# Russian Facebook Propaganda Detection with Classification Models

Paul Franklin, Donald Cooper, Jan Danel, and Tiger Hu

6/2/2020

## Table of Contents

**"We will never know whether the Russian intervention was determinative in such a close election. ... What does matter is this: The Russians successfully meddled in our democracy and our intelligence agencies have concluded they will do so again."**
***- Congressman Adam Schiff***

## I. Introduction

On February 16, 2018 Special Counsel Robert S. Mueller III indicted 13 Russian individuals and three Russian organizations for engaging in operations to interfere with U.S. political and electoral processes, including the 2016 presidential election. These organizations included the Internet Research Agency: the group responsible for disseminating thousands of advertisements designed to foment popular frustration toward the U.S. government and undermine its democratic institutions. Between 2015 and 2017, the Agency paid Russian actors to misuse online platforms and carry out their clandestine operations. In conjunction with this investigation, authorities ordered Facebook to locate and disclose data relating to the Russian Internet Research Agency sponsored Ads placed on their site.[1] This data is the basis of our analysis, for it (In combination with data recently gathered by Propublica) allows us to ask: "Is it possible to accurately predict whether or not a Facebook ad is Russian propaganda, given a combination of variables such as thematic content, syntax, target location, number of factual references, etc.?"

Recently, researchers released a fascinating new collection of Facebook political advertisement data. Specifically, ProPublica published a survey of Facebook ad metrics gleaned from the feeds of their users.[2] On top of this, we learned that NYU affiliates scraped and compiled ad data from Facebook's ad Library back in 2018.[3] These sets appeared very compatible with the Facebook ad propaganda data released by U.S. authorities back in 2017 (We will refer to this as IRA data). We devised that a binary response classification model could separate NYU and Propublica ads from IRA ones. Such a model would essentially look for differences between features of Facebook ads of domestic origin, and those written by IRA agents. With substantial differences, a model might accurately sort foreign malicious ads from domestic prosocial ones. The results of this analysis are presented in section VII. We believe such a model will reliably detect new propaganda since scholars agree that the IRA will continue several detectable tactics. For one, Russians have recently disseminated posts about police brutality toward Blacks.[4,5] They are also still masquerading as second amendment proponents. The Agency will undoubtedly ramp up its voter discouragement campaigns in the days leading up to the 2020 presidential election- making it particularly urgent that we identify and take down their pages. Disconcertingly, the most recent propaganda bust was over seven months ago. Of the 51 Russian accounts disabled that

---

[1] "Social Media Advertisements." U.S. House of Representatives Permanent Select Committee on Intelligence. Accessed April 28, 2020. https://intelligence.house.gov/social-media-content/social-media-advertisements.htm.
[2] ProPublica. "Political Advertisements from Facebook." ProPublica Data Store, March 19, 2019. https://www.ProPublica.org/datastore/dataset/political-advertisements-from-facebook.
[3] Edelson, Laura, Sikhar Sakhuja, and Damon McCoy, "FBPoliticalAds." Facebook Archive. shikhar394. Accessed April 28, 2020. https://github.com/online-pol-ads/FBPoliticalAds/blob/master/docs/Facebooks-archive.pdf.
[4] Kim, Young Mie. "New Evidence Shows How Russia's Election Interference Has Gotten More Brazen." Brennan Center for Justice, Analysis and Opinion. The Brennan Center for Justice at NYU Law, March 5, 2020. https://www.brennancenter.org/our-work/analysis-opinion/new-evidence-shows-how-russias-election-interference-has-gotten-more.
[5] Francois, Camille, Ben Nimmo, and C. Shawn Eib. "The IRA CopyPasta Campaign." Graphika Reports. Graphika, October 21, 2019. https://graphika.com/reports/copypasta/.

month, only one was a Facebook page.[6] Considering the scale of past Russian Facebook campaigns and Facebook's few new site restrictions[7], it is likely that many IRA pages are still out there. With the 2020 Presidential election looming, it is clear that another detection model is needed.

Admittedly, researchers are torn over whether Russian propaganda is detectable at all. Zannettou argued that the ever-adapting concealment efforts of the IRA prevent simple automatic detection.[8] Others have shown that syntactic and vocabulary flags of second language learners may betray a Russian born IRA author[9], and have constructed models to do so with reasonable accuracy. We side with the latter camp, and attempt to build a model to compete with these other detection models. While recent models exclusively analyze tweets[10,11], article text[12], or Instagram posts[13] this one is among the first public models to be trained on both Facebook and Instagram data. Other models have had some successes.[14] However, other models limit themselves to analyses of ad text, so they appear to have plateaued at about 88% predictive accuracy. However, the models proposed here include textual predictors as well as other publicly available ones (target location). Ultimately, we trained a random forest model which consistently correctly classifies 2015-2017 IRA ads as 'propaganda' with a precision of .9411, and a bagged model with .9503 recall. The first outperforms previously proposed detection models by Jane Im[15] and Nayeema Nasrin[16], who achieved precisions of .785 and .815, respectively. If our model is fed a constant stream of new ads (thanks to NYU's code scraping Facebook's ad archive), authorities and social media companies should be able to identify and take down disingenuous pages and posts much more quickly. Our model has also classified several active Facebook pages as 'propaganda', and we will disclose some of their names here.

---

[6] Gleicher, Nathaniel. "Removing More Coordinated Inauthentic Behavior From Iran and Russia" Facebook.com Newsroom. October 21, 2019. https://about.fb.com/news/2019/10/removing-more-coordinated-inauthentic-behavior-from-iran-and-russia/

[7] Young Mie Kim, section 'What should we do?"

[8] Savvas Zannettou, Tristan Caulfield, William Setzer, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. 2019. Who let the trolls out?: Towards understanding state-sponsored trolls. In *Proceed- ings of the 10th ACM Conference on Web Science*, pages 353–362. ACM.

[9] Spangher, Alexander, et al. "Analysis of Strategy and Spread of Russia-Sponsored Content in the US in 2017." Carnegie Mellon University, n.d. Accessed April 28, 2020.

[10] Im, Jane, Eshwar Chandrasekharan, Sargent, Jackson, Lighthammer, Denby, Bhargava, Hemphill, Jurgens, and Eric. "Still out There: Modeling and Identifying Russian Troll Accounts on Twitter." arXiv.org, Computer Science > Social and Information Networks. Cornell University, January 31, 2019. https://arxiv.org/abs/1901.11162.

[11] Stewart, Leo G., Ahmer Arif, and Kate Starbird. 2018. "Examining Trolls and Polarization with a Retweet Network." In Proceedings of WSDM workshop on Misinformation and Misbehavior Mining on the Web (MIS2). ACM, New York, NY, USA, 6 pages. https://doi.org/10.475/123_4

[12] Yoosuf, Shehel, Yin "David" Yang, "Fine-Grained Propaganda Detection with Fine-Tuned BERT", Proceedings of the 2nd Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda, Hong Kong, China, November 4, 2019 c 2019 Association for Computational Linguistics, pages 87–91.

[13] Young Mie Kim (See footnote 4)

[14] *Ibid*

[15] Jane Im et. Al 2019

[16] Nasrin, Nayeema. "How Many Users Are Enough? Exploring Semi-Supervision and Stylometric Features to Uncover a Russian Troll Farm." Proceedings of the 2nd Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda, pages 20–30, Association for Computational Linguistics. Department of Information Systems and Cyber Security, University of Texas at San Antonio, November 4, 2019. https://www.aclweb.org/anthology/D19-5003.pdf.

## II. Data and Collection

The data we selected was the same data collected during the House Intelligence Committee Minority Investigation by Special Counsel Robert S. Mueller who indicted 13 Russian individuals and three Russian organizations, including the IRA, for engaging in operations to interfere with U.S. political and electoral processes such as the 2016 presidential election. Facebook released "a total 3,519 total advertisements [that] were identified to have been purchased" with over "11.4 million American users exposed to those advertisements". The data has 25 different variables which include AdIDs, Adtext, Clicks, Impressions, Locations, CreationDate, EndDate, and AdSpend, but here we focus on target locations and text[17]. To interpret and use most of the data we convert the variables in character or string format to categorical data. Researchers began by row-binding the clean Propublica ad texts to the clean NYU ad texts. Together, they formed the domestic, or non-propaganda group. These ads were assigned a response value (variable 'y' in the code) of 0. Then, we attached the IRA ad texts to the vector of domestic ad texts. In the response, these ads were assigned a '1'. The same steps were taken to create the ad target location variables, but we then built binary indicator variables. For instance, if the ad's creator chose to target Baltimore, MD, the 'ifbalt' variable =1. Separate binary variables were built the same way: one for each of 'baltimore', "st louis", "cleveland", "atlanta", "texas", 'illinois', 'new orleans', 'milwaukee', 'maryland', 'charlotte', 'detroit', 'richmond', 'minneapolis', 'san francisco', 'houston', 'lancaster', 'madison', 'oakland', 'ferguson'. These choices are discussed in the next section. Ad Ids were preserved for recognition. For exploratory data analysis, researchers compared the entire ProPublica+NYU political ad set (137370 ads, with 63378 unique messages) to the foreign propaganda set (3255 ads). We chose not to analyze 107 benign ads posted by the IRA pages 'Memeopolis' and 'Facemusic'. We could then look for noticeable differences in the proportion of ads with certain textual themes in each set. Unfortunately, only a handful of ads (about 10) in the ProPublica data were on Facebook at the same time as the propaganda data. This is a major potential objection to this analysis, so one solution researchers sought was to select themes based on their consistency over time. Specifically, the variables analyzed in section VII (Violence, race, voting) were chosen because they appeared consistently for the whole duration of Russia's 2015-2017 Facebook propaganda campaign, or in a large portion of the ProPublica ads. Boyd's findings[18] aided significantly in this regard. The assumption is that time-invariant themes were present in political ads in both time periods under consideration. The final step of data preparation randomly sampled 3255 ProPublica or NYU ads so the training and testing sets would not contain more ProPublica ads than IRA ads. Later, researchers added 93 predictors generated by Linguistic Inquiry and Word Count (LIWC) software. On the advice of Boyd et. al[19], we decided to include many of the syntactic predictors in the LIWC 2015 dictionary.[20] The program conducted syntactic and stylistic diagnoses of every advertisement. For example, their software generated measures of positive emotion, cognitive processes, and analytical thinking for each ad. Descriptions of other variables may be found on their website.[21] This brought the total number of predictors to 113.

---

[17] Note that these variables were renamed from the original dataset to fit our needs in R-Studio.

[18] Boyd, Ryan L, et al. "Characterizing the Internet Research Agency's Social Media Operations During the 2016 U.S. Presidential Election Using Linguistic Analyses." University of Texas at Austin, n.d. Accessed April 28, 2020. https://files.osf.io/v1/resources/ajh2q/providers/osfstorage/5bb21c61717460001650732d?action=download&version=2&direct&format=pdf, page 3.

[19] Boyd, p.6.

[20] We purchased their text analysis software here: https://liwc.wpengine.com/ in order to generate the predictors.

[21] https://liwc.wpengine.com/compare-dictionaries/

## III. Exploratory Data Analysis (EDA):

The following quantitative and graphical exploratory analysis pertains to the categorical research question. These helped us identify predictors not generated by the LIWC.

*Figure 1.5: Number of Russian Advertisement Targets by City and State*

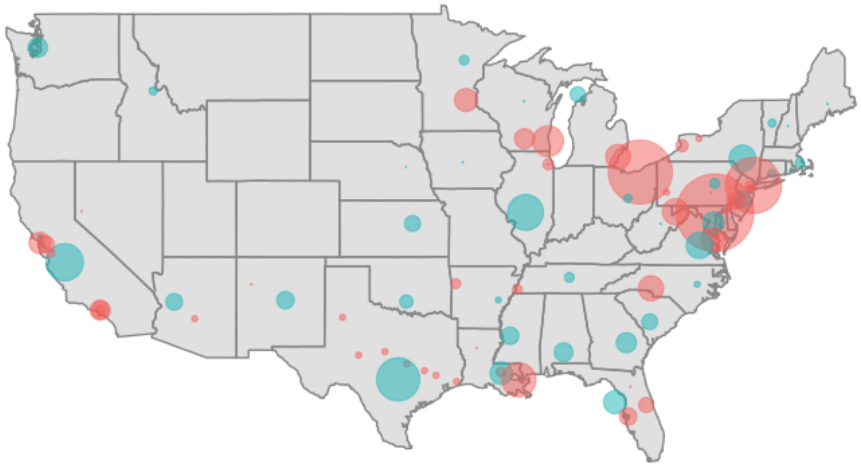


Figure 1.5: Each circle represents a city or state targeted by the Russian advertisements. Blue circles indicate the number of times that state was targeted, whereas red circles represent cities. The size of the circles are proportional to the number of times that location was targeted by an advertisement. One should note that while the largest bubble displayed here corresponds to Baltimore, MD with 244 targets, while our data contained 2566 ads that only specified "United States" as the target location. The map provides another interesting dimension of the distribution of our location dummy variables. Notice the target specificity of many Russian advertisements, which often appeared in northeastern cities. This reveals a concerted effort on the part of Russian agents to influence specific groups of people they believe to be vulnerable, as opposed to a 'land where it may' strategy. This graphic informed the choices of the location predictor categories, since this study assumes that locations frequently targeted in the past are good predictors of future propaganda. To view this interactive graph click the following link: https://pbf2tp.github.io/.

The preliminary search for propaganda-distinguishing categorical features proceeded as follows. Researchers began by randomly sampling propaganda messages, looking for patterns. A handful emerged: racial themes, voter suppression/discouragement language, police references, etc. We then compared the proportion of domestic ads containing these themes to the proportion of propaganda ads discussing the same. Starting with an indicator variable as to whether the thematic content was 'violent', the ProPublica & NYU political ads contained violent themes 28.16% of the time, whereas the propaganda data mentioned a violent word 32.5% of the time. This was not the largest margin of change, but it was decided to include it in the model anyway since it was a consistent theme throughout the length of the IRA's Facebook campaign. Researchers also noted a much different usage of exclamation points. Investigating further, the propaganda data used a lone exclamation point 46.07% of the time, whereas the domestic political ads only used it 12.96% of the time. This was a considerable difference, especially compared to that of the last variable, so it was tested. Similar methods explored whether ads referenced law enforcement officers (propaganda referenced police 15.58% more than the political ads), and whether the ad included racial words (this was the greatest difference, with propaganda mentioning race 28.725%

more often than ProPublica ads). Another important variable, labelled 'ifvote' indicated whether an ad discussed prosocial words such as 'elect', 'donate', or 'help'.

Researchers also created a measure of the number of facts in each ad. Specifically, a variable counted the number of times each ad included a digit, a link, proper nouns, or parenthetical notation, and other textual features associated with factuality. The typical political ad in the ProPublica set had 11.34 'facts', whereas the typical propaganda ad had only 2.69 'facts'. Finally, researchers constructed variables which indicated the presence of a location frequently targeted by propaganda. For example:

| AdID | AdCount | State | TX | VA | MD |
|---|---|---|---|---|---|
| 1643 | 124 | *Texas* | 1 | 0 | 0 |
| 23 | 25 | *Virginia* | 0 | 1 | 0 |
| 412 | 30 | *Maryland* | 0 | 0 | 1 |

The political ads targeted these and other locations 52.15505% of the time, whereas the propaganda targeted these places 98.89078% of the time. The margin was increased substantially upon the addition of a search for the "United States" as it seems that the Russians almost always specify which country they want to target, while domestic advertisers sometimes merely select the state they wish to target. Encouraged by these results, the analysis was undertaken.

## IV. Method of Analysis

Our first proposed model is a Recursive Binary Split classification tree, with split predictors chosen on the basis of mean decrease in Gini index. This model is our baseline, for it is less flexible (ergo probably more biased) than our other known classification algorithms: Bagging, Random Forests, and Boosting. We could then attempt the latter three, and decide whether decreasing the bias aids prediction accuracy. The tree also required fewer initial tuning parameters and high interpretability. Its decision tree would help us decide to cut predictors, should they be entirely unimportant. Every model was run 40 times with a different random sample (without replacement) of 3255 domestic ads. This allowed us to calculate a typical accuracy for each model, thereby stabilizing sample-to-sample accuracy variance. The second model, the bagged model, was built with the default settings of the 'randomForest' package in R. 113 predictors were considered at each split, since the St. Louis target location predictor was unimportant according to the results of the first tree model.

Next, we opted to build a Random Forest model with a minimum terminal node size adjustment. Since we had a couple of dominant predictors, like location, which could lead to highly correlated trees, we felt that limiting the number of predictors considered at each split could improve accuracy. It considered $\sqrt{113} = 10.63$ predictors at each split, according to the convention.

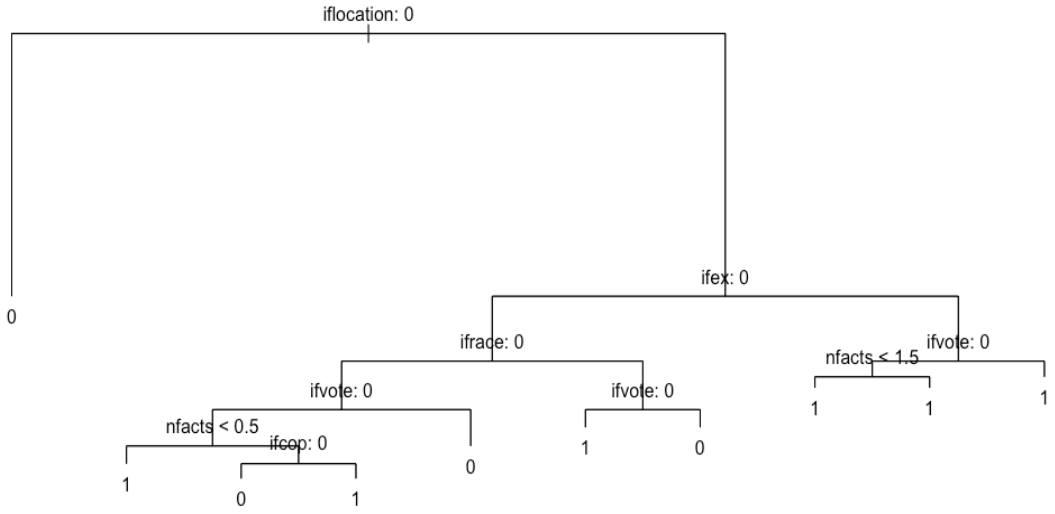
To obtain the node size setting for each of the 40 resamples of domestic ads, we recorded the out of bag errors for each of nodesize=1 through 20 while generating 300 trees at each iteration. The node size corresponding to the lowest out of bag error became the node size setting of the final prediction Random Forest. Again, we trained a forest with 300 trees, but this time set the minimum node size equal to one of 1 through 20. Therefore, we had 40 models, which made predictions on 40 different (possibly overlapping) resample and single fold test sets. We then calculated the average, median, and maximum overall accuracy, precision, true positive rate, and precision of the 40.

Finally, we calculated the same accuracy metrics for 40 boosted models. Even though boosted models run the risk of overfitting the data, we were curious about the effect of residual updating on overall accuracy. The three tuning parameters of said model were chosen in the following ways: interaction depth

was set to 1 every time, according to Gareth James et. al..[22] To control the learning rate, we changed the shrinkage parameter from .1 to .01, according to the same. .001 may also be a valid parameter, so future models may compare these. Finally, all 40 models generated 300 trees in order to be consistent with the forest models. Our choices of the number of trees to generate is admittedly somewhat arbitrary. Sometimes these models may need to generate as many as 2000 trees in order to find the global minimum error rate.[23] Our drive to find the absolute lowest error rate was somewhat hampered by a lack of computational power and time. We hope other researchers will explore the effect of more trees on error.

## V. Analysis & Results

An RBS classification tree model would initially provide an adequate prediction accuracy and allow for interpretation. Other classification models have been found to show greater test data accuracy, but the ability to interpret the relative importance of predictors was a priority for this study. This model grew a tree by RBS using the indicator variable = 1 if propaganda, 0 otherwise. Weights and all other settings were default according to the 'tree' package in R. The split criterion was Gini Index, as it is more sensitive to node purity than the classification error rate. Accordingly, the following tree diagram was generated. Note that the LIWC variables are not included here since doing so would cause the tree to reach maximum depth. Propaganda was coded as 1, so the first node reads "if the ad did not target a location, then the ad was classified as an organic/ProPublica post (the first terminal node to the left).




The first predictor to stand out was location. It appears that location is the most important variable for forming pure nodes, followed by the presence of exclamation points, and others. One interesting subcluster generated contains ads that referenced race (the only internal node labelled ifrace:0), then mentioned voting. This split reveals that among ads discussing race, voting is the next most important predictor of propaganda. Also, ads with fewer facts were always predicted to be propaganda. Finally, all ads with lone exclamation points were classified as propaganda. This could reflect the IRA's penchant for commands like "Follow!" or "Subscribe!".

Regarding the model's prediction accuracy, overall, it correctly predicted whether a test-set ad was propaganda 85.12514% of the time. Of greatest interest, however, is how often it successfully

[22] James, Gareth., Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. 7. Vol. 7. New York: Springer, 2017., p. 322

[23] *Ibid* p. 324

detected that an ad was indeed propaganda: *the model was 87.99776% accurate when the ad was propaganda*. In other words, if you showed the classification tree algorithm 100 pieces of propaganda, we would expect it to correctly label 87 of them as 'propaganda'. The true positive rate decreased slightly upon the application of a random forest model, to *86.26098%*. The random forest randomly sampled $\sqrt{7}$ = 2.64575 variables per split since we had 7 predictors. Based on node sizes 1-20, we found that the best size was 15. Resampling from the ProPublica data 40 times, the random forest model returned the following distribution of true positive rates:

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|
| 0.8305 | 0.8487 | 0.8571 | 0.8626 | 0.8778 | 0.9072 |

In every model attempted, the true positive rate drove up the overall error rate. Given a ProPublica ad, the random forest and classification tree correctly identified it 87.90723% and 82.16965% of the time, respectively. In all, the classification tree excels at identifying a propaganda advertisement, whereas a random forest model is better for predicting legitimate political ads.

Having built several variables ourselves, we decided to follow Boyd et. al[24], and include measures of text features generated by the LIWC. Curiously, the Authenticity measure (an index of deception) Boyd found useful for tweet detection was not important in our model. Instead, quantitative variables measuring the presence of 'Seeing' perceptual processes in the text, present-focused time orientation, and money-themed text led to great improvements in node purity at many nodes. Addition of the LIWC predictors led to accuracy increases across all models:

| Model | Measure | Mean | Median | Max. |
|---|---|---|---|---|
| Bagged | True Positive Rate | 0.9502971 | 0.9516101 | 0.9621690 |
| Random Forest | True Positive Rate | 0.9433209 | 0.9447645 | 0.9616111 |
| Boosted | True Positive Rate | 0.9364702 | 0.9363661 | 0.9509202 |
| Random Forest | True Negative Rate | 0.9407126 | 0.9421907 | 0.9528243 |
| Boosted | True Negative Rate | 0.9194309 | 0.9199307 | 0.9404836 |
| Bagged | True Negative Rate | 0.9106148 | 0.9101851 | 0.9287965 |
| Random Forest | Precision | 0.9411334 | 0.9420584 | 0.9530574 |
| Boosted | Precision | 0.9207653 | 0.9211075 | 0.9405573 |
| Bagged | Precision | 0.9139191 | 0.9137606 | 0.9334903 |
| Random Forest | Overall Accuracy | 0.9420123 | 0.9425499 | 0.9536098 |
| Bagged | Overall Accuracy | 0.9303943 | 0.9301075 | 0.9400922 |
| Boosted | Overall Accuracy | 0.9279263 | 0.9284178 | 0.9333333 |

Here, we built a bagged model instead of the Recursive Binary Split tree since the former could accommodate a greater number of predictors. The mean, median, and maximum were calculated across 40 resamples of 3255 from the Propublica+NYU ads. We then joined these ads to the 3255 propaganda ads to form a balanced set. During each of the 40 iterations, a single random fold split half of these 6510 ads into a training set, and saved the rest for testing. Note that the Random Forest model outperformed the other two in terms of precision, while the Bagged model had the superior True Positive Rate. In this way, if an investigator wished to limit the number of misclassified authentic American ads, they should employ the Random Forest model. Practitioners wishing to ensure that few pieces of propaganda go undetected should employ the Bagged model, as it had the highest true positive rate of those examined. The median nearly equaled the mean accuracies in each instance. This suggests that the accuracy distributions are very symmetric.

[24] Boyd, p.6.

Nasrin et. Al.[25] helped us realize that our control dataset (The Propublica+NYU ads) may not actually be propaganda-free. This could be driving up the false positive rates in more than one model. We researched many of these pages, and some are extremely suspicious.[26] We hope authorities or Facebook will investigate these further. Other false positives represent legitimate or verified pages led by immigrants from the former Soviet Union. These include PragerU and MoveOn. In this way, the model may be detecting the language of second-language learners, or classifying ads of pro-Russian organizations as propaganda. Still, many ads by legitimate news sources or political candidates were erroneously marked as 'propaganda', ostensibly for their propaganda-esque verb tenses or exclamation points. We discuss false positives further in the next section.

## VI. Potential Objections

One may question these models on the basis of the ProPublica data. These ads are very much a convenience sample, as ProPublica solicited its users to download a program to scrape the ads from their Facebook feeds. This means the ProPublica data is not representative of the population of all non-propaganda political advertisements which ran on Facebook at the time. We attempted to improve the representativeness of our 'domestic' ads by combining these with ads scraped by NYU. Their process was not random, however, so our control group of domestic ads is essentially two convenience samples combined. Therefore, one should not say that the results of the classification model totally reflect the true nature of all political ads on Facebook. For reference, in the third quarter of 2019, Facebook announced that more than 7 million advertisers used their platform.[27] Each advertiser surely had more than one ad running. Our accuracies derive from models trained on samples of 3255 'domestic' advertisements from the set of 137,370: a small fraction of the ads probably running on Facebook at the time. In the future, researchers might consider web scraping a random sample of ads from Facebook's Ad Library, using code similar to that written by NYU researchers.

Similarly, our set of 'domestic' political advertisements may not be propaganda-free. This means some false positives may actually be true positives. Further research must be done to verify the authenticity of these ads, but it is our hope that we have narrowed down the list for investigators. To prevent legitimate ads from being misclassified, however, future models should include a variable indicating whether the ad was placed by a 'verified' page (meaning Facebook has vetted the organization as not false or misleading[28]). This should drastically increase the accuracy, although building such a variable may be time-intensive.

This study's classification techniques could be questioned due to the date range of their data as well. Specifically, almost all the organic political advertisements (not propaganda) ran on Facebook 1-2 years after propaganda ads had concluded. The first propaganda ad started in 2015, and the last ended in August 2017, while most of the ProPublica ads ran from fall 2017 to 2019. Considering the fleeting nature of political issues, the topics discussed by organic political data in the span of the IRA's campaign may have been much different than the topics discussed by the ProPublica data only a year later. This could damage the validity of our thematic predictors, such as whether the ad contained 'racial' themes. It is very possible that real political ads running from 2015-2017 discussed race just as often as the

---

[25] p. 28.

[26] https://www.facebook.com/proudrightwinger/, https://www.facebook.com/pg/peoplestrumpet/about/?ref=page_internal, https://www.facebook.com/ReBuildUSANow/, https://www.facebook.com/countable.us/, https://www.facebook.com/antiseditionist/ among others.

[27] Clement, J. "Facebook Active Advertisers 2019." Statista, January 30, 2020. https://www.statista.com/statistics/778191/active-facebook-advertisers/.

[28] https://www.facebook.com/help/1288173394636262

propaganda. In this case, said predictor would be invalid. We would need Facebook to disclose earlier organic political advertisements if we are to prove such a predictor valid.

## VII. Applications

The classification model is especially practical in that anyone can pull up a Facebook ad on the Ad Library[29], enter their location and race, copy and paste the ad's text and it will tell you whether it thinks the ad is propaganda. Of course, these classifications are not 'hard and fast' so we would discourage the typical Facebook user from making decisions based upon these classifications. Instead, we hope social media teams and authorities will narrow their searches to the more suspicious pages identified by the model. This may be advised because of its low false positive rate, but we hope others will build upon this progress and build a model that can even more reliably identify disingenuous Facebook advertisements. Facebook and government officials (Such as the U.S. Cyber Command or the NSA[30]) could potentially employ this algorithm with a greater degree of accuracy since they have access to ad creation hour time stamps and can verify sponsors. In the past, they have easily identified propaganda by looking at receipts, since the Russians bought ads under their actual agency name and paid in rubles. After being exposed in 2017, however, they probably shifted their tactics and have begun funneling money through shell companies pretending to be legitimate U.S. political interest groups. In addition, the number of advertisers on Facebook has increased dramatically in recent years (from 3 million in 2016 to 7 million in late 2019[31]), so it is increasingly difficult to verify the legitimacy of each firm advertising on Facebook. This means Facebook and others will have to rely more heavily on propaganda characteristics and traits (thematic content, syntax, target demographics, etc.) if they are to successfully root out propaganda. The IRA will also likely expand its campaigns to include other social media platforms (They actually recently upgraded their office building), so unless these platforms gather detailed metrics on each advertisement, they will have to rely on an advertisement's facade to combat disinformation.

In the future, researchers might improve true positive prediction rates by adding predictors like post hour created (see Boyd p. 2), run time of the ad, number of totally capitalized words, or by increasing the number of target locations to detect. Future research should also select locations in which the Russians have historically invested much time and money (this could be measured in a rubles per minute online variable). The target location predictor could greatly benefit from such a change. Russian agents masquerade as individuals promoting Southern identity/nationalistic themes, so a predictor indicating the presence of these themes could increase node purity. New models should also indicate whether the target community recently suffered an instance of police force against African Americans (or whether the target community has an unusually high proportion of African American residents, and was posted after violence elsewhere). Between 2015 and 2017, Russians often targeted such communities in the immediate aftermath of violence.[32] Such a model might identify divisive advertising before it strikes. This research only examined target location and ad text (the text you normally see above a picture, typed by the poster), but page post name and image text should also be searched for thematic content data. Additional data could be gathered from NYU's scraped set of political advertisements. These were not included in the present study due to merging difficulties and lack of message data. Finally, it could be valuable to build a

---

[29] https://www.facebook.com/ads/library/?active_status=all&ad_type=all&country=US&impression_search_field=has_impressions_lifetime

[30] Nakashima, Ellen. "U.S. Cyber Command Operation Disrupted Internet Access of Russian Troll Factory on Day of 2018 Midterms." The Washington Post. WP Company, February 27, 2019. https://www.washingtonpost.com/world/national-security/us-cyber-command-operation-disrupted-internet-access-of-russian-troll-factory-on-day-of-2018-midterms/2019/02/26/1827fc9e-36d6-11e9-af5b-b51b7ff322e9_story.html.

[31] Clement, J. "Facebook Active Advertisers 2019." Statista, January 30, 2020. https://www.statista.com/statistics/778191/active-facebook-advertisers/.

[32] See ads 1558, 2379, or 2353 among many others. Ad 2353 targeted Cleveland, Ohio (which is 53.3% African American, compared to the national average of 12.7%) and discussed police force against Lamar Smith in St. Louis. We observed many similar ads.

model stating the probability that a given ad is propaganda. This way, Facebook users can decide for themselves whether a statement like “there is an 87.99776% chance that this ad is Russian Propaganda” is cause for concern.

## VIII. Conclusion

We found that classification models can accurately predict whether an ad is propaganda. This study also found that predictors like target location and thematic predictors (racial topics) are very important for separating observations into pure groups. We believe that the best classification models (Bagged or Random Forest) perform better than random guessing, but it appears that propaganda detection rate can be improved upon the addition of informative predictors and given augmented classification models. These may benefit from an increase in the number of trees generated. Finally, true positive rates could improve with the application of a pruned model or linear support vector machine.

## XI. Works Cited

Boyd, Ryan L, et al. "Characterizing the Internet Research Agency's Social Media Operations During the 2016 U.S. Presidential Election Using Linguistic Analyses." University of Texas at Austin, n.d. Accessed April 28, 2020.

Clement, J. "Facebook Active Advertisers 2019." Statista, January 30, 2020. https://www.statista.com/statistics/778191/active-facebook-advertisers/.

Dutt, Ritam, et al. "'Senator, We Sell Ads': Analysis of the 2016 Russian Facebook Ads Campaign." Indian Institute of Technology Kharagpur, India, n.d. Accessed April 28, 2020.

Edelson, Laura, Sikhar Sakhuja, and Damon McCoy, "FBPoliticalAds." Facebook Archive. shikhar394. Accessed April 28, 2020. https://github.com/online-pol-ads/FBPoliticalAds/blob/master/docs/Facebooks-archive.pdf.

Francois, Camille, Ben Nimmo, and C. Shawn Eib. "The IRA CopyPasta Campaign." Graphika Reports. Graphika, October 21, 2019. https://graphika.com/reports/copypasta/.

Gleicher, Nathaniel. "Removing More Coordinated Inauthentic Behavior From Iran and Russia" Facebook.com Newsroom. October 21, 2019. https://about.fb.com/news/2019/10/removing-more-coordinated-inauthentic-behavior-from-iran-and-russia/

Howard, Philip N, et al. "The IRA, Social Media and Political Polarization in the United States, 2012-2018," n.d. Accessed April 28, 2020.

"Impressions." Facebook Business Help Center. Accessed April 28, 2020. https://www.facebook.com/business/help/675615482516035.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. 7. Vol. 7. New York: Springer, 2017., p. 322

Kim, Young Mie. "New Evidence Shows How Russia's Election Interference Has Gotten More Brazen." Brennan Center for Justice, Analysis and Opinion. The Brennan Center for Justice at NYU Law, March 5, 2020. https://www.brennancenter.org/our-work/analysis-opinion/new-evidence-shows-how-russias-election-interference-has-gotten-more.

Long, Jacob A., et al. "Analysis and Presentation of Social Scientific Data." Package 'jtools'. R-Studio, April 21, 2020. https://cran.r-project.org/web/packages/jtools/jtools.pdf.

Nakashima, Ellen. "U.S. Cyber Command Operation Disrupted Internet Access of Russian Troll Factory on Day of 2018 Midterms." The Washington Post. WP Company, February 27, 2019. https://www.washingtonpost.com/world/national-security/us-cyber-command-operation-disrupted-internet-access-of-russian-troll-factory-on-day-of-2018-midterms/2019/02/26/1827fc9e-36d6-11e9-af5b-b51b7ff322e9_story.html.

ProPublica. "Political Advertisements from Facebook." ProPublica Data Store, March 19, 2019. https://www.propublica.org/datastore/dataset/political-advertisements-from-facebook.

Ripley, Brian, et al."Support Functions and Datasets for Venables and Ripley's MASS." Package 'MASS.' R-Studio, April 26, 2020. https://cran.r-project.org/web/packages/MASS/MASS.pdf.

Scott, David. Tukey Ladder of Powers. Accessed April 28, 2020. http://onlinestatbook.com/2/transformations/tukey.html.

Simpson, J. R. and Montgomery, D. C. (1998). "The Development and Evaluation of Alternative Generalized M-Estimation Techniques". *Communications in Statistics - Simulation and Computation*, 27, 999–1018.

"Social Media Advertisements." U.S. House of Representatives Permanent Select Committee on Intelligence. Accessed April 28, 2020. https://intelligence.house.gov/social-media-content/social-media-advertisements.htm.

Spangher, Alexander, et al. "Analysis of Strategy and Spread of Russia-Sponsored Content in the US in 2017." Carnegie Mellon University, n.d. Accessed April 28, 2020.

Stewart, Leo G., Ahmer Arif, and Kate Starbird. 2018. Examining Trolls and Polarization with a Retweet Network. In Proceedings of WSDM workshop on Misinformation and Misbehavior Mining on the Web (MIS2). ACM, New York, NY, USA, 6 pages. https://doi.org/10.475/123_4

Yoosuf, Shehel, Yin "David" Yang, "Fine-Grained Propaganda Detection with Fine-Tuned BERT", Proceedings of the 2nd Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda, Hong Kong, China, November 4, 2019 c 2019 Association for Computational Linguistics, pages 87–91.

Zannettou, Savvas, Tristan Caulfield, William Setzer, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. 2019. Who let the trolls out?: Towards understanding state-sponsored trolls. In *Proceed- ings of the 10th ACM Conference on Web Science*, pages 353–362. ACM.