The patternization of Russian Troll to the 2016 Presidential Candidates on Twitter

Billy Zhang

August 2, 2020

Abstract

Every four years, America holds its presidential election. The most recent presidential election has been under scrutiny for a variety of reasons, including aid from russian trolls. The primary purpose of this paper is to answer who the russian trolls mimic better in a smaller scope, Donald Trump or Hillary Clinton, by using metrics derived from Natural Language Processing and unsupervised learning. The dataset is split into multiple time periods. One metric used to measure the relevance between the three subjects is intercluster distance with tf-idf and K-Means clustering. The other metric is the overlaps present in each subject's tweets with wordvec. The results show that the results of these two metrics speculates that the trolls are more likely to model after Trump than Hillary.

1 Introduction

In the recent decade, there has been a boom in the usage of social media as well as our reliance on it. For instance, Twitter now boasts 330 million monthly users as of Twitter's Q1 Earning Report [1]. There is also undeniable evidence that twitter is being used to advance political agendas. For example, in the United States 2016 Presidential Elections, Russian Trolls were charged with interfering in the elections via Twitter[2]. They manipulated users on the platform through promoting and spreading misinformation popularly dubbed "fake news." Whoever the next few presidents may be, it is important to understand the threats or would be threats from Russian manipulation of American elections.

Past research on Russian Trolls largely investigates the classification of Twitter accounts as trolls and the classification of troll account ties(e.g., other troll accounts or websites). Badawy et al.[3] studied the Russian interference with the 2016 Presidential Election. They focused on the classification of twitter trolls with a bot detection program. They also used label propagation to classify users who interacted with the trolls as conservative or liberal to determine effects of the trolls on each group. In an event as relevant as America's 2016 election, Narayanan et al. [4] focused on the role of Russian Trolls tweets and their interaction and effect they had during the Brexit Referendum, whether or not Britain should leave the European Union. In their paper, they mainly focused on the classification of the Russian troll tweets. They classified 142,918 links referenced in tweets, a 10% sample of their total data, into either Professional News and Information, Professional Political Content, or Other Political Content. However, this breakdown did not provide much leads. Both studies provide better insight on the threat of Russian trolls.

This paper brings in relevant political figures into the discussion to draw connection with the trolls. Specifically, the scope of this paper focuses on the tweets of two prominent political figures of the 2016 American Presidential Election, Donald J. Trump and Hillary Clinton. In addition, there is evidence of Russian Troll Farms, namely the Internet Research Agency(IRA) based in Russia, meddling with politics of other countries, such as America. This paper serves to discover possible patterns between the russian troll tweets and the political figureheads through unsupervised learning. The main patterns found were intercluster distances through K-Means and overlaps between topics through word2vec.

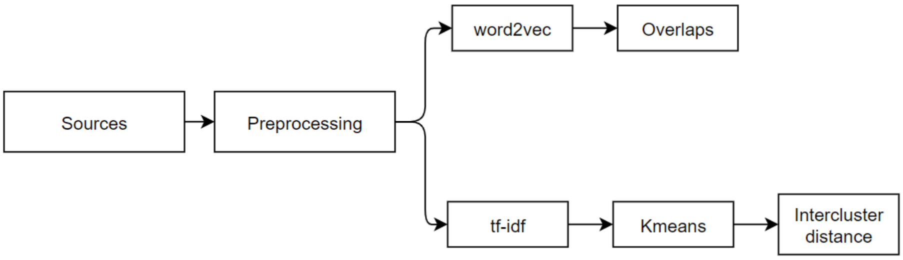


Figure 1: Flowchart of methodology

2 Methods

2.1 Sources

Two pre-existing datasets composed of tweets were used in this paper: Russian Troll Tweets[5] and American Presidential Candidates[6]. The dataset is premade by Fivethirtyeight, Russian Troll Tweets, and Ben Hamner of Kaggle, Presidential tweets. The Russian Troll Tweets from Fivethirtyeight contains roughly 3 million tweets from the time period of 2015 to 2017. It contains 2,848 users that have connections to troll factories, an institutionalised group of users designed to affect the politics and decision making of subjects. They were originally sourced by Clemson's Social Media Listening Center using Saleforce's Social Studio. These tweets are connected to the Internet Research Association, a Russian "Troll Factory". In the Kaggle dataset for the Presidential tweets, the range of time for Trump's tweets started early January of 2016, but Hillary's tweets did not start until mid April. For comparison sake, Trump's tweets before Hillary's tweets, April 17, were discarded. The dataset ended its collection of both political figurehead's tweets on September 27. It collected a total of 6,444 tweets, of which 3,226 tweets were Hillary's and 3,218 tweets were Trump's. However, because the data set was cut down, only 1451 Trump tweets and 2515 Hillary tweets were analyzed.

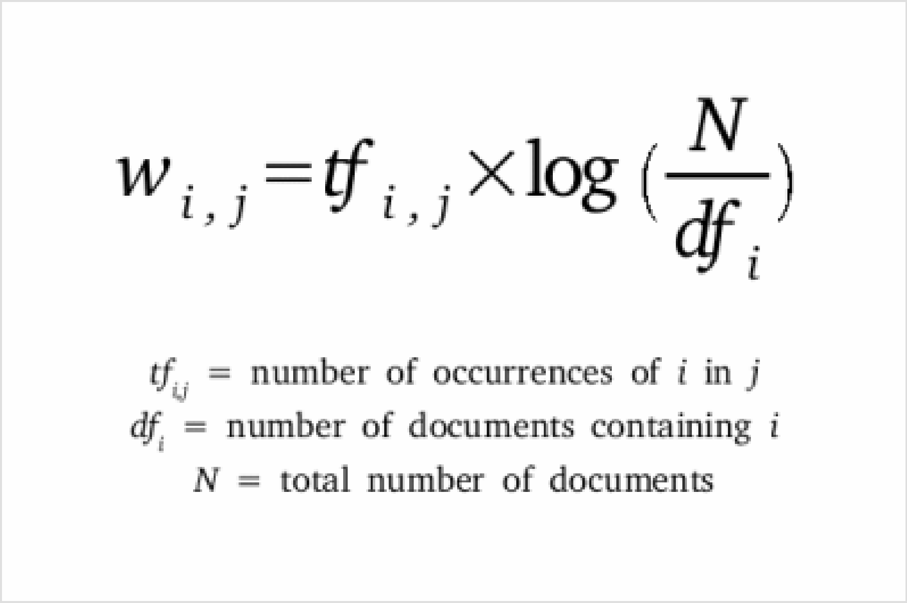
Similarly, for the comparison, for Russian Troll tweets, they were also cut so they would share the same time frame as that of the presidential tweet range. Thus, the 3 million tweets provided by FiveThirtyEight was cut down to 246,975 tweets. Furthermore, I decided that staggering the dataset may be important or influential to the findings. One part would be a dataset that contained all the troll tweets from April 16 to April 21, and another dataset would stagger April 21 to April 26. In total, the 6 months were broken down into 27 datasets.

2.2 Preprocessing

The preprocessing of data was achieved through the removal of punctuation, removal of URLs, tokenization the tweets, and removal of stopwords. The removal of punctuation and the urls were done with python Regex, regular expressions. Removing punctuation does not affect the language or processing of the language at all, and URLs were removed because they don't provide much information and are not the scope of this paper. Tokenization is the breakdown of language into individual words. For example, the string "I play football" would be split into [“I”, “play”, “football”]. With tweets broken down into simpler pieces, it will make the processing pipeline easier to clean. Stopwords are words that do not provide much information, such as “the”, ”a”, ”an”, etc. Thus, they will be removed because it takes up processing time without much benefit. The extraction and removal of these stopwords are accomplished by NLTK (Natural Language ToolKit), a python library. NLTK contains a corpus of stopwords, and if a token is in the corpus, it is removed.

2.3 Tf-idf and Kmeans for intercluster distance

Tf-idf was applied to convert the tweets to a vector representation. Tf-idf, term frequency inverse document frequency, scores words in a corpus based on how often the term appears while penalizing a word if it appears in multiple documents. Because tf-idf vectorizes data into multiple features, principal component analysis(PCA) reduces the dimensions; in this case, PCA reduced all the vectors to two dimensions.



$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$ = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

Figure 2: The formula for tf-idf [7]

K-Means, an unsupervised learning algorithm, was used to cluster the data. The elbow method, figure 3, was used to determine the optimal number of clusters to be used in K-Means. Sometimes, an elbow would not be found, so the date would be dropped. K-Means is a model that assigns K, the optimal number of clusters, number of centroids to a point. The model then calculates intercluster distance between points. Over multiple iterations, it updates a centroid if a better position is found. It will stop when all the centroids are in the best position. Figure 4 shows a visualization of clusters created.

After clustering the data, the cluster centroids were used to determine relationships between the figureheads and the Russian trolls. The K-Means model already calculated all the centroids of the clusters during its building. From the dimensional reduced data of the political figureheads, the distance of each tweet to its respective centroid is calculated through the distance formula.

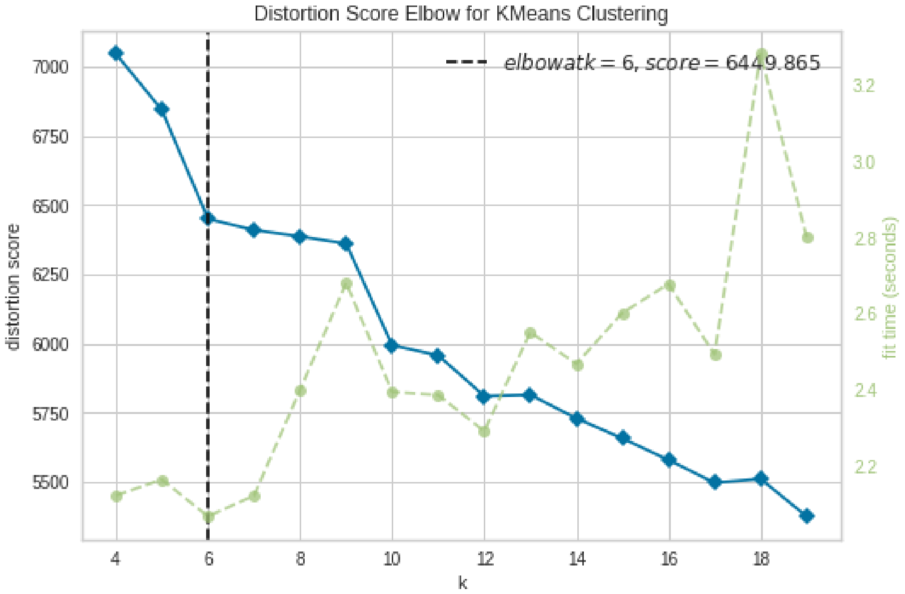
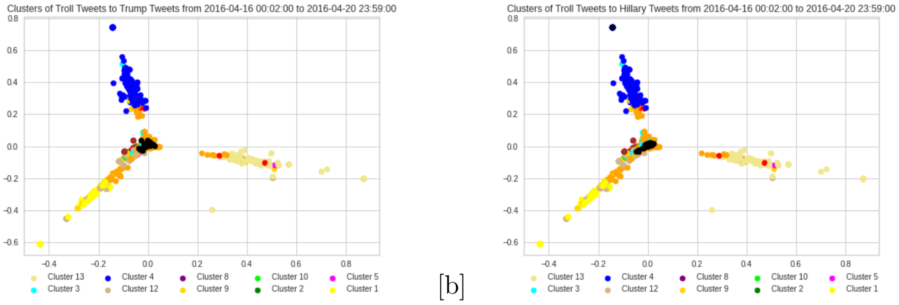


Figure 3: The Elbow Visualizer The blue shows the distortion score, and the elbow occurs at the dotted line. The green line is only a measure of time.



[a] [b]

Figure 4: Combined tweet distribution (Trump + Russian Trolls left, Hillary + Russian Trolls, right) within each cluster projected into a 2-D space. The black points represent the tweets of the respective figureheads

2.4 Word2Vec and wordclouds for overlaps

Word2Vec is used for vectorizing the individual words of tweets. Unlike tf-idf, word2vec can interpret language with more than just counting, such as semantics. This would prove useful because it can tell if two words are similar. The word2vec model differed from the tf-idf model because the word2vec model was created with the whole dataset rather than staggered. Unlike tf-idf, the word2vec model seemed to have a limited vocabulary; thus, to offset this problem, the model used the whole of the troll dataset.

WordClouds are a good way to visualize the words/topics relevant to the dataset. For example, figure 5 shows the most used words of respective parties from April 16-April 21. The visualization is not the main tool, but the list of top words/topics of the corpus. With all three wordclouds created, a script can count the overlaps of the figureheads to the trolls: Trump to trolls and Hillary to trolls. Overlaps are defined as words that are found in or similar in two vocabularies. The script counted the same words as one or words that were deemed similar by the word2vec model. If one figurehead has “America” as a top word and the troll has either “America” or a similar word, the script would count that as an overlap.

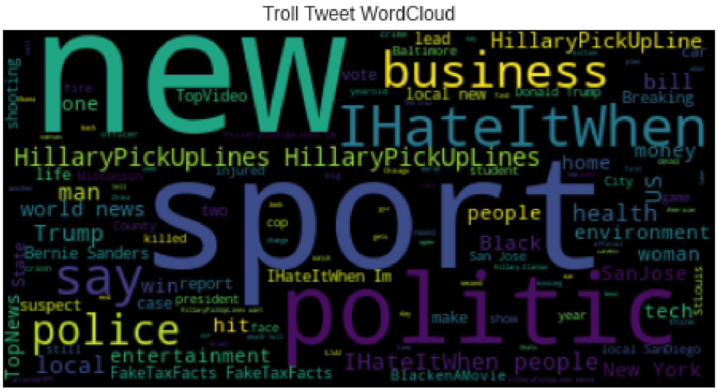


Figure 5: WordClouds created from April 16-April 21 tweets These wordclouds show the most used words/topics of the time period

3 Results

3.1 Tf-idf and Kmeans for intercluster distances

		Trump				Hillary		
	date	distance	tweets	average		distance	tweets	average
0	4/16-4/21	1.7352	64	0.02711316893		1.859	68	0.0273394784
1	4/21-4/26	6.0088	65	0.09244360439		8.7594	96	0.09124461304
2	4/26-4/30	4.2752	71	0.06021413179		10.7412	93	0.1154974241
3	5/6-5/11	4.9306	78	0.06321300538		9.763	85	0.1148594156
4	5/11-5/16	4.4478	74	0.06010595903		4.1511	71	0.05846660722
5	5/16-5/21	4.9685	114	0.04358404162		3.5634	82	0.04345611118
6	5/21-5/26	2.3437	73	0.0321059837		8.367	98	0.08537824998
7	6/1-6/6	6.5467	78	0.08393260868		7.1638	102	0.07023342323
8	6/6-6/11	5.6224	71	0.07918987906		11.4534	132	0.0867685822
9	6/16-6/21	1.9588	48	0.04080919285		8.2974	93	0.08921993736
10	6/26-6/30	4.9023	62	0.07906995281		7.442	104	0.07155836438
11	7/1-7/6	6.8703	71	0.09676540991		6.7626	86	0.07863603382
12	7/6-7/11	4.2812	52	0.0823314286		6.3079	105	0.06007617165
13	7/16-7/21	6.7183	55	0.122151629		9.5392	154	0.06194341208
14	7/21-7/26	4.6453	91	0.05104738789		15.3666	249	0.061713257
15	7/26-7/31	14.4133	141	0.1022225927		81.4602	417	0.1953484331
16	8/1-8/6	14.4462	90	0.1605144016		19.1256	115	0.166310192
17	8/6-8/11	9.671	70	0.1381579892		19.55	118	0.1656785787
18	8/11-8/16	4.433	61	0.0726737671		12.6213	136	0.09280409509
19	8/16-8/21	5.0162	47	0.1067278102		11.1134	117	0.09498668172
20	8/21-8/26	8.5065	60	0.1417762548		9.2358	119	0.07761213222
21	8/26-8/31	6.8661	71	0.09670584611		20.5931	113	0.1822404348
		133.6074	1607	0.08314088363		293.2364	2753	0.1065152198

Table 1: Political Figurehead Tweet's Distance From Centroid The average distance away from the centroid is described on the final row and final column

By calculating the distance from a tweet to its cluster centroid, we're able to get additional context about how "relevant" the tweet is to the cluster. The further away the tweet is from the cluster centroid, the less similar the tweet is from its cluster.

In table 1, Trump's tweets are overall more similar (average distance = 0.0831) to the russian troll tweets than Hillary's tweets(average distance = 0.1065). However, there are also ten time periods where Hillary's tweets have a closer relationship with the russian troll tweets: 4/21-4/26 (difference of 0.0011), 5/11-5/16 (0.0016), 5/16-5/21 (0.0001), 6/1-6/6 (0.0136), 6/26-6/30 (0.0075), 7/1-7/6 (0.0181), 7/6-7/11 (0.0222), 7/16-7/21 (0.06020821689), 8/16-8/21 (0.0117) and 8/21-8/21 (0.0641). Further runs with different random seeds of these dates also have verified that 5/11-5/16, 6/26-6/30, 7/1-7/6, 7/16-7/21, and 8/21-8/26 share the same pattern.

	date	hillary overlap	trump overlap	difference
0	4/16-4/21	41	53	12
1	4/21-4/26	48	61	13
2	4/26-4/30	37	56	19
3	5/1-5/6	40	62	22
4	5/6-5/11	42	46	4
5	5/11-5/16	45	44	-1
6	5/16-5/21	51	59	8
7	5/21-5/26	55	61	6
8	5/26-5/31	39	55	16
9	6/1-6/6	53	56	3
10	6/6-6/11	41	72	31
11	6/11-6/16	49	63	14
12	6/16-6/21	41	57	16
13	6/21-6/26	47	58	11
14	6/26-6/30	59	66	7
15	7/1-7/6	50	59	9
16	7/6-7/11	44	62	18
17	7/11-7/16	46	70	24
18	7/16-7/21	53	69	16
19	7/21-7/26	52	72	20
20	7/26-7/31	60	78	18
21	8/1-8/6	40	60	20
22	8/6-8/11	37	52	15
23	8/11-8/16	43	60	17
24	8/16-8/21	47	66	19
25	8/21-8/26	33	58	25
26	8/26-8/31	39	68	29
		1232	1643	

Table 2: Overlaps of words between Trump to Trolls and Hillary to Trolls The last row is the summation of all the overlaps present in the respective datasets

3.2 Word2Vec and wordclouds for overlaps

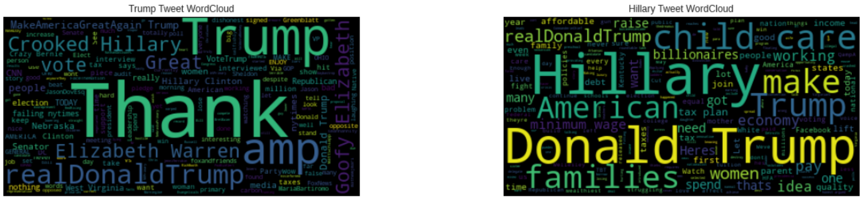
The overlaps of the data is summarized in table 2. Hillary and the Russian trolls share a total of 1232 words, while Trump and the Russian trolls share 1643 words, a difference of 411 words.

4 Discussion

Overall the data seems to skew towards Trump being more of the example for the trolls. While Trump's tweet during the period did have a lower number of tweets, his average exceeded Hillary's in both metrics. Overlap counting is the more substantial metric because it shows a bigger difference. If both figureheads had the same number of overlaps in relation to how much they tweet, Hillary should have more overlaps, but she does not. Hillary has 1064 more tweets, but shares 411 less words than Trump, table 2. This would most likely mean that the words/topics the trolls tweeted about were more likely to come from Trump than Hillary. The intercluster distance between Trump to Trolls and Hillary to Trolls also supports this claim. Trump's average distance was smaller than Hillary's. A tweet that is 0.0831 units away from the centroid is more relevant than a tweet that is 0.1065 units away.

Over the every time span, Hillary only once shared more words, one more word, than Trump with the trolls on 5/11-5/16. This date is also an overlap date where trolls also had a closer intercluster distance to Hillary(4.1). In table 1, it summed up the total tweets that both users tweeted during this period, Trump had 74 tweets and Hillary had 71 tweets. Looking at the wordclouds, figure 6 for this time period, Trump did not tweet about himself as much as Hillary tweeted at or about him. The top word for Trump during this period is "Thank", while Hillary's is "Hillary". Another deeming factor is on the lower right side of Trump's word cloud where "Donald Trump" is smaller than Hillary's "Donald Trump". This heavily differs from most other wordclouds during that month because "Trump" or a derivative is always the few most used. This time period may be an outlier where the trolls might have modelled after Hillary.

The outlier dates(5/11-5/16, 6/26-6/30, 7/1-7/6, 7/16-7/21, and 8/21-8/26) can also be outliers like 5/11. One possible explanation could be the volume tweeted by Hillary. Hillary tweeted substantially more, over 50% more, on all dates except for 5/11-5/16 and 7/1-7/6. The overlaps only counted frequent words and did not merge the amount of times the word was used.



[a]

[b]

Figure 6: WordClouds created from May 11-May 16 tweets These wordclouds show the most used words/topics of the time period

5 Limitation

The primary purpose of this paper is to speculate which of the two political figureheads the russian trolls are more likely to be modeled off of. However, results are not conclusive enough to determine exactly who the trolls are being modelled after. The metrics used here can be applied to other figureheads to answer that question for further studies.

6 Conclusion

The main purpose of this paper is to find patterns and relations between the Russian Troll Tweets and that of the Political Figureheads, Trump and Hillary between the two metrics. The results seem to show that the Russian Trolls seem to be more relevant to Trump than Hillary. The K-Means model calculated that the intercluster distance between the two political figureheads skewed more towards Trump. The overlap counter using word2vec also had similar conclusions. Understanding more about the Russian trolls can help better facilitate political forums present in Twitter along with reducing the greater implications trolls can cause.

References

[1] "Q1 2019 Earning Report," tech. rep., Twitter Company., 2019.

[2] "Justice Department drops plans for trial over Russian interference in 2016 U.S. election."

[3] A. Badawy, E. Ferrara, and K. Lerman, "Analyzing the digital traces of political manipulation: The 2016 russian interference twitter campaign," in *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 258–265, 2018.

[4] V. Narayanan, P. N. Howard, B. Kollanyi, and M. Elswah, "Russian involvement and junk news during brexit," *The computational propaganda project. Algorithms, automation and digital politics. https://comprop. oii. ox. ac. uk/research/working-papers/russia-and-brexit*, 2017.

[5] "fivethirtyeight/russian-troll-tweets," July 2020. original-date: 2018-07-30T23:35:18Z.

[6] "Hillary Clinton and Donald Trump Tweets."

[7] 351shares and 11kreads, "TF-IDF: Can It Really Help Your SEO?," Oct. 2019.