Community Effects From Misinformation Flags on Twitter

Nigel Doering, Raechel Walker, Tanuj Pankaj

Abstract. Recent events including the 2016 election, COVID-19 pandemic, the 2020 election, and the development of a COVID-19 vaccine has laid bare the essential need to prevent misinformation from spreading uncontrollably on social networks. Social media companies have developed systems for preventing the further spread of misinformation. Most notably, some companies have begun placing flags that warn a user of the misinformative content of the post. Research has addressed a way to analyze Twitter users on how conservative versus liberal, moderate versus extreme, and pro-science versus anti-science they are based on their tweet history [2]. We detail a novel machine learning approach to classify users based on three similar dimensions. We then conduct an analysis comparing Twitter users who retweeted flagged tweets to those who retweeted unflagged tweets, with the tweets coming from high profile conservative Twitter users, such as Eric Trump. Results from the analysis suggest that users who are sharing these flagged tweets tend to be slightly more liberal and more moderate than the users who are sharing unflagged tweets. We propose possible explanations, as well as future work to better understand the impact of misinformation flags.

Introduction. The spread of misinformation among social networks is a critical problem facing societies today. The world has faced rapid and widespread dissemination of misinformation through social networks. Recently, users on social networks have sewn fear and mistrust regarding the US election process by spreading false information about the voting process [5]. As well, after the election, doubt was further spread about the validity of the election based on incorrect and intentionally misleading information [6]. With the threat that misinformation poses to societies, social networks, doubtlessly encouraged by public dismay, have sought to find ways to rein in the spread and prevalence of misinformation within their networks. Twitter, in particular, has developed a method of flagging tweets that contain potential misinformation as a way to inform users of the malicious content while deterring its continued spread. Current understanding of the effects of this system is not well understood, however, preliminary research has shown that there may be real benefits in deterring some types of misinformation [1]. There is a need to understand the effects these systems are having in order to improve upon them. Specifically, given flagged and unflagged tweets from the same users, are different types of users interacting with the flagged tweets compared to the unflagged tweets? By exploring this question, more work can be done to understand how to improve the effectiveness of misinformation prevention systems. We combine machine learning and a user's polarity, a quantitative measure of the users preferences for information, to give us comprehensive three dimensional polarity scores. We collect a sample of unflagged and flagged tweets coming from several different popular conservative Twitter users. We then conduct an analysis on the differences and similarities of a community of users that interacted with the unflagged tweets and

a community that interacted with the flagged tweets. We are then able to better understand how misinformation flags change the makeup of the users who interact with flagged tweets.

Related Work. In *Echo Chambers Surrounding Misinformation on Twitter,* researchers developed a methodology for calculating a user's polarization, i.e. whether they tend to like misinformation [3]. The researchers demonstrated that using an analysis of hashtags that are correlated with misinformation they can get a reasonable idea of whether a user, based on their use of hashtags, tends to like misinformation. First, they gathered a few hashtags that are indicative of misinformation amongst COVID-19 tweets, for instance #hoax. Second, they gathered hashtags indicative of scientific understandings of COVID-19 like #wearamask. Using these several marker hashtags they then defined a concept of hashtag polarity which describes how much a new hashtag tends to correspond with misinformation related hashtags or scientific ones. Using these hashtag polarities the team analyzed users' hashtag history in order to then describe the polarity of that user. This concept of polarity allowed the researchers to analyze echo chambers surrounding COVID-19 misinformation. While we found this methodology unstable for our analysis, it nonetheless helped lay the groundwork for understanding a user's preferences from a polarity score.

Furthering the idea of user polarities, Rao et al. in *Political Partisanship and Anti-Science Attitudes in Online Discussions about Covid-19* designed a methodology for giving a user a polarity score for three separate dimensions. In this study, the researchers, in order to quantify a user's attitudes, defined a three dimensional polarity score based on a user's attitudes regarding science, politics, and moderacy [2]. Each user was then defined based on their tweet history in regards to how pro-science versus anti-science, conservative versus liberal, and moderate versus extreme they are. Using these scores, the researchers were able to understand how polarized users in one dimension are related to those in another dimension. For instance, they found that users with politically moderate views tend to hold more pro-science views, while those with hardline views, on the left and right, tend to hold more anti-science views [2]. While Rao et al. use more extensive techniques in order to define polarities, we nonetheless adopt their three dimensional polarity score in our own methodology. We design a machine learning approach that uses a classifier to calculate the user's polarities. We find that our approach works well and some results are able to be confirmed intuitively. Our dimensions change slightly, focusing instead on how conservative versus liberal, moderate versus extreme, and credible versus incredible a user's preference for information is. Using these three dimensional polarity scores, we then conduct our analysis regarding the change of community characteristics between flagged tweets and unflagged tweets.

While there is doubt whether warning labels are sufficient enough to stop a user from spreading misinformation, there is evidence of their effectiveness from a study done by the China Media Project on the change in engagement with Twitter users that Twitter decided to label as state

affiliated with the Chinese Communist Party. Essentially, Twitter made the decision that media profiles on Twitter that are state affiliated, discounting independent government sponsored organizations like the BBC in Britain and NPR in the U.S., should be flagged as state affiliated. Their choice was based on the idea that users have a right to know the context of the information that they are consuming. The study found that there was a dramatic decrease in terms of favorites and retweets with the users that had the label placed on them [1]. Furthermore, users with the flag noticed dramatic changes in the amount of new followers they were receiving. This may be a result of a change in Twitter's algorithms so that flagged users no longer appear in the search bar when users type words related to their username. Interestingly to our study, this work demonstrates that although there is a change in engagement after a user has had a flag placed on them there is still a reduced community of users that are interacting with the flagged tweets. We suggest that this reduced community portrays a similar set of characteristics that can be partially understood along the three dimensions of polarity we defined. While this group's work focused more on the effectiveness of misinformation flags in this particular context, we continue their work but with a focus on understanding the types of users that are ignoring misinformation flags. We also note the differences of their study with ours, as the state affiliated flags were attached to all tweets of a user, whereas misinformation flags are attached to only a small minority of tweets posted by even the most controversial users.

Modeling & Data. In order to conduct our analysis, we laid out a methodology for calculating the polarities of users that retweeted flagged and unflagged tweets. We first attempted to use a hashtag analysis approach, as was used in *Echo Chambers Surrounding Misinformation on Twitter,* but applied to three dimensions instead of one. However, this approach proved unstable as many users do not use the collection of popular hashtags that our analysis was conducted on. Moving forward, we designed a more robust and novel machine learning approach that allows us to categorize user's polarities along a spectrum, giving us a more nuanced understanding of each user's preferences. To train our models, we took advantage of the TweetSets data repository managed by the George Washington University which contains Twitter activity of many political and government related users, including news organizations. We curated a list of news organizations across the ideological spectrum and used the independent bias rating website Media Bias/Fact Check to assign scores regarding credibility, political bias, and moderacy for each news organization. See Figure 1 for an example of the rating system used by Media Bias/ Fact Check. For the political polarity dimension we assigned a political polarity score as follows: -3: Extreme Left, -2: Left, -1: Center Left, 0: Least Bias, 1: Center Right, 2: Right, and 3: Extreme Right. Likewise for the credibility dimension we assign a score using the factual reporting section of Media Bias/ Fact Check, as follows: 0: Very Low, 1: Low, 2: Mixed, 3: Mostly Factual, 4: Highly Factual, and 5: Very Highly Factual. Lastly, the moderacy score is calculated by: 0: Least Bias, 1: Left Center, Right Center, 2: Left, Right, and 3: Extreme. A breakdown of the scoring system is seen in Table 1. With these scores, we then get an accurate sense of how liberal or conservative, credible or not credible, and moderate or extreme a news

organization is, which we will assume is reflected in the content of that organization's tweets, which we had obtained from TweetSets.

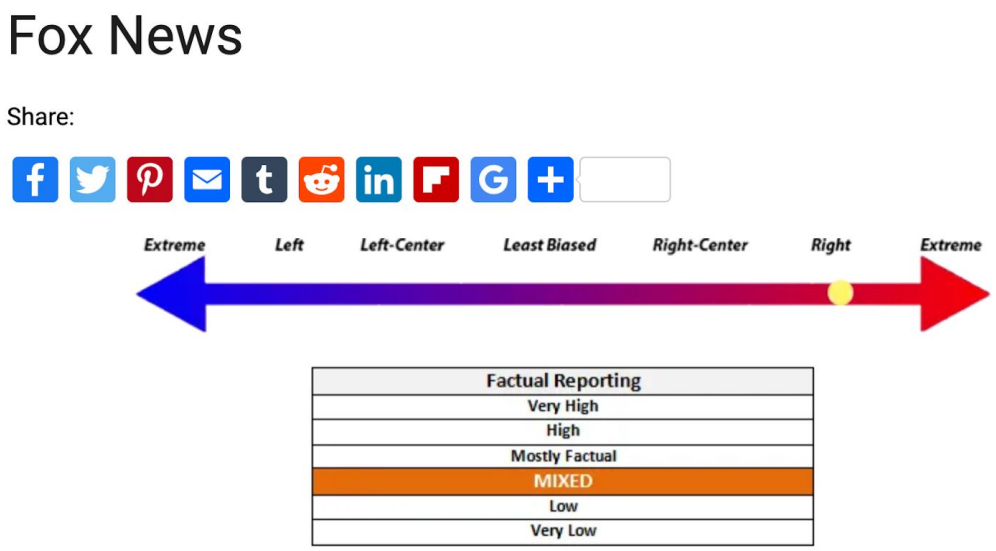

Figure 1.

MBFC Bias Rating	Assigned Political Rating	MBFC Factual Rating	Assigned Credibility Rating	MBFC Moderacy Rating	Assigned Moderacy Rating
Extreme Left	-3	Very Low	0	Extreme Left	3
Left	-2	Low	1	Left	2
Center Left	-1	Mixed	2	Center Left	1
Center	0	Mostly Factual	3	Center	0
Center Right	1	High	4	Center Right	1
Right	2	Very High	5	Right	2
Extreme Right	3			Extreme Right	3

Table 1.

Continuing the methodology, a tweet history for each news organization was downloaded from the TweetSets repository and then hydrated, a process in which you download a full tweet based

on an ID number. Once hydrated, every tweet from an organization is assigned a score for each of the three dimensions using the system we laid out above. We then extract only the tweet text for every tweet and we use every tweet as the instances making up our training set and the polarity scores being our three targets when training three separate classifiers. We split the dataset into training and validation portions and used a TF-IDF vector representation of every tweet's text as the dependent variable. We then trained three separate Naive Bayes models for each dimension and achieved accuracy scores of 53%, 62%, and 62% for the political, credibility, and moderacy dimensions respectively. Note that while these scores do not appear very good, our task at hand more closely resembles a regression task, however, we found Naive Bayes to have the best performance in terms of mean average error (MAE). In this case, the MAE scores of the three dimensions were 1, .6, and .4, respectively. For each of the three dimensions these are relatively low errors and reflect well on the accuracy of our model. We must also mention that while the models perform well between the training and validation sets, we must assume that the distributions of tweets from normal users must be similar to the distributions of tweets from news organizations. Under this assumption the models should perform similarly when classifying users based on their tweet history, although we acknowledge the possibility that news organizations tweets are semantically different from those of regular users. We feel our assumption is credible. Since Twitter limits the amount of characters within its tweets it is likely our bag of words technique picked up on most words that are shared between both news organizations and users.

Our team then collected a sample of flagged and unflagged tweets from the users Tomi Lahren, Eric Trump, Donald Trump Jr, Rudy Guiliani, Adam Laxalt, Maria Bartimoro, and Paduch. These users were chosen because they each have several flagged tweets surrounding the events of the January 6th riot at the U.S. Capitol and the 2020 U.S. election. We focus on these two events as they are relatively recent and Twitter was very active in flagging misinformation around the two events. However, we do note that the flagged tweets have a general right leaning skew due to the political nature of the two events. This is somewhat unavoidable as Twitter's API does not have a property for detecting flagged tweets, thus it is difficult to construct a randomly sampled set of flagged tweets. For our initial study, using only the 7 users we outlined will be sufficient as it will still give us an understanding of whether there is a change in the type of user that is interacting between flagged tweets and unflagged tweets. Since we are focusing on users within a similar ideology, we assume that the change in the types of users engaging in flagged and unflagged tweets will be somewhat similar across the several different users. Our study is thus particularly focused on how flagged and unflagged right wing tweets differ in types of users that are retweeting the information.

Furthermore, for each flagged tweet selected we chose an unflagged tweet that was tweeted relatively close in time to the flagged tweet, i.e. within a few days, and from the same user. This is to prevent large changes in the makeup of the followers of the users that we selected above.

We then collected up to 100 users, as permitted by the Twitter API, that retweeted the original tweets. This came out to be 669 users in total. Using Tweepy, we downloaded the history of their tweet timelines. We extracted only the text of their tweets and then ran every tweet through each of our respective models. For every user, we averaged each predicted polarity score for every one of their tweets from our three different models, thus giving us our three scores. The outcome of this can be seen in Figure 2, where every row is a user and the Flagged column indicates whether that user retweeted a flagged or unflagged tweet. Using this methodology we are able to get a more substantial and non-biased idea of a user's scores along the political, credibility, and moderacy dimensions, thus allowing us to perform our analysis.

	Flagged	Political	Credibility	Moderacy
0	True	0.46	2.00	1.80
1	True	-0.31	2.00	1.93
2	True	-0.16	2.01	1.83
3	True	0.50	2.04	1.99
4	True	0.11	2.01	1.64

Figure 2.

Results. Our analysis focused on detecting if there were any significant differences between the group of users who retweeted a flagged tweet and the group who retweeted unflagged tweets. We first do a simple exploratory analysis comparing the distributions of the two groups along each of the three dimensions. We then conduct a series of permutation tests to detect if there are statistically significant differences between the two groups. Beginning with analysis of the political polarization scores, note in Figure 3, that the distribution of polarization scores from the flagged group actually appears slightly more to the left, indicating a more liberal polarization score than the users who retweeted the unflagged tweets. We also note for the flagged distribution that there is a hump on the right side of the normal curve, indicating a still sizable amount of users with a more right leaning polarization score. Both plots show a long right tail stretching to the upper extreme right end of the polarization score, something we would expect given the right leaning nature of the tweets we are examining. The results of the permutation test, run between the two groups of users and their political polarization scores, confirms what we see in the graph. Specifically, that the political polarization of users retweeting the flagged posts are significantly more liberal than the group of users retweeting the unflagged posts. This result is unexpected, as the flagged tweets were all aligned with right wing talking points, such as claiming fraud regarding the 2020 election. The evidence suggests that flagging, at least right leaning tweets, has a more complicated outcome than just attracting further right wing users, an

outcome we ourselves predicted. Possible explanations for this are explored more thoroughly in the Discussion section.

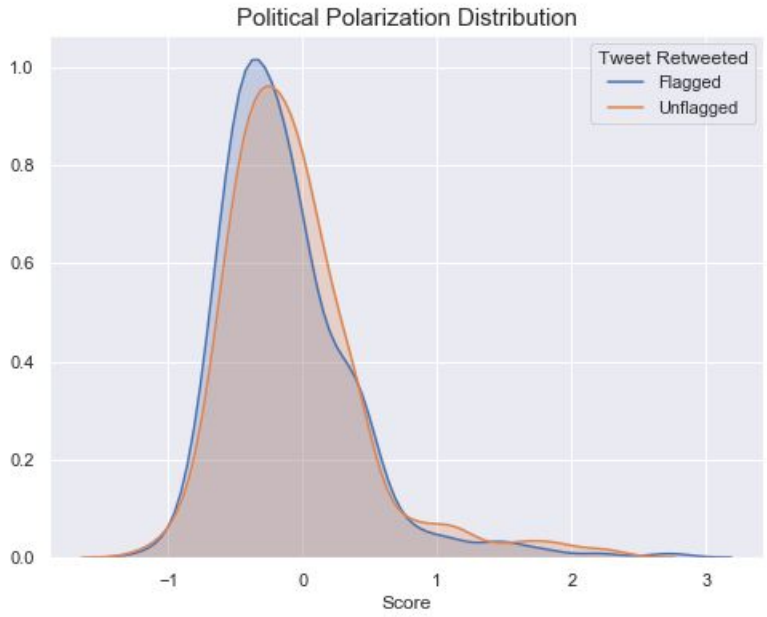

Figure 3.

In Figure 4, comparing the distributions of the moderacy scores, we note a similar trend to the political distributions. The group of users retweeting the flagged tweets show a tendency to be more moderate, rather than extreme. This is somewhat the opposite of the result we would expect. We would expect that only the most extreme users are the ones retweeting tweets that have been flagged. The visual trend is further validated with a permutation test indicating that the group of users retweeting the flagged tweets are indeed more moderate than those retweeting unflagged tweets. However, the users retweeting unflagged posts actually show a larger tail on the right side, indicating a larger share of extreme users are retweeting the unflagged tweets. We do not necessarily suggest that flagged posts are deterring the most extreme users, only that the evidence suggests they are attracting more moderate users. Combining the outcome of the permutation test for the political dimension with the outcome for the moderacy dimension paints a clearer picture of the effect of the flags. The data suggests that more moderate, center left to left wing users are retweeting flagged tweets from right wing political figures in larger proportions than unflagged tweets. However, we do not see this trend continue along the credibility dimension. For credibility, as can be seen in Figure 5, the two groups share very similar distributions of scores and a permutation test validates that suspicion. This outcome is worrying in and of itself as it possibly suggests that while flags are causing a difference in users along a political and moderacy dimension, users show no change in their preference for credible

information, hinting that there may be some confusion among users as to what information is actually credible. Without speculating too far, we wonder whether this points to a greater crisis brewing on social media, a blurring of lines between credible information and misinformation.

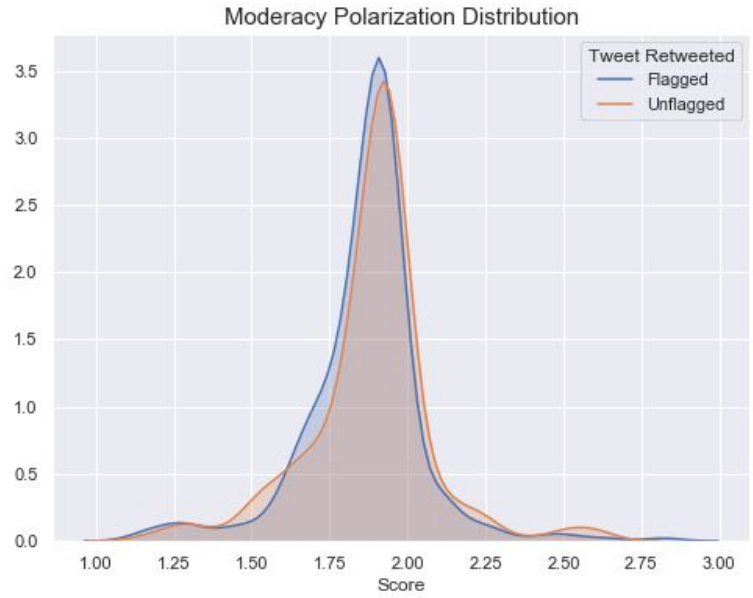

Figure 4.

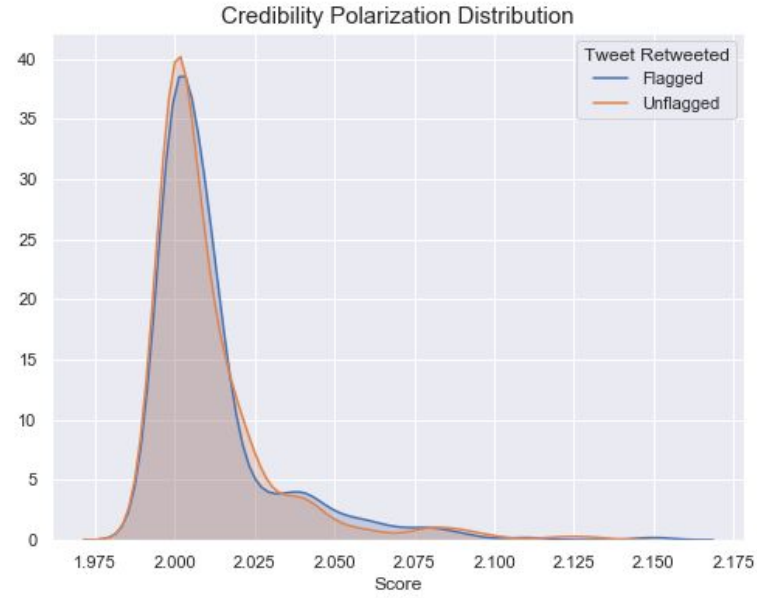

Figure 5.

Discussion & Future Work. With the data we have currently it is difficult to draw definitive conclusions, but we are nonetheless able to observe an effect of misinformation flags,

specifically when placed on tweets of right wing Twitter users. Our analysis revealed that misinformation flags change the makeup of the users retweeting tweets compared to users retweeting similar tweets that are unflagged. With confidence, our data suggests that misinformation flags are attracting more liberal and moderate users that are then sharing these posts. This outcome is opposite of what we had predicted, that misinformation flags of the tweets we collected would attract more extreme and right wing users. A larger analysis would need to be done to understand whether, while more liberal and moderate users are attracted to these flagged tweets, there is a greater amount of extreme and right wing users as well. Since our study was limited by the Twitter API only allowing access to 100 retweeters of a tweet, we have to speculate whether our findings apply across the entirety of users retweeting the original flagged and unflagged tweets. We also must note that Twitter does not tell us how the 100 retweeters are collected. For instance, are they the last 100 retweeters? In this case it is very probable that the misinformation flag had already been placed on the tweet prior to users retweeting it. Or is the 100 retweeters a random sample of all retweeters? This would make it more difficult to understand the effects that the flag is actually having because we do not know whether users retweeted the posts before the flag was placed on it. Given the significant differences in the makeup of the users retweeting the flagged tweets versus the unflagged tweets, as well as the similar content of the flagged tweets and unflagged tweets, we assume that users must have been aware of the flag before they retweeted. Otherwise, we would expect almost no difference in the types of users retweeting these tweets.

We further suggest that the reason for the increase in liberal and more moderate users is not because they agree with the contents of the flagged tweets. Rather, given the controversial nature of tweets that are flagged, we suspect that these users were more likely to want to retweet what they viewed as outrageous in a mocking sense. A logical extension to our analysis would then be a sentiment analysis of any retweets that include a comment attached to it. Analyzing these sentiments for our two groups of users would help reveal whether the increase in more liberal and moderate users also corresponds to more negative replies to the original tweet itself, rather than positive and supportive replies as would be expected. Further analyzing changes in the community of users that are following the original users who posted the tweets being analyzed, would also be helpful in understanding more long lasting effects of misinformation flags. For instance, since the beginning of the placement of misinformation flags by Twitter, has the amount of moderate and liberal users following these profiles increased? This would then explain and further validate why there is an increase in left and moderate users who retweeted flagged tweets.

With the observed results it is difficult to say whether Twitter's misinformation flags prove to be effective or not. Undoubtedly, more research needs to be done. However, we can say that if our findings can be extrapolated across the entirety of the unobserved retweeters that we could not access, then it is likely that the misinformation flags proved somewhat useful in breaking down

echo chambers around right wing misinformation. The following assumption would need to be validated, for instance through the sentiment analysis we outlined above. We speculate that the more liberal and moderate users retweeting flagged tweets are most likely doing so in a ridiculing way. With this in mind, then the misinformative nature of the tweets is being shared among communities of users that are more skeptical and critical of the information they are viewing. Thus, it is likely that the flagged tweets sparked more debate and scrutiny compared to unflagged tweets, a positive outcome. We are hopeful that this increase of debate would help users who traditionally consume extreme information unchecked be prompted to have to defend their views, whether externally against other Twitter users or internally with themselves. We do not express the opinion that these misinformation flags are the end all for preventing the spread of misinformation, but we acknowledge that there is evidence suggesting misinformation flags can lead to the breakdown of echo chambers and hopefully greater dialogue between users of different polarizations. We thus hope our research helps better understand and contributes to the prevention of the spread of misinformation on social media.

Works Cited

[1] Schoenmakers, Kevin, and Claire Liu. “China's Telling Twitter Story.” *China Media Project*, 18 Jan. 2021, chinamediaproject.org/2021/01/18/chinas-telling-twitter-story/.

[2] Rao, Ashwin, et al. "Political Partisanship and Anti-Science Attitudes in Online Discussions about Covid-19." *arXiv preprint arXiv:2011.08498* (2020).

[3] Nigel Doering, and Mark Chang. "Echo Chambers Surrounding Misinformation on Twitter" 2020.

[4] Vicario, Michela. “The Spreading of Misinformation.” 2015. https://www.pnas.org/content/pnas/113/3/554.full.pdf.

[5] Fessler, Pam. “Robocalls, Rumors And Emails: Last-Minute Election Disinformation Floods Voters.” 2020. https://www.npr.org/2020/10/24/927300432/robocalls-rumors-and-emails-last-minute-election-disinformation-floods-voters.

[6] Wakabayashi, Daisuke. “Election misinformation continues staying up on Youtube.” 2020. https://www.nytimes.com/2020/11/10/technology/election-misinformation-continues-staying-up-on-youtube.html.