Classification and influence of Twitter bots in 2019 Canadian federal elections at British Columbia.

Ana Cecilia Leon Morales

This work was developed at The University of British Columbia under the supervision of professor Ruben H. Zamar.

Summary

Social bots can affect online communication among humans. Previous research suggests that bots are present on political discussions and social movements [4] [6] [2]; a qualitative study of bots in Canadian politics found that these bots appeared to have a limited influence on the 2015 Canada federal election. We aim to analyze the presence of Twitter bots, their classification, and influence on the political discussions about the 2019 Canadian federal elections at British Columbia. To accomplish this goal, we collected data from three main topics which are: elections, political parties, and party leaders from June 21st to October 21st of 2019 by accessing the Standard Twitter API, and we performed two clustering analyses according to the characteristics of suspicious accounts and the content of their issued tweets.

1 Motivation

The popularity of Twitter, which positions it within the top 20 most popular social networks worldwide as of October 2019 with 330 million monthly active users [1], has attracted many automated programs which are known as "Bots"; these programs have appeared to be a double-edged sword for the social network.

Bots can autonomously carry out actions such as tweet or re-tweet, as well as follow, un-follow, or directly message other Twitter accounts. Beneficial uses of bots include the transmission of useful information such as global news, the automatic generation of interesting or creative content, and the automatic response to Twitter users through direct messages. Conversely, malign bots spread spam or malicious content and distort the shared information.

As Twitter became an important tool for protests, political conversations, and people mobilizations, the damage to shared information in this social media increased significantly. In September 2017, Twitter's executives informed the Intelligence Committee of the United States that they have found evidence that signaled Russia was behind some fake and automatized accounts that were active during the 2016 presidential election. Studies [5] have found that about 20% of the activity on Twitter related to such elections came from bots suspicious accounts.

In contrast to the United States, it appears that the influence of Twitter bots on Canada is limited. A study from 2014 [5] found that just under half of Canadians had visited a government website. Even

fewer had friended or followed a political actor on Facebook (6 percent) or Twitter (4 percent). In this study, it appears that not only do Canadians avoid politicians online, but they also avoid politics of all sorts since only 18 percent of Canadians had signed a petition, posted a political message on Facebook (14 percent) or had retweeted political content (3 percent).

This previous qualitative research [5] on the usage of political bots states that Twitter bots had a limited impact in the social media discourse during 2015 Canada federal election; it identifies four main types of bots in the Canadian political ecosystem which are presented in the following scheme.

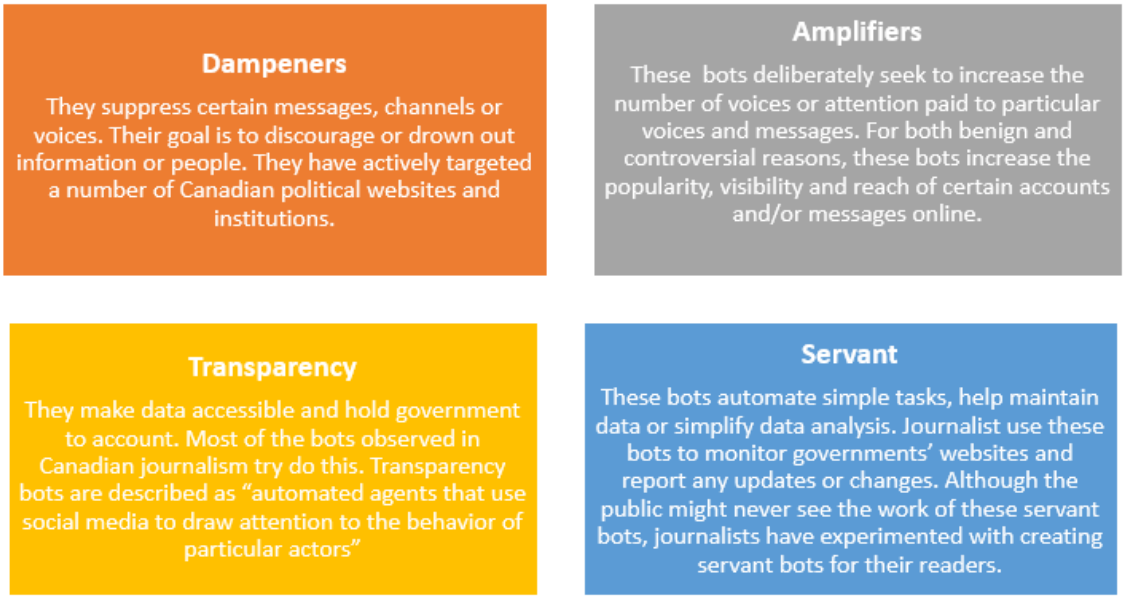

Figure 1: Identified types of bots that were present in political discussions about the 2015 Canada federal elections according to McKelvey and Dubois [5].

A more detailed description of each type of bot, as well as lists of suspicious accounts that were identified through this study are provided in [5]. In what follows we present some examples of these types of bots that were presented in [5].

- **Dampener.** An example of these bots behavior is told by a professor of law at The University of British Columbia. His testimonial at [5] tells that after issuing a tweet about the trending #GoodRiddanceHarper, which celebrated the resignation of ex-Prime Minister Stephen Harper, he received a negative reply that quickly received over 1,000 likes in few hours. He found out that bots amplified this negative reply to discourage him from tweeting.

- **Amplifier.** According to [5], the account *hashtag_cdnpoli* is an example of an amplifier bot. This account retweets Canadian politics news with the #cndpoli hashtag, and it issues tweets more than 10 times per day. Another example of these bots was observed during the 2017 provincial elections in British Columbia, when a suspicious amplifier account, *ReverendSM*, issued content mainly about the incumbent Christy Clark of the Liberal Party with accusations of corruption.

- **Transparency.** The automated account *gccaedits* is a known example of this type of bot. This account issues a tweet every time an anonymous edit to Wikipedia is made from a Government of Canada IP address. This type of bots was scarcely found in [5].

- **Servant.** These bots are different from the others in the sense that they are referred to automated programs that perform tasks beyond the publishing of information through social media. For example, Kathleen Wynne, who was the Communications Office of Ontario Premier until 2018, managed her Facebook page so that it automatically removed posts that contained any word from a list of banned words. Another application of this kind of bots out of the political context in Canada is by providing assistance to manage the problem of online child exploitation by identifying suspicious images on public websites.

2 Objectives

We aim to analyze the presence of Twitter bots and the information they disseminated at British Columbia during the 43rd Canadian general election through the following goals:

- To describe the time evolution of tweets with political discourse in the months leading to the 2019 Canadian federal elections.

- To study the presence of bots, classify, and characterize them according to their account characteristics.

- To analyze the types of tweets that are issued by each type of bot according to their content, and investigate if bots can be classified under the classes: 'Dampeners', 'Amplifiers' and 'Transparency', which were proposed in the qualitative research of 2015 Canada federal elections [5].

3 Data description

From June 21st to October 21st of 2019, we collected a sample of tweets related to the Canadian federal election which were issued by Twitter accounts registered in British Columbia and surrounding areas according to the information provided by users. Data were accessed by consulting the Standard Twitter API through three queries that contain terms related to three main topics: the Canadian federal elections, political parties in the House of Commons of Canada, and leaders of these parties.

In addition to the information related to each tweet, the probability of Twitter accounts being a bot was estimated by Botometer, which is the most popular random forest classifier of Twitter accounts to detect Bots based on hundreds of features. These features are categorized into six main classes which are related to networks of the account, user, friends, temporal patterns in issued tweets, their content and sentiment.

4 Exploratory data analysis

We present an exploratory analysis of the previously described data.

The number of tweets that were collected from 113,960 different Twitter accounts through our three queries is as follows:

Query	Numer of tweets
Elections	247, 116
Parties	196,774
Leaders	663,610

Table 1: Number of total collected tweets by query.

4.1 Historic trends

4.1.1 Elections

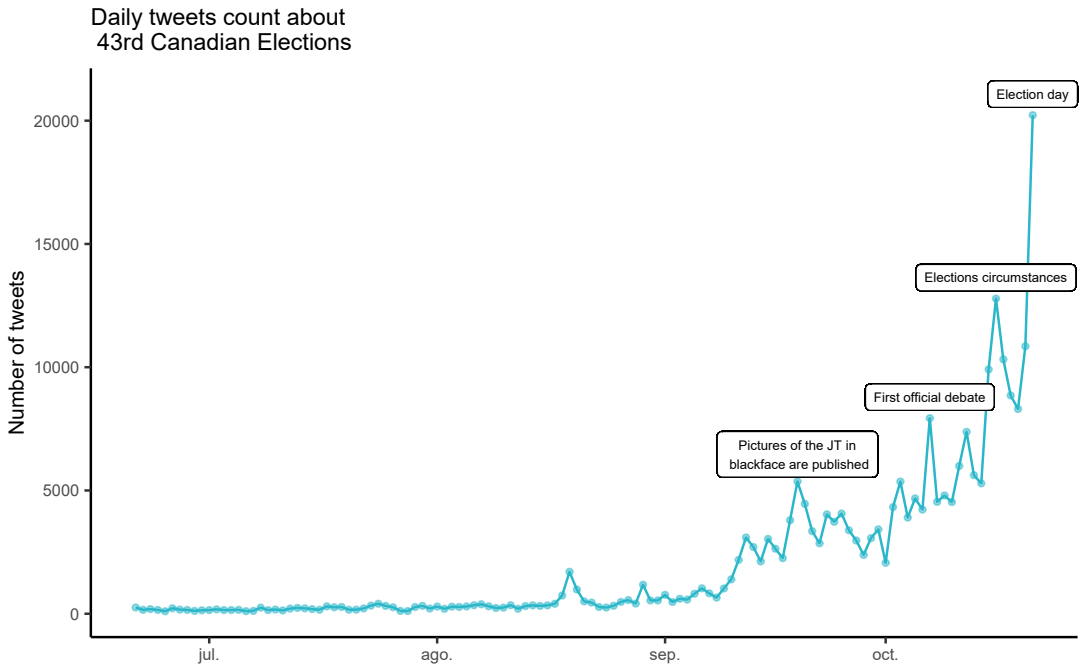

Figure 2: Trend of the number of tweets in British Columbia about the 2019 Canadian federal elections that were sampled.

A total of 247,116 tweets with content related to the 2019 Canadian federal elections in British Columbia (BC) was sampled, this quantity represents 22.31 % of the total number of collected tweets. As is noted in Figure 2, the number of tweets has an increasing trend. The number of tweets

starts increasing at the beginning of September and reaches its maximum of 20,225 on the election day; during this period, peaks match remarkable situations: on September 19th, controversial pictures of Prime Minister Justin Trudeau in blackface were published, while on October 7th the first official debate was held.

4.1.2 Parties

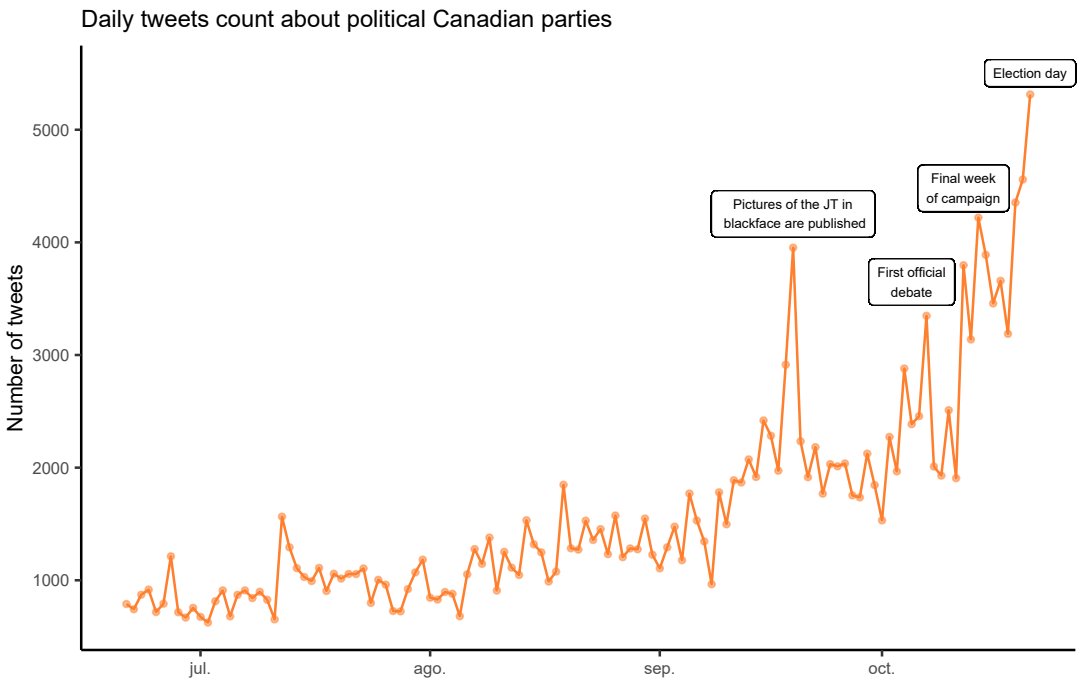

Figure 3: Trend of the number of tweets in British Columbia about the political Canadian parties that were sampled.

The "Canadian parties" topic corresponds to the less popular with a total of 196,774 tweets of related content, which represents 17.78 % of the total number of sampled tweets in BC. Similarly to the tweets about elections, according to Figure 3, the number of tweets about political parties has an increasing trend that reaches its maximum on the election day with 5,313 tweets, peaks of this trend match the same remarkable situations of the tweets about elections such as the controversial pictures of Justin Trudeau and the first official debate, as well as the final week of the campaign.

4.1.3 Leaders

The largest quantity of collected tweets corresponds to content related to the six leaders of each party at the House of Commons on 2019, which are Justin Trudeau from liberals, Andrew Scheer from conservatives, Jagmeet Singh from New Democratic Party, Yves-François Blanchet from Bloc Québécois, Elizabeth May from Green Party, and Maxime Bernier from People's party.

As it is noted in Figure 4, the number of collected tweets has an increasing trend, where Justin Trudeau is the most popular leader followed by Andrew Scheer, with an exception on October 3rd and 4th when the number of tweets about Andrew Scheer is bigger than Justin Trudeau; these dates match with the first French-language debate in which Scheer was questioned about his personal views on abortion rights and his party's environmental policies. The third place in popularity is interchanged among Elizabeth May, Maxime Bernier, and Jagmeet Singh, while the leader of Bloc Québécois was the less mentioned in tweet contents issued by users registered at British Columbia.

Similarly to tweets about elections and parties, there are some peaks that match dates corresponding to the first official debate and the day when pictures of Justin Trudeau were published, which is the day when this leader presents its maximum number of tweets with a total of 15,892.

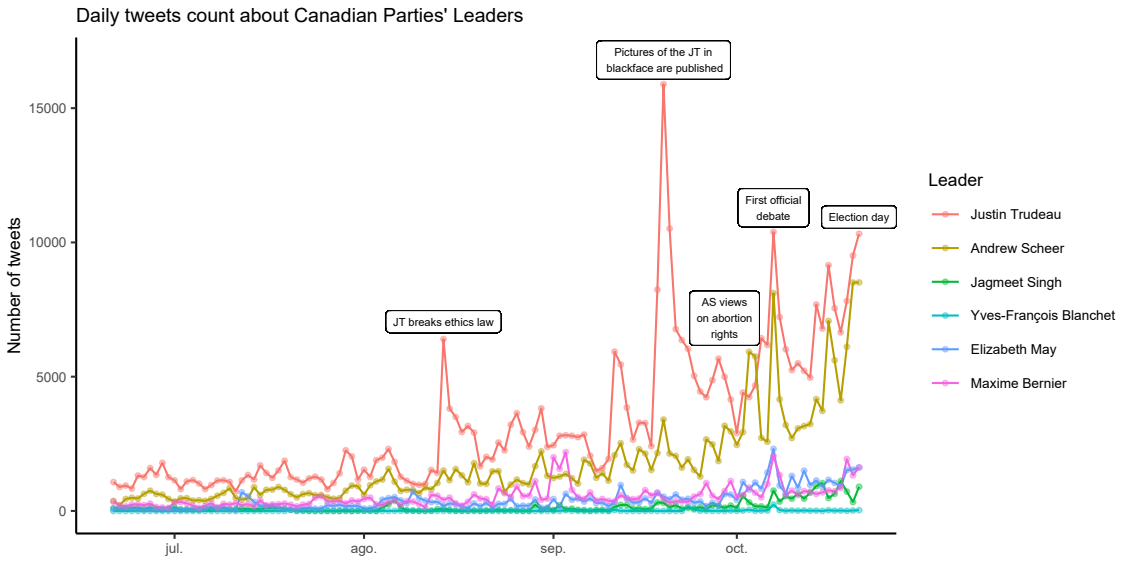

Figure 4: Trend of the number of sampled tweets in British Columbia with content about the leaders of political Canadian parties.

4.1.4 Twitter accounts

The number of different Twitter accounts, which issued content about one or more of the considered topics, has an increasing trend as it is shown in Figure 5. Data about Twitter accounts were collected on a weekly basis according to the norm ISO 8601 where each week starts on Mondays and finishes on Sundays. It is worth to remark that the week 2019-43 only contains information about the 21st October, which is the day of the election and that has a total of 31,251 accounts, while the number of accounts of previous weeks corresponds to the cumulative counts of the whole week.

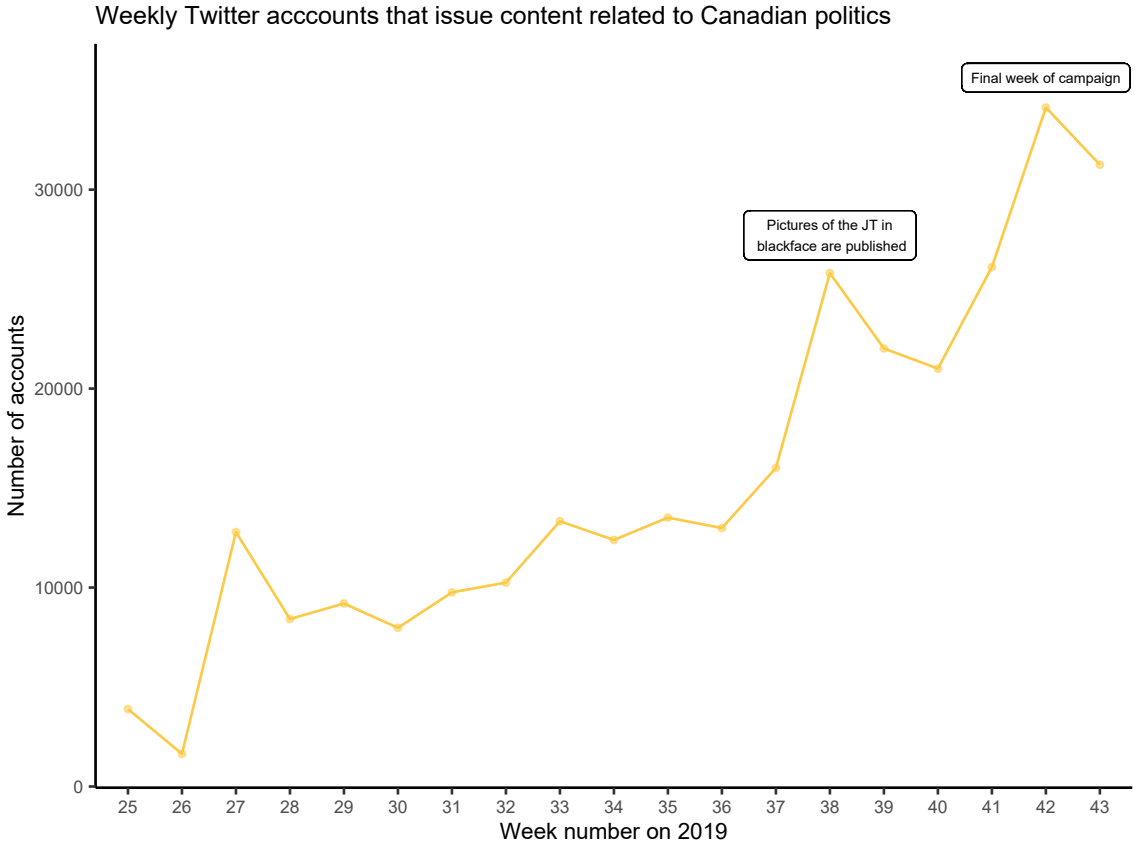

Figure 5: Trend of the number of users in British Columbia that issued content about 2019 Canadian elections, parties and leaders.

4.2 Presence of bots by query

The probabilities of Twitter accounts to be a bot were estimated at the end of each week through the public API of Botometer (`https://botometer.iuni.iu.edu/#!/api`), which is the most well known publicly available classier for bot detection, using the packages `httr`, `xml2`, `RJSONIO`, `mongolite`, `jsonlite`, and taking the function `botcheck()` provided by Joey Marshal (`https://github.com/marsha5813/Botcheck`) as an initial reference to create our own code.

Botometer [3] is a machine learning classifier that is the result of a joint project of the Network Science Institute and the Center for Complex Networks and Systems Research at Indiana University. This classifier was trained on more than 5.6 million tweets to classify Twitter accounts according to their estimated probability of being a bot based on approximately 1,200 features that feed seven different classifiers. These classifiers correspond to the overall classifier and six classifiers for each subclass of features, which are categorized as follows:

- **Network**. It is comprised of features that capture information diffusion patterns. It consists of statistical features like degree distribution and centrality measures of various networks based on retweets, mentions, hashtags, etc.
- **User**. This set of features consists of data related to the Twitter account including language, geographic locations, creation time, etc.
- **Friends** It is comprised of features that are built from descriptive statistics relative to social contacts of the account, such as the median, moments, and entropy of the distributions of their number of followers, friends, and so on.
- **Temporal** This set of features captures timing patterns of content generation and consumption, such as the inter-tweet time distribution.
- **Content** It is comprised of features that are based on linguistic cues computed through natural language processing, especially part-of-speech tagging.
- **Sentiment** This set of features is built using general purpose and Twitter specific sentiment analysis algorithms including algorithms that consider happiness scores, emoticon scores, the three dimensions that describe emotional states according to a psychological model, which are called arousal (how pleasant emotion is), dominance (what is the dominant nature of the emotion) and valence (the intensity of emotion), etc.

The reason to estimate the probabilities at the end of each week is because of the classifier extracts features from both the analyzed Twitter account and from up to the last 200 issued tweets, so probabilities change along time. As Figure 6 illustrates, accounts with lower probabilities of being a bot keep consistent along the period of study while accounts with higher estimated probabilities tend to be less stable, although there are accounts with consistent high probability of being a bot such as the account *MichelleCasti*.

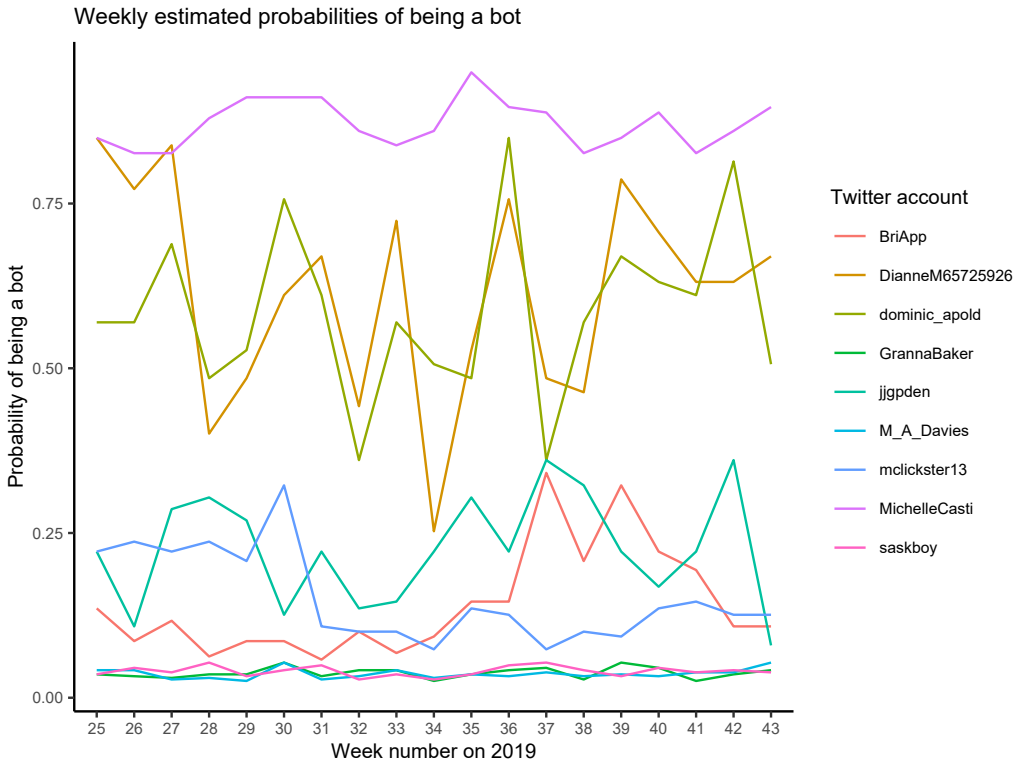

Figure 6: Trend of the estimated probabilities of being a bot of some sampled Twitter accounts that issued content related to elections, parties, or leaders every week during the period of study.

We classify Twitter accounts into five categories according to their estimated probability of being a Bot, where the class "missed" represents accounts that could not be analyzed because its information was restricted or because the account had been eliminated before the end of the corresponding week.

4.2.1 Elections

During the period from July 22nd to October 21st, as Figure 7 shows, the most of tweets about elections are issued by accounts with a probability of being a bot lower than 25%. Around 5% of the tweets come from accounts with a probability of being a bot greater than 50%, while tweets issued by the category of accounts with an estimated probability greater than 75% represent the smaller percentage of the total collected tweets. This distribution is consistent along this period of study both in a monthly basis as table 2 shows, and in a daily basis as Figure 21 (see Appendix) illustrates.

To describe the content of the tweets by month, we split tweets by words and give the *Term frequency-inverse document frequency* score to each of their words. Tf-idf is a commonly used score for extracting features in a bag-of-words model, and it represents the importance of a word for a document over a corpus of documents, which in this case corresponds to the set of tweets about elections. We provide a more detailed description of this score calculation in Section 5.1.2

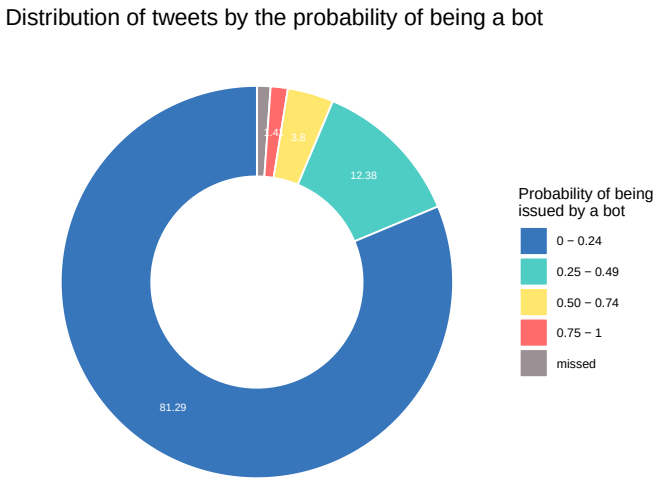

Figure 7: Proportion of tweets about elections that are issued by accounts classified according to their probability of being a bot.

Category of issuing account	July - August	August - September	September - October
Probability < 0.25	82.23	81.58	81.15
Probability $\in [0.25, 0.50)$	11.78	12.40	12.42
Probability $\in [0.50, 0.75)$	3.74	4.08	3.73
Probability $\in [0.75, 1)$	1.50	1.31	1.43
Missed	0.75	0.63	1.27

Table 2: Proportion of tweets about elections that are issued by accounts classified according to their probability of being a bot by month.

Figure 8 shows the highest scored words of tweets with content about elections according to two categories which are "Likely bot" and "Likely human"; these classes correspond to the probability of being issued by a bot greater or equal to 75% and less than 75%, respectively. During the period from July 22nd to August 21st, most of these words are similar between categories, however, there are some terms, like "bernier',' that appear only in the category "Likely bot". In the following periods from the end of August to the day of elections, "truedau" is consistently one of the most important terms in both classes of accounts.

In general, similar terms are present into the content of tweets generated by both types of account; this fact raises the question if the accounts are well classified, or if there is a spreading information pattern from bots content to the rest of accounts, or if bot accounts emulate the content that is issued by "likely human" accounts.

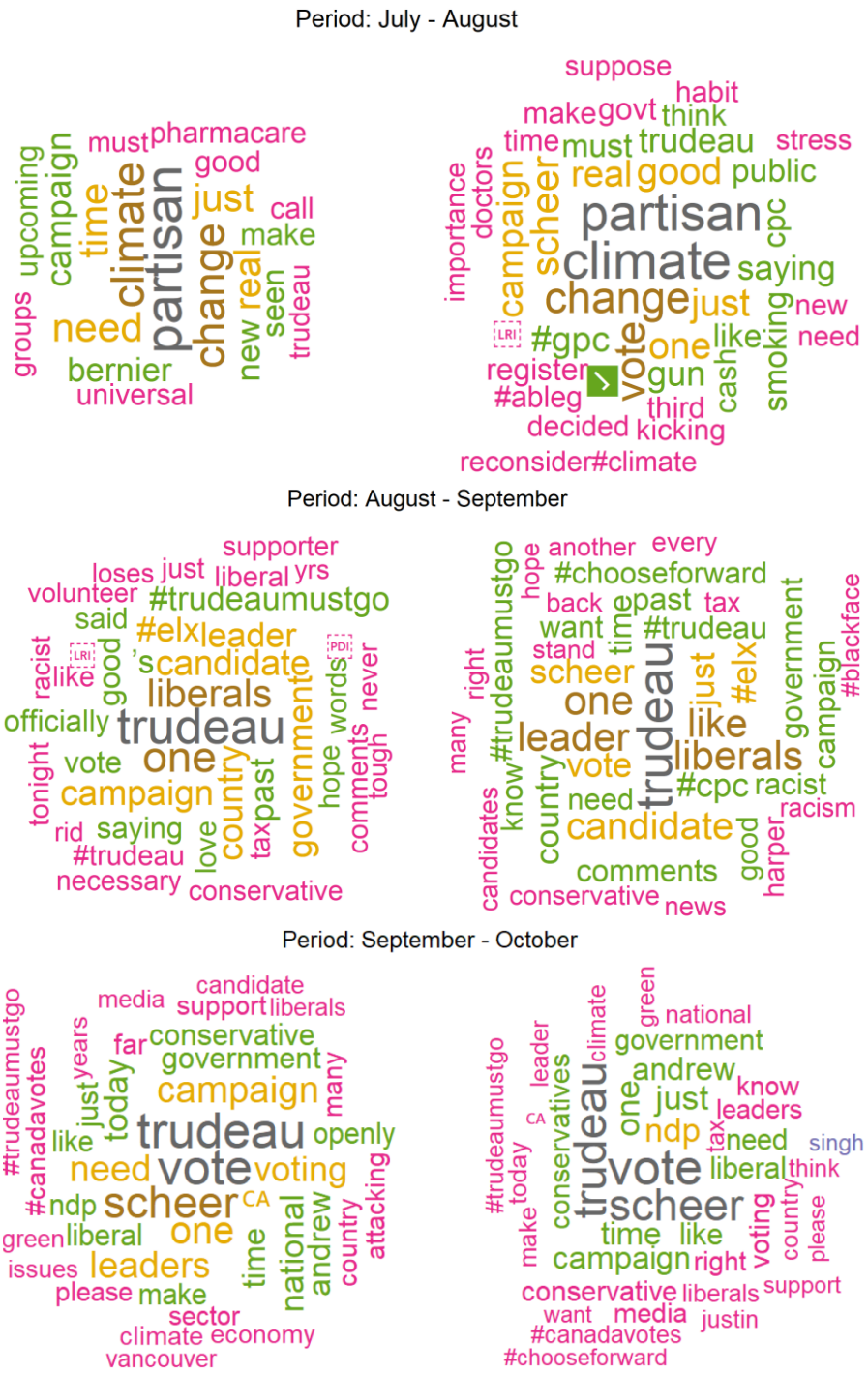

Figure 8: Cloud of highest scored words into tweets about elections from July 22nd to October 21st. Left-hand side clouds correspond to tweets issued by accounts belonging to category “Likely bot”; Right-hand side clouds correspond to accounts belonging to the “Likely human” class.

4.2.3 Leaders

Figure 11 shows the percentages of tweets with content about Canadian parties leaders during the period from July 22nd to October 21st according to the type of issuing account. It is noted that the most of tweets about the six leaders are issued by accounts with a probability of being a bot lower than 25%; around 5% of tweets come from accounts with a probability of being a bot greater than 50%, while tweets issued by accounts with an estimated probability greater than 75% represent the smallest percentage of collected tweets. This distribution is similar to the corresponding percentages of tweets about elections and Canadian parties, and it is consistent in a daily basis from three months before election day as Figure 23 (see Appendix) shows.

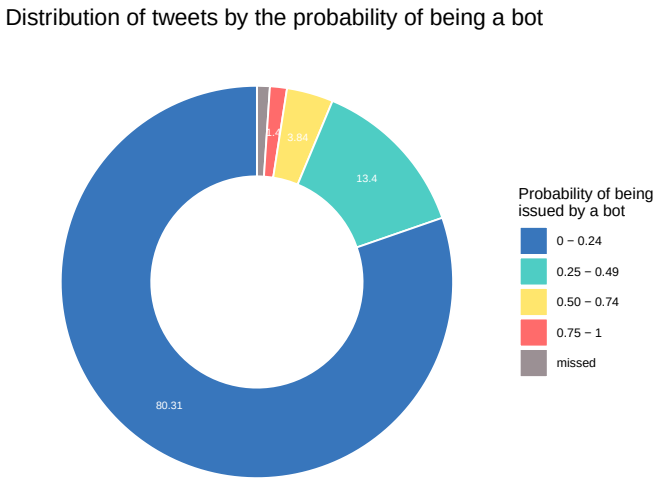

Figure 11: Proportion of tweets about leaders of Canadian parties that are issued by accounts, classified according to the probability of being a bot.

We calculate the percentage of tweets according to four categories of issuing accounts during the last month of election campaigns including the election day, and we display these percentages according to the mentioned leaders into tweets as Figure 12 shows. It is noted that tweets with the higher probabilities of being issued by a bot have content about the leader of the Peoples' Party, Maxime Bernier, while the leader of Conservative Party, Andrew Scheer, is mentioned by accounts that mostly have a small probability of being a bot: 82.3% of his tweets belong to accounts with the lowest probability of being a bot.

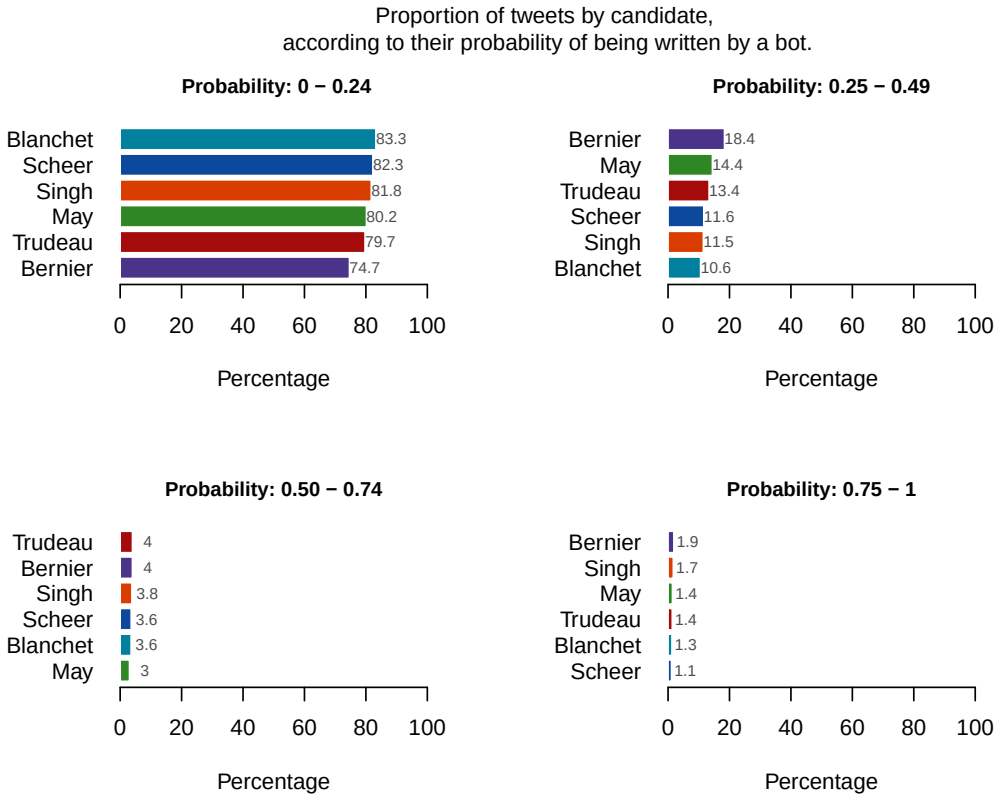

Figure 12: Percentages of tweets issued by account type and mentioned leader from September 22nd to October 21st.

5 Clustering

To accomplish our goal of investigating if accounts that are denominated by us as "likely bots", and also as 'bots' or 'suspicious accounts' in what follows, can be classified into the categories that were proposed by McKelvey and Dubois in 2017 (omitting the servant type), we perform a clustering analysis to the set of tweets about elections that were issued from September 22nd to October 21st and their corresponding accounts.

We apply the standard k-means to our first set of features, which consists of a subset of 18 variables that describe account characteristics, these characteristics are provided by the standard Twitter API (see the Appendix to know the set of all characteristics). Then, we study the content of tweets that are issued by each type of account. We do this through the application of the sparse hierarchical clustering proposed in [8] to our second set of features, which captures the content of tweets. In this way, we characterize the kind of content that is issued by each type of account and compare our findings to the results of McKelvey and Dubois.

5.2.2 Results

After performing the three previously proposed cluster analyses by considering three, four, and six groups, we find that the results with four clusters provide a clearer characterization and a more evident separation of groups than the obtained results for three and six clusters. For this reason, in what follows we present and describe in detail the results for $k = 4$ groups of suspicious accounts, while results for $k = 3$ and $k = 6$ clusters are presented in the Appendix.

We classify a total of 1,264 suspicious accounts that issued content about elections from September 22nd to October 21st according to the 18 variables described in Section 5.1.1. The size of each cluster is reported in Table 4.

Group	Type I	Type II	Type III	Type IV
Number of accounts	257	31	174	802

Table 4: Sizes of clusters that were obtained through the application of k-means with $k = 4$ to suspicious Twitter accounts that issued content about elections from September 22nd to October 21st of 2019.

To apply the k-means algorithm to the our first set of features, which consists of 18 variables, we standardize the variables and we use the Euclidean distance as the measure of distance to the clusters' centers. In order to analyze how the variables characterize each of the obtained clusters we calculate the means of each standardized variable by cluster and plot them, as Figure 14 shows. Bellow we describe the prototype of each cluster, which is characterized by the mean of these 18 variables.

Type I This type of account is the only that issues quoted tweets, replies tweets to other accounts and does not emit retweets. In general, accounts of this type are not verified, do not have so many followers or friends, and they have the greatest amount of liked tweets. Because of these characteristics, we denominate this type as **"repliers"**.

Type II This group mostly consists of accounts that are verified and emit the most quantity of tweets, including retweets. Accounts belonging to this group are the most popular since they have the biggest quantity of followers, friends and lists memberships. Because of these characteristics, we denominate this type as **"popular amplifiers"**.

Type III Accounts of this type are mostly not verified, and their content consists mainly of retweets that have been commonly retweeted and that are generated by accounts that issue the biggest amount of likes. Thus, we denominate this type as **"retweeter amplifiers"** because they spread information by retweeting popular retweets.

Type IV This group of accounts is similar to *retweeter amplifiers* in the sense that its accounts are mostly not verified and its content mainly consists of retweets. However, this group has two main differences: *i*) its content is not retweeted as much as the content issued by *retweeter amplifiers*, and *ii*) its content comes from accounts that do not have issued so many likes. Because of these characteristics, we denominate this type as **"not amplified retweeters"**.

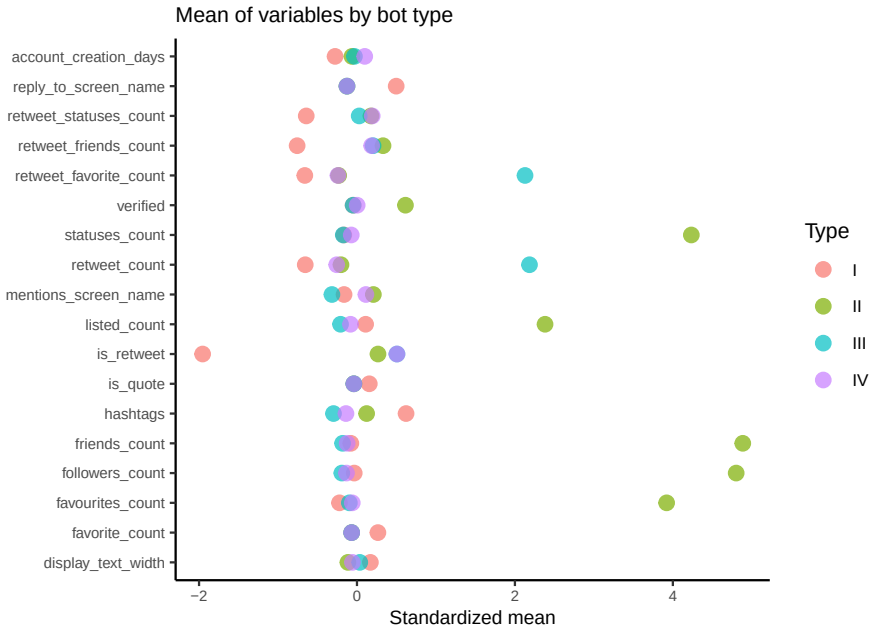

Figure 14: Means of standardized variables by group for $k = 4$ of suspicious accounts that issued content about elections from September 22nd to October 21st of 2019.

To visualize how the analyzed accounts are clustered according to the assumption that there are four groups of bots, we use the MDS technique to reduce the dimension of the 18 variables to three dimensions, which are called "DM1", "DM2" and "DM3". Figure 15 shows how the accounts, which are represented in the three dimensional space, are separated.

5.3 Cluster of tweets according to their content

We have characterized four types of accounts through the clustering analysis that considers our first set of features with characteristics of accounts. Both "Popular amplifiers" and "Retweeter amplifiers" appear to match the description of amplifiers, while the group of "Repliers" shares some characteristics with dampeners since this group issues replies and its content has the biggest amount of likes. However, this analysis does not describe the content that each type of account tends to generate, which is necessary to compare our findings to the obtained results in [5].

In order to describe the content that is issued by each type of suspicious account, we use our second set of features. For each group of accounts, we consider the set of words in their tweets that has a tf-idf score greater or equal to the median of the scores. Through this approach, for each type of account, we create clusters of tweets that share similar content according to their most important words. After performing the clustering of tweets, we summarize the content that each type of account tends to issue.

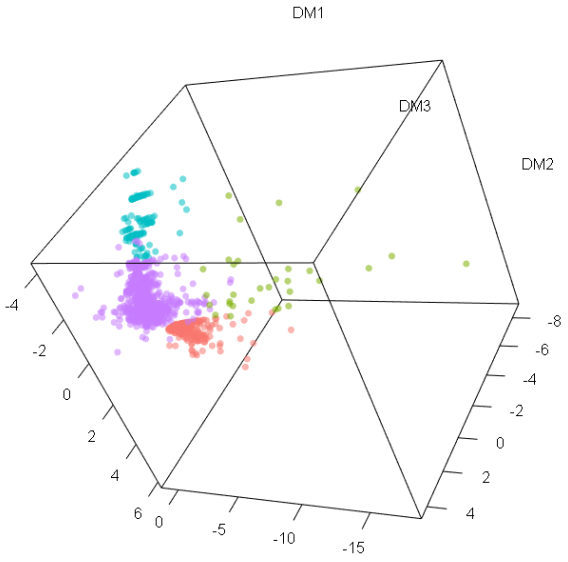

Figure 15: Representation of the analized suspicious accounts in 3D by reducing the dimension of their characteristics using MDS. Each color represents a different group: Type I (red), Type II (green), Type III (blue) and Type IV (purple).

5.3.1 Sparse hierarchical clustering

The method of clustering that we apply to the set of scored words of each type of account is called sparse hierarchical clustering, it was proposed in 2010 at [8], and we applied it to our data by using the R package `sparcl`. This algorithm provides sparse weights to features by applying the L1 regularization into the optimization problem that aims to maximize the dissimilarities between clusters with some additional constraints. More specifically, the said optimization problem is stated as follows:

$$\begin{aligned} \max_{w,U} \quad & \left\{ \sum_j w_j \sum_{i,i'} d_{i,i',j} U_{i,i'} \right\} \\ \text{subject to} \quad & \sum_{i,i'} U_{i,i'}^2 \leq 1, \quad ||w||^2 \leq 1, \quad ||w||_1 \leq s, \quad w_j \geq 0 \ \forall j, \end{aligned} \tag{2}$$

where:

- w is the vector of weights of features.
- w_j denotes the weight of the j-th feature.
- U corresponds to the matrix of dissimilarities between tweets.
- $U_{i,i'}$ represents the dissimilarity between the i-th and the i'-th tweet.
- $d_{i,i',j}$ denotes the distance between the i-th and the i'-th tweet in feature j.
- s is the hyperparameter that controls the level of regularization; smaller values of s generate more sparse solutions.

The solution to this optimization problem results in both a sparse vector of weights w^* and a dissimilarity matrix U^*, which is formed by a subset of the original set of features. Then, applying the hierarchical clustering to the matrix U^* results in a sparse hierarchical clustering. Additionally, the vector w^* makes interpretable the clustering results because "it is possible to determine what features are responsible for the observed differences between clusters " [8].

According to [8], a method to select the best value of s consists in calculating the gap statistic for several values of s and pick such value that generates the biggest gap statistic. The package `sparcl` has implemented this method in the function `HierarchicalSparseCluster.permute`.

In what follows we present the results that were obtained by separately applying the sparse hierarchical clustering with the squared Euclidean distance and the complete linkage, to tweets that were issued by each type of suspicious account: "Repliers", "Popular amplifiers", "Retweeter amplifiers", and "Not amplified retweeters".

5.3.2 Results

In general, dendrograms that result from the sparse hierarchical clustering form one big group of tweets and another one or two small groups of tweets according to tweets' content. This happens because of the previous clustering by type of issuing account, and because of the threshold we use to determine the set of important words to be considered into the sparse hierarchical clustering. The former provides homogeneity to each set of tweets because their issuing accounts are similar, so their content might be similar too, while the latter provides homogeneity to each set of tweets by comparing them through a set of important words instead through all the words in their content. These facts facilitate the interpretation of results, which are presented by type of suspicious account in what follows.

Tweets from Cluster 1
1. Attended an all Candidates Debate Meeting in East Vancouver. In British Columbia Canada. Major Topics include Housing, Health Care, Indigenous Community Support. I was there to show, My Support, of the National Political Party I'll be Voting for in Our upcoming Federal Election.
2. I'm asking my friends across Canada , to ask your friends across Canada. Why can't we support each other? Coast to Coast to Coast. Support pipelines, create jobs. #ItsOurVote #elxn43 #elxn2019 #CDNPoli #pipelines #oilandgas
3. This election in Canada is hard. NDP and liberal will support each other if conservatives get elected good chance it's a minority which leaves no real support for the conservatives.
Sampled tweets form Cluster 2
1. I voted! - voting in the 2019 Canadian federal election
2. voting in the 2019 Canadian federal election
3. Oh goody, four years of political turmoil and mudflinging between the #Liberals and the #Conservatives. #CanadaElection2019 #cdnpoli
4. Reminder! Today is Election Day in Canada. Polls are open 7am to 7pm. Require information on where to go to vote ... here is a link to Elections Canada.
5. voting in the 2019 Canadian federal election
6. VOTE NDP FOLKS - voting in the 2019 Canadian federal election
7. Democracy is an important value for our Credit Union. Today is federal election day in Canada and we encourage each of our members to get out and vote. (Ballot Box emoji) #elxn43 #democracy
8. Monday 21 October 2019 is the day! Don't forget to vote (Canadian flag emoji) Your voice matters, let it be heard by voting. #voting #vote #ElectionsCanada #government #votingrights #electionday #yourvotematters #ivoted #govote #votingday #votingmatters #votingrights #CanadaElection2019 #spgill
9. 'No one has ever asked': Andrew Scheer explains why he never mentioned his dual Canadian-U.S. citizenship
10. The Federal election is happening on October 21st, 2019 and your Library has many political titles available. Come in and check out the Politics of Canada display! Find these titles and many more here:
Tweets from Cluster 3
1. NDP sees surge of support from B.C. voters in election's final days, poll finds | CBC News
2. Tell your local election candidates to support a diabetes strategy for Canada now! Learn more and take action at ... (3 times)
3. No matter who you support, be sure to get out there and vote today! Make your voice count. #elxn43 #Canada-Election2019 #ElectionDay
4. Tell your local election candidates to support a diabetes strategy for Canada now! Learn more and take action at $\cdots$
5. NDP sees support jump at expense of Liberals, Tories one week before election

Table 5: Sample of tweets that belong to each cluster according to the de dendrogram of tweets that were issued by Repliers.

Popular amplifiers

According to Table 4, this group consists of 31 accounts. This group of accounts issued 119 tweets; we use the content of these tweets to build a set of features based on a bag of words model, which has a total of 37 included words that reached a tf-idf score of at least 6. We pick the tuning parameter for regularization $s = 1.45$, this generates the highest Gap statistic among different values of s and a set of 4 words with non-zero weights as Figure 16 shows.

Table 11 (see Appendix) presents words that are used as features for clustering the tweets belonging

to this type of account. As we illustrate in Figure 16, there are four words with non-zero weights that separate tweets according to their content. It is noted that three words with non-zero weights correspond to the first and last name of the leader of the Liberal Party, Andrew Scheer, and to the first name of Justin Trudeau.

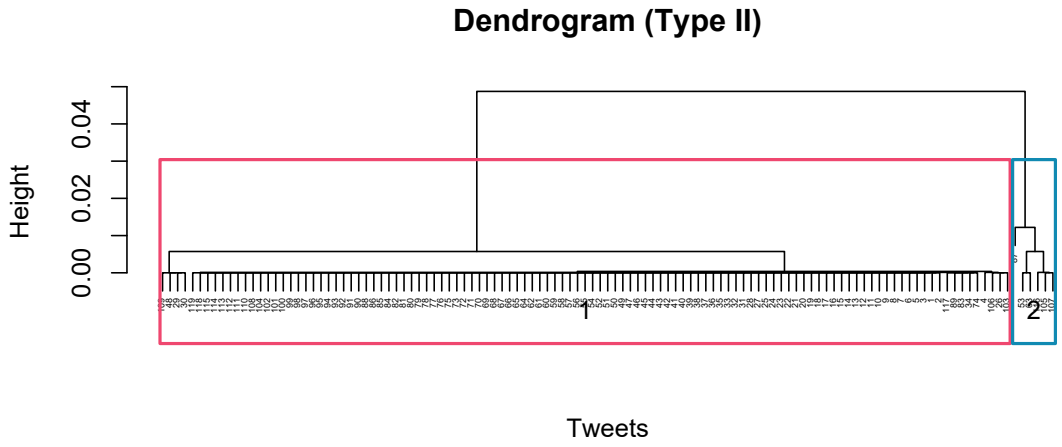

Figure 18: Dendrogram generated by the sparse hierarchical clustering of tweets that were issued by accounts of Type II ("Popular amplifiers").

We separate tweets into two clusters of sizes 113, and 6, respectively, according to the dendrogram that is shown in Figure 18. In general, these accounts issued content related to the leaders of Liberal and Conservative parties, where the hashtags #trudeaumustgo (against Trudeau) and #scheerpm (supporting Scheer) are among the most important terms in this set of tweets. In what follows we show a sample of tweets belonging to each cluster of tweets; it is noted that Cluster 1 mainly consists of tweets against Justin Trudeau and in favor of Andrew Scheer and conservatives, while Cluster 2 is comprised of tweets against Andrew Scheer.

Tweets from Cluster 1
1. #cdnpoli #TrudeauMustGo #TrudeauDictatorship Our only hope is #Scheer4PM this is the most important election in Canadian History. Our kids are depending on sane votes. Don't throw it away on a party guaranteed to lose, this election is too important!
2. Canada cannot afford four more years of Justin Trudeau's #LPC, let alone four more years with the #NDP pulling the strings. (stop sign emoji) Only the @CPC_HQ will lower taxes; balance the budget so that you can get ahead. (blue circle emoji) #cdnpoli #elxn43
3. This isn't a plan to make life more affordable; it's an election ploy. If you want more money in your pocket, vote @CPC_HQ: Universal Tax Cut Tax credits for children's activities (man artist emoji)(softball emoji) Taking the GST off home-heating (house with garden emoji) #cdnpoli #elxn43
4. Great speaking at the Alliance of Senior's Centres Forum today. Only 2 Vancouver-Centre candidates showed up. Seniors are a priority for the Conservative Party. #CPC #cdnpoli #elxn43
5. Rosemary Barton should stand down and decide if she wants to be a journalist or an activist. This campaign has damaged her brand, perhaps irreparably. #elxn43
6. #TeamCavey out in full force this morning on the Burrard Street Bridge! It's time for YOU to get ahead! #CPC #cdnpoli #elxn43
7. Liberal apologists today: stop sharing that #TrudeauBlackface video because you're just being a bully. Also it's racist to show this racism. THAT'S WHY YOU'RE LOSING, LIBS! #elxn43 @gmbutts @cathmckenna 8. @gmbutts @fordnation Good morning Buttons! I see you're in a feisty mood today. The news poll results got you worried? Nobody likes your campaign theme song? Your team only rolled yours out, like 48 hrs ago, what's the rush? Most of #Scheer4PM platform is already out there. #TrudeauMustGo #Elxn43
8. @CBCNews @AaronWherry Correction: From citizenship to same-sex marriage to pre-politics resumés, Liberals are struggling to redefine Scheer's image in desperate attempts to deflect from the damage #KokaneeGroper #BlackfaceTrudeau #TwoPlaneTrudeau has done to their own. #FixedIt #elxn43 #cdnpoli
9. (Police car light emoji)(Police car light emoji)(Police car light emoji) Trudeau rumoured involved in an explosive sex scandal percolating at the highest echelons of Canada's media establishment (face with open mouth emoji) #leaderdebate2019 #elxn43
10. Trudeau is a man much better able to apologize for the failures of those long dead than recognize where his own conduct has fallen short. And he has fallen short
Tweets from Cluster 2
1. Ahead of #elxn43, I invited federal leaders to discuss the key issues of housing, transit; opioids with me. All agreed except for Andrew Scheer. When I read his platform I knew why. It's clear to me that Andrew Scheer would be worse than Stephen Harper. #cdnpoli #vanpoli
2. Tomorrow we have our election in Canada. Andrew Scheer and his Conservative party aren't much different from the GOP. Its tax breaks for the rich, limiting immigration, questioning #ClimateChange, opposing abortion and same sex marriage. I'll be voting for Justin Trudeau.(Grinning Face with Smiling Eyes Emoji)(Thumbs up emoji) (2 times)
3. Hello American Twitter Peeps: Canada is in the middle of a Federal Election. Some of my tweets reflect this. Andrew Scheer is running 4 PM against our current PM, Justin Trudeau. Mr. Scheer is trump 2.0. I will outline in this thread how they compare. We need everyone's help. (2 times)
4. Federal Election 2019: Andrew Scheer says Surrey Newton will make the difference this time, vote for Harpreet Singh #harpreet4surreynewton #Surrey #surreybc #surreynewton #victoryisknocking #elxn43 #cdnpoli #surreypoli #conservative @CPC_HQ @AndrewScheer

Table 6: Sample of tweets that belong to each cluster according to the dendrogram of tweets that were issued by Popular amplifiers.

6 Conclusion

To analyze the presence of bots into the politic discourse at BC during the period of study that starts three months before the day of Canadian Federal elections and which ends on October 21st of 2019, we have collected a sample of tweets that were issued during this period and whose content is related to three main topics: the Federal Canadian elections, the political parties, and their leaders. We have applied an implemented random forest classifier called Botometer, which is the most known machine learning algorithm to detect bots in Twitter. Using this tool we have estimated that the percentage of accounts with a probability of being a bot greater or equal to 50% that issued content about these topics is around 5%, while the estimated percentage of accounts with a probability of being a bot greater than 74% corresponds to less than 2% of the number of sampled issuing accounts. In addition, by weekly estimating these probabilities, we have found that the probabilities estimated by Botometer that are smaller than 25% are less variable than higher probabilities.

We have characterized suspicious accounts, which are denominated by us as "likely bots", into four groups that we have named as "Repliers", "Popular amplifiers", "Retweeter amplifiers" and "Not amplified retweeters" according to their characteristics through the analysis of 18 variables that describe characteristics of Twitter accounts. We have summarized the type of content that each group of accounts tends to issue through the application of the sparse hierarchical cluster proposed in [8], and we have found that three of these groups, which are the "Repliers", the "Popular amplifiers" and the " Retweeter amplifiers", exhibit the same behavior as the "Amplifiers" of McKelvey and Dubois. However, we have not detected groups of accounts that follow the same behavior as "Dampeners" or " Transparency bots".

Twitter accounts classified as "Repliers" resemble some characteristics of "dampeners" like the fact that the tweet they issue are replies to other accounts, and that their content is the most liked among the four classes of accounts; however, after analyzing the content of tweets that were issued by "Repliers" we have not found negative or discouraging discourse, we have actually found that their content encourages Twitter users to vote during the election day. Regarding the content issued by "Popular amplifiers", we have found that it mainly consists of negative discourse against Justin Trudeau, and against Andrew Scheer to a lesser extent. We have detected that the main topics in the discourse of "Retweeter amplifiers" are related to the debate. For this category, we also have found an amplified tweet that was originally issued by Jason Kenney. Moreover, we have found that most of the accounts belong to the category of "Not amplified retweeters", and although we have not detected a general pattern to describe the content of these tweets, we have found that the presence of the Green, People's and New Democratic parties in this set of tweets is more notable.

Through this analysis, we have found that bots exhibit different complexity levels in their operating mechanism. For example, accounts belonging to "Repliers" are mostly verified and issue their own content, which might correspond to a more complex mechanism that the used by accounts of type "Retweeter amplifiers", which do not emit their own created content. Additionally, "Retweeter amplifiers" appear to have a more complex mechanism than the "Not amplified retweeters" because the former tends to retweet content that has been commonly retweeted, while latter does not. This suggests that the "Not amplified retweeters" are inefficient bots regarding the task of amplifying their content.

Determining if the scarce presence of bots into the discussion of Canadian Federal elections impacted the results of the election is a complex task. We have reached our goal of comparing our results to those obtained by McKelvey and Dubois, but our findings do not provide enough evidence to state that there was or there was no impact of Twitter bots into the results of elections at BC. We have found through the exploratory analysis that the content generated by "likely human" and "likely bot" accounts tends to be similar, but it would be necessary to conduct a deeper analysis to explain if this happens because the content of bots is spread to human accounts or because the bots try to resemble human content to keep themselves unnoticed. We have also found that much of content about Justin Trudeau and the Liberal party has a negative connotation on all types of analyzed Twitter accounts, however, a deeper analysis should be carried out to explain how this relates to the fact that the Liberal party lost 6 seats in BC during the Canadian Federal elections of 2019.

7 Further work

The methodology employed in this analysis is not unique and can be carried out with other variations. Some of our ideas in this regard consist of the application of other clustering methods to determine the number and characteristics of types of suspicious accounts according to our first set of features, such as the application of DBSCAN, which does not require knowing the number of clusters and which is capable of detecting outliers. Additionally, with respect to the hierarchical grouping method, there are different variations to construct the matrix of features, such as the use of n-grams instead of words, or the election of a different threshold to determine words to be included. In addition, the sparse hierarchical clustering method has different variations, such as the choice of the measure to calculate the distance between observations, which in our case is the squared Euclidian distance, as well as the choice of the measure to calculate the distance between clusters, which in our analysis corresponds to the complete linkage.

References

[1] https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/.

[2] Alessandro Bessi and Emilio Ferrara. Social bots distort the 2016 us presidential election online discussion. 2016.

[3] Clayton Allen Davis, Onur Varol, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. Botornot: A system to evaluate social bots. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 273–274. International World Wide Web Conferences Steering Committee, 2016.

[4] Yuriy Gorodnichenko, Tho Pham, and Oleksandr Talavera. Social media, sentiment and public opinions: Evidence from# brexit and# uselection. Technical report, National Bureau of Economic Research, 2018.

[5] Fenwick McKelvey and Elizabeth Dubois. Computational propaganda in canada: The use of political bots. *The Computational Propaganda Project Working Paper Series*, 2017.

[6] Pablo Suárez-Serrato, Margaret E Roberts, Clayton Davis, and Filippo Menczer. On the influence of social bots in online protests. *International Conference on Social Informatics*, pages 269–278, 2016.

[7] Robert Tibshirani, Guenther Walther, and Trevor Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001.

[8] Daniela M Witten and Robert Tibshirani. A framework for feature selection in clustering. *Journal of the American Statistical Association*, 105(490):713–726, 2010.