Predicting Media Partisanship in Relation to the 2020 U.S. Presidential Election

Marisa Papagelis and Natalie Reid[1]
Wellesley College, USA
mpapagel@wellesley.edu nreid@wellesley.edu

Abstract

Facebook is becoming an increasingly popular distributor of political news, especially among younger generations. As the 2020 U.S. presidential election approaches, news outlets are reporting on a variety of political topics, and their readers are interacting with them on a widespread scale. This project uses these ideas to conduct an investigation into the most frequently discussed political topics across partisan and nonpartisan news sources on Facebook and builds a classification system to predict partisanship of news posts. To accomplish this, the code for the project uses Selenium with Chromedriver and BeautifulSoup to scrape, store, and parse publicly accessible posts on Facebook, and the NLTK package for Python to create a supervised classification system.

1 Introduction

Since the results of the 2016 presidential election, many researchers have shown that the use of Facebook as a news source had a significant impact on voters' attitudes towards the two major-party candidates [Vogels *et al.*, 2020]. The prevalence of fake and biased news was widespread, and Facebook responded to this misinformation epidemic by creating new tools to combat it [Facebook, 2017]. Sharing news on Facebook still remains a very popular activity and as the 2020 presidential election approaches, engagement with news sources on the platform is increasing. By scraping the content of these pages, one can determine the most frequently discussed issues and gain insight into the most relevant topics of the election cycle, from both a partisan and nonpartisan perspective. A significant difference between the frequency of topics discussed on right- and left-wing pages could indicate a difference in priority of issues between the two parties and could provide a window into the opinions and motivations of prospective voters. With the understanding that many Facebook users seek out partisan information that aligns with their political beliefs, it follows that understanding and identifying partisanship may be a growing concern among digital citizens. With this in mind, we investigated the possibility of creating a classification system that could accurately label partisanship of Facebook posts by popular news pages. To help guide our thinking, our three research questions were written as follows:

RQ1: Which topics (keywords) are most commonly discussed on partisan Facebook news pages leading up to the 2020 U.S. presidential election, and how do they compare between pages?

RQ2: Can we build a classifier that predicts the partisanship (left-wing, center, or right-wing) of a given post from one of three predetermined partisan news pages on Facebook?

RQ3: Can the classifier, trained on the three predetermined partisan news pages, be generalized to predict partisanship of given posts from other partisan news pages that were not represented in the initial training data set?

2 Related Work

The following subsections explain the themes of this study and provide context for topics such as web scraping, biased news, censorship, and others.

2.1 Legality and Ethical Issues

Web scraping on Facebook brings about many legal and ethical concerns to ensure that research does not interfere with user privacy in a harmful or negative way. It is the moral responsibility of the researcher to preserve data privacy wherever possible to ensure integrity in their research [Mancosu and Vegetti, 2020]. Data should not be identifiable, and only the minimum relevant data should be collected. Additionally, web scraping procedures are still problematic in terms of TOS (terms of service) compliance [Mancosu and Vegetti, 2020]. Currently, the act of performing screen scraping on Facebook violates the TOS, potentially causing the researcher to be susceptible to actions such as "immediate ban" or "injunctive relief" from Facebook. Additionally, in March 2020, ACLU convinced a court to say that violating the TOS for auditing purposes is

[1] M.P. developed the script to collect the data, developed the script to explore/analyze the keyword frequency data, developed the script to explore/analyze the data modeling/classifier data, created the visuals/tables shown in this paper, and annotated/organized both notebooks using Markdown. N.R. developed the script for the classifier creation/implementation and wrote the abstract, introduction, discussion, and conclusion. We worked on the other sections and edited the paper together.

not criminal [ACLU, 2020]. Many researchers are making use of this court decision.

2.2 Political News Source Bias

Traditionally within the news and media industry, many news outlets are known to fit into three categories, with plenty of space on the spectrum in between. These three pinpoint categories consist of left-wing biased, right-wing biased, and neutral. Researchers have examined biases in online news sources looking at selection bias, coverage bias, and statement bias [Saez-Trumper *et* al., 2013]. It is important to note how each of these three biases fit in context with politics. In politics, selection bias is the preference for selection stories from one party. Coverage bias is the preference for giving a larger amount of coverage to stories about one party. Lastly, statement bias is the preference for expressing more favorable statements for one party [Saez-Trumper *et* al., 2013]. Demonstrating selection bias, this study concluded that the number of different *stories* a news source publishes is correlated (r-squared = 0.83) with the number of *articles* that are published [Saez-Trumper *et* al., 2013]. Furthermore, researchers discovered users typically perceive news outlets to be biased in a way that is dependent on their own partisanship, and so will continue to share articles that they feel are more objectively truthful [An *et al*., 2014]. This finding holds true for both high and low-activity users regardless of political affiliation [An *et al*., 2014]. In a networked environment like Facebook, these partisan interactions create a feeling of validation and a sense of involvement - two key attributes that encourage the behavior to continue [Oeldorf-Hirsch and Sundar, 2014].

2.3 Political Viewpoint Censoring

How would an individual user know whether certain viewpoints were being censored on Facebook? Roughly 75% of Americans believe that social media sites intentionally censor political viewpoints that they find objectionable [Vogels *et al.*, 2020]. Most prevalent among conservative users, this belief leads to decreased trust in social media sites such as Facebook. In May of 2016, Facebook was publicly accused of censoring conservative viewpoints through manual manipulation of the 'trending topics' page. While Facebook is not itself a news network and therefore has no journalistic obligation to report the objective truth, the site purports to have no political affiliation and technically have no reason to censor any viewpoints unless it serves the site's business interests. At a time when 62% of U.S. adults receive news from social media and 44% of that activity is localized to Facebook, political censorship is clearly evolving into an issue that has a deep impact on our daily lives [Carlson, 2018].

2.4 Fake News and Misinformation

The prevalence of fake and misleading news on Facebook is a known issue - especially as that news relates to current topics in politics. Often hard to spot, Facebook has developed tools that aim to disrupt the cycle of misinformation spread and provide users with facts about their news consumption [Facebook, 2017]. However, it is important to understand that older age groups are most likely to believe fake news because they have not experienced the kind of digital literacy training that has been given to the younger generations [Guess *et al.*, 2019]. Older age groups never learned to examine news source credibility by looking at qualities such as suspicious web addresses and misuse of quotations. It has been demonstrated that age is by far the best predictor of the likelihood for a user to share misinformation. Age holds constant over other attributes, such as ideology and party identification [Guess *et al.*, 2019].

2.5 Classification Systems

Classification systems are an important tool in studying news values, public opinion, negative campaigning, or political polarization in a digital setting. Using sentiment analysis as an example, researchers have presented a computer-based procedure for collecting sentiment scores through crowd-coding to build a negative sentiment dictionary [Haselmayer and Jenny, 2016]. Crowd-coding enables the calculation of the tonality of sentences and of a measurement of sentiment. In a political setting, media negativity is used in order to measure the tonality of media stories and the degree of conflict or confrontation in the news [Haselmayer and Jenny, 2016]. The researchers concluded that there is a strong correlation (0.82) between a computer automated crowd-coding score, like shown in their procedure, and an expert score of a human being performing the same procedure by hand [Haselmayer and Jenny, 2016]. Using a dictionary to cite keywords can be helpful in parsing through additional stories to measure sentiment in the future.

3 Methods

This section contains methods used to collect data, create variables, clean data, and perform analysis. Below is a description of the methods used.

3.1 Data Collection

Reuters, MSNBC, and Fox News are three reputable news sources with official pages on Facebook that were chosen for the exploration. Using the Ad Fontes media bias chart, we classified the three sources as center/neutral (Reuters), left-wing (MSNBC), and right-wing (Fox News). Selenium was used to automate scrolling through these pages and to scrape and save the content to an HTML file. Python's BeautifulSoup library was used to parse the HTML files for the post text in each post. Pandas, another Python library, was used to format the scraped information into a data frame, and the data frames were saved to JSON files for future analysis. For a further exploration, 200 posts each

were collected from six more news pages: AP and Bloomberg (center), Wonkette and CNN (left-wing), and The Daily Caller and The Washington Times (right wing). It is important to note that the Ad Fontes media bias chart classifies Wonkette and The Daily Caller as providing misleading and/or incomplete information, often with bias.

3.2 Variable Creation

The list of keywords counted for frequency is as follows: coronavirus/covid, president, government, police, economy, supreme court, race, climate change, congress, healthcare, crime, guns, foreign policy, immigration, inequality, abortion. 'coronavirus' and 'covid' were counted as one category because they refer to the same topic. This list of keywords was developed from a Pew Research Article that discussed the most important issues relating to the 2020 presidential election [Pew, 2020].

3.3 Data Cleaning

Data was stored in JSON files labeled with the name of the corresponding news page. Some posts did not contain any text and therefore could not be used for our data exploration, so they were manually removed from the JSON files before any analysis or exploration was performed. In order to ensure consistency, each file was a subset within the Python notebook and included only the first 800 posts with text. For the section that counted keyword frequency, all text was converted to lowercase for accuracy.

3.4 Analytic Methods

To understand media partisanship in relation to the 2020 U.S. presidential election, analytic methods to collect and analyze keyword frequency as well as classification systems to predict media partisanship were implemented.

3.4.1 Keyword Frequency

In order to measure keyword frequency, each news source was given a dictionary containing the predetermined keywords as keys. The text of the posts was looped through, and each time a post contained a keyword, the value of the key in the dictionary was increased to reflect the number of posts containing each keyword. Posts containing multiple keywords were counted multiple times, once towards each key.

3.4.2 Supervised Classification

A supervised classification system was built, based on training data from the three original news sources, to predict the partisanship of a post given its text. The training labels on the data were "left-wing," "right-wing," and "center" to correspond with MSNBC, Fox News, and Reuters respectively. The testing data set originated from the three training news sources, and then the testing data set was generalized to six unique news sources for investigation.

4 Results

The results section will focus on three research questions that examine news sources to discover relevant 2020 U.S. presidential election topics and partisanship bias using text classification systems.

4.1 RQ1: Presidential Election Keywords

We calculated the frequency of our predetermined topics relating to the 2020 presidential election across all three news sources as well as within each news source individually. We used percentages to calculate relative frequency in relation to our dataset. 83% of posts contained topics relating to the 2020 U.S. presidential election and 17% of posts did not. Further breaking this information down into our three partisan news sources as seen in Fig 1, Fox News had the highest overall keyword frequency, with roughly 95% percent of all scraped text posts containing at least one keyword. Roughly 83% and 71% percent of posts from MSNBC and Reuters respectively contained keywords, potentially indicating that Fox News had more posts overall relating specifically to the 2020 presidential election.

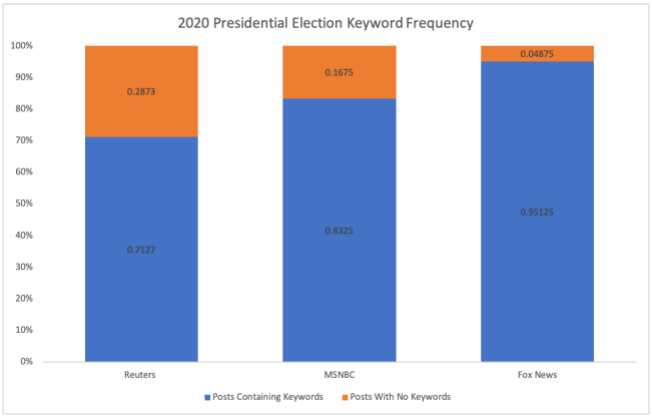

Fig 1 Frequency of posts that contained at least one election keyword, separated by partisan news source

Fig 2 displays the percentage of posts containing the top keywords per news page out of all of the posts from that page which contained at least one keyword. Fig 2 is visualized as a heat map. The two most common keywords across all three news sources were "coronavirus/covid" and "president." This data was to be expected because at the time of the data collection, most U.S.-based news sources were reporting on President Trump's stay at the Walter Reed Hospital following his COVID-19 diagnosis. The fourth most common keyword across all three sources was "police," and MSNBC was the only source to have "race" appear in its top five most frequent keywords.

Heat Map of Most Common Keywords					
		News Page			
		Reuters	MSNBC	Fox News	Overall
Keyword	coronavirus/covid	54%	35%	24%	36%
	president	17%	50%	56%	43%
	government	10%	1%	1%	3%
	police	6%	3%	6%	5%
	economy	4%	1%	1%	2%
	supreme court	3%	7%	10%	7%
	race	2%	2%	1%	2%

Fig 2 A heat map of the top 5 most common keywords from each news source

4.2 RQ2: Partisan Classification System

The supervised classifier was given 700 text posts to train on and 100 posts to test from each Facebook news page. The training set was constructed as a list of tuples within indices containing the entire post text and a label of either "Center," "Right-wing," or "Left-wing," indicating whether they originated from Reuters, Fox News, or MSNBC respectively.

4.2.1 Accuracy

The overall accuracy of the classifier was approximately 80%, which is much higher than chance for a 3-class classifier (33%), indicating that there is high reliability. To further understand the results, we created a confusion matrix to show the breakdown of the classification. Fig 3 shows the confusion matrix as a colorized heat map, with superimposed values representing the frequency of that specific classification or misclassification. The left to right diagonal contains the true positives from the classifier testing, and we can see that the classifier most frequently correctly classified the right-wing posts 90 out of 99 times. It also correctly classified left wing and center posts 75 and 72 times out of 99, respectively. On average, the classifier misclassified posts as right-wing the most frequently, with 24 center and 19 left-wing posts misclassified as right-wing. This contributes to the higher overall true positive value for right-wing sources.

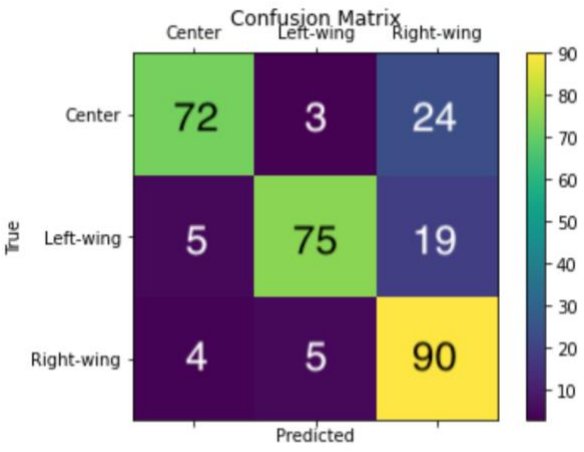

Fig 3: A confusion matrix for the output of the classifier, colorized as a heat map

4.2.2 Interpreting Accuracies

To interpret the results of our confusion matrix, and furthermore, to investigate why Fox News had the greatest classification accuracy, we perform analysis of post length and percentage of words in train/test set to find possible answers.

Post Length We calculated post length in order to see whether there was a correlation between the number of words in a post and the accuracy of a classifier. As shown in Fig 4, MSNBC had the largest average post length with 31.4 words followed by Reuters with 30.5 and Fox News with 27.5, so the hypothesis is proven incorrect.

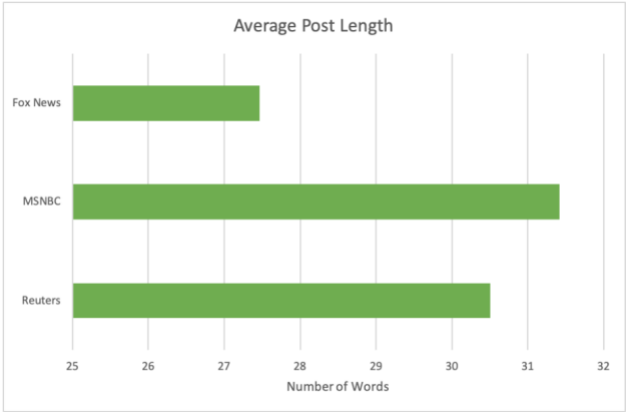

Fig 4: Average length of posts scraped from Fox News, MSNBC, and Reuters

Percentage of Words in Testing and Training data sets We answered the question regarding what percentage of *unique* words each test data set was seen in its corresponding training data set by using a scikit count vectorizer. Fig 5 contains the resulting percentages. Overall, 70.9% of words in the whole testing data set were also seen in the whole training data set. Furthermore, Reuters saw the greatest percentage of unique words (54.3%) from its testing data set inside its training data set, followed by MSNBC (48.7%) and Fox News (49.6%). There appears to be a correlation between Reuters percentage score and its classifier accuracy. The fact that Reuters test data set contained the most unique words from the training data set could explain why the Reuters classifier was the least accurate.

Percentage of Words in Test Data Set Seen in Corresponding Training Data Set			
Total (all 3)	*Reuters*	*MSNBC*	*Fox News*
70.9%	54.3%	48.7%	49.6%

Fig 5 A table answering the question: what percentage of *unique* words in each test data set was seen its corresponding training data set?

4.3 RQ3: Generalization of Classification System

As a further exploration, we decided to scrape six more Facebook news pages and test whether our trained classifier could accurately label their partisanship. A subset of 200

posts per page was used to create new testing data sets, which were then tested both together and individually using the classifier. A summary of the accuracy scores is shown in Fig 6. The additional sources were selected based on their position on the Ad Fontes media bias chart, with two being rated as moderate and factual, two as left leaning, and two as right leaning. One source each from the left and the right wings were rated as providing false, inaccurate, or misleading information. These two sources were included to explore how the classifier behaved when presented with unreliable information. From Fig 6, we can see that the classifier was less accurate when using these six sources, resulting in an accuracy only slightly greater than chance (33%). Interestingly, the classifier was more frequently accurate when labeling Wonkette and The Daily Caller—the two news sources that were rated the least reliable in the new set. In other words, our classifier appears to be predicting less than or equal to chance, with the exceptions of Wonkette and The Daily Caller, which are deemed as unreliable and containing misinformation. Our classifier appears to be predicting inaccurately or guessing. Our classifier has not been trained enough to be generalized to additional news sources outside of the chosen three (Reuters, MSNBC, Fox News).

Partisanship and Accuracy of Additional Selected News Sources

News Source	*Partisanship*	*Accuracy*
AP	Neutral	0.33
Bloomberg	Neutral	0.33
Wonkette*	Left-wing	0.62
CNN	Left-wing	0.42
The Daily Caller*	Right-wing	0.55
The Washington Times	Right-wing	0.39
All 6 sources run against the tester	*2 neutral, 2 left-wing, 2 right-wing*	*0.44*

*Ad Fontes media bias chart marks this source as unreliable, containing misleading information, incomplete story, opinionated, etc. https://www.adfontesmedia.com/interactive-media-bias-chart-2/

Fig 6: A table showing the accuracy of the classifier on six selected partisan news sources individually as well as combined

5 Discussion

While not all of our hypotheses were supported, the wealth of data we collected provides some valuable insights into the partisanship of Facebook posts from news pages.

5.1 RQ1: Presidential Election Keywords

The findings of our keyword analysis showed that the most commonly found keywords across the three original pages were “president,” “coronavirus/covid,” “supreme court,” “police,” “government,” “economy,” and “race,” with “president” and “coronavirus/covid” being the most common. There was no significant difference in frequency between the pages, indicating that the issues in question are being discussed with the same regularity across party lines. While these findings do not account for sentiment towards the issues being investigated, we show that there is consistency in mentions with a high confidence level.

5.2 RQ2: Partisan Classification System

As shown above, we were successfully able to build a supervised classifier with training data from the Reuters, MSNBC, and Fox News Facebook pages. The classifier had an 80% accuracy score when run on testing data from the three pages, and more frequently labeled posts as “right-wing” for reasons stated earlier. The accuracy score is much higher than a chance score, so we consider this to be an overwhelming success for the specific set of testing data. In order to interpret these results, we experimented with two hypotheses. First, we studied post length and concluded that the range in average number of words was ~4 words, which would not make much impact on the classifier. We also saw no pattern or correlation between post length and accuracy in our three news sources. Secondly, we calculated the percentage of unique words in each testing data set that were also seen in each corresponding training data set. We did not have much insight into Fox News or MSNBC, but we speculated that Reuters accuracy score, the least out of the three news sources, may have been linked to the fact that Reuters percentage score was 54.3%, which was the highest of the three news sources. The two hypotheses we studied did not show us why the accuracy of Fox News was so large compared to those of MSNBC and Reuters. Lack of this information is a shortcoming of our research project brought up by the limitation of time. In a setting with a relaxed time constraint we would further hypothesize that the frequency of words has an impact on the accuracy of the classifiers. That is, we would aim to answer the question: of the unique words in each test data set that are also in the corresponding training data set, does the number of times each word appears relate to each classifier’s accuracy, and if so, how?

5.3 RQ3: Generalization of Classification System

We concluded that our classifier, trained on the three original news sources, was not generalizable to other sources that were not represented in the training data set. While all individual accuracy scores for each news source were at or above chance, the wide variation shows that the classifier is much less effective than before. We found it interesting that the classifier had the highest accuracy scores for the two historically misleading sources. Because of this, it may be true that our classifier, using the Naive Bayesian approach, is picking up on linguistic quirks that are more often found in misleading news sites, which in turn gives them a higher accuracy score. We would need further exploration to confirm or deny this theory.

5.4 Reflections

In total, data was collected from nine pages on Facebook. Although it was evenly distributed in partisanship, this was not an entirely representative sample of all news coverage related to the 2020 U.S. presidential election. We consider this to be a notable limitation of our data, and if this project were to be continued, we would seek to add more sources to

our training and testing data sets for both the supervised classifier and keyword frequency explorations. Additionally, the labels "left-wing", "center", and "right-wing" do not completely capture the diversity of political views in the United States. Even within the two major political parties, there are a wide variety of attitudes towards the issues we considered in our keyword exploration.

We know from our research that Facebook users tend to see and interact with posts only from news networks that share their partisan beliefs, and this effect is known as an echo chamber [An *et al*., 2014]. In order to combat this, Facebook could implement a recommender system that incorporates keyword frequency and partisan classification to show users articles with similar topics from other news pages. For example, if a user interacts with several posts from left-wing pages that all have to do with the second amendment, Facebook could suggest articles on that same topic from neutral and right-wing pages. This could be accomplished by matching posts with a high frequency of specific words related to a politically dividing issue across partisan lines, and building a classification system that can correctly label the partisanship of such posts with a high accuracy. This would provide users with opportunities to engage with beliefs they are unfamiliar with and could begin to mitigate the echo chamber effect.

6 Conclusion

This project served as a preliminary exploration into the partisanship of news sources on Facebook through a data mining and classification-based approach. With the 2020 U.S. presidential election quickly approaching, the distribution of partisan news on social media has the potential to affect millions of voters, including those who vote strictly within the lines of their party and those undecided. As we have shown, the frequency of election-related keywords is consistent across news pages, indicating that left- and right-wing pages are discussing the same election-related topics at roughly the same frequency. A supervised classification system using the Naive Bayes approach is able to correctly classify the partisanship of a given news post with high accuracy when the training and testing data consist of posts from the same news sources. However, the accuracy falls dramatically when other sources are included in the testing data set. A sensible follow-up to this project would include larger data samples from quite a few more sources which could train a more accurate generalized classifier. We hope that this paper inspires readers to design better systems that use multiple approaches and analyses to predict partisanship.

Acknowledgments

The authors are very grateful to Professor Eni Mustafaraj and Junita Sirait for their guidance and contributions throughout the research process and writing of this paper.

References

[ACLU, 2020] ACLU. (2020) Sandvig v. Barr – Memorandum Opinion. Retrieved October 16, 2020, from https://www.aclu.org/sandvig-v-barr-memorandum-opinion?fbclid=IwAR0O3jrMJx1EVyRDJzAHYlQhpWdT5KWEl_QGiNfXdLvd16gBsPWg2wR6NhU

[An *et al.*, 2014] An, J., Quercia, D., & Crowcroft, J. (2014). Partisan sharing. *Proceedings of the Second Edition of the ACM Conference on Online Social Networks - COSN 14*. doi:10.1145/2660460.2660469

[Carlson, 2018] Matt Carlson. Facebook in the News. (2018) Retrieved September 27, 2020, from https://www.tandfonline.com/doi/abs/10.1080/21670811.2017.1298044

[Facebook, 2017] Working to Stop Misinformation and False News. (2017, April 7). Retrieved September 27, 2020, from https://www.facebook.com/facebookmedia/blog/working-to-stop-misinformation-and-false-news

[Guess *et al.*, 2019] Guess, A., Nagler, J., & Tucker, J. (2019, January 01). Less than you think: Prevalence and predictors of fake news dissemination on Facebook. Retrieved September 27, 2020, from https://advances.sciencemag.org/content/5/1/eaau4586?rss=1&fbclid=IwAR0AnmmBOuikMvGya9AUs3Zd0418CI4aeKLocjhhPfZIQJcgXbGyw3Ix-nE

[Haselmayer and Jenny, 2016] Haselmayer, M., & Jenny, M. (2016). Sentiment analysis of political communication: Combining a dictionary approach with crowdcoding. *Quality & Quantity, 51*(6), 2623-2646. doi:10.1007/s11135-016-0412-4

[Mancosu and Vegetti, 2020] Mancosu, M., & Vegetti, F. (2020). What You Can Scrape and What Is Right to Scrape: A Proposal for a Tool to Collect Public Facebook Data. *Social Media Society, 6*(3). doi:10.1177/2056305120940703

[Oeldorf-Hirsch and Sundar, 2014] Oeldorf-Hirsch, A., & Sundar, S. S. (2014, December 10). Posting, commenting, and tagging: Effects of sharing news stories on Facebook. Retrieved September 27, 2020, from https://www.sciencedirect.com/science/article/pii/S0747563214006232

[Pew, 2020] Important issues in the 2020 election. (2020, September 04). Retrieved September 27, 2020, from https://www.pewresearch.org/politics/2020/08/13/important-issues-in-the-2020-election/

[Saez-Trumper *et* al., 2013] Saez-Trumper, D., Castillo, C., & Lalmas, M. (2013). Social media news communities.

Proceedings of the 22nd ACM International Conference on Conference on Information & Knowledge Management - CIKM 13. doi:10.1145/2505515.2505623

[Vogels *et al.*, 2020] Vogels, E. A., Perrin, A., & Anderson, M. (2020, September 18). Most Americans Think Social Media Sites Censor Political Viewpoints. Retrieved September 27, 2020, from https://www.pewresearch.org/internet/2020/08/19/most-americans-think-social-media-sites-censor-political-viewpoints/