The First 100 Days: A Corpus Of Political Agendas on Twitter

Nathan Green, Septina Larasati
Marymount University, Charles University
Arlington Virginia, Prague Czech Republic
ngreen@marymount.com, septina.larasati@gmail.com

Abstract

The first 100 days corpus is a curated corpus of the first 100 days of the United States of America's President and the Senate. During the first 100 days, the political parties in the USA try to push their agendas for the upcoming year under the new President. As communication has changed this is primarily being done on Twitter so that the President and Senators can communicate directly with their constituents. We analyzed the current President along with 100 Senators ranging the political spectrum to see the differences in their language usage. The creation of this corpus is intended to help Natural Language Processing (NLP) and Political Science research studying the changing political climate during a shift in power through language. To help accomplish this, the corpus is harvested and normalized in multiple formats. As well, we include gold standard part-of-speech tags for selected individuals including the President. Through analysis of the text, a clear distinction between political parties can be found. This analysis shows the important item of their political agendas during the first 100 days of a new party in power.

Keywords: corpus construction, politics, social media

1. Introduction

Political communication in the United States has changed dramatically over recent years. Gone are the fireside chats on the radio, open press conferences on television are becoming less frequent. The current form of communication of politicians to their constituents and to other politicians is Twitter (Conway et al., 2015; Perlmutter, 2008). A new administration's public agenda would have once been in the purview of the media but now it is in more control of the campaign using new media such as Twitter (L. Towner and Dulio, 2012). While we will examine the 2017 USA government, this same phenomenon has been shown in other countries as well during recent elections (Dang-Xuan et al., 2013)

2. Background

Studying politics and political power shifts has been a difficult problem due to limited resources (Menini and Tonelli, 2016) and lack of examples during the age of social media. Shifts in political ideology tend to happen over longer periods of time. In the last United States election, the executive branch (President) switched from a Democratic President to a Republican administration. This is considered a switch from the political left to the political right. Two years prior in 2015 the United States' Senate switch in the same direction. This was the first time since 2009 that all branches of government were under the same party's power when all branches were run by the Democratic party. The previous time this had happened for Republicans was in 2003, the first election after the 9/11 terror attacks and before that we would have to go all the way back to 1953. Given these long gaps, this is the first time we can examine their ability to push new agendas via social media.

There has been previous research in the area of social media and politics in the USA dealing with topic classification (Glava et al., 2017) and Summarization (Egan et al., 2016). These have lead to successful attempts to build models to determine someone's political ideology. Prior to social media analysis, research was conducted on modeling political agendas from Senate press releases (Grimmer,).

Twitter has been used as a source of corpus (Johnson and Goldwasser, 2016) and model building for some time (Cieliebak et al., 2017; Bermingham and Smeaton, 2011) but to the best of our knowledge, the 100 day corpus is the first to offer resources for tracking agendas in social media driven by a political power change.

3. Methodology

3.1. Corpus Construction

In order to produce the corpus, Twitter data was scrapped during the first 100 days of the new political term. This ranges from January 20, 2017 to April 29, 2017. The first 100 days in the USA is traditionally used to evaluate what is important to a new administration to evaluate if they can accomplish their main goals. For this reason, politicians are very active on social media trying to push what they consider to be most important to the public and their constituents.

The corpus is a collection of tweets of the new President along with 100 Senators. A tweet is a singular post made to the social media site. For each tweet we collect its:

- ID
- Tweet Text
- Date and Time Posted
- Number of times it was favorited
- Number of times it was retweeted (shared)
- From what software it was posted

All 101 Twitter accounts that are scrapped are labeled with their respective political party. In our release the corpus contains multiple variations (text only, lower cased, tokenized) and different formats. The corpus is available in raw data, comma separated, text only, and available with part-of-speech (POS) annotation. The scripts used to generate, convert, and process this corpus will be made available with the data.

Along with the corpus, we include a list of each Senator and their approximate GPS coordinates for their home region that they represent. This information can be used to study which parts of the country are most active online and which Senators converse with which regions of the USA according their their @ references.

3.2. Corpus Annotation

The Corpus currently contains two annotations. Each folder contains 1 individual Twitter user (Senator or President) and has been annotated with their appropriate political party (Republican, Democrat, Independent). After that we hand annotated the POS tags for the President.

POS annotation was done using Brat (Stenetorp et al., 2012). For the time being all Senators are automatically annotated with their POS using python and the Natural Language Tool Kit(NLTK) (Loper and Bird, 2002). These will be hand annotated in future releases of the Corpus. Scripts have been included to convert back and forth from NLTK to Brat's stand off annotation format so researchers can convert the data into the format of their choice.

3.3. Metrics

To evaluate differences in Twitter users' language we use the log-likelihood ratio statistic (Dunning, 1993; Rayson and Garside, 2000). We only take results that are significant with a $p < .01$. We run this both on unigram and bigram phrases comparing the political parties as a whole.

To evaluate communications we form a network graph where each node is a Twitter user and each edge is a reference. The thicker an edge the heavier the weight and the more the communication was used via the @ symbol. The degree of a node indicates how many different Twitter users with which the politician communicates.

4. Results

Republicans have control of the Senate with 52 members. We include the President in the total for Republicans with 53 members. Democrats have the minority position with 46 members. There are two Independent Senators who both caucus with the Democrats so we have included them in the Democrat's contingent with 48 members.

In Table 1 we break down the basic statistics behind the corpus dividing most per political party with Republicans symbolized with an R and Democrats with a D. The corpus has a whole has over half a million tokens with Democrats providing the majority of the text despite being in the political minority.

In a 100 day period the Democrats provided 27.37% more tweets than the Republicans. Despite the Republican President being famous for the amount he tweets, the opposition party seems to be the most vocal. The breakdown in Table 2 and Table 3 seems to give us more hints at why this might be the case. Democrats did not only tweet more but they added more vocabulary to the corpus as well with a 12.29% larger vocabulary. This could be attributed to the complexity of their message or it could be attributed to being "off" message where the Republicans seem to use similar language across members.

Stats	counts
Tokens	574,095
Tokens (R)	233,547
Tokens (D)	340,548
Members (R)	53
Members (D)	48
Number of Tweets (R)	25,803
Number of Tweets (D)	33,986
Vocab Size (R)	38,498
Vocab Size (D)	43,540
Avg Number of Tweets (R)	486
Avg Number of Tweets (D)	708
Avg Length of Tweet (R)	9.05
Avg Length of Tweet (D)	10.02

Table 1: Corpus Stats for the First 100 days corpus

4.1. Political Word Usage

Republican	Type	Democrat	Type
gorsuch	NE	health	N
enjoyed	V	trump	NE
#senate	N	gop	NE
#jobs	N	must	AV
meeting	N	care	N
great	JJ	fight	V
obamacare	NE	stand	V
hearing	N	climate	N
foxnews	NE	muslimban	NE

Table 2: 10 Most Overused words per political party according to the log-likelihood ratio statistic. Type is our categorization of the term. NE (Named Entity), N (Noun), AV (Auxiliary Verb), V (Verb), JJ (Adjective)

Table 2 shows the most overused words by each political party during the first 100 days. This is calculated using the log-likelihood ratio statistic. Basic stop words are removed from the list so the end result indicates a truer objective of each party. The Republican party was coming off of an election win and seems to be focused on governing objectives. "gorsuch" refers to the Republicans first governing task of nominating and confirming a new Supreme Court Justice, Judge Neil Gorsuch. Democrats rarely used his name, while it was the Republicans most overused word, and in the end it was their first legislative achievement in the new Senate. Republicans continued with the governing theme with "#senate", "#jobs","meeting","hearing",and "obamacare" all possibly indicating they were trying to move past the election and get to work. There are some sentiment words included such as "enjoyed" and "great" indicating some post victory celebration. "foxnews" is the more conservative of the news stations (Devaney, 2013) in the USA so presumable the Senators were referencing stories in their favor.

Democrats seem to be overusing words that indicate their platform from the previous President that they want to keep. One achievement of the previous administration is the Affordable Care Act (ACA), commonly refereed to as Obamacare. The top overused word from Democrats, "health", seems to indicate they are ready to protect their legacy on this issue or they are fearing changes to it. The two other issues that come up in their list are "climate" and "muslimban" both which, during the campaign, they were opposed to the new President's promises on these issues. While Republicans seem to move on from the election, Democrats seem to be referencing the opposing part via overusing "trump" and "gop", the nickname of the Republican Party". Democrats overuse some language associated with opposition such as "fight" and "stand" as well. We observed that the more overused words on this list tended to be Verbs, possibly indicating their willingness for action.

We observed that Republicans tended to overuse words that were Nouns, possibly indicating that they were pushing their overall objectives and agenda items. Meanwhile Democrats almost exclusively overused Verbs. This is likely an attempt to show their supporters that they are ready for action in response to their election loss.

Republicans	Democrats
look forward	we must
i enjoyed	health care
pleased to	millions of
the #senate	fight for
hearing on	this is
neil gorsuch	we need
judge gorsuch	climate change
meeting with	cuts to
congratulations to	will fight

Table 3: 10 Most Overused bi-grams per political party according to the log-likelihood ratio statistic

To go a little deeper into what each party is stressing we next looked at the bigrams overused by each party, which can be seen in Table 3. Once again we can see the winning party is very forward looking and overall with a positive sentiment. The bigrams still show they are pushing their main agenda item of confirming a new Supreme Court Justice, Neil Gorsuch. The Democrats, on the other hand, use language that indicates they are ready to take a stand for certain issues, "climate change" and "health care" in particular. As it turns out, these were two of the larger fights in the first 6 months of the new President's term.

In both Table 2 and Table 3, we evaluated overused terms. This indicates that both sides mentioned the term at least once. To see agenda items that one side said that the other one never mentioned we examine Table 4

We feel Table 4 shows the true polarizing views of each party. Republicans again are focusing on their first achievement on the Supreme Court and celebrating their win with "#inauguration2017". They also bring up issues that are purely on the conservative spectrum in the USA such as

Spoken by Republican	Spoken by Democrat
#confirmgorsuch	#trumpcare
#neilgorsuch	#aca
well-qualified	#nobannowall
#inauguration2017	#womensmarch
#iran	#broadbandprivacy
#marchforlife	#climatechange
#repealandreplace	#stopgorsuch

Table 4: Agenda words spoken by one party but not the other

- "#iran" : opposing the previous administration's deal with Iran.
- "#marchforlife" : traditionally a march against abortion
- "#repealandreplace" : a rallying cry for those who want to get rid of the Affordable Care Act (Obamacare)

The more polarizing agenda items on the Democrats side includes mostly alternative views to the previous views such as "#stopgorsuch" and "#climatechange". Along with these some new issues that did not previously come up appeared in the form of hashtags such as:

- "#trumpcare" and "#aca" : Democrats attempt to defend the Affordable Care Act while labeling new proposals as Trumpcare, a similar tactic done with the previous administration to label health care Obamacare
- "#nobannowall" : One January 27th the President signed an executive order that was consider by the Democrats to be a ban on Muslims. Additionally during the campaign, the President often declared he would build a wall between the USA and Mexico. This hashtag is in opposition to the new administration.
- "#womensmarch" : Also on January 27th a Woman's March took place where approximately 440,000 women took to the streets in DC and an estimate 5 million women marched around the world. In the USA this was seen as an opposition to the President who had made many remarks towards women in the campaign that were seen as negative.
- "#broadbandprivacy" : On April 3rd the President signed into law Senate Joint Resolution 34 which effectively allows telecommunication companies to sell private data of there customers. This hashtag appears to try to bring attention to that bill.

  One noticeable difference across Table 2,3, and 4 is the consistency of the Republican party. Across all three segments (unigrams, bigrams, and agenda words), the Republican party consistently repeats the same themes and language. The democrats seem to change their focus over the three areas and have a wider amount of topics that they cover. This corresponds to what we saw in Table 1 with Republicans using less vocabulary, less tweets, and less words per tweet. Their message appears to be short and on topic across most of the Senators. Alternatively, the Democrats seems to cover more topics , tweet more, and use more vocabulary to get their agenda across.

4.2. Mentions and Dialogs

The Twitter handles in this corpus represents real public figures with real public agendas that they want to push. We picked Democratic and Republican political party's tweets to analyze because of their opposing nature and it is the way they try to push their agenda and construct dialogs. This corpus allows us to analyze how the two sides connect and communicate using a social media platform.

Here we use Twitter mentions as a sign of a reference to initiate a conversation or dialog with another Twitter user. We collect all the mentions in the tweets and see how the communications are done in the senate.

Total	**To (D)**	**To (R)**	**To Other**
By (D)	548 / 11.42	767 / **15.98**	6101 / 127.10
By (R)	197 / **3.72**	468 / 8.83	6128 / 115.62

Unique	**To (D)**	**To (R)**	**To Other**
By (D)	328 / 6.83	201 / 4.19	3662 / 76.29
By (R)	132 / 2.49	220 / 4.15	3652 / 68.91

Table 5: Mention Count / Mention Average in 100 days period

There are total of 14,209 mentions in the corpus. Table 5 shows that on average each Democrat uses 15.98 mentions while Republican only uses 3.72 mentions to their opposing side in the 100 days period.

We assume a dialog happens when there is a mutual reference where senator A mentions Senator B and also the other direction (This may include a self-reference in a tweet).

Table 6 shows there are total of 239 dialogs that happened across party lines. The data also shows Democrats have more bi-directional conversation among Democrats compared to Republicans among Republicans.

	Democrat	**Republican**
Democrat	128	44
Republican	44	67

Table 6: Number of Dialog in the Tweets by Party

We draw these connections as graphs using an open source graph visualization tools, Gephi (Bastian et al., 2009). We represent the Twitter user as a node and a mention of a user B from a user A as an directed edge from node A to B. The edges are also weighted as how many mentions occur.

Since we also have the data of the states the Senators represent, we visualize the node based on their location. This way we can see geographically how the conversation happened in the 100 days. In the visualization, we can see that the mentions directed to other senators regardless of where they are and not bound geographically, although there are still some Senators that only have dialogs among the Senators from the same state, e.g. senators from Alaska and Mississippi.

The graph can also focus on a particular node and show how they are being mentioned and how they mention other senators as seen in Figure 1

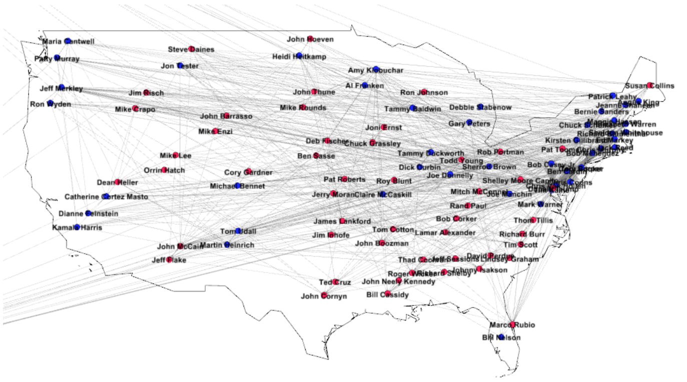

Figure 1: The caption of the figure.

We provide the .gephi Graph file and .csv node and edge files as part of the distributes corpus.

5. Conclusion

The 100 day corpus is unique in its slice of time and in its overall purpose. It is the first corpus dedicated to showing the change in language between opposing political factions during a change in power. Through corpus analysis we are able to pinpoint key agenda items, as well as differencing linguistic styles towards their persuasion.

The corpus includes over half a million tokens annotated for political party, location, and part-of-speech. The release of the corpus additionally contains the python scripts necessary to recreate the corpus, evaluate the corpus, and convert the corpus into a variety of working formats. The corpus will be available at the Marymount Data Fusion Center. With its availability we hope the community can gain further insight into the use of social media in the political realm.

6. Future Work

In the future we plan on continuing work on the corpus by making all Senators have gold standard POS annotations. We believe annotating dependency structure we will add additional insight into how linguistics can be used to push a political agenda. Annotations of sentiment will also aid in researchers ability to draw conclusions about this particular slice in political time.

7. Bibliographical References

Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks.

Bermingham, A. and Smeaton, A. F. (2011). On using twitter to monitor political sentiment and predict election results.

Cieliebak, M., Deriu, J., Egger, D., and Uzdilli, F. (2017). A twitter corpus and benchmark resources for german sentiment analysis.

Conway, B., Kenski, K., and Wang, D. (2015). The rise of twitter in the political campaign: Searching for intermedia agenda-setting effects in the presidential primary. 20, 05.

Dang-Xuan, L., Stieglitz, S., Wladarsch, J., and Neuberger, C. (2013). An investigation of influentials and the role of sentiment in political communication on twitter during election periods. 16:795–825, 06.

Devaney, H. (2013). Perceptions of media bias: viewing the news through ideological cues, April.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Comput. Linguist.*, 19(1):61–74, March.

Egan, C., Siddharthan, A., and Wyner, A. Z. (2016). Summarising the points made in online political debates. In *ArgMining@ACL*.

Glava, G., Nanni, F., and Ponzetto, S. P. (2017). Cross-lingual classification of topics in political texts.

Grimmer, J. (). A bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. In *In Proceedings of the First Workshop on Social Media Analytics, SOMA ?10*.

Johnson, K. and Goldwasser, D. (2016). "all i know about politics is what i read in twitter": Weakly supervised models for extracting politicians' stances from twitter. In *COLING*.

L. Towner, T. and Dulio, D. (2012). New media and political marketing in the united states: 2012 and beyond. 11:95–119, 01.

Loper, E. and Bird, S. (2002). Nltk: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, ETMTNLP '02, pages 63–70, Stroudsburg, PA, USA. Association for Computational Linguistics.

Menini, S. and Tonelli, S. (2016). Agreement and disagreement: Comparison of points of view in the political domain. In *COLING*.

Perlmutter, D. D. (2008). Political blogging and campaign 2008: A roundtable. *The International Journal of Press/Politics*, 13(2):160–170.

Rayson, P. and Garside, R. (2000). Comparing corpora using frequency profiling. In *Proceedings of the Workshop on Comparing Corpora - Volume 9*, WCC '00, pages 1–6, Stroudsburg, PA, USA. Association for Computational Linguistics.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J. (2012). Brat: A web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 102–107, Stroudsburg, PA, USA. Association for Computational Linguistics.