Lauren Shade, Data Profile
Data Curation_LIS545
02/10/2020

Dataset and URL: Political Advertisements from Facebook

1. Overview

Context:

Spatial coverage: Always English - USA
Temporal coverage: Date released Feb 2020 - con't
Topics: political advertisements, politics, advertisements, targeted ads

The database I chose to create a data profile on is named Political Advertisements from Facebook. I found this database on the ProPublica Data Store. According to the ProPublica website the Political Advertisements from Facebook dataset "contains ads that ran on Facebook and were submitted by thousands of ProPublica users from around the world. We asked our readers to install browser extensions that automatically collected advertisements on their Facebook pages and sent them to our servers. We then used a machine learning classifier to identify which ads were likely political and included them in this dataset." The producers for this dataset are vast as there are many different types of businesses, NGOs, politicians, and others who use Facebook's Ad Manager.

Because ProPublica is an independent, non-profit newsroom with a focus on investigative journalism, we can assume some context regarding this dataset. More and more, companies and persons are using social media channels like Facebook to send targeted advertisements with malicious intent. ProPublica's dataset tries to shine an objective light onto the types of political advertisements available to its data collectors on Facebook, which may help inform studies, research, and literature around Facebook advertising in local and national politics, international affairs, and more.

Data collectors: ProPublica lists no criteria for data collectors. Data collectors do not have to be a certain age or have a certain educational background. There's some implicit knowledge we can gather from the collection method, however. Since data is collected on Facebook, it can be deduced that data collectors are 16 years of age or older per Facebook's rules and regulations on platform age-requirements. Data collectors must have access to a computer and internet in order to collect data using ProPublica's Political Ad Collector Chrome or Firefox extension. Due to the nature of the data collection process (anyone with access to a computer, internet, and Facebook can collect for this dataset), this dataset is considered to be crowdsourced.

2. Methods

Variables: 23

Sampling strategy: As stated in the overview of this dataset, ProPublica does not have criteria for who can collect data but its methods are less ambiguous. The methods used to produce this dataset can be broken down into two categories--crowdsourcing and machine learning classification.

The method ProPublica uses for data collection is crowdsourcing, which I discussed in the overview. There are inherent limitations to crowdsourcing online in this context. People without access to a computer, internet, and a Facebook account cannot use the Political Ad Collection extension. The Political Ad Collector extension can also not be used outside of Chrome or Firefox. These criteria, though not explicit, already limit the sorts of folks who can be data producers. And limiting the types of folks who can use the political ad collector extension, limits the types of advertisements being collected. So those who may be socioeconomically disadvantaged may not have access to Facebook on a computer (maybe they only use the Facebook mobile app where the extension is unavailable) and therefore can not take part in this dataset's collection methods despite these users likely having unique targeted political advertisements of their own (say on the Facebook mobile app). In general, the software and hardware limitations will also limit the diversity of political advertisements used in this dataset.

To better understand the crowdsourcing data collection done in this dataset, we have to explore the Political Ad Collection extension deeper. The Political Ad Collection extension collects, according to their website, the text and links in an ad, the picture in an ad, information Facebook provides about the ad's targeted audience, the time and date the ad was seen, the number of times the ad has been seen, and the ad's language of origin. The extension works in the background as users browse Facebook. In particular it "places a content script on every Facebook page you visit. That script scans for ads, which it then stores on your computer. These ads are also sent to The Globe and Mail to support research and journalism."

After the data is collected, ProPublica uses machine learning to sort it before archiving it into this dataset. Users--data collectors or folks interested in this dataset for research or other purposes--do not have access to the machine learning algorithm ProPublica used to classify ads as "likely political." The term "likely political" brings its own set of questions. Mostly, without understanding how ProPublica defines "likely political", we can't fully measure the effectiveness of this dataset. What data may be missing from this dataset? What advertisements fall into the cracks because of the vagueness surrounding the "likely political" classification.

ProPublica does have a way for users to help train their critical intelligence on capturing political ads. This can also be done manually by users after the Political Ad Collection extension is installed on Google Chrome or Firefox. When a user sees an ad on their Facebook, they are prompted to answer the question "Which type of ad is this?" or "Is this a political ad?". Using this method, users can determine

(based on their tactic knowledge of political advertisements) if their ads are political or not. Though this option does encourage trust (clearly ProPublica is proactive in teaching the machine learning algorithm that collections its ads), added transparency around these processes can only strengthen its usefulness to end-users.

Privacy: Must consent to use collection extension tool. ProPublica states they do not collect your Facebook ID number, who has liked and shared a post, your name, birthday, friends list, etc. and identifying information about who saw what ad.

Other constraints: Limited number of users downloading extension can affect the diversity of the political advertisements collected in the dataset.

3. Data Dictionary

Data Dictionary		
Variable label	Allowed values	Variable definition
id	Numeric	post id number on facebook
html	String	HTML of the ad as collected by the Political Ad Collector
political	Numeric	number of Political Ad Collector users who have voted that the ad is political
not_political	Numeric	number of Political Ad Collector users who have voted that the ad is not political
title	String	ad title
message	String	ad content
thumbnail	String	link for a thumbnail of the profile image (of the advertiser)
created_at	Numeric	date ad was first collected by the Political Ad Collector
updated_at	Numeric	the most recent time that it got an impression OR the most recent time it was voted on
lang	String	language of the ad. always en-US.
images	String	link for images included in the ad
impressions	Numeric	number of times the ad has been seen by the Political Ad Collector
political_probablity	String	calculated by the classifier. data only includes ads with a probability >=0.7
targeting	String	Facebook's "Why am I seeing this?" disclosure provided to Political Ad Collector users
supressed	String	value is false. suppressed ads are excluded from this data set because they were misclassified.
targets	String	a parsed version of targeting
advertisers	String	the account that posted the ad
entities	String	named entities mentioned in the ad, extracted using software
page	String	the page that posted the ad
lower_page	String	the Facebook URL of the advertiser that posted the ad (the "page" column, lowercased)
targetings	String	an array of one or more of Facebook's "Why am I seeing this?" disclosures provided to Political Ad Collector users
paid_for_by	String	for political ads, the entity listed in Facebook's required disclosure as having paid for the ad
targetedness	Numeric	an internal metric for estimating how granularly an ad is targeted, used for sorting in the ProPublica search interface

4. Reuse Potential

At projects.probulica.org/politcal-ad-collector, ProPublica writes “Online ads are usually seen only by the audience the advertiser wants to target, and then they disappear. This makes it difficult for the public to monitor them and hold advertisers, including political groups, accountable.” This data can be a beneficial instrument in understanding our current political landscape and holding politicians and businesses to a standard of transparency and honesty.

This data may be used by researchers and/or scholars in social sciences and political science or journalists to evaluate, for instance, the rise of deceptive or misleading ads. It can be used to evaluate the role of social media and paid advertisements in the upcoming election(s).

Limitations: Advertisers target by audience whether it be behaviors, education status, relationship status, interests, or all a combination of these. If you’re not using the extension tool, your voice may not be accounted for--those ads for you may also not be accounted for. Researchers and journalists must keep this in mind while using or reusing this dataset.

5. Curation Needs

The dataset is well-curated though there are some limitations on data collectors that may affect the dataset’s future use. The following curation activities can add value to the dataset:

- Information about data collectors (demographics, for instance) can be beneficial to helping researchers and future users make informed decisions about the dataset
- Information about IRB approval is missing from the dataset
- Shedding more light on the quality control and data processing aspects of the collection brought up above can bring more transparency to the dataset, and help researchers and journalists evaluate any limitations and/or constraints the database has--and what they should take into consideration when using or reusing the dataset
- Information about how classifier calculates the probability of political ads - what is the range and meaning (see poltical_probablity in data dictionary)

[1] Political Ad Collector Retrieved February 1, 2020, from https://projects.propublica.org/political-ad-collector/

[2] Political Advertisements from Facebook. Retrieved February 1, 2020, from https://www.propublica.org/datastore/dataset/political-advertisements-from-facebook