|  |
| --- |
| **Investigating Russian “Troll Tweets”** |
|  |
| **Tabitha Kaiser Angelina Campbell** |
| TLK937 AMC6452 |
| kaiser.tabitha@yahoo.com amc6452@utexas.edu |
|  |
|  |

Abstract

|  |
| --- |
| 000  001  002  003  004  005  006  007  008  009  010  011  012  013  014  015  016  017  018  019  020  021  022  023  024  025  026  027  028  029  030  031  032  033  034  035  036  037  038  039  040  041  042  043  044  045  046  047  048  049 |

Russian involvement with the United States 2016 presidential election is no longer just a conspiracy theory, but an undeniable fact supported by millions of data. This paper attempts to address the relevant topics tweeted by individuals employed by the Russian Internet Research Agency through the Latent Derelict Allocation algorithm. Our goal is to prove our hypothesis that Russian “trolls” targeted specific events or issues to cause division among Americans. Furthermore, we apply our LDA model to both “Left” and “Right” leaning tweets to reveal potential differences in priorities in political ideologies. Unfortunately, we are unsuccessful to fully show these differences as we had hoped for. We recognize now that LDA and our data did not produce clear results.

Introduction

|  |
| --- |
| 050  051  052  053  054  055  056  057  058  059  060  061  062  063  064  065  066  067  068  069  070  071  072  073  074  075  076  077  078  079  080  081  082  083  084  085  086  087  088  089  090  091  092  093  094  095  096  097  098  099 |

Twitter is an extensive micro-blogging tool that publishes millions of posts daily. The users of this website “tweet” about a multitude of topics such as business, politics, and personal life. This range of topics provides easily accessible data to collect and share. Twitter has been found to be involved in interference in the 2016 American election due to the influx of millions of “Russian Troll Tweets”. The data we are analyzing includes just under 3 million tweets and are organized into “left” or “right-leaning” political views. The tweets were collected and categorized by Darren L. Linvill and Patrick L. Warren (2018). These tweets come from accounts associated with the Internet Research Agency, a Russian “troll factory”.

Initially, we encountered difficulties when trying to concatenate our nine CSV files into a larger CSV file. By the recommendation and guidance of Laura Manor, we decided to use the Pandas package to make a data frame out of our CSV files. We also opted out majority of the data, as mentioned above, to a smaller sample size of 9000 tweets.

We believe that this project is important because not only did Russian involvement change politics in the United States, but also revealed that the internet can be seen as an security flaw, and should be treated as such. We need to continue to study this data to fully understand the results of Russia’s involvement. In order for us to understand the impact made by the Russian Internet Research Agency, we need to understand how these troll accounts were able to influence voters and the political climate. Learning what issues were targeted could reveal how political division was created or intensified over the years. Ultimately, when we better understand how much social media affects politics now, we can potentially protect future elections from being tampered with.

Data

The data we used for this project consists of nine separate .csv files shared from Linvell and Warren (2018), a working paper and data source from Kaggle. Linvell and Warren used the Social Studio social media listening platform to collect tweets that were traced back to Russia’s Internet Research Agency. It should be mentioned that a large majority of the data was not used for this project, since LDA has a tendency to take a long time to process data of this size. Interestingly, in Linvell and Warren’s (2018) paper they mention how atypical the data is for the tweets respective party views. The “Right troll”, for example, was described as distinctly different from traditional Republican themes such as abortion, taxes and level of government involvement. The “Left troll” focused mostly on Hillary and had a strong preference for Bernie before the primaries ended.

The .csv files contain nearly 3 million tweets collected from 2012 to 2018. In addition to the content of the tweets themselves, each tweet contains a significant amount of metadata, including the name of the twitter account, the region the tweet was tweeted from, the language the tweet is in, the date and time the tweet was published, the date and time it was harvested, the number of accounts the user was following, the number of other users following this account, the number of “update actions”, i.e. likes and retweets, if the tweet itself was a retweet or not, and labels as assigned by Linvill and Warren (2018). These labels are discussed in-depth below.

3 Methodology

The structure of this project was formed around several different machine learning algorithms. Initially, we attempted to concatenate our CSV files into a single combined CSV, but ultimately decided to use Pandas, as we are more familiar with using data in a data frame. We converted the CSV files into a Pandas data frame and used Glob to effectively combine the files. Following Dr. Li’s process in the ipython notebook “LDA Newsgroup”, we used NLTK to import a stemmer as part of our preprocessing. We utilized Gensim to assist in preprocessing and begin our Latent Derelict Allocation (LDA). At the suggestion of Dr. Li, we used pyLDAvis to build interactive visual representations of the resulting topics. As we will explain in the following paragraphs, we have come to the understanding that the data we have does not fit well in an LDA algorithm. Nevertheless, we continue to work with what we have and acknowledge the shortcomings of this algorithm using mass tweet data.

**3.1 Errors We Overcame**

We encountered several errors along the way, such as getting our .csv files into a compiled form, which was solved by using a more familiar dataframe, pandas. As mentioned in our presentation, we also had difficulty with bag of words because we originally did not use iterrows, which caused many problems with our lists. However, these problems only became apparent when we attempted to implement Bag of Words. At the same time as the iterating through rows error occurred, we were unable to get the function to create a list of popular tweets. We initially deleted popular tweets from our code, but eventually came back to it once we understood what had gone wrong. Once the iterrows error was corrected, it was simple to form a list of tweets from popular users. However, we encountered another error with popular tweets. This was an error with preprocessing, and it is discussed in the section of this paper titled preprocessing. As expected with case-sensitive code, we also came across spelling errors and capitalization changes.

|  |
| --- |
| 150  151  152  153  154  154  155  156  157  158  159  160  161  162  163  164  165  166  167  168  169  170  171  172  173  174  175  176  177  178  179  180  181  182  183  184  185  186  187  189  190  191  192  193  194  195  196  197  198  199 |
|  |

**3.2 Categories of Data**

Our “Russian Troll Tweets” data is organized by five different categories, each with their own unique patterns and behaviors. Among these categories, each as a unique number of handles, tweets, median (M) and standard deviation (SD). The five categories are organized by “Right Troll”, “Left Troll”, “News Feed”, “Hashtag Gamer”, and “Fearmonger”. In addition to these categories assigned by Linvill and Warren (2018), each tweet is labeled with its author’s account type: left, right, or Russian. We elected to perform topic modeling on four different sets of tweets: tweets in English, tweets labeled “left”, tweets labeled “right”, and tweets from users with 10,000 followers or more. Our goal was to determine common topics across the board, and then compare those to common topics for right-wing and left-wing trolls. By performing LDA on tweets from popular users, we hoped to discover which topics would attract more attention and make a troll account more successful.

4 Preprocessing

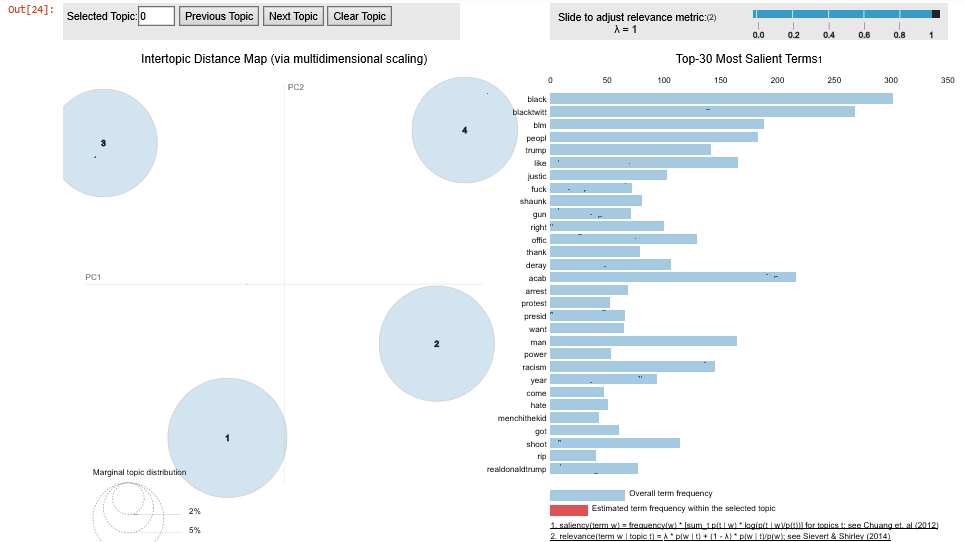
|  |
| --- |
| 100  101  102  103  104  105  106  107  108  109  110  110  111  112  113  114  115  116  117  118  119  120  121  122  123  124  125  126  127  128  129  130  131  132  133  134  135  136  137  138  139  140  141  142  144  145  146  147  148  149 |
|  |

We used the same preprocessing as Dr. Li did in her LDA example. We used gensim’s simple preprocessing and removed stopwords and tokens two characters or shorter. In Dr. Li’s code, the limit is three characters, but we decided to restrict length to two due to the prevalence of three letter acronyms, such as DNC and GOP, in political news. Additionally, we used a stemmer.

Once we had used gensim to convert our lists of preprocessed tweets into dictionaries, we filtered those dictionaries for extremes with the same parameters that Dr. Li used in “LDA Newsgroup”. However, this filter resulted in our dictionary of popular tweets becoming empty, so we did not perform this step on that dictionary.

**5 Visual Representation**

After successfully performing bag of words and LDA on each set of tweet (using the same process as in Dr. Li’s LDA Newsgroup), we used pyLDAvis to create an interactive visual representation of the resulting topics of each LDA model. pyLDAvis creates two graphs. The first is an Intertopic Distance Map (via multidimensional scaling) organizing topics (1-4). When hovering over the topic circle, the right graph will present the most relevant terms within the topic. If none of the topic circles are selected from the Intertopic Distance Map, then the “Top-30 Most Salient Terms” are displayed. Here is an example from the English tweets:



In an attempt to produce coherent results, we used the Intertopic Distance Map to find the highest number of topics that did not overlap for each LDA model. It was our hope that this would reduce the number of words that occurred in multiple topics.

6 Results

Overall, we have not found any coherent pattern with the topics we have reviewed. We tried different slices of the data frame and sometimes they ended up with very different topics and subject matter. We believe that there is some way that the data ended up being ordered in a way that it is not ideal for representing the entire data set. Perhaps, some users were more likely to tweet about particular subjects than others, and since the tweets were grouped by user, this skewed the results. With such a large data set, it is extremely difficult to know whether we used a slice that is representative of the rest of the data.

However, it is also likely that the subject matter might not be ideal for LDA because how interconnected everything is politically. The topics tended to have the same words with different weights. It proved difficult to label any particular topic because they all tended to be very similar. We did not know that this data would prove difficult to interpret until we began coding. In an attempt to try and produce a coherent result, we used the Intertopic Distance Map to attempt to pick a number of topics that were not overlapping. This does not seem to have been very successful.

In conclusion, this has been a great learning experience. Trying to force our data to fit LDA has proven to be very tedious and not as successful as expected. Either our data was not suited to this task, or topic patterns were too subtle for LDA to capture. Regardless, it still has been interesting to go through the data and analyze commonly used words by the Russian “trolls” who so successfully tampered with our country’s election. Although our project did not prove our hypothesis thoroughly, we can confirm that these were manipulative, yet subtle tactics used by the Russians. It is evident that social media will continue to be a part of politics, and politicians will not be limited to those with political experience. In this new age of internet and social media involvement, anyone can share their opinions and although this share of information can be largely beneficial, it is a breeding ground for blatant lies and misinformation. The best way to combat this looming threat is to further the analysis of things such as this, and text data across all platforms.

**6.1 Division of Work**

Most of this project was done together. We consulted with each other over most aspects of the work, but most notably Tabitha was able to make office hours so that we could further our work and iron out some of the errors we ran into. Angelina wrote the bulk of this paper.

7 References

Not including Dr. Li’s example code assignments, we used two other points of reference for this project. As mentioned above, we are more familiar with the pandas data frame, and when we tried to implement the same method of concatenation through .txt files with our .csv files, we encountered errors. Using the website, Stack Overflow, we found a solution to reading multiple .csv files in our directory. Although we had not used Glob in class, we decided it was the best course of action to use for our code. In addition to Stack Overflow, we also used pyLDAvis to illustrate our data into two graphs. pyLDAvis analyzes the relevance of a term to a topic and then presents those results through a ranking system of topics that is easily viewed.

Linvill, D. L., & Warren, P. L. (2018). Troll Factories: The Internet Research Agency and State-Sponsored Agenda Building. Retrieved from <http://pwarren.people.clemson.edu/Linvill_Warren_TrollFactory.pdf>

<https://stackoverflow.com/questions/20906474/import-multiple-csv-files-into-pandas-and-concatenate-into-one-dataframe>

<https://www.kaggle.com/fivethirtyeight/russian-troll-tweets>

<https://nbviewer.jupyter.org/github/bmabey/pyLDAvis/blob/master/notebooks/pyLDAvis_overview.ipynb>