Natural Language Processing and Machine Learning for Political Facebook Ads

Matthew Borelli

4/27/2020

Abstract

The main goal of this project is to understand political ads on FaceBook solely on their textual content and compare the models we create to ProPublica's machine learning model for predicting the likelihood that an ad is political. By utilizing natural language processing, we analyze the term-frequency weights for unique words that appear in more than 1% of FaceBook advertisements. We utilize principal components analysis and random forest modelling in order to understand the variations of the most common words between political advertisements and create a model that predicts the likelihood that an ad is political. The first two principal component vectors show that positive opinion words generally make an add less likely to be political. The random forest model generated predictions for an advertisement's political nature that match the ProPublica model's predictions approximately 85% of observations in the testing set. We conclude that advertisements have a high prevalence of "call-to-action" words that solicit users to perform some specified action, usually donating to a political cause or voting for/against a specific politician or political party. The random forest model generated for this project and ProPublica's model disagree on advertisements of extremely short or extremely long length, but generally agree to a high level that could be improved with larger models.

Introduction

If you look closely at your Facebook timeline, you might see the word "Sponsored" at the top of the post. Integrated facebook ads are advertisements that appear as part of your normal news feed. These ads could range over anything, from retail merchandising, ESPN videos, even news articles. As 2020 is an election year, many of these ads are political. In the lead-up to the election, you should expect to see many political ads on the internet either strongly for or strongly against politicians, political movements, or policy positions. Thousands of firms want to influence national political opinions through the use of advertising. Thus it makes sense to study political ads to understand what I can learn about the broader political sphere.

Classifying political ads will not be an easy task however. "Politics" isn't some singular entity with well-defined characteristics, rather it is a wide field that covers a variety of topics, industries, and motivations. For example, say that a environmental non-profit firm uses an ad on Facebook to promote a message about the construction of the Keystone Pipeline. While this is likely political, it does not mean that every ad from the same entity will necessarily be political. As such, the focus of political advertisement classification in this project will not be based on the advertiser. A more effective method for classifying facebook ads would be to analyze their content, in this case the text of an advertisement. This project has two main goals: 1- Generate data-backed insights on the different types of political ads. 2- Create a model to predict the likelihood that an ad is political. These two research goals will help us to understand political Facbook ads using both quantitative and qualitative methods. Textual analysis, even for advertisements, is an inherently subjective process as the meaning applied to words depends on surrounding context to a degree that computers cannot

handle. Therefore, it is important in this project that we utilize statistical methods to bring us as close to empirical statistics as possible and then create well-reasoned interpretations based on those results.

Methods

This project utilizes ProPublica's "Political Advertisements from Facebook" data set. The ProPublica data set contains 162324 advertisements gathered through the use of a browser extension. ProPublica users download the extension, which sends information on all Facebook ads they encounter to ProPublica. These ad collectors also vote on whether they believe an ad is political, but not all ads have votes. ProPublica then uses their own machine learning classifier to determine which ads were likely political, and enters those ads into the data set. Therefore, most (if not all) of the advertisements in this data set are determined to be political already. To give us a baseline, ProPublica's model predicts that 95.92% of ads are political (under the assumption that a political probability of greater than 0.5 is a political ad) From the whole data set, the following variables are important to our analysis:

- *political*: Number of users who voted that an ad is political.
- *non_political*: Number of users who voted that an ad is not political.
- *message*: The plain text written with the ad (not anything in an image).
- *political_probability*: ProPublica's prediction of the probability that an ad is political using machine learning.

Using the variables *political* and *non-political*, I create the variable *pol_vote_pct* = *political* / (*political* + *non-political*) to calculate the fraction of users who voted that a post is political. Our goal is to predict *pol_vote_pct* using the most common words amongst the documents as feature variables, then compare that model's predictions to ProPublica's probability predictions and analyzing which documents they agree and disagree on. As well, in order better to understand the general tone of political advertising, I want to try to quantify their messages' tone. To do that, I use text files from Minqing Hu and Bing Liu's paper "Mining and Summarizing Customer Reviews." that compile thousands of words generally considered as having positive or negative connotations. Then I associate those with the advertisements in our model.

After performing data pre-processing on ProPublica's data including removing stop words and stemming words, we then transform our data into a document-term matrix (DTM) that associates each document with a row and each unique word to a column. Then we trim down to words that appear in over 0.5% of observations, approximately 811 advertisements. This collapses our DTM from 86294 unique words to 467 unique words without needing to remove any documents. As well, we apply term-frequency (TF) weights to the words of each document. TF weights, $TF_{ij} = X_{ij} / \sum_{j=1}^{d} X_{ij}$ normalize the document terms in a row to sum to 1, which controls for the length of the document. For general natural language processing projects, it would usually be preferable to utilize inverse-documentfrequency (IDF) weights as well, as they heavily weight more specific words that don't appear in many documents. However, since most of our advertisements are political, we actually want to analyze the more frequent terms, which would make TF weights preferable to IDF weights in this case.

Using the variables *political* and *non-political*, I create the variable *pol_vote_pct* = *political* / (*political* + *non-political*) to calculate the fraction of users who voted that a post is political. Our goal is to predict *pol_vote_pct* using the most common words amongst the documents as feature variables, then compare that model's predictions to ProPublica's probability predictions and analyzing which documents they agree and disagree on. In order to test predictions, we will utilize a train/test split, where the training data consists of any advertisements that have been voted on by users and the testing data consists of any advertisements without user votes. While this isn't necessarily random, there doesn't seem to be any consistent reason behind which ads don't have votes, so it is still at least quasi-random.

The first statistical model I will use to analyze the text is principal components analysis (PCA). This process will find linear combinations of the TF weights that best explain the variation amongst the political

advertisements in the data. Each linear combination acts as a vector in N-dimensional space where N is the number of feature variables, in this case words. This allows us to analyze all of the words in our DTM not just individually, but also how they relate to each other and affect the likelihood of a message being political. PCA is an unsupervised method, meaning that there is no prediction built from this model. Rather, we can analyze the principal component vectors by looking at which words are weighted the highest. Words with highest weighting are the most important in determing the outcome of that As a note, I did not have to do any rescaling of the TF weights because they are already normalized to 1 in a row, meaning that there should be no scaling issues. The other statistical model that I will utilize is random forest tree modelling. A single decision tree is a series of binary splits of feature variables that reduce a loss function. We then aggregate this over random samples of the data set with the same number of rows and resampling. Random forest is a particular type of decision tree process where variables to be added to the model are randomly selected and then aggregated together over all of the trees. This model will use the TF word frequencies to predict the political vote percentages in the training set, then apply that model to the testing set to generate predictions for the probability that a Faceook ad is political.

Results

Most Common Words

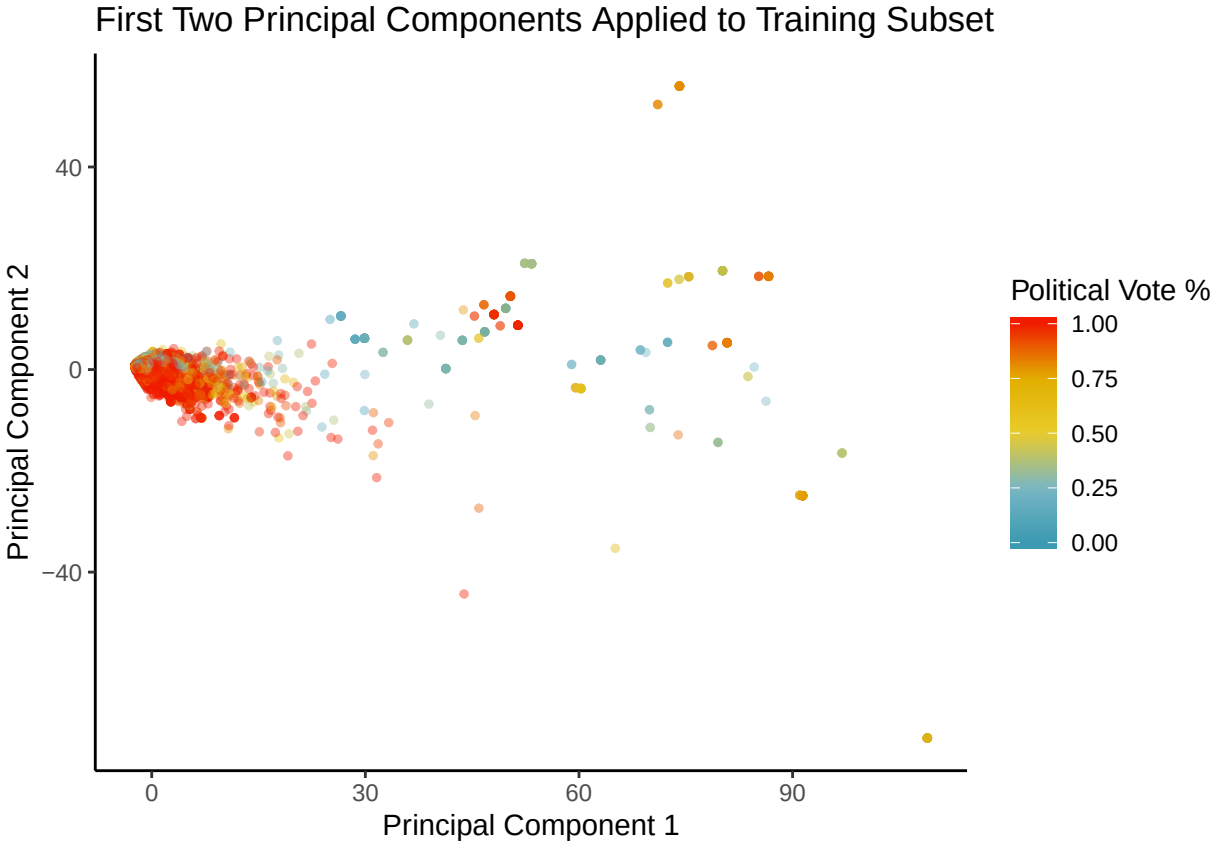
Before the complicated analyses that will follow, we start with a simple look at the most frequent words in the training set of facebook ads. The following word cloud has the most frequently found words in the center, decreasing in frequency as you move outwards.

Without any additional context, the most frequent words seem vague and unrelated to politics: “will”, “help”, “now”, and “can”. However, they all share a common tie: fundraising. All four of those words are

commonly used in fundraising pitches, such as "Can you help us with your donation?" or "We need your contribution now. Will you help us?" As we move from the center, we see more specifically-political words, such as "campaign", "state", "trump", and "vote". Overall, there are a mix of fundraising and political terms that frequently appear in political ads on Facebook.

Principal Components Analysis

For the principal components analysis, we computed the first 10 principal component vectors of the training set of data. Most of the important variation in user's political vote percentage is captured with the first two principal component vectors however, so we will discuss those two in more detail within this section. The following graph displays the space spanned by the first two principal component vectors, with a color gradient denoting the user political vote percentage for each advertisement.



We see that there is a lot of variation along the x-axis, which represents the first principal component, and less variation along the y-axis, which represents the second principal component. There is a large cluster around the same area of this PC space, gradually lowering in political vote percentage as you increase principal components 1 & 2. Generally, less political advertisements as voted on by users tend to have higher values of both principal components, although this isn't a perfect separation by any means. In order to gain any inference from these, we next look at the terms with the 5 highest and lowest word weights in each principal component to understand which terms are being associated as "more political".

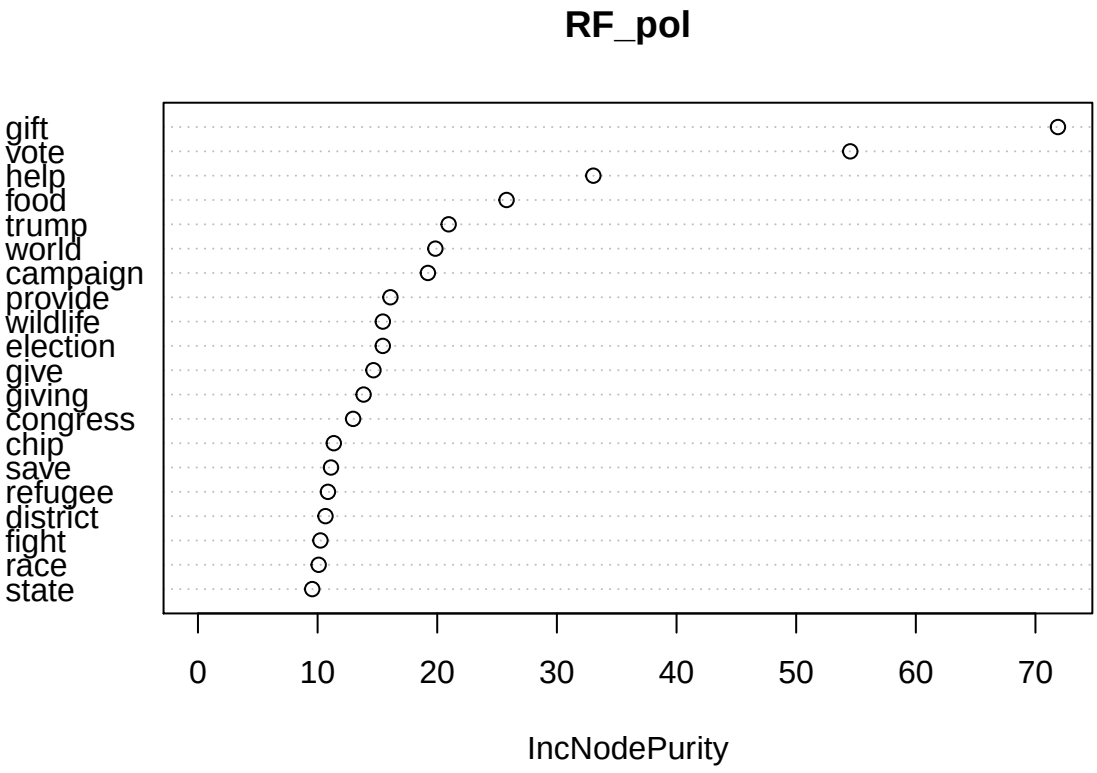
The appendix contains Table 1 and Table 2, which list the 5 highest and lowest weighted words for the first two principal component vectors. For the first principal component, the most positively weighted words all seem to be somewhat tonally positive. Three of them are positive adjetives, which means we have reason to believe that positive language is negatively correlated with user's political votes. For the negative weights, three of the terms are related to the pro-choice/pro-life debate: abortion; reproductive, as in reproductive

rights; and parenthood, as in the health services organization Planned Parenthood. For the second principal component vector, we see that the most positively weighted words are seem to be tonally positive or neutral, while the most negatively weighted words are more abjectly political like "political" and "republicans". There seem to be neutral words amongst both the high and low weight, which might mean that the second principal component separates more nuanced variations in sentences by heavily weighting words that don't carry significant emotional sentiment with them.

Random Forest Model

Model Selection

We run the random forest model with 25 randomly generated decision trees, as beyond that amount of trees we don't see significant decreases in the model error. The model predicts the percentage of ProPublica users who voted that an ad is political with the 486 TF word weights in the training set of data. After aggregating the 25 trees, the variations caused by differing word weights are aggregated over the trees into the final predictive model. The following is a variable importance plot showing the 20 most important words in the branches within the aggregated decision trees. These weights are calculated by summing the average amount of variation that a word accounts for in the political vote percentage across the 20 decision trees.



One of the words on this list has an obvious meaning, as *trump* generally refers to Donald Trump, current U.S. President. It makes sense that many political ads would mention his name. Some of the other words can be generally considered together as "calls-to-action": words or phrases that implore the reader to make some kind of a choice. In this case the words *gift*, *help*, and *provide* are generally used to solicit donations towards an organization, while *vote* is commonly used to compel an individual to vote for or against something in U.S. elections. While many of the 20 most important words are either "calls-to-action" or inherently political like *congress* and *president*, there are some words that don't have any obvious political meaning. *World*,

food, and *wildlife* are potentially less important in deciding that an ad is political and more important in determining that an ad is not political. Since variable importance measures absolute change, it is possible that these words are actually the most important words that decrease the possibility of an ad being perceived as political by ProPublica users.

Random Forest Predictions

Now that we have formed the random forest model, we fit the predict the probabilities of ads in the testing data set and compare our predictions to both the user's vote percentages and ProPublica's machine learning model. But first, we want to see what, if any, improvement the random forest model created in prediction error. To do this, we take the root mean squared error (RMSE) of the predictions for the testing set and compare it to a null model. For the null model, we simply "predict" that each advertisement will have the average user political vote percentage. The null gives an RMSE of 0.2555372. For the random forest model, we compare the predictions of the random forest model against the actual vote percentages in the data. The random forest predictions gives us an RMSE of 0.1548194. This means that the random forest model gives a 39.41% decrease in the root mean squared error. Our random forest model makes significant improvements over the null model in predicting the percentage of users who voted that an ad is political and more extensive tree models could reduce this error even more.

To compare the random forest model and ProPublica's machine learning model, take a two-step process for both models:

1. Generate predictions for the probability that an ad is political.
2. Create a binary dummy variable set to 1 if an ad has more than a 50% probability of being political, and 0 otherwise.

For ProPublica's model, the first step has already been done and we just create the dummy variable based on their listed probability. For the random forest model, we predict the probability that an ad is political for the 55430 observations in the training subset, then create the dummy variable. Comparing the results of the binary variables for each model gives us this confusion matrix.

```
Random_Forest
```

ProPublica 0 1 0 714 4870 1 3580 46266

Our random forest model model and ProPublica's machine learning models agree on 84.76 which is a fairly good amount. For 8.79% of the advertisements in the training set, our random forest model predicts that an ad is political while ProPublica does not. The remaining 6.46% of advertisements are ones that the random forest model did not predict were political while ProPublica did.

Conclusion

Model Agreements and Disagreements

There are three main categories that we want to examine in order to compare the models:

- Random Forest predicts a higher probability of an ad being political than ProPublica
- Random Forest predicts a lower probability of an ad being political than ProPublica
- Random Forest and ProPublica agree on the advertisement.

Rather than use the binary variables for this, we will use the differences between the raw predicted values from the random forest model and ProPublica's model and display some of the the advertisements with the widest and closest gaps in predicted political probability. As a note, these are the raw texts of the advertisements and include include JSON and HTML tags that make it harder to read the sentence. For integrity, I have not changed the raw message from ProPublica's database. Therefore, these are not necessarily the most extreme differences, rather they are the cleanest examples in the top 20 in each of the following categories

Random Forest Predicts More Political

- Lets fight this. Purchase a 4Ocean bracelet to remove one pound of trash. (**Percentage Point Difference:** 83.499)
- Create your own union-printed, campaign mail, collateral, and apparel with our online tools. Use promo code WELCOME10 for 15% off. (**Percentage Point Difference:** 87.327)
- This campaign really ties the country together. Order Here: <a href="“https://goo.gl/ubisLU”“>https://goo.gl/ubisLU”“This aggression will not stand, man."" (**Percentage Point Difference:** 84.775)

Random Forest Predicts Less Political

- BREAKING NEWS: The ASPCA is on the ground in the Florida Panhandle to help rescue and care for animals impacted by Hurricane Michael. Were finding animals stranded by the storm, without food, water or shelter. We are working as quickly as possible to save lives alongside the Florida State Animal Response Coalition. Please make an urgent gift today to help us to continue all our lifesaving efforts. (**Percentage Point Difference:** -77.813)
- This time of year is all about celebration, abundance and giving–or it should be. Todays 2x match is your chance to double your contribution. Vulnerable kids living on the streets desperately need warmth, nourishment, and love this time of year. Will you help? (**Percentage Point Difference:** -76.88)
- To all our readers in the U.S., We will get straight to the point: we need our Facebook readers to protect Wikipedia's independence. We depend on donations averaging about $15. Only a tiny portion of our readers give. If everyone reading this gave $3, we could keep Wikipedia thriving for years to come. The price of your coffee today is all we need. When we made Wikipedia a non-profit, people warned us we'd regret it. But if Wikipedia became commercial, it would be a great los<span class="“text_exposed_hide”“>... <span class=”“text_exposed_show""">s to the world. Wikipedia is a place for you to learn, not a place for advertising. It unites all of us who love knowledge: contributors, readers and the donors who keep us thriving. The heart and soul of Wikipedia is a community of people working to bring you unlimited access to reliable, neutral information. Please take one minute to help us keep Wikipedia growing. Thank you. (**Percentage Point Difference:** -71.242)

Random Forest and ProPublica Predict the Same

- Real connections. Real people. Real independence. Real leadership. Vote Ben Walsh for Mayor and <a href="“https://www.facebook.com/hashtag/riseabove”" class="" _58cn"“><span class="" _5afx"“><span class="" _58cl _5afz"“>#<span class="" _58cm"">RiseAbove. (**Percentage Point Difference:** 0.31)
- Cut through the Democrats' big lies and expose the truth! Pre-order now. (**Percentage Point Difference:** -1.283)
- Commissioners Val Arkoosh and Ken Lawrence are getting real results. Let's keep the momentum going! In the May 21st primary vote for both to keep our Democratic team working for us. (**Percentage Point Difference:** -0.453)

Political Ad Content

Both the principal components analysis and the advertisements in the section above give us a good idea of the language commonly used in political advertisements as well as the difference between the random forest model and ProPublica's model. From principal components analysis, we gathered that many of the words that are highly associated with the politicalness of an ad are either "call-to-action" words or words that deal with politics directly such as names of politicians or terms centered around elections. The prevalence of "call-to-action" words in political advertising demonstrates that one the primary motivations is for organizations to convince FaceBook users to give said organizations money in order to fulfill the promises made in their ads. We also see that ads are utilized in order to sway the opinions of users who come across them. All three of the advertisements that the random forest model and ProPublica agree on invoke words like "real", "truth", and "lies" to convince people that a certain politician or political group is trustworthy/untrustworthy.

Interestingly, both models seem to overweight "call-to-action" words, but they sometimes disagree on which words. Both models disagree on advertisements that clearly read like apolitical fundrasing solicitations, such as the 4Ocean bracelet and Wikipedia ads in the above section. However, there is a starker distinction that can be made: the random forest model might be inclined to predict that longer advertisements are less political and potentially overestimate the political nature of shorter advertisements. Each of the ads that random forest predicts significantly less political than ProPublica are quite long, including advertisements not displayed here. Differences between the models aside, they do tend to agree on ads that the average person would consider quite political. The three advertisements displayed where the models agreed quite closely (within a thousandth of a percentage point) either mention political parties or politicians such as Ben Walsh, Val Arkoosh, and Ken Lawrence. Overall, both of these models tend to generally perform well at predicting the political nature of FaceBook advertisements even though future improvements could be made. A model like the one created here could be extended to predict the likelihood of a political ad containing false or misleading information. Considering the amount of political advertising set to take over FaceBook as the 2020 general U.S. election approaches, using machine learning models and natural language processing to classify ads as political becomes an important task.

Limitations and Future Research

The ProPublica data set contains 162324 observations at the time of download, but some of the ads were repeated observations that were not cleaned out. While I did not remove them this time because they did have unique vote amounts, future research projects could try to condense repeated advertisements as well as aggregate the user votes from the different observations. Another potential problem that can't be definitively solved is the pre-processing steps used for this data. Cleaning text using natural language processing methods is an imperfect process that one could always spend more time on. Choices made during the text cleaning can impact the analyses. For instance, we removed punctuation marks in this analysis, but that makes the words "we're" and "were" the same, even though they could have The most important limitation of this project was that we could only analyze a small portion of the unique words found amongst all of the advertisements due to processing limits. The training set included 92383 unique words but we could only analyze approximately 500 of them due to storage limits, approximately 0.54% of the words. Future research conducted with more storage space could utilize more of the words in the training set for the random forest model, allowing us to lower the RMSE further and generate more accurate and fleshed-out predictions.

Another potential research avenue using this same data would be to perform sentiment analysis on the political ads in the data set. Sentiment analysis is a process that quantifies the words used in a text as either positive opinions, negative opinions, or neutral, and then performs machine learning analysis to determine the overall tone, or sentiment, of a text. In this context, we would be utilizing sentiment analysis to determine if political ads contain more negative opinions than non-political ads. While there is a general conception that political ads tend to be attack ads and therefore more negative, a proper statistical analysis could reveal by how much that is.

Table 1: Highest and Lowest Weighted Words for First Two Principal Components

PC 1				PC 2			
Highest	**Weight**	**Lowest**	**Weight**	**Highest**	**Weight**	**Lowest**	**Weight**
really	0.1352	parenthood	-0.00206	thank	0.16723	thing	-0.14667
much	0.1293	mike	-0.00183	love	0.16426	political	-0.14402
just	0.12847	survey	-0.00127	click	0.15437	things	-0.13317
good	0.12735	reproductive	-0.00109	makes	0.15158	republicans	-0.12556
time	0.12277	abortion	-9.4×10^{-4}	link	0.12269	taking	-0.12278

Citations

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

Appendix