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 corpora of Russian twitter data containing over 200,000 deleted tweets. This corpora 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 seperated 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.

Data

Layout

References

Alfred. V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling, volume 1*. Prentice-Hall, Englewood Cliffs, NJ.

American Psychological Association. 1983. *Publications Manual.* American Psychological Association, Washington, DC.

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. [Alternation](http://dl.acm.org/citation.cfm?doid=322234.322243). *Journal of the Association for Computing Machinery*, 28(1):114-133. https://doi.org/10.1145/322234.32224.

Association for Computing Machinery. 1983. *Computing Reviews*, 24(11):503-512.

James Goodman, Andreas Vlachos, and Jason Naradowsky. 2016. [Noise reduction and targeted exploration in imitation learning for abstract meaning representation parsing](http://aclweb.org/anthology/P16-1001). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 1–11. <https://doi.org/10.18653/v1/P16-1001>.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

Mary Harper. 2014. [Learning from 26 languages: Pro- gram management and science in the babel program](http://aclweb.org/anthology/C14-1001). In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin City University and Association for Computational Linguistics, page 1. <http://aclweb.org/anthology/C14-1001>.

Alexander V. Mamishev and Murray Sargent. 2013. *Creating Research and Scientific Documents Using Microsoft Word*. Microsoft Press, Redmond, WA.

Alexander V. Mamishev and Sean D. Williams. 2010. *Technical Writing for Teams: The STREAM Tools Handbook*. Wiley-IEEE Press, Hoboken, NJ.

|  |  |
| --- | --- |
| 600  601  602  603  604  605  606  607  608  609  610  611  612  613  614  615  616  617  618  619  620  621  622  623  624  625  626  627  628  629  630  631  632  633  634  635  636  637  638  639  640  641  642  643  644  645  646  647  648  649 | 650  651  652  653  654  655  656  657  658  659  660  661  662  663  664  665  666  667  668  669  670  671  672  673  674  675  676  677  678  679  680  681  682  683  684  685  686  687  688  689  690  691  692  693  694  695  696  697  698  699 |