Who Falls for Online Political Manipulation?

Adam Badawy
Information Sciences Institute,
University of Southern California
abadawy@usc.edu

Kristina Lerman
Information Sciences Institute,
University of Southern California
lerman@isi.edu

Emilio Ferrara
Information Sciences Institute,
University of Southern California
emiliofe@usc.edu

ABSTRACT

Social media, once hailed as a vehicle for democratization and the promotion of positive social change across the globe, are under attack for becoming a tool of political manipulation and spread of disinformation. A case in point is the alleged use of trolls by Russia to spread malicious content in Western elections. This paper examines the Russian interference campaign in the 2016 US presidential election on Twitter. Our aim is twofold: first, we test whether predicting users who spread trolls' content is feasible in order to gain insight on how to contain their influence in the future; second, we identify features that are most predictive of users who either intentionally or unintentionally play a vital role in spreading this malicious content. We collected a dataset with over 43 million election-related posts shared on Twitter between September 16 and November 9, 2016, by about 5.7 million users. This dataset includes accounts associated with the Russian trolls identified by the US Congress. Proposed models are able to very accurately identify users who spread the trolls' content (average AUC score of 96%, using 10-fold validation). We show that political ideology, bot likelihood scores, and some activity-related account meta data are the most predictive features of whether a user spreads trolls' content or not.

KEYWORDS

Political Manipulation, Russian Trolls, Bots, Social Media

ACM Reference Format:
Adam Badawy, Kristina Lerman, and Emilio Ferrara. 2019. Who Falls for Online Political Manipulation?. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3308560.3316494

1 INTRODUCTION

The initial optimism about the role of social media as a driver of social change has been fading away, following the rise in concerns about the negative consequences of malicious behavior online. Such negative outcomes have been particularly evident in the political domain. The spread of misinformation [29, 33] and the increasing role of bots [4] in the 2016 US presidential elections has increased the interest in automatic detection and prediction of malicious actors.

In this study, we focus on the role of Russian trolls in the recent US presidential election. Trolls are usually described as users who intentionally "annoy" or "bother" others in order to elicit an emotional response. They post inflammatory messages to spread discord and cause emotional reactions [25]. In the context of the 2016 US election, we define trolls as *users who exhibit a clear intent to deceive or create conflict.* Their actions are directed to harm the political process and cause distrust in the political system. Our definition captures the new phenomenon of paid political trolls who are employed by political actors for a specified goal. The most recent and important example of such phenomenon is the Russian "troll farms"—trolls paid by the Russian government to influence conversations about political issues aimed at creating discord and hate among different groups [2, 12].

Survey data from the Pew Research Center [13] show that two-thirds of Americans get their news from Social Media. Moreover, they are being exposed to more political content written by ordinary people than ever before. Bakshy et al. [3] report that 13% of posts by Facebook users—who report their political ideology—are political news. This raises the question of how much influence the Russian trolls had on the national political conversation prior to the 2016 US election, and how much influence such trolls will have in the upcoming elections. Although we do not discuss the effect that these trolls had on the political conservation prior to the election, we focus our efforts in this paper on two questions: **RQ1**: *Can we predict which users will become susceptible to the manipulation campaign by spreading content promoted by Russian trolls?* **RQ2**: *What features distinguish users who spread trolls' messages?*

The goal of these questions is, first, to test whether it is possible to identify the users who will be vulnerable to manipulation and participate in spreading the messages trolls post. We refer to such users as *spreaders* in this paper. Our second goal is to better understand what distinguishes spreaders form non-spreaders. If we can predict who will become a spreader, we can design a counter-campaign, which might stop the manipulation before it achieves its goal.

For this study, we collected Twitter data over a period of seven weeks in the months leading up to the election. We obtained a dataset of over 43 million tweets generated by about 5.7 million distinct users between September 16 and November 9, 2016. First, we cross-referenced the list of Russian trolls published by the US Congress with our dataset and found that 221 Russian trolls exist in our data. Next, we identified the list of users who retweeted the trolls. We gather important features about the users and use a machine learning framework to address the questions posed earlier.

We used different machine learning classifiers on different models (each model includes a subset of the features, with the full model including all the features). We are able to achieve over 90% average AUC score in the full model using Gradient Boosting. In terms of

WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316494

feature importance, political ideology is the most prominent feature; number of followers, statuses (no. tweets), and bot scores were also in the top most predictive features. We finally discuss robustness and sanity checks that we carried out to generalize our results.

2 RELATED LITERATURE

The use of trolls and bots in political manipulation campaigns around the globe is well documented through an array of reports by mainstream media outlets and academics (*cf.* Tucker et al. [32] for a comprehensive review). This phenomenon is not entirely new: researchers warned about the potential for online political manipulation for over a decade [15, 17]. Reports tracking and studying this phenomenon date back to the early 2010s [21, 27, 28]. Since then, an increasing account of such events has been recorded in the context of several elections, both in the United States [4, 19, 20, 35, 38, 39] and all over the world, including the U.K. [16], and Italy [6].

Although trolls do not necessarily need to be automated accounts, in many cases bots play a substantial role in exerting influence on social media [22]. Bessi and Ferrara [4] estimated that up to 400K bots were responsible for posting 3.8 million tweets in the last month of the 2016 US presidential election, which is one-fifth of the total volume of online conversations they collected. Russian political manipulation campaigns did not only target the US: there is evidence of Russian interference in German and British elections, and the Catalonian referendum [30]. Russian interference was also reported in the 2017 French presidential election, where bots were detected during the so-called *MacronLeaks* disinformation campaign [9]. Moreover, a recent NATO report claims that around 70% of accounts tweeting in Russian and directed at Baltic countries and Poland are bots.

3 DATA COLLECTION

3.1 Twitter Dataset

We created a list of hashtags and keywords that relate to the 2016 U.S. Presidential election. The list was crafted to contain a roughly equal number of hashtags and keywords associated with each major Presidential candidate: we selected 23 terms, including five terms referring to the Republican Party nominee Donald J. Trump (#donaldtrump, #trump2016, #neverhillary, #trumppence16, #trump), four terms for Democratic Party nominee Hillary Clinton (#hillaryclinton, #imwithher, #nevertrump, #hillary), and several terms related to debates.

By querying the Twitter Search API continuously and without interruptions between September 15 and November 9, 2016, we collected a large dataset containing 43.7 million unique tweets posted by nearly 5.7 million distinct users. Table 1 reports some aggregate statistics of the dataset. The data collection infrastructure ran inside an Amazon Web Services (AWS) instance to ensure resilience and scalability. We chose to use the Twitter Search API to make sure that we obtained all tweets that contain the search terms of interest posted during the data collection period, rather than a sample of unfiltered tweets. This precaution we took avoids known issues related to collecting sampled data using the Twitter Stream API that had been reported in the literature [23].

Table 1: Descriptive statistics of Twitter data

Statistic	Count
# of Tweets	43,705,293
# of Retweets	31,191,653
# of Distinct Users	5,746,997
# of Tweets/Retweets with a URL	22,647,507

Table 2: Descriptive statistics on Russian trolls

	Value
# of Russian trolls	2,735
# of trolls in our data	221
# of trolls wrote original tweets	85
# of original trolls' tweets	861

Table 3: Descriptive statistics of spreaders

	Value
# of spreaders	40,224
# of times retweeted trolls	83,719
# of spreaders with original tweets	28,274
# of original tweets	>1.5 Million
# of other tweets and retweets	>12 Million

3.2 Russian Trolls

We used a list of 2,752 Twitter accounts identified as Russian trolls that was compiled and released by the U.S. Congress.[1] Table 2 offers some descriptive statistics of the Russian troll accounts. Out of the accounts appearing on the list, 221 exist in our dataset, and 85 of them produced original tweets (861 tweets). Russian trolls in our dataset retweeted 2,354 other distinct users 6,457 times. Trolls retweeted each other only 51 times. Twitter users can choose to report their location in their profile. Most of the self-reported locations of accounts associated with Russian trolls were within the U.S. (however, a few provided Russian locations in their profile), and most of the tweets were from users whose location was self-reported as Tennessee and Texas (49,277 and 26,489 respectively). Russian trolls were retweeted 83,719 times, but most of these retweets were for three troll accounts only: 'TEN_GOP', received 49,286 retweets; 'Pamela_Moore13', 16,532; and 'TheFoundingSon', 8,755. These three accounts make up for over 89% of the times Russian trolls were retweeted. Overall, Russian trolls were retweeted by 40,224 distinct users.

3.3 Spreaders

Users who rebroadcast content produced by Russian trolls, hereafter referred to as *spreaders*, may tell a fascinating story, thus will be the subject of our further investigation. Out of the forty thousand total spreaders, 28,274 of them produced original tweets (the rest only generated retweets). Overall, these 28K spreaders produced over 1.5 Million original tweets and over 12 Million other tweets and retweets—not counting the ones from Russian trolls (cf., Table 3).

4 DATA ANALYSIS & METHODS

In order to answer the questions posed in this paper, we gather a set of features about the users to *(i)* predict the spreaders with the highest accuracy possible and *(ii)* identify feature(s) which best distinguish spreaders from the rest. Table 4 shows all the features we evaluated in this paper, grouped under the following categories:

[1] See Recode.net: https://www.recode.net/2017/11/2/16598312/

Table 4: Features used to describe users in our study

Metadata	LIWC	Engagement	Activity	Other
# of followers	Word Count	Retweet variables	# of characters	Political Ideology
# of favourites	Postive Emotion	Mention variables	# of hashtags	Bot Score
# of friends	Negative Emotion	Reply variables	# of mentions	Tweet Count
Status count	Anxiety	Quote variables	# of urls	
Listed count	Anger			
Default Profile	Sadness			
Geo-enabled	Analytic			
Background-image	Clout			
Verified	Affection			
Account Age	Tone			

Figure 1: Feature correlation for all users in the dataset.

Metadata, Linguistic Inquiry and Word Count (LIWC), Engagement, Activity, and Other variables.

To understand what each variable in the Metadata and LIWC categories means, see the Twitter API documentation and [24], respectively. The Activity variables convey the number of characters, hashtags, mentions, and URLs produced by users, normalized by the number of tweets they post. Tweet Count, under Other, is the number of user's tweets appearing in our dataset. The remaining variables are more involved and warrant a detailed explanation: we explain how Political Ideology, Bot Scores, and Engagement variables were computed in the following sections. One may wonder how much the features evaluated here correlate with each other, and whether they provide informative signals in terms of predictive power about the spreaders. Figure 1 shows that, besides Engagement variables, most of the features are not highly correlated among each other (Pearson correlation is shown, results do not vary significantly for Spearman correlation). There are however a few notable exceptions: Word Count and Tweet Count, LIWC Positive Emotion and Affection, Anxiety and Anger—these pairs all show very high correlation. This is not surprising, considering that these constructs are conceptually close one another. As for the Engagement variables, we can see a "rich get richer" effect here, where users who have higher scores in terms of some of the sub-features in the Engagement category, are also higher in other sub-features. For example, by construction the *Retweet h-index* will be proportional to the number of times a user is retweeted, and similarly for replies, quotes and mentions—all these Engagement features are explained in great detail in a section §4.3.

Table 5: Liberal & Conservative domain names

Liberal	Conservative
www.huffingtonpost.com	www.breitbart.com
thinkprogress.org	www.thegatewaypundit.com
www.politicususa.com	www.lifezette.com
shareblue.com	www.therebel.media
www.dailykos.com	theblacksphere.net

4.1 Political Ideology

4.1.1 Classification of Media Outlets. We classify users by their ideology based on the political leaning of the media outlets they share. We use lists of partisan media outlets compiled by third-party organizations, such as AllSides[2] and Media Bias/Fact Check.[3] The combined list includes 249 liberal outlets and 212 conservative outlets. After cross-referencing with domains obtained in our Twitter dataset, we identified 190 liberal and 167 conservative outlets. We picked five media outlets from each partisan category that appeared most frequently in our Twitter dataset and compiled a list of users who tweeted from these outlets. The list of media outlet domain names for each partisan category is in Table 5.

We used a polarity rule to label Twitter users as liberal or conservative depending on the number of tweets they produced with links to liberal or conservative sources. In other words, if a user had more tweets with links to liberal sources, he/she would be labeled as liberal and vice versa. Although the overwhelming majority of users include links that are either liberal or conservative, we remove any users that had equal number of tweets from each side[4]—this to avoid the conundrum of breaking ties with some arbitrary rule. Our final set of labeled users include 29,832 users.

4.1.2 Label Propagation. We used *label propagation*[5] to classify Twitter accounts as liberal or conservative, similar to prior work [5]. In a network-based label propagation algorithm, each node is assigned a label, which is updated iteratively based on the labels of the node's network neighbors. In label propagation, a node takes the most frequent label of its neighbors as its own new label. The algorithm proceeds updating labels iteratively and stops when the labels no longer change (see [26] for more information). The algorithm takes as parameters *(i)* weights, in-degree or how many times node i retweeted node j; *(ii)* seeds (the list of labeled nodes). We fix the seeds' labels so they do not change in the process, since this seed list also serves as our ground truth.

We construct a retweet network where each node corresponds to a Twitter account and a link exists between pairs of nodes when one of them retweets a message posted by the other. We use the 29K users mentioned in the media outlets sections as seeds, those who mainly retweet messages from either the liberal or the conservative media outlets in Table 5, and label them accordingly. We then run label propagation to label the remaining nodes in the retweet network.

To validate results of the label propagation algorithm, we applied stratified 5-fold cross validation to the set of 29K seeds. We train

[2]See AllSides: https://www.allsides.com/media-bias/media-bias-ratings
[3]See Media Bias Fact Check: https://mediabiasfactcheck.com/
[4]We use five categories, as in left, left center, center, right center, right, to make sure we have a final list of users who are unequivocally liberal or conservative and do not fall in the middle. The media outlet lists for the left/right center and center were compiled from the same sources.
[5]We used the algorithm in the Python implementation of IGraph [7]

Table 6: User breakdown by political ideology

	Liberal	Conservative
# of users	>3.4 M	>1 M
# of trolls	107	108
# of spreaders	1,991	38,233

the algorithm on four-fifths of the seed list and test how it performs on the remaining one-fifth. The averge precision and recall scores are both over 91%.

To further validate the labeling algorithm, we notice that a group of Twitter accounts put media outlet URLs as their personal link/website. We compile a list of the hyper-partisan Twitter users who have the domain names from Table 5 in their profiles and use the same approach explained in the previous paragraph (stratified 5-fold cross-validation). The average precision and recall scores for the two validation methods we use is 91% and 93% respectively, cementing our confidence in the performance of the labeling algorithm.

4.2 Bot Detection

Determining whether either a human or a bot controls a social media account has proven a very challenging task [10, 31]. We use an openly accessible solution called Botometer (a.k.a. BotOrNot) [8, 40], consisting of both a public Web site (https://botometer.iuni.iu.edu/) and a Python API (https://github.com/IUNetSci/botometer-python), which allows for making this determination with high accuracy. Botometer is a machine-learning framework that extracts and analyses a set of over one thousand features, spanning six sub classes:

User: Meta-data features that include the number of friends and followers, the number of tweets produced by the users, profile description and settings.

Friends: Four types of links are considered here: retweeting, mentioning, being retweeted, and being mentioned. For each group separately, botometer extracts features about language use, local time, popularity, etc.

Network: Botometer reconstructs three types of networks: retweet, mention, and hashtag co-occurrence networks. All networks are weighted according to the frequency of interactions or co-occurrences.

Temporal: Features related to user activity, including average rates of tweet production over various time periods and distributions of time intervals between events.

Content: Statistics about length and entropy of tweet text and Part-of-Speech (POS) tagging techniques, which identifies different types of natural language components, or POS tags.

Sentiment: Features such as: arousal, valence and dominance scores [36], happiness score [18], polarization and strength [37], and emotion score [1].

We utilize Botometer to label all the spreaders, and we get bot scores for over 34K out of the total 40K spreaders. Since using Botometer to get scores all non-spreaders (i.e., over 5.7M users) would take an unfeasibly long time (due to Twitter's restrictions), we randomly sample the non-spreader user list and use Botometer to get scores for a roughly equivalent-size list of non-spreader users. The randomly-selected non-spreader list includes circa 37K users. To label accounts as bots, we use the fifty-percent threshold which has proven effective in prior studies [8]: an account is considered to be a bot if the overall Botometer score is above 0.5. Figure 2 shows the probability distribution for spreaders vs. non-spreaders. While most of the density is under the 0.5 threshold, the mean of spreaders (0.3) is higher than the mean of non-spreaders. Additionally, we used a t-test to verify that the difference is significant at the 0.001 level (p-value).

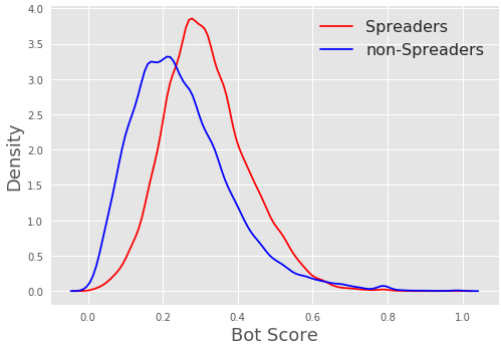


Figure 2: Probability density distributions of bot scores assigned to spreaders (red) and non-spreaders (blue).

As for the plots in Figure 3, it is evident that the spreaders are different on almost all the Botometer subclass scores, except for the temporal features. The differences in all plots are statistically significant ($p < 0.001$). Besides, looking at the distributions, we can see that the difference in user characteristics (metadata), friends, and network distributions, are substantively different as well. Moreover, the mean of spreaders is higher in all the subclass features.

4.3 Engagement

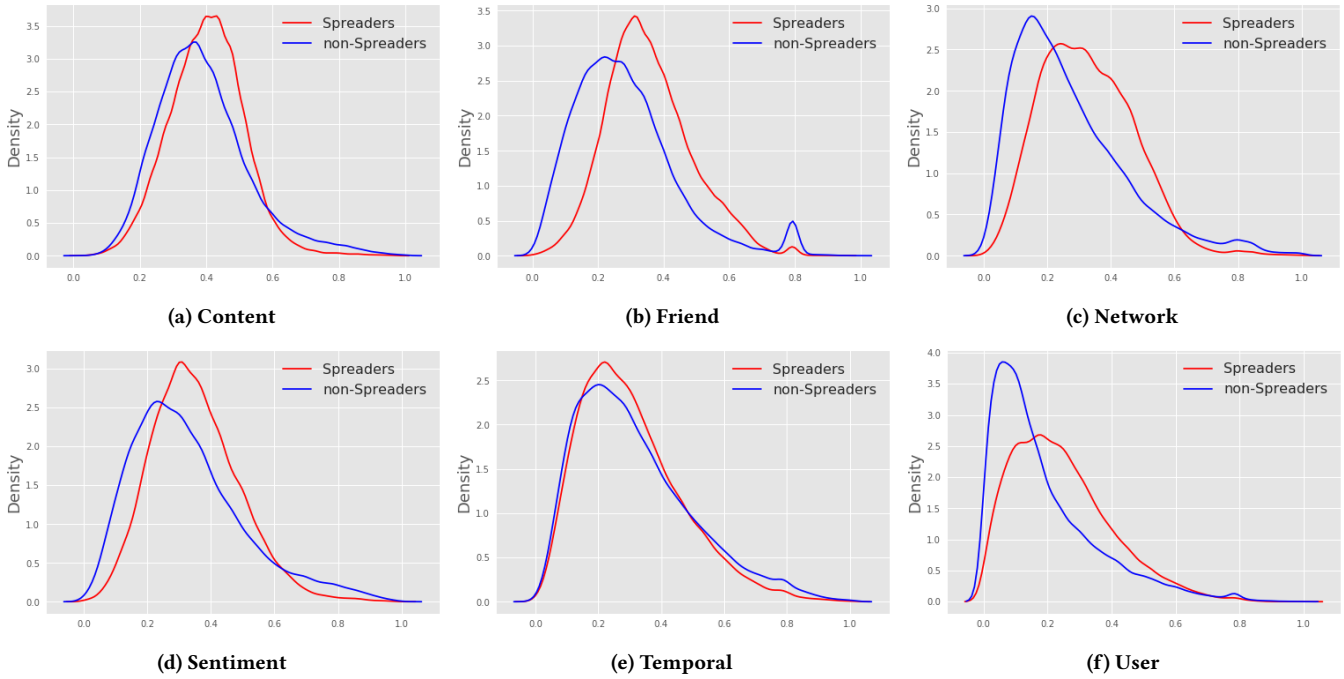
We plan to measure user engagement in four activities: retweets, mentions, replies, and quotes. Engagement of a user is measured through three components: the quantity, longevity, and stability in each activity. For instance, for a set of N users, this measure would calculate the engagement index score of user $i \in N$ by including the following:

1) number of retweets, replies, mentions, and quotes by $N - i$ users for user i;
2) time difference between the last and the first quote, reply, and retweet per tweet;
3) consistency of mentioning, replying, retweeting, and quoting by $N - i$ users for user i across time (per day);
4) number of unique users who retweeted, commented, mentioned, and quoted user i

Item three is measured using h-index [14]. The measure captures two notions: how highly referenced and how continuously highly referenced a user is by other members in the network. This measure was originally proposed to quantify an individual's scientific research output. In this context, a user has index h if for h days, he/she is referenced at least h times and in all but h days no more than h times.

5 RESULTS

Predicting spreaders on the original dataset may be considered a daunting task: only a relatively small fraction of users engaged



(a) Content **(b) Friend** **(c) Network**

(d) Sentiment **(e) Temporal** **(f) User**

Figure 3: Distribution of the probability density of Botometer subfeature scores for spreaders vs. nonspreaders.

with Russian trolls' content (about 40K out of 5.7M users). However, for the same reason, if a model were to trivially predict that no user will ever engage with Russian trolls, the model would be accurate most of the time (i.e., most users won't be spreaders), even if its recall would be zero (i.e., the model would never correctly predict any actual spreaders)—provided that we want to predict spreaders, this model would not be very useful in practice. In other words, our setting is a typical machine-learning example of a highly-unbalanced prediction task.

To initially simplify our prediction task, we created a balanced dataset that is limited to users who have bot scores. Will get back to the original prediction task on the highly-unbalanced dataset later. This balanced dataset has about 72K users, with 34K spreaders and 38K non-spreaders. To test our ability to detect spreaders and to see which features are most important in distinguishing between the two groups, we leverage multiple classifiers and multiple models: the first model serves as a baseline with each model including more variables until we reach the full model. Since our goal was not that to devise new techniques, we used four off-the-shelf machine learning algorithms: Extra Trees, Random Forest, Adaptive Boosting, and Gradient Boosting. We train our classifiers using Stratified 10-fold cross-validation with the following preprocessing steps *(i)* replace all categorical missing values with the most frequent value in the column *(ii)* replace missing values with the mean of the column.

Table 7 shows all the models we evaluate, from the simplest baseline model (Metadata) to the full model that includes all the features we present in Table 4.

For Gradient Boosting, which is the best performing classifier among the four we evaluate, we obtained average AUC scores for the 10 folds that range from 85% to 96%. Figure 4 shows the ROC curve plots for each model (using the fold/model with the highest

Table 7: Models: from Baseline (Metadata) to Full

Model	Features
1	Metadata
2	Metadata + LIWC
3	Metadata + LIWC + Activity
4	Metadata + LIWC +Activity + Engagement
5	Metadata + LIWC +Activity + Engagement + Other

AUC score among the trained ones). The jump from 89% to 96% for the AUC scores from Model 4 to 5 shows that the addition of bot scores and political ideology are meaningful in distinguishing spreaders from non-spreaders (the legend in Figure 4 shows the average AUC score for each model). To better understand the contribution of the features in predicting the target values (i.e., spreader vs. non-spreader), we look at the variable importance plot of the Gradient Boosting results for Model 5. The *Variable Importance* plot (cf., Figure 5) provides a list of the most significant variables in descending order by a mean decrease in the Gini criterion. The top variables contribute more to the model than the bottom ones and can discriminate better between spreaders and non-spreaders. In other words, features are ranked based on their predictive power according to the given model. Figure 5 shows that, according to Model 5 and Gradient Boosting, political ideology is the most predictive feature, followed by number of followers, statuses/tweets count (obtained from the metadata), and bot score, in a descending order of importance. The plot does not show all the features, since the omitted features contribute very little to the overall predictive power of the model.

Feature importance plots reveal which features contribute most to classification performance, but they do not tell us the nature of the relationship between the outcome variable and the predictors. Although predictive models are sometime used as black boxes,

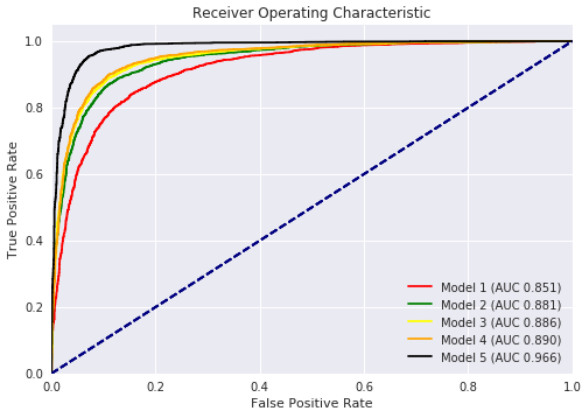


Figure 4: Area under the ROC curve plot for five GB models

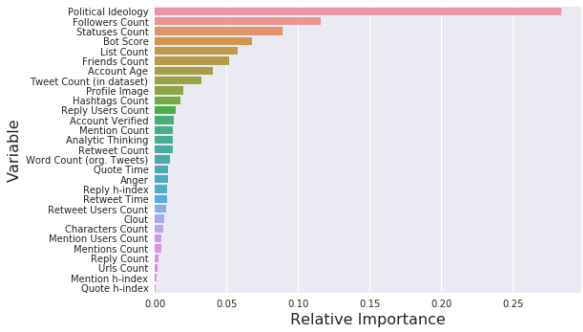
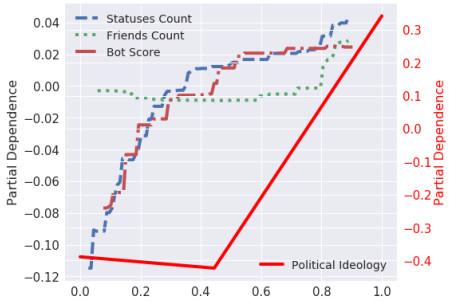


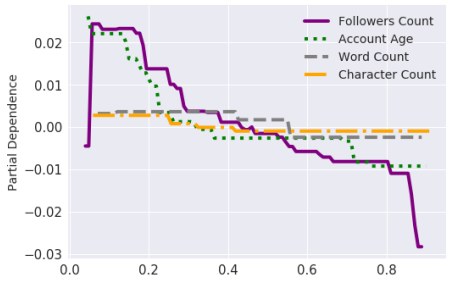
Figure 5: Relative importance of the features (GB full model)

Partial Dependence plots (cf., Fig. 6) can tell us a lot about the structure and direction of the relationship between the target and independent variables. They show these relationships after the model is fitted, while marginalizing over the values of all other features. The dependency along the x-axis captures the range of each feature, with that feature values normalized between 0 and 1. Political ideology should be considered in the range from 0 (to identify left-leaning users), to 1 (for right-leaning ones).

Using Partial Dependence, we illustrate that the target variable (spreader) has positive relationships with the following features: political ideology, statuses count, bot scores, and friends count. Figure 6a visualizes these relationships (we put political ideology on a different y-axis in order to show that its magnitude of influence on the target variable is significantly higher compared to all other features, including downward trend features in Figure 6b). This suggests that moving from left to right political leaning increases the probability of being a spreader; larger number of posts, more friends, and higher bot scores are also associated with higher likelihood of being a spreader. On the other hand, we can see that the outcome variable has a negative relationship with followers count, account age, characters count, and word count, as shown in Figure 6b. This means that having fewer followers, having a recently-created account, posting shorter tweets with fewer words, are all characteristics associated with higher probability of being a spreader.



(a) Upward Trends



(b) Downward Trends

Figure 6: Partial Dependence plots for some important features (GB full model). Plots show the dependence of the outcome variable (spreader) on each feature, marginalizing over all the other features (x-axis values are CDF-normalized).

5.1 Robustness and Generalization Analysis

Original vs. Rebalanced Data: Going back to the original highly-unbalanced dataset, we aim to validate the results above using two strategies: *(i)* we run Gradient Boosting (with the same preprocessing steps) on the whole dataset of 5.7M for the five models we outlined in table 7; *(ii)* we run Gradient Boosting on models without imputations and with all missing observations deleted. For the first approach, the average AUC scores ranged form 83% for the baseline model to 98% for the full model. In terms of feature importance, bot score and political ideology are the most predictive features as expected in the full model prediction. For the second approach, due to the sparsity of some features, the overall number of observations decreases significantly when these features are added. Putting the overall number of observations aside, the average ROC scores for a 10-fold validation for the roughly same set of models specified earlier range from 84% to 91%. In terms of feature importance, political ideology is the most important feature in the full model, with status count and bot scores following. In summary, the results above remain consistent when validating on the highly-unbalanced prediction.

Excluding bots: Removing Bots from the balanced dataset then predicting spreaders using Gradient Boosting yields the same AUC scores for models 1-5, ranging from 85% to 96%. In terms of feature selection, political ideology is the best predictive feature by a wide margin then list, followers, posts (tweets), and friends count in this exact order. Although the ordering is different for the balanced dataset with bots, these features are the top 4 (if we exclude bot scores) most predictive features with or without the inclusion of bots.

Error Analysis: Few trends emerge when we compare correctly labeled users vs the ones that are misclassified in the balanced dataset. Correctly labeled users, in terms of metadata, have more followers, favourites, friends, and posts. Moreover, they have higher variance in terms of word count and anger words and they use more analytical words.

6 DISCUSSION AND CONCLUSION

The results in previous section show that *(i)* with some insight on users who spread or produce malicious content, we are able to predict those that will spread trolls' messages; *(ii)* in the case of the 2016 US presidential election, political ideology was highly predictive of who is going to spread trolls' messages. Fig. 3b, 3c, and 3f show that spreaders and non-spreaders are significantly different on the dimensions of friends, network, and user metadata, with spreaders having higher bot scores on all three. Basic metadata features give a strong signal in terms of differentiating spreaders from non-spreaders, along with the bot score (cf. Fig. 5).

Looking at the partial dependence plots, we can deduce that spreaders write a lot of tweets (counting retweets as well), have higher bot scores, and tend to be more conservative (conservative is labeled as the highest numerical value in the political ideology feature). Also, since the range of the y-axis tells us about the range of influence a feature has on the target value, it is evident that political ideology has by far the most influence on distinguishing between spreaders and non-spreaders. On the other hand, we can also deduce that spreaders do not write much original content, tend not have that many followers, and have more recently established user accounts. In the downward trends in Figure 6b, we can see that followers count and account age have more influence on the target value in comparison to the other features in this plot. To conclude, this paper focused on predicting spreaders who fall for online manipulation campaigns. We believe that identifying likely victims of political manipulation campaigns is the first step in containing the spread of sponsored content [11, 34]. Access to reliable and trustworthy information is a cornerstone of any democratic society. Declining trust of citizens of democratic societies in mainstream news and their increased exposure to content produced by ill-intended sources poses a danger to democracy.

Acknowledgements. This work was supported by the Air Force Office of Scientific Research (#FA9550-17-1-0327).

REFERENCES

[1] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, et al. 2011. Sentiment analysis of twitter data. In *ACL*. 30–38.

[2] Adam Badawy, Emilio Ferrara, and Kristina Lerman. 2018. Analyzing the Digital Traces of Political Manipulation: The 2016 Russian Interference Twitter Campaign. In *ASONAM*. 258–265.

[3] Eytan Bakshy, Solomon Messing, and Lada A Adamic. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science* (2015).

[4] Alessandro Bessi and Emilio Ferrara. 2016. Social bots distort the 2016 US Presidential election online discussion. *First Monday* 11 (2016).

[5] M Conover, B Gonçalves, J Ratkiewicz, et al. 2011. Predicting the Political Alignment of Twitter Users. In *Social Computing*. 192–199.

[6] Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, et al. 2017. The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race. In *WWW*. 963–972.

[7] Gabor Csardi and Tamas Nepusz. 2006. The igraph software package for complex network research. *InterJournal, Complex Systems* (2006).

[8] Clayton Allen Davis, Onur Varol, Emilio Ferrara, et al. 2016. Botornot: A system to evaluate social bots. In *WWW*. 273–274.

[9] Emilio Ferrara. 2017. Disinformation and social bot operations in the run up to the 2017 French presidential election. *First Monday* 22 (2017).

[10] Emilio Ferrara, Onur Varol, Clayton Davis, et al. 2016. The rise of social bots. *Commun. ACM* 59, 7 (2016), 96–104.

[11] Emilio Ferrara, Onur Varol, Filippo Menczer, et al. 2016. Detection of promoted social media campaigns. In *ICWSM*. 563–566.

[12] Theodore P Gerber and Jane Zavisca. 2016. Does Russian propaganda work? *The Washington Quarterly* 39, 2 (2016), 79–98.

[13] Jeffrey Gottfried and Elisa Shearer. 2016. *News Use Across Social Medial Platforms 2016*. Pew Research Center.

[14] Jorge E Hirsch. 2005. An index to quantify an individual's scientific research output. *PNAS* 102, 46 (2005), 16569.

[15] Philip Howard. 2006. *New media campaigns and the managed citizen*.

[16] Philip Howard and Bence Kollanyi. 2016. Bots,# strongerin, and# brexit: Computational propaganda during the uk-eu referendum. (2016).

[17] Tim Hwang, Ian Pearce, and Max Nanis. 2012. Socialbots: Voices from the fronts. *Interactions* 19, 2 (2012), 38–45.

[18] Isabel M Kloumann, Christopher M Danforth, Kameron Decker Harris, et al. 2012. Positivity of the English language. *PloS one* 7, 1 (2012).

[19] Bence Kollanyi, Philip Howard, and Samuel Woolley. 2016. Bots and automation over Twitter during the first US Presidential debate. (2016).

[20] Alice Marwick and Rebecca Lewis. 2017. Media manipulation and disinformation online. *Data & Society Research Institute* (2017).

[21] Panagiotis T Metaxas and Eni Mustafaraj. 2012. Social media and the elections. *Science* 338, 6106 (2012), 472–473.

[22] Bjarke Mønsted, Piotr Sapieżyński, Emilio Ferrara, et al. 2017. Evidence of Complex Contagion of Information in Social Media: An Experiment Using Twitter Bots. *Plos One* 12, 9 (2017), e0184148.

[23] Fred Morstatter, Jürgen Pfeffer, Huan Liu, et al. 2013. Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. In *ICWSM*. 400–408.

[24] James Pennebaker, Ryan Boyd, Kayla Jordan, et al. 2015. The development and psychometric properties of LIWC2015. (2015).

[25] Whitney Phillips. 2015. *This is why we can't have nice things: Mapping the relationship between online trolling and mainstream culture*.

[26] Usha Nandini Raghavan, Réka Albert, and Soundar Kumara. 2007. Near linear time algorithm to detect community structures in large-scale networks. *Physical review E* 76, 3 (2007), 036106.

[27] Jacob Ratkiewicz, Michael Conover, Mark Meiss, et al. 2011. Truthy: mapping the spread of astroturf in microblog streams. In *WWW*.

[28] Jacob Ratkiewicz, Michael Conover, Mark R Meiss, et al. 2011. Detecting and tracking political abuse in social media. *ICWSM* (2011).

[29] Samantha Shorey and Philip Howard. 2016. Automation, Algorithms, and Politics: A Research Review. *Int. J Comm.* 10 (2016).

[30] Massimo Stella, Emilio Ferrara, and Manlio De Domenico. 2018. Bots increase exposure to negative and inflammatory content in online social systems. *Proceedings of the National Academy of Sciences* (2018).

[31] VS Subrahmanian, Amos Azaria, Skylar Durst, et al. 2016. The DARPA Twitter bot challenge. *Computer* 49, 6 (2016).

[32] Joshua Tucker, Andrew Guess, Pablo Barberá, et al. 2018. Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature. (2018).

[33] Joshua A Tucker, Yannis Theocharis, Margaret E Roberts, et al. 2017. From liberation to turmoil: social media and democracy. *Journal of democracy* 28, 4 (2017), 46–59.

[34] Onur Varol, Emilio Ferrara, Filippo Menczer, et al. 2017. Early Detection of Promoted Campaigns on Social Media. *EPJ Data Science* 6, 13 (2017).

[35] Yu Wang, Yuncheng Li, and Jiebo Luo. 2016. Deciphering the 2016 US Presidential Campaign in the Twitter Sphere: A Comparison of the Trumpists and Clintonists. In *ICWSM*. 723–726.

[36] Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior research methods* 45, 4 (2013), 1191–1207.

[37] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *HLT-EMNLP*.

[38] Samuel C Woolley. 2016. Automating power: Social bot interference in global politics. *First Monday* 21, 4 (2016).

[39] Samuel C Woolley and Philip Howard. 2016. Automation, Algorithms, and Politics: Introduction. *Int. Journal of Commun.* 10 (2016).

[40] Kai-Cheng Yang, Onur Varol, Clayton A Davis, et al. 2019. Arming the public with artificial intelligence to counter social bots. *Human Behavior and Emerging Technologies* (2019), e115.