# Exploring Political Polarization Through Russian Troll Tweets

**Avinash Boppana**
Princeton University
aboppana@princeton.edu

**Michael Gao**
Princeton University
mg25@princeton.edu

**Josh Gardner**
Princeton University
jg41@princeton.edu

**Michelle Yuen**
Princeton University
mjyuen@princeton.edu

## Abstract

Russian interference in the 2016 election continues to be a source of national controversy. It is unclear what strategies the Russian trolls are employing. Predicting the efficacy of trolls may help researchers to better understand the Russian strategy, why it seems to be working, and how to combat it effectively. FiveThirtyEight has partnered with Clemson researchers to release a dataset of roughly 3 million troll tweets [2]. This paper applies LDA and k-means clustering to identify latent structures in the tweets. After identifying these structures, we attempt to construct regressors (linear, ridge, elastic net) to predict the number of followers on Twitter. This paper also performs binary classification on the content of tweets to predict whether an account is a left troll or right troll. We found that both left and right-leaning trolls consistently focused on social identity, and depicted identity as a binary "us" versus "them" issue.

## 1 Introduction

Political polarization in the US has been increasing, and many political analysts believe that the polarization is being at least partially driven by deliberate Russian efforts to manipulate public opinion on social media sites like Twitter [1]. However, analysts have failed to describe in detail specific strategies used by the trolls. Identifying these strategies may help researchers better understand the causes of polarization. Moreover, this may help policymakers better combat polarization by flagging divisive and inflammatory information. One of the purposes of this study was to identify keywords and topics used by effective political trolls on both the left and right. We also seek to determine the strategies used by the most effective trolls by applying regression to predict number of followers and then examining the highest weighted features.

Using k-means clustering, we were able to partition each tweet, already put into bag-of-words format, into clusters based on content. In addition, we used LDA to reveal latent structures in the troll tweets. We trained categorical classifiers such as Gaussian Naive Bayes, AdaBoost, and Random Forest to distinguish between the tweets using features from bag-of-words encoding for each tweet. Finally, we predict the number of followers for tweets using linear regression, ridge regression, and elastic net regression.

We found that both liberal and conservative trolls focused overwhelmingly on social identity, and moreover that liberal tweets as well as tweets containing videos of police and protesters clashing had higher numbers of followers than conservative tweets.

## 2 Related Work

Researchers in data science and political science fields have become increasingly concerned with identifying troll Twitter accounts online. Researchers from the University of Michigan and Georgia Tech used high-accuracy classifiers to separate accounts into troll and non-troll accounts [**?**]. Prior researchers have also focused on tweets pertaining specifically to news outlets [4]. In their study, significant results were achieved using LDA with 26 topics on a dataset of over 200,000 tweets. This suggests LDA as a viable tool for analyzing the dataset.

Previous work has not focused on the effectiveness of trolls, as measured by followers and likes. Previous work gives no indication what makes a troll effective. Moreover, work done on left-wing and right-wing trolls only set out to establish categories of trolls. Relatively little has been done to examine the specific language used by trolls in each category. Accordingly, this paper sets out to identify specific words and phrases that are disproportionately used by effective trolls on the left and the right, as well as to examine the relationships between tweet categories and follower counts.

## 3 Description of Data and Processing

We obtained our data from FiveThirtyEight, which published the dataset in collaboration with Clemson University [5]. The dataset contains 2,973,371 tweets generated by 2,848 Twitter accounts traced to the Russian Internet Research Agency, a governmental cyber-warfare initiative, focused on inciting political division through social media [2]. The data, collected by Clemson University Researchers Linvill and Warren, contains all tweets from these accounts from June 19, 2015 to June 2018. For each tweet, the tweet content, user handle, date posted, number of interactions (likes, retweets, etc.), number of followers and following for the account, and "account type"(a rough marker of political affiliation or account behavior as coded by Linvell and Warren) are recorded.

Given the variety of analyses we perform, we processed and partitioned that data into two datasets, one for binary classification and one for unsupervised learning and regression. First, we amalgamated 10 tweet datasets from FiveThirtyEight's GitHub repository into one super dataset, and removed samples that contained non-English content for the sake of comprehension. Then, we removed samples that were classified as re-tweets, as we wanted the content of a tweet to be original, rather than duplicated. We restricted the binary dataset to tweets that were classified as "left troll" or "right troll", as well as tweets that were classified as "fearmonger." Finally, to reduce the size of the dataset and improve computational speed, we randomly sampled 30,000 tweets each from the "left troll" and "right troll" tweets partitions for the sake of computational efficiency, however since there were only 9671 fearmonger tweets, we just took all of them. To generate the dataset for unsupervised learning and regression, we created a subset of the entire data set that only contained tweets from authors with at least 20 tweets, since we wanted only tweets from authors who consistently interacted on Twitter.

## 4 Methods

### K-means Clustering and Latent Dirichlet Allocation

We used K-means clustering using a Euclidean distance function to group together similar tweets into functional clusters based on the bag-of-words representations of tweet content. Optimization of cluster number was performed by evaluating mean distance from the centroid of each cluster. Latent Dirichlet Association was similarly performed on the bag-of-words representation of each tweet, and for each latent variable each feature is given a pseudocount indicating the likelihood of that latent variable to generate that feature.Perplexity scores were used to optimize the number of topics used in LDA, resulting in an optimum of five(See Appendix). LDA was performed using Scikit Learn's LatentDirichletAllocation with the hyper-parameter n_components set to 5.

### Binary Categorical Classification

For the tweets labelled "left troll" and "right troll", we performed supervised learning to predict based on tweet content, whether the account was a left or right troll. Prior to classification, all tweets were converted to bag-of-words representations. To mark a statistical baseline, since the training set had equal instances of left and right trolls, we examined the performance of a baseline predictor that randomly assigned predictions with equal chance. Then, we trained three categorical

classifiers: Gaussian Naive Bayes, AdaBoost, and Random Forest. After assessment of these models, we selected the most accurate(Random Forest) and finetuned its hyperparameters to improve performance. Then, we used this model to classify tweets labelled "fearmonger" to predict whether they were left or right-leaning.

**Regression on Follower Count**

We wanted to determine which features are associated with highly effective accounts(as measured by followers). To accomplish this, we performed linear regressions attempting to predict follower count. We performed this using regular linear regression, ridge regression,and elastic net regression. Ridge regression reduces overfitting by selectively altering the weights of the terms in our regression. Moreover, we also performed elastic net regression. By combining L1 regularization and L2 regularization, elastic net regression reduces the number of features and asjusts the weights of the remaining features. While this could avoid overfitting and yield a more concise model, dropping key features could also lead to loss of important variations causing a decrease in accuracy. We wanted to determine which model was the most effective. To do this, we calculated MSE (mean-squared error) to determine the deviation of our predicted follower counts from the actual follower counts in testing sets.

## 5 Results and Discussion

**K-means Clustering Results**

We conducted k-means clustering to find hidden structures in the data. First, we wanted to determine which value of k yielded the most appropriate structures as measured by silhouette scores. To determine this, we ran k-means clustering with values of k ranging from 2 to 20, and averaged the silhouette coefficients for 1000 randomly selected points in each cluster for each value of k. We repeated this process 10 times for each value of k and took the average of the averages over each sample. Further, we calculated the average distance of a sample to its assigned cluster for a held-out test set. We found that the best best scores consistently occurred for a cluster number of k=2 (please see figures 5 and 6 in the Appendix for the corresponding graphs). This indicates that for k=2, the data points are assigned to more appropriate clusters than they were for the other values of k.

We wanted to determine whether there are certain features which are uniquely important to each of the clusters, and whether these features seem to be semantically related. To do this, we ran k-means clustering with k=2 and constructed a "disproportionality value" by subtracting the base proportion of tweets that contain the feature for the entire dataset from the proportion of tweets that contain the feature in the cluster of interest. Note that both proportions have values between 0 and 1 and thus the "disproportionality value" must range between -1 and 1.

The only semantic pattern we observed was that cluster 1 contained lots of references to debates both Democratic and Republican, whereas cluster 2 did not and instead contained a disproportionately high number of links, particularly to new articles. Since we found relatively few semantic patterns, we decided to try to further subdivide each cluster in hopes of revealing more clear semantic patterns. To do this, we ran k-means clustering with k=2 on each of the clusters to yield four new sub-clusters, in order to further investigate the sub-structure within the higher-level clusters.

| All Troll Tweets | | | | | | | |
|---|---|---|---|---|---|---|---|
| Cluster 1 | | | | Cluster 2 | | | |
| Cluster 1-1 | | Cluster 1-2 | | Cluster 2-1 | | Cluster 2-2 | |
| Feature | Factor | Feature | Factor | Feature | Factor | Feature | Factor |
| islamkil | 0.005455 | demndeb | 0.903822 | rt | 0.039047 | co | 0.696609 |
| blacklivesmatt | 0. 004242 | demdeb | 0.571992 | kochfarm | 0.023584 | http | 0.691616 |
| rt | 0.004030 | carson | 0.088262 | turkey | 0.014553 | trump | 0.130711 |
| gopdeb | 0.003977 | hillari | 0.063608 | blacklivesmatt | 0.011572 | ’ | 0.084598 |
| vegasgopdeb | 0.003228 | pari | 0.058663 | usda | 0.010011 | ” | 0.061992 |
| love | 0.003111 | ben | 0.054151 | peopl | 0.008768 | “ | 0.061448 |

Figure 1: Highest Significance Features for Clusters Based on Disproportionality Values)

We once again isolated features which occurred at disproportionate rates for each sub-cluster and looked for patterns. Cluster 1 was split into clusters 1-1 and 1-2, where cluster 1-1 seemed to focus on Republican debates, ”gopdeb”, ”vegasgopdeb” and politically charged words like ”islamkil”, ”blacklivesmatt”, ”stopthegop” and ”stopislam.” Cluster 1-2 focused on democratic debates ”demndeb”, ”demdeb”, ”debat”, and ”tonight”, strangely also focused on Republican primary candidate Ben Carson(features ”ben” and ”carson”) and generally controversial terms such as those that appeared in cluster 1-1. Notably, the disproportionality values were much higher for cluster 1-2 (with the top values approaching the maximum of 1), suggesting that cluster 1-2 is very distinct from the rest of the dataset whereas cluster 1-1 is only somewhat distinct with disproportionality values much closer to 0.

Cluster 2-1 had the following features in its top 10: ”kochfarm”, ”turkey”, ”usda”, ”foodpoison”, ”thanksgiv.” Upon further investigation, there was a known hoax perpetrated by Russian trolls to incite public panic around a fictional food poisoning incident involving Delaware-based Koch Farms [3]. Strangely enough this cluster also included ”blacklivesmatt” in its top five features.

Cluster 2-2 had the following features: ”co”, ”http”, ”trump”, ”break”,”video”, and various quotation marks. This suggests that this cluster contains links to new videos and external websites, many of which are related to Donald Trump. Ostensibly, ”break” could be related to breaking news. Moreover, the quotations suggest the possibility of citing news sources or quoting politicians. Some of the features related to links (”co” and ”http”) in cluster 2-2 had high disproportionality values close to 0.7 whereas the features in cluster 2-1 generally had disproportionality values much closer to 0, suggesting that cluster 2-2 is more distinct from the rest of the dataset and moreover that it is distinct because it contains links. See figure 8 in the Appendix for a high-level categorization of the clusters and sub-clusters.

**Latent Dirichlet Allocation**

We wanted to determine if there was some semantically meaningful latent structure in the dataset. To do this we performed Latent Dirichlet Allocation. We used n_components=5 because perplexity was lowest for 5 topics, in a range from 1 to 10 topics (See figure 7 in the Appendix for the corresponding graph). The analysis appears to have yielded semantically interpretable results, which are displayed in the following table:

| Latent Tweet 1 | | Latent Tweet 2 | | Latent Tweet 3 | | Latent Tweet 4 | | Latent Tweet 5 | |
|---|---|---|---|---|---|---|---|---|---|
| Feature | Pseudocount | Feature | Pseudocount | Feature | Pseudocount | Feature | Pseudocount | Feature | Pseudocount |
| co | 8579.117990 | peopl | 1811.573281 | demndeb | 2790.198885 | co | 20325.080784 | http | 14065.74011 |
| http | 8559.475078 | get | 1685.977597 | demdeb | 2001.198948 | http | 20293.981639 | co | 13702.81327 |
| kochfarm | 2321.199674 | black | 1463.123377 | islamkil | 1892.199059 | trump | 1595.595500 | rt | 5893.518407 |
| turkey | 1459.199526 | love | 1391.963715 | trump | 1587.205031 | ‘ | 1159.844030 | ’ | 2338.266991 |
| trump | 1281.922497 | like | 1293.498590 | gopdeb | 1375.198673 | ” | 1083.690445 | trump | 2192.970318 |
| ’ | 1264.446652 | life | 1200.106230 | refuge | 1073.221940 | “ | 1075.395005 | tcot | 1415.216335 |
| usda | 961.199897 | make | 1073.236914 | brussel | 1048.199155 | video | 925.013642 | blacklivesma | 1381.252632 |
| thanksgiv | 911.199511 | one | 1021.255912 | vegasgopdeb | 1041.198673 | polic | 805.835641 | pjnet | 1336.716321 |
| foodpoison | 816.199873 | day | 922.249122 | u | 900.299440 | break | 783.408555 | … | 1271.143784 |
| happi | 807.877726 | staywok | 916.197896 | need | 897.115607 | protest | 585.803642 | ‘ | 1002.192277 |

Figure 2: Latent Tweets Created Using LDA

Latent tweet one appears to be focused on links (”co” and ”http”) to news sources surrounding a food poisoning incident around Thanksgiving (”thanksgiv”, ”usda”, ”foodpoison”, ”turkey”). Moreover, the feature ”kochfarm” appears to be tied to Koch Farms, a large turkey farm in Schuylkill, Pennsylvania. Upon further investigation, it became clear that these tweets are referencing a food poisoning hoax perpetrated by Russian Trolls surrounding Koch Farms in 2015.

Latent tweet two appears to use words that may be associated with liberalism. The features ”peopl”, ”black”, ”love”, ”like”, ”life”, and ”staywok” seem to suggest social justice oriented content with positive sentiment. The emphasis on wokeness, racial identity, and love strongly

suggests that the tweets are left-leaning.

Latent tweet three seems to focus on debates(”demndeb”, ”demdeb”, ”gopdeb”, ”vegasgopdeb”) and moreover on topics potentially related to islamophobia (”islamkil”, ”refuge”). Furthermore, the feature ”brussel” is ostensibly related to the 2015 terrorist attacks in Brussels for which ISIS claimed responsibility. Taken with the feature ”refuge” which is likely associated with Syrian refugees, this suggests a general focus on issues surrounding Islam and Syrian migrants. The fact that these features have been grouped with both the Republican and Democratic debates may suggest that the trolls were trying to tie migrants and terrorism to the debates in public consciousness.

Latent tweet four has features involving links (”co”, ”http”) and features such as ”police” and ”protest” which suggest generally civil unrest and potential conflicts with law enforcement either as a cause of the unrest or as a result. Moreover, three of the top 6 features are forms of quotation mark, and the seventh most significant feature is the word ”video.” Taken with the link features, this suggests the use of quotations and videos surrounding police and protesters. Ostensibly, these tweets contain news content surrounding police brutality or surrounding protesters clashing with police.

Latent tweet five similarly seems to also be focused on links (”co” and ”http”) and quotations however seems to be more conservative. Features such as ”tcot” which is an acronym for ”top conservatives on twitter” and ”pjnet” which stands for patriot journalist network suggest that the links are to conservative media. Moreover, the feature ”blacklivesma” clearly refers to Black Lives Matter. This suggests that these tweets may be linking to conservative media sites which are discussing Black Lives Matter in some way. See figure 9 in the Appendix for a high-level categorization of the 5 latent variables.

**Binary Classification of Left and Right Trolls**

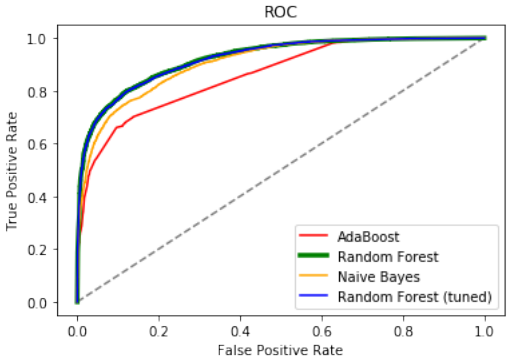
In order to see if the text of the tweet could be used to predict whether a troll was left or right-leaning, we examined a baseline predictor, and trained three classifiers: AdaBoost, Naive Bayes, and Random Forest. We used a bag-of-words representation in this case with a vocabulary of 2000 words. Additionally, we performed an 80:20 training and test split on the original dataset after reducing the account category label in the dataset to two possible labels: left troll and right troll. The resulting accuracies and scores are as follows.

| | **Accuracy** | **Precision** | **Recall** | **Specificity** |
|---|---|---|---|---|
| **Baseline** | 0.5073 | 0.5028 | 0.5028 | 0.5055 |
| **AdaBoost** | 0.7792 | 0.8381 | 0.6971 | 0.8653 |
| **Naive Bayes** | 0.8126 | 0.8177 | 0.8038 | 0.8208 |
| **Random Forest** | 0.8228 | 0.8444 | 0.8261 | 0.8478 |
| **Random Forest (tuned, n=95)** | 0.8393 | 0.8532 | 0.8310 | 0.8572 |

Because the baseline predictor would guess whether a tweet was left or right with equal probability, it was to be expected that its accuracy was lower than trained models. We note that the trained classifiers all had relatively similar performance in regards to accuracy, but the Random Forest had the highest initial accuracy and scores, and AdaBoost had the lowest. For all initially trained classifiers, we used their default settings without any tweaking of hyperparameters. Then, we wanted to see if we could improve the performance of Random Forest, and used grid search with cross-validation to finetune its hyperparameters. From this we obtained the optimal setting of n=95, which represents the number of trees in a forest. All other settings were kept as default. We saw some very slight improvement in accuracy and other scores from this finetuning.

We see similar relative performance when plotting the ROC curves against one another. Because the tuned Random Forest model had very minimal improvement, their ROC curves are very closely aligned. We then further examined what words for the AdaBoost model and the tuned Random Forest model were identified as most important for identification.

The majority of words are commonly associated with politically-charged tweets, notably ”trump”, ”hillary”, and movements like ”blacklivesmatter”. We also noticed that ”rt” for retweets was a




| Binary Classification of Most Important Words | |
|---|---|
| **AdaBoost** | **Random Forest** |
| trump | rt |
| https | https |
| rt | blacklivesmatter |
| vegas | black |
| vegasgopdebate | trump |
| foxnews | blacktwitter |
| cop | hillary |
| liberal | staywoke |
| antifa | breaking |
| blacktwitter | police |

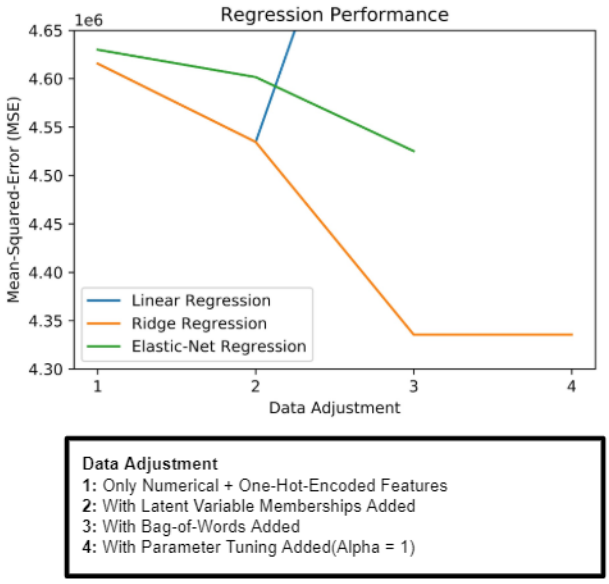
top feature, which means some tweets were misidentified as not being retweets since this dataset intended to remove all retweeted content. Furthermore, another top word for classification was ”https” which is commonly associated with links. This can go to show that troll tweets often try to further their narrative at other sites and redirecting their followers to another page. Overall, we can see relatively accurate prediction of whether an account is a left or right troll based on the content of their tweet. To take this classification a step further, we took labelled Fearmonger tweets (which were not specifically labelled as left or right trolls), and used our finetuned Random Forest model to attempt to classify the Fearmonger tweets as either more left or more right-leaning. We then saw that 4996 of the 9671 Fearmonger tweets were identified as right trolls. This aligns with results obtained by the Clemson researchers who noted that a slight majority of Fearmonger accounts were right-leaning, and some accounts even started out tweeting like right trolls before switching over to a fearmonger approach.

**Results of Regression on Number of Followers**

We wanted to determine which factors predict the number of followers for each tweet. To do this, we ran a variety of linear models achieving the following results: We wanted to determine which regression model had the highest performance based on MSE. To this, we started by performing all three regressions on a dataset restricted to only include numerical features and one-hot encoded features. We found that ridge regression and standard regression achieved virtually identically results, and elastic net under-performed both. This suggests that elastic net is removing important features and decreasing predictive accuracy.

Furthermore, we wanted to determine whether the Latent Dirichlet Allocation we performed earlier could be used to improve the performance of our models. To test this, we added the five LDA latent tweet topic proportions for each tweet as additional features. We then re-ran all three models (ride, elastic net, linear) again on this expanded dataset. This improved the performance of all three models suggesting that LDA captured latent structures associated with troll effectiveness as measured by number of followers.

We wanted to determine information about whether the specific wording of each tweet would




improve the performance of the models. To do this, we created binary features for each word in tweets using a bag-of-words representation and added them to the data. For the bag-of-words representation, we removed standard stopwords and used a high threshold count (40) to increase variance by ensuring that only the most significant words were included in the bag-of-words. After adding the new features, we ran all three forms of regression again and compared the computed MSEs. We found that MSE increased substantially for the linear regression model suggesting that the new words added noise rather than making signals clearer. However, MSE decreased significantly for both ridge and elastic net suggesting that the L1 and L2 parameters were able to remove some of the noise from the new features and were able to get a stronger signal. Ridge had the lower MSE of the two.

Accordingly, we decided to tune its alpha hyper-parameter to optimize performance. We performed grid search from an alpha = 0.001 to an of alpha = 1, over a list of 6 distinct values. As the grid search instantiated 6 different Ridge Regression Models, the performance was evaluated using 5-fold cross validation of the the train set, measuring performance with an averaged $r^2$ coefficient of determination. The best performing model had an alpha value of 1, which suggests that there were many features in our feature-set that need to be weighted minimally and were not important for the sake of regressing follower count. The result of hyper-parameter tuning further reduced the mean-squared-error (see figure 3 above).

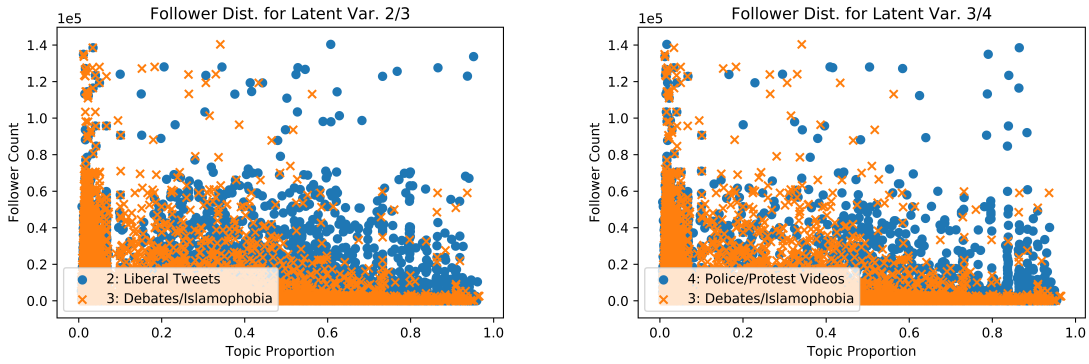
| **Heavily Weighted Regression Features** | |
|---|---|
| **Feature** | **Weight** |
| altright | 2706.698687 |
| amber | 2624.121964 |
| berkeley | 2610.878235 |
| daca | -2125.574106 |
| donlemon | 2737.293083 |
| irma | -2262.143364 |
| manchest | 2218.450844 |
| nytim | 2879.226464 |
| sandra | -3393.529541 |
| steeler | -2009.780956 |

The majority of the heavily weighted regression features clearly relate to controversial issues or major public events. Obviously, ”altright” refers to far-right white nationalist movements. ”Berkeley” refers to the University of California, Berkeley, which is well-known for protests and political controversy. ”Daca” suggests an interest in DACA, Deferred Action for Childhood Arrivals, a Fed-

eral policy granting citizenship to the children of undocumented immigrants which has generated substantial controversy. "Irma" refers to hurricane Irma, a category 5 hurricane which devastated Puerto Rico in 2017, and which lead to the Trump administration being heavily criticized for failing to support Puerto Rico. "Manchest" likely refers to the Manchester Arena Bombing in which 23 people were killed in a suicide bombing on May 22, 2017 by an Islamic terrorist. "Sandra" appears to refer to Sandra Bland, the African American woman who was found hanged in her cell on July 13, 2015 three days after being arrested for a traffic violation. These features all suggest that posting about controversial issues relates to having a high number of followers. Possibly, controversial posts get more attention.

**Analysis of Latent Tweets and Tweet Popularity**

We wanted to determine if there was a relationship between latent tweets and follower counts. To do this we plotted the distribution of followers for latent tweets and looked for patterns. We found two distinct patterns which are displayed here:




Liberal tweets(latent tweet two) and police/protest videos(latent tweet four) appear to have significantly more followers than tweets focused on debates and islamophobia(latent tweet three). Assuming that latent tweet three reflects alt-right conservatism, this suggests that trolling related to police brutality, protests and liberalism is actually more effective, at least on Twitter. Notably, there were roughly 2.5 times as many right leaning tweets in our dataset than there were left leaning tweets(16,328 left leaning to 40,352 right leaning). Nevertheless the left leaning ones appear to be much more popular. Ostensibly, this reflects the fact that younger people are both more likely to use Twitter and to be liberal.

## 6 Conclusion

We set out to determine key strategies used by the Russian trolls as well as to identify which strategies are particularly effective. We identified distinct categories of troll tweets using unsupervised learning methods like LDA and k-means clustering. Based on these, we determined that across all tweet categories, trolls are focusing on social identities to create division. Similarly, using regression methods, we found that political controversy is associated with high follower counts. Moreover, we found that trolls often use links to existing US media to disseminate divisive narratives. Notably, links to videos involving police and protesters have vastly higher follower counts then links to conservative media and fake news. This suggests that trolling using videos, particularly potentially violent or conflict-focused videos, may be especially effective. Moreover, the higher follower counts for liberal troll tweets as opposed to conservative troll tweets suggests that Twitter-based Russian information/misinformation campaigns targeted at liberals may be as effective or more effective than those targeted at conservatives. On the whole, our analysis suggests that controversial posts particularly relating to issues of interest to liberals may be a critical part of the Russian strategy. While misinformation campaigns directed at conservatives have been a major focus of US media, information/misinformation campaigns targeted at liberals have received less attention. In future work, more advanced natural language processing techniques could be used to shed more light on tweet features. The bag of words representation fails to capture the effects of negation and sarcasm. However there are novel techniques which could yield better results, such as Google Research's BERT.

## Acknowledgments

We would like to thank the COS 424 course staff for their help with this project, and all their teaching this semester.

## References

[1] Adam Badawy, Emilio Ferrara, and Kristina 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)*, pages 258–265. IEEE, 2018.

[2] FiveThirtyEight. Russian troll tweets, Aug 2018.

[3] Dave Fusaro. Before meddling in u.s. elections, the russians picked on food.

[4] Michael Jensen. Russian trolls and fake news: Information or identity logics? *Journal of International Affairs*, 71(1.5):115–124, 2018.

[5] Oliver Roeder. Why we're sharing 3 million russian troll tweets, Jul 2018.

## Appendix

### Additional Scores for Binary Classification

| | Baseline | AdaBoost | Naive Bayes | Random Forest | Random Forest (tuned) |
|---|---|---|---|---|---|
| **TN** | 3001 | 5193 | 4926 | 5088 | 5120 |
| **FP** | 2935 | 808 | 1075 | 913 | 866 |
| **FN** | 3096 | 1817 | 1177 | 1043 | 1011 |
| **TP** | 2968 | 4182 | 4822 | 4956 | 5003 |

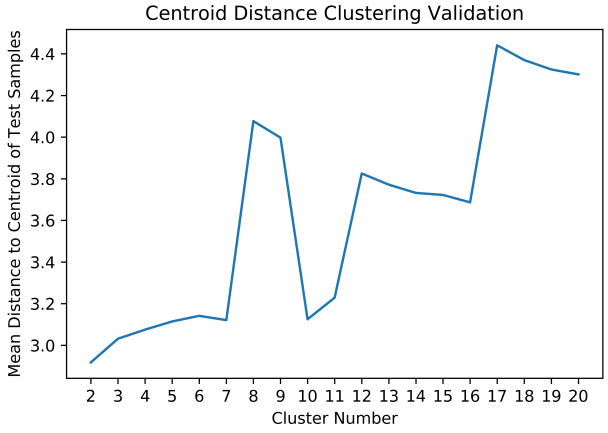



Figure 3: Cluster Centroid Distances

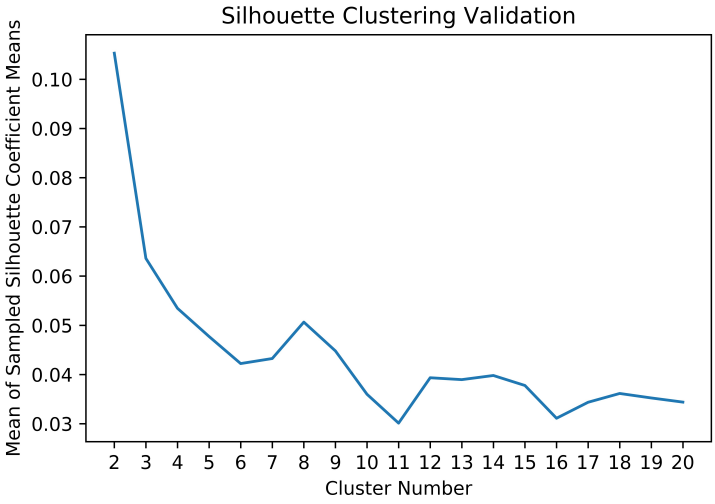



Figure 4: Cluster Silloutte

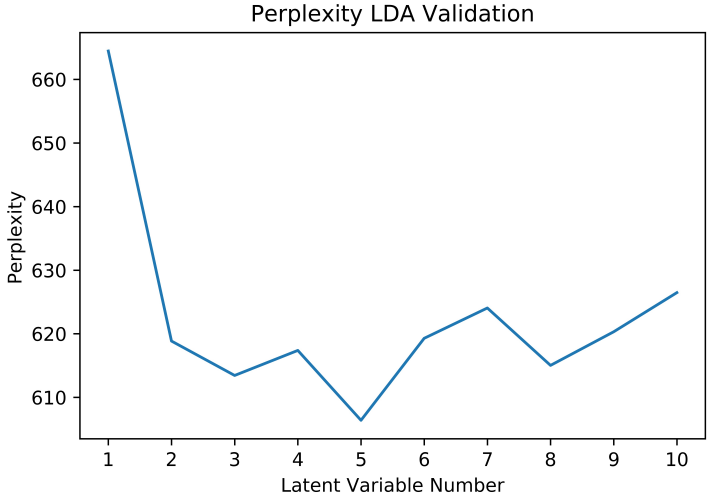



Figure 5: Perplexity Values for LDA

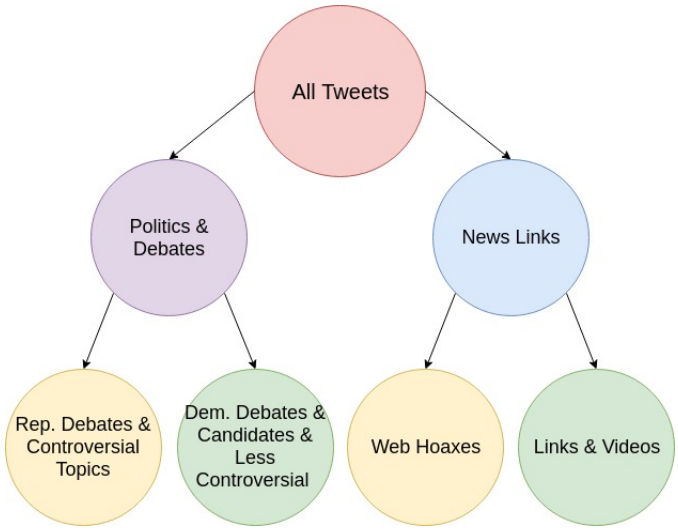



Figure 6: Cluster Tree

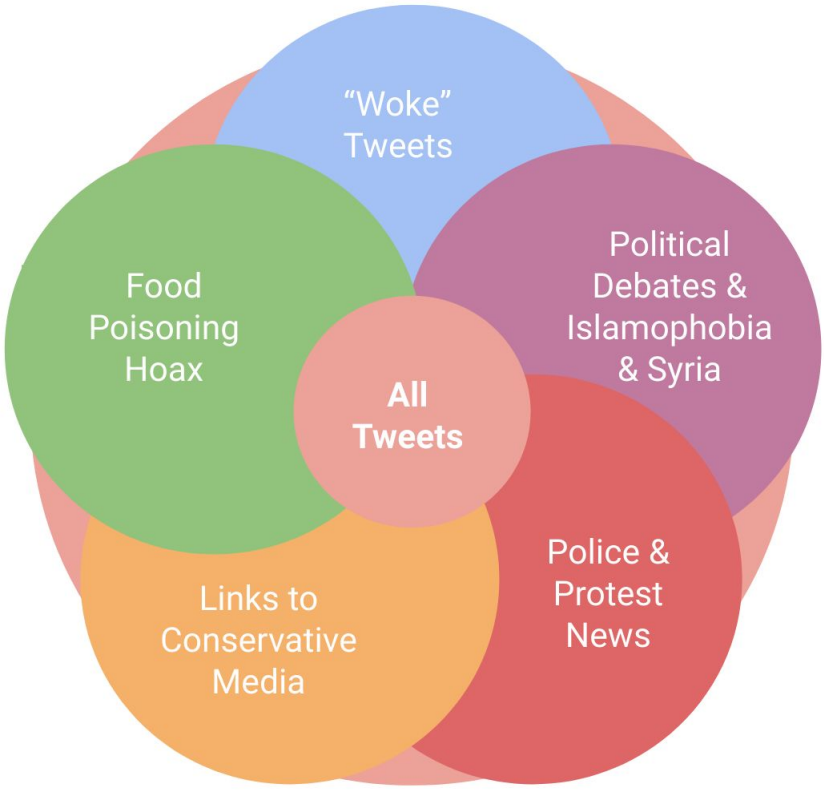



Figure 7: Latent Variables