Topic Modeling for Classification of Russian Tweets

|  |  |
| --- | --- |
| 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 | 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 |

|  |
| --- |
|  |
| Edward Burns, Anthony Trasatti, Sarah Whalen |
|  |
|  |
|  |
|  |
|  |

Introduction

Our project will focus on Russian propaganda tweets from the 2016 American presidential election. We find this particularly interesting because it is a relatively recent phenomenon (and one likely to reoccur in upcoming elections) that poses a threat to our democratic process. We plan to build a topic model, and use the topics found as features in a classifier for Russian vs. American tweets centered around the election in question.

We will have a topic model that is built on a combination of the Russian and the non-Russian tweets. We plan to then use these topics as features in a classifier, to determine if the inclusion of topics significantly improves the classifier. If the classifier is improved by the inclusion of topic as a feature, we can assume that the topics of Russian propaganda tweets are significant, and draw corresponding conclusions about the nature of the Russian propaganda machine, given the topics that prove to be significant.

Background

Badawy et. all (2016) were investigating the dissemination of the information on twitter that the Russian Troll accounts spread as well as its effects. They were focused on the question of who was engaging with the Russian trolls on twitter and what was the role of the political ideology. This included looking at how often liberal and conservatives accounts retweeted the Russian Troll tweets, the total number of tweets by these users, and the number bots on each side. The data used for this included 43 million elections-related posts retrieved from twitter between the dates September 16 and October 21, 2016, which included 5.7 million distinct users. One of the first things they did with the data was classify it into liberals and conservatives. The classifier was built on labeled propagation to classify the tweets into liberal or conservative with over 90% precision and recall. The results showed that significantly more of the trolls were producing conservative content and over 25 times as many conservatives retweeted content from the trolls. One critique is that one thing that may need to be accounted for was if there were many more conservative Russian troll accounts than that could lead to why so many more conservatives were engaging in activity with the trolls. However, the classifier of liberal vs. conservative sentiment is very cool.

Russian propaganda has historically been categorized according to the persuasive editing techniques associated with valorizing films such as  "Battleship Potemkin," however, the emergence of new communication media have spurred a contemporary brand of Russian propaganda, characterized by Paul et al.’s four distinctive features: (1) a high-volume and multi-channel approach, (2) rapid, continuous, and repetitive messaging, (3) a lack of commitment to objective reality, and (4) a lack of commitment to consistency. (2016) Paul et al. argues for the effectiveness of the contemporary Russian propaganda model using literature on influence and persuasion, as well as experiment-based psychological research. Results indicated the effectiveness of such a propaganda model, and pulled upon several phenomena, including the   "Multiple Sources Effect " (Harkins et al. 1981; Harkins et al. 1987) in which the presence of multiple sources espousing the same information increase the perceived truthfulness of said news story and the "Sleeper Effect " (Pornpitakpan 2004; Henkel et al. 2011) in which "low-credibility sources manifest greater persuasive impact with the passage of time, " as a result of their being dissociated from the original source.

(Paul et al. 2016) Paul et al. corroborate the effectiveness of Russian propaganda and suggest, given the nature of persuasion employed, that Russian propaganda be countered by ether (a) cutting of access to it or (b) producing messaging that undermines the ultimate goal of the propaganda, (which is often unknown). Their results,

while informative regarding the likely defining features of Russian propaganda (as well as informative regarding the means by which those features persuade), do not provide quantitative evidence to prove these characteristics as most important in the Russian propaganda machine.

|  |  |
| --- | --- |
| 100  101  102  103  104  105  106  107  108  109  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  143  144  145  146  147  148  149 | 150  151  152  153  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  188  189  190  191  192  193  194  195  196  197  198  199 |

Data

NBC released a corpus of Russian twitter data containing over 200,000 deleted tweets. This corpus contains both tweets that were generated by the Russian trolls and those from original users that were retweeted by the trolls. All the tweets are in the same domain because they were all tweeted or retweeted by the Russian Troll accounts. We have separated the tweets into two groups, Russian troll generated and Non-Russian troll generated to see if we can build a classifier that identifies if the tweets were generated by the Russian troll accounts.

The preprocessing performed removed a number of tweets and text anomalies in order to ensure that no data was double counted, and that non-word characters as part of hashtags and usernames were removed. These transformations are detailed below.

After performing pre-processing, we found that there were a number of duplicate tweets. After removing those, we found that there were 121,202 retweets of non-Russian trolls (which we will be using as our “real” tweets) and 52,734 original tweets from the Russian accounts, accounting for 69.7% and 30.3% of tweets respectively of the 173, 936.

|  |
| --- |
| **Preprocessing Steps** |
| Remove duplicates. |
| Remove any tweets that are retweets from other Russian trolls. |
| Remove “RT @<username>:” from tweets |
| Remove “@” only from usernames |
| Remove “#” only from hashtags |

Note that we also performed the experiment with both stemmed and unstemmed data.

Methods

Classification

Classification of tweets as Russian or non-Russian was first performed without the inclusion of topics as feature. This classifier thus acted as our baseline. We used a neural net classifier from scikit-learn, Python’s machine learning library, and a training set of Russian and non-Russian tweets transformed into feature vectors. The features included were:

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature Name** | **Description** | **Type** | **Instances** |
| Retweeted | A boolean flag to indicate if the tweet has been retweeted. | Boolean | 1 |
| URL | A boolean flag to indicate if the tweet contains a URL. | Boolean | 1 |
| Mentions | A boolean flag to indicate if the tweet contains a ‘mention,’ or reference to another user account. | Boolean | 1 |
| In reply to | A boolean flag to indicate if the tweet is in reply to a preexisting tweet. | Boolean | 1 |
| Bag of words | Many boolean flags (the size of the vocabulary) to indicate if the word in question (which acts as an individual feature) is present in the tweet. | Boolean | Size of tweet vocabulary (excluding hashtags, urls, and usernames) |
| Hashtags | Many boolean flags (the size of the hashtag vocabulary) to indicate if the hashtag in question (which acts as an individual feature) is present in the tweet. | Boolean | Size of hashtag vocabulary |
| Reply count | The number of replies to this tweet. | Integer | 1 |
| Retweet count | The number of times this tweet was retweeted. | Integer | 1 |
| Favorite count | The number of users who favorited this tweet. | Integer | 1 |

This classifier was then tested on a held-out data set to determine precision, recall, and f-1 score. These measures were later used to evaluate our final results, as discussed in section 4.3.

Topic Modeling

In order to determine the importance of topic in the detection of Russian politically-driven tweets, we created a topic model to predict common abstract topics of our training data as a whole. Our topic model, a word2vec model trained on Google News, created tweet embeddings through the summation into a single vector of the individual token embeddings in the body of each tweet. Those vectors, consisting of summed word embeddings for each token in the tweet, were mapped onto a semantic space by the word2vec model. The semantic space was then partitioned in order to determine topic groups, or clusters using K-Means clustering, a partitioning technique which determines k clusters of n objects by grouping a given object with the cluster with the nearest mean.

It is important to note that in building the topic model we did not differentiate Russian from non-Russian tweets. In creating a topic model from the dataset as a whole (with the exception of held out testing data) topic clusters are independent of tweet origin, thus the inclusion of topic as a feature in the classifier will only be successful if the k clusters naturally divide between Russian and non-Russian tweets. A cluster made almost entirely of Russian tweets in a model including both Russian and non-Russian tweets would indicate that the Russian tweets can be characterized, and thus classified by topic.

In order to improve the model, the number of clusters, k, was tweaked to achieve greater improvements in the classifier from section 4.3 as compared to the classifier from section 4.1.

\*While we are still experimenting with the number of topics, or number of clusters, to be pulled from our topic model, the topics found will be used as additional features in a replica of the neural net classifier from section 4.1 in order to determine topic significant in the classification of Russian vs. non-Russian tweets.

Classification with Topics

Our second round of classification replicated our first classifier exactly with the exception of the inclusion of topic as a feature. In the creation of tweet vectors for the classification of any given tweet, we determined the nearest topic cluster, using the model from section 4.2, and demarcated the tweet as belonging to a specific topic cluster. For example, if a tweet were determined by the topic model to belong to the i cluster out of k clusters, the feature vector for that tweet would have the integer i in the index associated with tweet topic.

The classifier was then evaluated using a held-out data set of Russian and non-Russian tweets to determine precision, recall, and f-1 score. These measures were compared with those of the classifier from section 4.1 to determine improvement based upon the inclusion of topic as a feature. The resulting measures were improved through the shifting of the number of topic clusters considered, as is discussed in section 4.2.

# Results

# Conclusions and Future Work

References

Adam Badawy, Emilio Ferrara, and Kristina Lerman. 2018. *Analyzing the Digital Traces of Political Manipulation: The 2016 Russian Interference Twitter Campaign.* In Proceedings of The Web Conference (WWW’18). ACM, New York, NY, USA, 8 pages.

Asborn Otteson Steinskog, Jonas Foyn Therkelsen, Björn Gambäck. 2017. Twitter Topic Modeling by Tweet Aggregation. In Proceedings of the 21st Nordic Conference of Computational Linguistics. Pages 77-86.

Henkel, Linda A., and Mark E. Mattson, "*Reading Is Believing: The Truth Effect and Source Credibility*, " Consciousness and Cognition, Vol. 20, No. 4, December 2011, pp. 1705–1721.

Paul C, Matthews M. *The Russian "Firehose of Falsehood " Propaganda Model.* RAND Corporation. 2016.

Pornpitakpan, Chanthika. "*The Persuasiveness of Source Credibility: A Critical Review of Five Decades’ Evidence,* " Journal of Applied Social Psychology, Vol. 34, No. 2, February 2004, pp. 243–281.

Stephen G. Harkins and Richard E. Petty. "*The Multiple Source Effect in Persuasion:  The Effects of Distraction*, " Personality and Social Psychology Bulletin, Vol. 7, No. 4, December 1981

Stephen G. Harkins and Richard E. Petty, "*Information Utility and the Multiple Source Effect, "* Journal of Personality and Social Psychology, Vol. 52, No. 2, 1987.

|  |  |
| --- | --- |
| 200  201  202  203  204  205  206  207  208  209  210  211  212  213  214  215  216  217  218  219  220  221  222  223  224  225  226  227  228  229  230  231  232  233  234  235  236  237  238  239  240  241  242  243  244  245  246  247  248  249 | 250  251  252  253  254  255  256  257  258  259  260  261  262  263  264  265  266  267  268  269  270  271  272  273  274  275  276  277  278  279  280  281  282  283  284  285  286  287  288  289  290  291  292  293  294  295  296  297  298  299 |