707	0.000000

Table 24: Pairwise Party JS Divergence of BERTopic Topics in 2008

	Bloc Québécois	NDP	Conservative	Liberal	Independent
Bloc Québécois	0.000000	0.025566	0.044160	0.031442	0.027244
NDP	0.025566	0.000000	0.037214	0.032977	0.038650
Conservative	0.044160	0.037214	0.000000	0.033029	0.055834
Liberal	0.031442	0.032977	0.033029	0.000000	0.038478
Independent	0.027244	0.038650	0.055834	0.038478	0.000000

Table 25: Pairwise Party JS Divergence of BERTopic Topics in 2009

	Conservative	Liberal	Bloc Québécois	NDP	Independent
Conservative	0.000000	0.033587	0.037516	0.036119	0.070289
Liberal	0.033587	0.000000	0.030474	0.035548	0.053811
Bloc Québécois	0.037516	0.030474	0.000000	0.031911	0.028775
NDP	0.036119	0.035548	0.031911	0.000000	0.044859
Independent	0.070289	0.053811	0.028775	0.044859	0.000000

Table 26: Pairwise Party JS Divergence of BERTopic Topics in 2010

	NDP	Liberal	Conservative	Bloc Québécois	Independent
NDP	0.000000	0.038098	0.058547	0.044738	0.071728
Liberal	0.038098	0.000000	0.044637	0.032363	0.056805
Conservative	0.058547	0.044637	0.000000	0.052512	0.092904
Bloc Québécois	0.044738	0.032363	0.052512	0.000000	0.045591
Independent	0.071728	0.056805	0.092904	0.045591	0.000000

Table 27: Pairwise Party JS Divergence of BERTopic Topics in 2011

	NDP	Conservative	Liberal	Bloc Québécois	Green Party	Independent
NDP	0.000000	0.040288	0.041826	0.056690	0.049224	0.049224
Conservative	0.040288	0.000000	0.042679	0.075795	0.084714	0.084714
Liberal	0.041826	0.042679	0.000000	0.058377	0.065073	0.065073
Bloc Québécois	0.056690	0.075795	0.058377	0.000000	0.075376	0.075376
Green Party	0.049224	0.084714	0.065073	0.075376	0.000000	0.000000
Independent	0.049224	0.084714	0.065073	0.075376	0.000000	0.000000

Table 28: Pairwise Party JS Divergence of BERTopic Topics in 2012

	NDP	Liberal	Green Party	Conservative	Bloc Québécois	Independent
NDP	0.000000	0.031946	0.084304	0.054838	0.074713	0.090278
Liberal	0.031946	0.000000	0.088904	0.053568	0.088700	0.101968
Green Party	0.084304	0.088904	0.000000	0.092524	0.109937	0.088530
Conservative	0.054838	0.053568	0.092524	0.000000	0.106399	0.133122
Bloc Québécois	0.074713	0.088700	0.109937	0.106399	0.000000	0.114518
Independent	0.090278	0.101968	0.088530	0.133122	0.114518	0.000000

Table 29: Pairwise Party JS Divergence of BERTopic Topics in 2013

	Conservative	NDP	Bloc Québécois	Liberal	Green Party	Independent
Conservative	0.000000	0.046407	0.110585	0.043202	0.099573	0.143990
NDP	0.046407	0.000000	0.092235	0.040353	0.112843	0.119599
Bloc Québécois	0.110585	0.092235	0.000000	0.118447	0.151149	0.155884
Liberal	0.043202	0.040353	0.118447	0.000000	0.123293	0.129281
Green Party	0.099573	0.112843	0.151149	0.123293	0.000000	0.202753
Independent	0.143990	0.119599	0.155884	0.129281	0.202753	0.000000

Table 30: Pairwise Party JS Divergence of BERTopic Topics in 2014

	Conservative	NDP	Liberal	Independent	Green Party	Bloc Québécois	Forces et Démocratie
Conservative	0.000000	0.046612	0.048699	0.111463	0.112755	0.243507	0.107365
NDP	0.046612	0.000000	0.039538	0.069654	0.059164	0.209013	0.053832
Liberal	0.048699	0.039538	0.000000	0.099252	0.097797	0.238233	0.091122
Independent	0.111463	0.069654	0.099252	0.000000	0.039402	0.193611	0.023680
Green Party	0.112755	0.059164	0.097797	0.039402	0.000000	0.188924	0.016527
Bloc Québécois	0.243507	0.209013	0.238233	0.193611	0.188924	0.000000	0.178126
Forces et Démocratie	0.107365	0.053832	0.091122	0.023680	0.016527	0.178126	0.000000

Table 31: Pairwise Party JS Divergence of BERTopic Topics in 2015

	Conservative	Liberal	NDP	Bloc Québécois	Green Party	Independent	Forces et Démocratie
Conservative	0.000000	0.061195	0.059597	0.205304	0.116998	0.113321	0.198698
Liberal	0.061195	0.000000	0.063183	0.201179	0.101585	0.091553	0.146049
NDP	0.059597	0.063183	0.000000	0.201916	0.080148	0.125502	0.171514
Bloc Québécois	0.205304	0.201179	0.201916	0.000000	0.217399	0.238796	0.265889
Green Party	0.116998	0.101585	0.080148	0.217399	0.000000	0.140840	0.176605
Independent	0.113321	0.091553	0.125502	0.238796	0.140840	0.000000	0.108509
Forces et Démocratie	0.198698	0.146049	0.171514	0.265889	0.176605	0.108509	0.000000

Table 32: Pairwise Party JS Divergence of BERTopic Topics in 2016

	Liberal	NDP	Conservative	Bloc Québécois	Independent	Green Party
Liberal	0.000000	0.066381	0.050867	0.097207	0.197512	0.114181
NDP	0.066381	0.000000	0.040764	0.059682	0.159583	0.084978
Conservative	0.050867	0.040764	0.000000	0.078137	0.166450	0.086566
Bloc Québécois	0.097207	0.059682	0.078137	0.000000	0.173338	0.070952
Independent	0.197512	0.159583	0.166450	0.173338	0.000000	0.169140
Green Party	0.114181	0.084978	0.086566	0.070952	0.169140	0.000000

Table 33: Pairwise Party JS Divergence of BERTopic Topics in 2018

	Liberal	Conservative	NDP	Green Party	Independent	GPQ (ex-Bloc)	Bloc Québécois	Québec debout
Liberal	0.000000	0.089684	0.104005	0.192120	0.252736	0.404287	0.230782	0.216468
Conservative	0.089684	0.000000	0.120466	0.149905	0.205636	0.400556	0.194124	0.163566
NDP	0.104005	0.120466	0.000000	0.165639	0.214987	0.383550	0.204006	0.174092
Green Party	0.192120	0.149905	0.165639	0.000000	0.139585	0.435767	0.216009	0.088890
Independent	0.252736	0.205636	0.214987	0.139585	0.000000	0.427765	0.196627	0.061555
GPQ (ex-Bloc)	0.404287	0.400556	0.383550	0.435767	0.427765	0.000000	0.255047	0.410219
Bloc Québécois	0.230782	0.194124	0.204006	0.216009	0.196627	0.255047	0.000000	0.153413
Québec debout	0.216468	0.163566	0.174092	0.088890	0.061555	0.410219	0.153413	0.000000

Table 34: Pairwise Party JS Divergence of BERTopic Topics in 2019

	Conservative	NDP	Liberal	Bloc Québécois	Green Party	Independent
Conservative	0.000000	0.084633	0.109335	0.211656	0.109403	0.137587
NDP	0.084633	0.000000	0.092089	0.210928	0.116740	0.154883
Liberal	0.109335	0.092089	0.000000	0.225323	0.158732	0.158175
Bloc Québécois	0.211656	0.210928	0.225323	0.000000	0.213782	0.153196
Green Party	0.109403	0.116740	0.158732	0.213782	0.000000	0.131839
Independent	0.137587	0.154883	0.158175	0.153196	0.131839	0.000000

D Appendix D: Generalized Jensen-Shannon Divergences by Topic Model

Table 35: Generalized Jensen-Shannon Divergence Using LDA Topics 2004-2019

Year	**GJS Div.**
2004	0.027989
2005	0.024593
2006	0.018134
2007	0.026649
2008	0.024047
2009	0.019707
2010	0.021022
2011	0.031438
2012	0.025465
2013	0.027248
2014	0.031065
2015	0.042333
2016	0.029861
2017	0.027091
2018	0.042706
2019	0.037878

Table 36: Generalized Jensen-Shannon Divergence Using BERTopic Topics 2004-2019

Year	**GJS Div.**
2004	0.102015
2005	0.141048
2006	0.056902
2007	0.058936
2008	0.060724
2009	0.063524
2010	0.088381
2011	0.083962
2012	0.109614
2013	0.134849
2014	0.132482
2015	0.192646
2016	0.124999
2017	0.093258
2018	0.278334
2019	0.273140