Investigation on Intra-party Factionalization via Tweets

Etienne Gagnon and **Jason Z.S. Hu** and **Bide Xu**
McGill University
etgagnon@icloud.com
zhong.s.hu@mail.mcgill.ca
bide.xu@mail.mcgill.ca

Abstract

Politicians in the same political party usually share common political preferences. However, these preferences might not be perfectly aligned. Minor disagreements could create factionalization of politicians even within the same party. In this paper, we apply natural language processing techniques to an investigation on the intra-party structure of the Conservative Party of Canada (CPC) based on the tweets created by politicians in the party. We developed a process to cluster the politicians into groups and evaluate the effectiveness of the clustering.

1 Introduction and Related Work

Can social media give us insight into the internal workings of political parties? Recent conceptualizations of political parties have empathized that they consist of broad coalitions of various groups with divergent interests, as opposed to unitary actors with a single focus (McCarty and Schickler, 2018). Unfortunately, this phenomenon of intra-party competition is rarely studied. This is mainly due to the fact that intra-party groups are a characteristic of the party's social network that is hard to observe and capture directly (Dodeigne and Pilet, 2019). Previous attempts at observing intra party politics include Bäck (2008) and Bernauer and Bräuninger (2009), who ask politicians to reveal the degree of competition within their party through a survey. Recent approaches such as Ceron (2015) and Schwarz et al. (2017) analyze the text of speeches given by politicians to infer their ideological positioning compared to other members of the party. In this paper we extend this literature by assessing the potential of social media content posted by politicians as a mean of observing intra-party competition. Social media content may be informative since politicians often make spontaneous comments using it that reveal their true positions (Ceron, 2017).

We investigate intra-party competition in the Conservative Party of Canada (CPC). The CPC is the result of a 2003 merger between the fiscally conservative Progressive-Conservative Party of Canada and the socially conservative Canadian Alliance. As a result, most accounts of its' intra-party politics note that members are divided between a fiscally conservative wing and a socially conservative wing. This dissension has grown especially prevalent following the party's defeat in the 2019 election. Numerous news outlets mention that the social conservative wing is calling for the party chief, Andrew Scheer, to resign (Walsh, 2019; Levitz, 2019; Palkin, 2019). Based on this, we postulate three hypotheses:

H1 We will find at least two distinct clusters within the CPC.

H2 Members of Parliament (MPs) in one of the clusters will have tweets mostly about taxes and the economy, while the MPs of another cluster will have tweets about social-conservative issues (e.g. abortion, same-sex marriage, etc.).

H3 Pairs of members from the same cluster will exhibit more social proximity than pairs of members from different clusters.

In the rest of the paper, we present our methods and experiments to verify these hypotheses.

2 Algorithm

The proposed approach consists of first embedding politicians in a continuous vector space based on their tweets, and then clustering politicians to see if we can detect different groups in the internal party structure. Politician clustering is achieved with a 3-step process:

1. For each tweet in the dataset, each constituent

word is turned into a 50-dimensional GloVe word vector pre-trained on a large corpus of tweets, forming a matrix. GloVe is a type of word embedding model broadly analogous to the famous word2vec model (Pennington et al., 2014) (Section 3.2).

2. For each tweet, we convert the word matrix into a vector to represent the tweet (Section 3.3).
3. All tweet embeddings for a single politician are averaged along each dimension, yielding a politician level representation. This type of averaging is a commonly used technique when creating social media user embeddings (Pan and Ding, 2019) (Section 3.4).

Clustering is done through the k-means algorithm (Arthur and Vassilvitskii, 2007), using the squared embedding distance as a dissimilarity measure. We evaluate the clustering result in Section 4. We discuss some limitations of our approach in Section 5. We briefly mention other trials in Section 6 and conclude in Section 7.

3 Experiment

3.1 Data

We analyze a corpus of all tweets produced by Canadian Conservative MPs between January 1st 2018 and October 21st 2019, on which there was a federal election in Canada[1]. Canadian politicians tweet in both English and French. We only consider tweets in English, to avoid clustering politicians on the simple basis of which languages they use[2]. This has the unfortunate effect of removing 9 of the 12 Conservative MPs from the Province of Québec, who did not have a single tweet in English in the time period. We are left with 89 MPs to analyze. In addition, we filter out retweets so we only consider the original tweets. The resulting corpus is composed of 88561 unique tweets. Figure 1 shows the distribution of the total tweets made over the time period.

3.2 Preprocessing

In step 1, our goal is to convert each tweet into a word matrix, in which each row is a 50-dimensional word vector. For each tweet, we remove stop words, lemmatize each word, and only keep the words corresponding to word vectors in the GloVe model. This last step effectively removes URLs and emojis. This generates a $n \times 50$ matrix for each tweet, where n represents the number of remaining words in the tweet after the preprocessing.

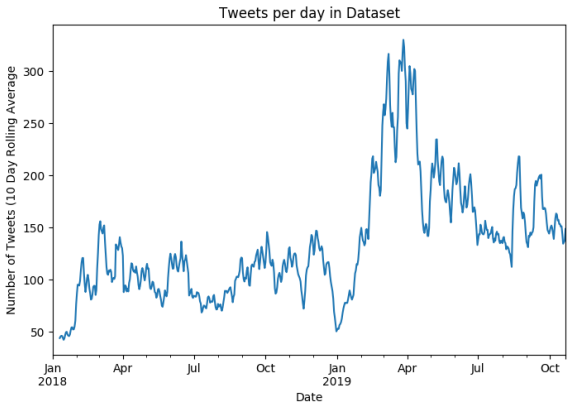


Figure 1: Tweets per day in our dataset

3.3 Tweet Embeddings

In step 2, we convert each matrix of word vectors into a vector representing the tweet. To achieve this, we use PyTorch (Paszke et al., 2019) to train a long short-term memory model (LSTM) (Hochreiter and Schmidhuber, 1997) that takes a sequence of the word embeddings as an input and predicts which politician posted the tweet as an output. We extract the last hidden layer as the embedding vectors of the tweets, which is essentially a set of derived features linking the tweets' content to the politicians. This LSTM model outputs a 64-dimensional vector for each tweet.

3.4 Clustering

In step 3, we aggregate the tweet embedding vectors obtained in Section 3.3 by their authors and average these vectors to generate embedding vectors for Conservative politicians.

Subsequently, we feed the vectors to the k-means algorithm. In order to apply the algorithm, we need to decide the number of clusterings, k. To determine this value, we first apply the elbow method (Thorndike, 1953). We look for a number where the sum of squared distance of the clusters decreases the most. From Figure 2, we can see that when k is 3 or 4, the sum of squared distance has relatively sharp drops.

We further apply the silhouette method (Rousseeuw, 1987) to cross-check

[1] We do not consider newly elected politicians as we consider that they might not have enough time to integrate the parliament's social network.

[2] Machine translating the French tweets was considered but it is commercial and quite costly.

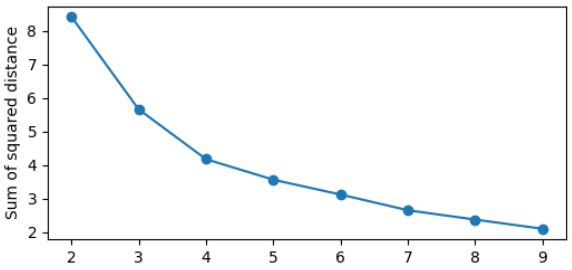
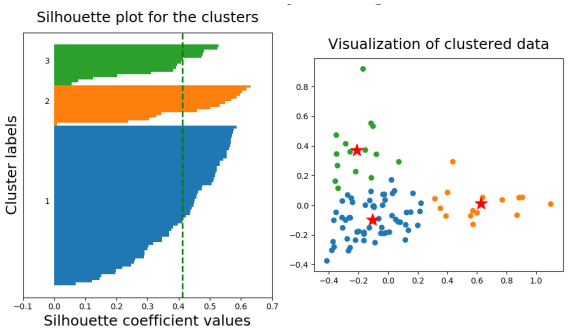


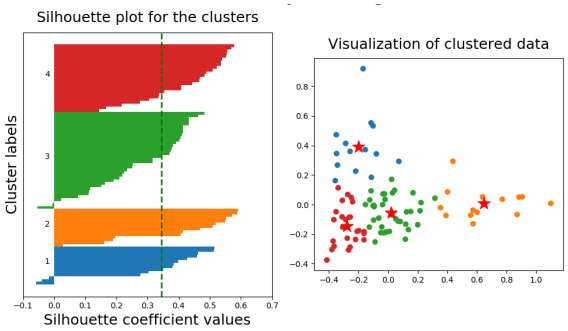
Figure 2: The elbow method for the politician embeddings

this conclusion. Intuitively, the silhouette value of a node compares its average distances from nodes in the same cluster and from nodes in the closest neighbor cluster. The larger the silhouette value is, the closer a node is to the center of the cluster, thus the better. In Figure 3, we show plots for three to five clusters. The plots on the left show the silhouette values of the clusters. The thickness represents the size of each cluster. The green dotted vertical lines are the averages of all silhouette values and we want this value to be large. The plots on the right visualize the clustering of the politician vectors (after reducing the dimension to two via principal component analysis). From the plots, we can see that when $k = 3$, the green line is the largest and there is no politician whose silhouette value is negative (in which case they are closer to the center of another cluster).

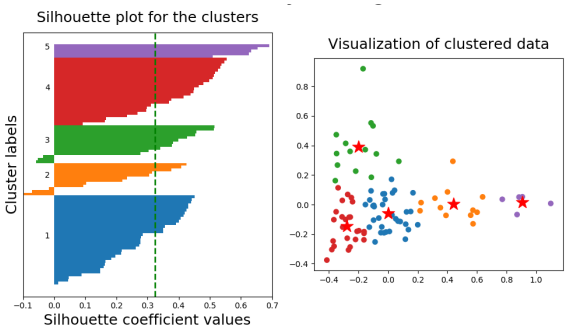
The agreement between both methods implies that the politicians in the CPC can be clustered into three distinct groups, confirming **H1**.



(a) Plots for three clusters



(b) Plots for four clusters



(c) Plots for five clusters

Figure 3: The silhouette method for 3-5 clusters

4 Evaluation

In the previous section, we obtained clusters of politicians. We want to check whether these clusters indeed reflect the relations between politicians, so we examine the clusters using topic modelling (**H2**) and fixed effects least squares regression (**H3**).

4.1 Topic Modelling

H2 states that there are factions in the CPC with different policy focuses. We thus apply topic modelling techniques to study the differences of topics of these groups.

For each cluster of Conservative politicians, we concatenate their tweets by date, creating a document composed of all tweets posted on the same day. This type of aggregation has been shown to give better results for topic analysis of social media data (Hong and Davidson, 2010).

We considered latent dirichlet allocation (LDA) (Blei et al., 2003) and non-negative matrix factorization (NMF) (Lee and Seung, 2000) for topic modelling. Since the corpus size is relatively small for each cluster after concatenation, around 700 (notice that the time span is less than two years), we empirically realized that NMF produces more meaningful results for this corpus size. We also found that setting number of topics to 40 gives the best defined topics.

Politicians tweet about various topics, which are often not of political nature. Since our analysis is only interested in political topics, we follow Barberá et al. (2019) and qualitatively identify the topics that are political in nature. The detected topics

Group 1		Group 2		Group 3	
Topic id	Topic Label	Topic id	Topic Label	Topic id	Topic Label
1	Snc-Lavalin Scandal	5	Energy Sector	7	Snc-Lavalin Scandal
2	Tax Cuts	11	Veteran Support	8	Budget
5	Appeal to Vote	15	Budget	11	Crime
15	Trade	17	Peace-Keeping Operations	14	Trudeau Judicial Interference
20	Snc-Lavalin Scandal	18	Transmountain Pipeline	17	Gun Control
21	Budget	20	Mental Health	23	Familial Tax Credit
28	Gun Control	22	Subsidy for School Material	32	Legal Injection Sites
37	Taxes	23	Trudeau Judicial Interference	33	Illegal immigration
39	Job Protection	24	Trade		
		27	Crime		
		30	Catholic Voters		
		32	Russia Criticism		
		33	Terrorism		

Table 1: Political topics detected per group

	Mentions		Retweets		Sentiment	
	k=3	k=4	k=3	k=4	k=3	k=4
Same Group Pair	−0.077	0.011	0.641	0.929	0.011	0.012
Standard Error	(0.096)	(0.085)	(0.604)	(0.532)	(0.007)	(0.006)
Fixed Effects	yes	yes	yes	yes	yes	yes
Observations	8,372	8,372	8,372	8,372	8,372	8,372
Note:						*p<0.05; **p<0.01

Table 2: Validation performance of the model

for each cluster are reported in Table 1. Though it is also reasonable to cluster the politicians into four groups, due to space limitation, we only show and analyze the topics of three.

In Group 1, which contains the party leader Andrew Scheer, most topics are related to the economy: taxes, trade, the budget and job protection. Another example of a MP in Group 1 is Scot Davidson, holder of a bachelor's degree in economics and a business owner from Ontario. This is congruent with our expectation that one group will mainly focus on the economy. Group 2 and 3 show a broader variety of topics. In Group 2 we find a mixture of social, international relations and economic issues. An example MP from this group is Stephanie Kusie, a former diplomat from Alberta who has a degree in Political Science. Group 3's topics are focused on social issues, such as legal injection sites and illegal immigration, as well as criticism of Prime Minister Trudeau's administration. Larry Miller is classified in this group, a career politician and farm owner from Ontario who made controversial comments regarding niqab wearing women (O'Malley, 2015).

Overall, we find a clear separation of topics between a group focusing on the economy and the rest of the MPs. That Andrew Scheer is in Group 1 is consistent with accounts stating that he is being opposed by the socially conservative wing of the party. The question of whether Group 2 and 3 are two parts of the socially conservative wing or whether they hold no relationship is difficult to answer on the basis of our evidence. In brief, **H2** is generally confirmed.

4.2 Fixed Effects Least Squares Regression

The clusters we detect are based on the content produced. For the detected clusters to represent some real-life association, they should be predictive of some social proximity between the MPs in order to verify **H3**. Due to the lack of real-life data, we make the assumption that people's online social network is a faithful approximation of their actual social interactions. We use the number of direct interactions on twitter, meaning retweets and mentions, between the MPs as a measure of social proximity. We look at whether being part of

the same cluster predicts a higher number of interactions between the MPs.

We estimate this effect with least squares regression. Since twitter interactions are typically a unilateral action (Retweeting does not require the author's consent), we use directed MP pairs as the unit of analysis. The dependent variable is the number of interactions of one type going from MP_1 to MP_2, regressed on a binary variable indicating if they belong to the same group. Because different users have different levels of twitter usage, we use fixed effects (Gelman and Hill, 2006), an econometric technique that accounts for variation caused by the observed unit's identity, on both MPs in a pair to account for this effect[3]. We in addition apply VADER sentiment analysis (Hutto and Gilbert, 2014) to the text of tweets that mention other MPs to see if being in the same group predicts more positive language. The sentiment measure ranges from -1 to 1, and higher values indicate more positive sentiment.

Table 2 shows results for these measures when clustering with either 3 or 4 groups. If we cluster the MPs in 3 groups, being in the same group predicts 0.08 less mentions between users, 0.64 more retweets and a sentiment score increase of 0.01. For 4 groups, we find 0.01 more mentions, 0.93 more retweets and a 0.01 increase in sentiment. Since these biases are in small scales, none of these effects is statistically significant at the 95% confidence level. Being clustered in the same group does not imply more interactions, which thus leaves **H3** unverified.

5 Limitations

We acknowledge certain limitations. First, rejecting French tweets means that we did not capture all the content produced by the MPs, and that many MPs from Québec were not included in the study. Second, though we presented evidence that Conservative MPs produce content that is consistent with the typical accounts of their intra-party politics, We did not find causal evidence with regards to what is behind different patterns in their production of content, or with regards to what the groupings that we detected imply for Canadian politics. Third, we did not manage to verify **H3**. Our speculation of the reason is that retweets and mentions do not capture the intentions of the interactions, which are especially important in measuring politicians' intra-party interactions. It would be interesting to gather other forms of interactions between Conservative MPs and see if a larger dataset helps to overcome our limitations.

[3]Concretely this means that a positive relationship will be detected if the level of interactions is higher than the mean level of interactions for the MPs, see Angrist and Pischke (2009).

6 Other Trials

In this paper, we focus our discussion on the combination which we found works the best. In this section, we briefly mention other approaches we attempted.

1. Different from Section 3.3, we also tried a simpler tweet embedding, which is computed by the weighted average of word vectors over a normalized TF-IDF vector of the tweet (Salton and Buckley, 1988).
2. As mentioned in Section 4.1, we tried LDA for topic modelling.
3. We also tried different aggregation strategies of tweets in topic modelling, including no aggregation and aggregation by author.

7 Conclusion

In this paper, we presented our investigation on the intra-party structure of the CPC based on tweets from the MPs in the party. We generated a vector in continuous space for each tweet via a LSTM model and aggregated the vectors to obtain vectors to represent the MPs. We applied the k-means algorithm to cluster the MPs into three groups. We evaluated the clusters via topic modelling and Least-Squares regression and found mixed evidence for the ability of social media to give insight into the internal workings of political parties. On one hand, we found fairly well separated clusters of the MPs based on the content they produce, and found that politicians in the different clusters focus on different policy issues, which is consistent with qualitative descriptions of the CPC. On the other hand, we did not discover indications of higher social dynamics within the same group.

In this paper, Gagnon contributed to the general design of the paper, the LSTM model, the least-squares validation, the analysis of the results and the writing. Hu contributed to the writing, embedding generations, k-means clustering and topic modelling. Xu contributed to the writing and k-means clustering.

References

Joshua D. Angrist and Jorn-Steffen Pischke. 2009. *Mostly Harmless Econometrics*, princeton university press edition.

David Arthur and Sergei Vassilvitskii. 2007. K-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '07, pages 1027–1035, Philadelphia, PA, USA. Society for Industrial and Applied Mathematics.

Pablo Barberá, Andreu Casas, Jonathan Nagler, Patrick J. Egan, Richard Bonneau, John T. Jost, and Joshua A. Tucker. 2019. Who Leads? Who Follows? Measuring Issue Attention and Agenda Setting by Legislators and the Mass Public Using Social Media Data. *American Political Science Review*, 113(4):883–901.

Julian Bernauer and Thomas Bräuninger. 2009. Intra-Party Preference Heterogeneity and Faction Membership in the 15th German Bundestag: A Computational Text Analysis of Parliamentary Speeches. *German Politics*, 18(3):385–402.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Hanna Bäck. 2008. Intra-Party Politics and Coalition Formation: Evidence from Swedish Local Government. *Party Politics*, 14(1):71–89.

Andrea Ceron. 2015. Brave rebels stay home: Assessing the effect of intra-party ideological heterogeneity and party whip on roll-call votes. *Party Politics*, 21(2):246–258.

Andrea Ceron. 2017. Intra-party politics in 140 characters. *Party Politics*, 23(1):7–17.

Jérémy Dodeigne and Jean-Benoit Pilet. 2019. Centralized or decentralized personalization? Measuring intra-party competition in open and flexible list PR systems. *Party Politics*.

Andrew Gelman and Jennifer Hill. 2006. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Analytical Methods for Social Research. Cambridge University Press.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Liangjie Hong and Brian D. Davidson. 2010. Empirical study of topic modeling in Twitter. In *Proceedings of the First Workshop on Social Media Analytics*, Washington D.C.

C. Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text.

Daniel D. Lee and H. Sebastian Seung. 2000. Algorithms for non-negative matrix factorization. In *Proceedings of the 13th International Conference on Neural Information Processing Systems*, NIPS'00, pages 535–541, Cambridge, MA, USA. MIT Press.

Stephanie Levitz. 2019. 'You can have a personal view' on issues like abortion, same-sex marriage and still be PM, says Scheer | CBC News.

Nolan McCarty and Eric Schickler. 2018. On the Theory of Parties. *Annual Review of Political Science*, 21(1):175–193.

Kady O'Malley. 2015. Larry Miller, Conservative MP, recants inflammatory niqab-ban comment.

Steve Palkin. 2019. Do social conservatives still have a place in Canadian politics?

Shimei Pan and Tao Ding. 2019. Social media-based user embedding: A literature review. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, IJCAI'19, pages 6318–6324. AAAI Press.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, 12:2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Peter J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53 – 65.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage.*, 24(5):513–523.

Daniel Schwarz, Denise Traber, and Kenneth Benoit. 2017. Estimating Intra-Party Preferences: Comparing Speeches to Votes. *Political Science Research and Methods*, 5(2):379–396.

Robert L Thorndike. 1953. Who belongs in the family. In *Psychometrika*, volume 18, pages 267–276.

Marieke Walsh. 2019. Social conservative groups call for Andrew Scheer to resign.