Group:

Marek Hulva

Jeremy Glebe

Literature Review Plan

When searching for articles, we will be looking for information which corresponds with the deeper concepts of our research project. (As opposed to surface-level similarities like scraping of Reddit/Twitter data) Papers of interest include: existing efforts towards detecting political biases/position, articles which contain information on detecting sentiment, articles which have addressed anomalies in text posts such as ironic comments, and articles addressing the extraction of topics/subjects from text posts.

Each article that we are reviewing will have three days dedicated to it. Day one will consist of reading the article and writing out a general summary of the article, with a focus on it's aspects that may correlate or pertain to my research project. We're looking for details here. What algorithms are used in the paper? How can those (potentially) be applied to our project? We'll be trying to determine the state of their work and whether it can be improved upon. On day two we will write a draft review of the article based on that initial reading.

Day three will be for cleaning up the review of the article, editing, as well as re-reading the article and looking for any insights we may have skimmed over that should be included. Here's where I will be looking for abstract concepts. (This may seem backwards, but we'd really like to check for substance first. Articles that turn out to only really be conceptual will not be useful) Is there a different way of looking at or understanding the problem that we should account for? At this pace, 6 articles can be completed in eighteen days, or about two and a half weeks.

One major issue is that political affiliation might be difficult to detect for a machine when online posts (especially Reddit) are rife with people speaking insincerely. "Oh wow, her speech was just brilliant" could be read as a compliment or could be understood as an attack against a speaker. Additionally, while determining the subject of a post should not be difficult (we have seen several projects that do this in passing), we will need to investigate how this is accomplished as it is not likely to be trivial. (Perhaps the initial establishing of subject *is somewhat trivial* but the process of correlating subjects, especially when expressed in different forms, so that our results can be compiled into meaningful rules, one would expect to be much more complex) We do not yet have a basis to solve these issues and thus need to investigate them in the literature review.

Data Acquisition Plan

In order to perform big data collection, we will utilize Apache Flume which is a data collection system that will retrieve data and move it all to a central store. Apache flume is a great choice for this project because it comes prebuilt with a Twitter data source (thus streamlining one of our target platforms), it can move and store data in safe and fast forms, and it allows for easy parallelism because it can be adapted for multiple data channels that all move data to the same sink. (as well as the ability to chain agents//collections together) We will use memory channels for Flume because they have very high throughput which is good for larger velocity sources. We will opt for the spillable memory channel instead of the normal memory channel just for the added protection and fault tolerance. We will use the HDFS sink really just because the textbook we are using goes over Hadoop Distributed File System in great detail and it will likely be a tool we will become familiar with and have more experience in as the semester goes on.

Twitter data can be collected in an ongoing manner using the Twitter data source for Flume. We will additionally separate and categorize tweets using their own tagging system (hashtags). Twitter does not keep any databases or data dumps of existing tweets that I can find. (Not any public ones, anyway, as they certainly have to store the tweets) So, to obtain existing data we will need to find already compiled datasets which collected older Twitter data as it came out, or scrape for additional data. For efficiency's sake, we are opting for the first option. "awesome-twitter-data" (license CCO) is a great source for Twitter datasets, many of which are political in nature and some of which are already classified. (https://github.com/shaypal5/awesome-twitter-data) We will utilize this public project to gather existing Twitter data for training.

We will be using PRAW for scraping data from Reddit. PRAW stands for Python Reddit API Wrapper. It is a Python package that allows us to access Reddit API. The Reddit

API encompasses many use cases and will almost certainly be sufficient for the collection of both new and existing data from Reddit. We will be scraping data that is related to political topics such as the covid-19 situation in the United States. It is worth noting that in order to keep a consistent design pattern in our project, the usage of PRAW should be integrated using a custom data source for Apache Flume. That data source will output objects in Avro format (just as the Twitter source does) in order to keep consistency and make processing more uniform.

To ensure that our training and testing data sets are independent, we will simply set up multiple data stores when we begin collection. (or one that is divided in two sections) We will collect data for them without overlap. No single item from a source will be sent to multiple stores. To ensure that we have identical distribution, we can compile data into spreadsheets and generate control charts with excel/sheets. Comparison of these control charts should highlight any glaring problems with distribution. Afterwards, we can try utilizing a time series analysis for further examination.

Literature Review

Example of literature review draft

Partisan Selective Sharing: The Biased Diffusion of Fact-Checking Messages on Social Media

https://academic.oup.com/joc/article/67/2/233/4082394

Jieun Shin, Kjerstin Thorson, Partisan Selective Sharing: The Biased Diffusion of Fact-Checking Messages on Social Media, *Journal of Communication*, Volume 67, Issue 2, April 2017, Pages 233–255,

I chose this study because it is directly related to fact checking I am trying to do in my study. My takeaway from this study is that if I will decide to work with Twitter posts instead of Reddit posts, or maybe both, I will be using the same method of collecting Twitter data.

In this study, there were almost 300 000 000 tweets collected for the data analysis during the 2012 presidential election. All of these tweets included at least 1 out of 427 keywords that could be typically found in political Twitter conversation. Examples of the types of keywords were names of candidates, political parties and also terminology that was specific to a typical political issue.

Data for this study was collected using Gnip PowerTrack service, which has access to the twitter firehose that allows access to all public Tweets, as well as metadata about those tweets.

Fact checking websites were used to determine which party was the fact checking more favorable for. They collected face checking for three variables:

- 1) Political party that gained a relative advantage from the fact check
- 2a) valence of the fact check towards Barack Obama
- 2b) valence of the fact check towards Mitt Romney

The values of the three variables were following:

- advantageous to the republican party
- advantageous to the democratic party
- neutral

Each fact checking tweet was coded for its valence towards each candidate: **positive, negative or neutral.**

SentiStrength was used as an analysis tool to assign the valence of the given text from Twitter posts and it was supplemented by labeling it’s relatability by determining if the tweet was retweeted by accounts that democrats or republicans tend to follow.

New Data Sources in Social Science Research: Things to Know Before Working With Reddit Data

ew Data Sources in Social Science Research: Things to Know Before Working With Reddit Data

Ashley Amaya, Ruben Bach, Florian Keusch, Frauke Kreuter

First Published December 18, 2019 Research Article

This paper talks about collecting data from Reddit, its advantages and disadvantages and methods.

Most researchers are using Reddit for their data collection, mostly collecting comments, posts and votes, along with a tag that serves as a keyword to the post that will be associated with the content the user is posting about.

There are two ways to access the Reddit database. You can either download them as Reddit content or use Reddit API.

The downloadable data sets can be downloaded as large compressed json files, between 6 - 9 gigabytes large. Disadvantage of this method is that the file might be too large and researchers might not be able to open the entire file at once.Reddit API is better for researchers, because it can be used to target specific content. Researchers can choose which users to target and which subreddits and threats to focus on, as well as which date. Reddit API also allows to structure data in multiple ways. Therefore using Reddit API is superior then to the other method.

Popularity information is also available, based on how many views and how many likes did the comment received. The paper also talks about the importance of cleaning the data, as some comments may just be duplicates. Also, deleted posts are also contained in the datasets. It was found that 5.5% of all posts were deleted by their user. There are also posts in foreign languages. Another aspect of messy data on Reddit is typos and social media language, for example “LMAO, LOL, and FYI” etc…

Disadvantage of using Reddit data for research is that all users are anonymous, that means there will be no demographic information, which might be an issue for some political studies that are focused on determining demographics of political supporters and affiliates.

Echo Chamber or Public Sphere? Predicting Political Orientation and Measuring Political Homophily in Twitter Using Big Data

Elanor Colleoni, Alessandro Rozza, Adam Arvidsson

https://academic.oup.com/joc/article/64/2/317/4085994

Journal of Communication, Volume 64, Issue 2, April 2014, Pages 317–332,

I chose this study because of the method they determined Republican and Democrat supporters based on their posts from Twitter accounts. This study was using a Big Data approach combining machine learning-based analysis of textual content to analyze homophily of Twitter users. They identified whether users were either Democrats or Republicans by analyzing the content they shared through their Tweets.

In this study, researchers scraped data off political tweets in order to measure political homophily on Twitter. Data was obtained by scraping all the political tweets of users based on if they follow either Democrat or Republican accounts. The users who are following both parties were not counted.

They identified 1,683 users as Democrats from 28,167 Tweets and 8,898 users from 189,933 Tweets. They used a corpus of text labeled to fit a category (Political/Non-political and Democrat/Republican for) as training data. Training set was then used to classify the algorithm, then they tested the algorithm by measuring its ability to classify each example.

To distinguish Republican from Democrat content, they trained an algorithm to compute a frequency of words that were contained in *DemRepTrainingSet.* They reached 79% accuracy on DemRepTrainingSet.

To evaluate the overall political orientation of a user, they counted the number of tweets classified as Democratic and Republican.

When they compared the political orientation of users based on following either Republican or Democratic accounts, and compared them to users whose political orientation was determined based on the content of their tweets, they found that the number of Republicans that follow Republican politicians is 10 times higher than with Democrats following Democrats.

The Effect of Fact-Checking on Elites: A Field Experiment on U.S. State Legislators

Brendan Nyhan Jason Reifler

https://onlinelibrary.wiley.com/doi/full/10.1111/ajps.12162

This article examines to what extent fact-checking in politics has an impact on political discourse.

This article helps to highlight the importance of fact-checking that will be crucial for our research, as we examine the spreading of fake news regarding fake news and promises of Covid-19 situation in the United States.

This study found that possible political fact-checking decreases chances of legislators to make inaccurate statements during their political campaign.

These findings are huge for our study and future studies as well, because the more fact-checking of political statements will be present, the more will politicians be motivated to make more accurate statements and voters will then have the choice to make decision based on factual statements rather than fake news.

As increased frequency of political fact-checking is found to have effect on political behavior, if the findings of our study will be posted, politicians and especially Reddit and Twitter users might be more careful to spread fake news because they will now that they will be held accountable for the credibility of their tweets.

Social media competitive analysis and text mining: A case study in the pizza industry

WuHe, ShenghuaZha, LingLi

https://www.sciencedirect.com/science/article/pii/S0268401213000030?via%3Dihub

This study performed analysis for three large pizza chains on social media by looking at each chain's pattern that they use on their Twitter and Facebook accounts and looked for differences in between them.

The text mining of the chain's social media started with pre-processing - formatting the data into usable format from raw data by cleaning and assigning attributes to the data. SPSS Clementine and Nvivo 9 were used for text mining and analysis. SPSS Clementine was used mostly for generating key categories and exploring key components for the textual data. NVivo 9 was used for query searches to find patterns and connections in the data.

As in this study, looking for similarities in data will be crucial for our study to determine a user's political affiliation to determine whether they are Democrats or Republicans, as well as making connections between their posts and fake-news stigmas to determine factuality of the content they are posting about covid-19 situation in the United States.

An investigation of data and text mining methods for real world deception detection

Christie M.Fuller, David P.Biros, DursunDelenc

https://www.sciencedirect.com/science/article/pii/S0957417411000522

In this study, researchers analyzed statements to find commonality in deceptive speaking amongst stories of different persons of interest. They determined 30 deceptive cues based on previous research to analyze these statements (similar process that we will be doing in our study, where we determine typical lies democrats say about covid-19 situation in the United States).

The statements were processed by two software packages: General Architecture for Text Engineering (GATE) and Linguistic Inquiry and Word Count (LIWC) to identify and calculate desired features.

They used three detection methods: artificial neural networks, decision trees and logistic regression.

The one that we would be interested in the most for the purpose of our study are Decision Trees. Decision Tree is a data mining technique used to build classification models in the form of a tree-like structure into root nodes, branches, and leaf nodes. This technique could be also used for our project to fact check statements about covid-19 situations in the United States.