Methods Supplement for "Twitter put warning labels on hundreds of thousands of tweets during the election period. What can we learn about the platform's interventions?"

Megan A. Brown

December 4, 2020

1 Data Collection

Every day, we collect a 10% sample of all tweets on Twitter, including retweets. For this research, we are specifically interested in tweets by a set of politicians and political organizations that could potentially be labelled by Twitter according to Twitter's public interest exception policy. The list of the 3,842 accounts in this dataset can be found here. The politicians list contains members of Congress, candidates for federal office in 2020, prominent political organizations, and prominent members of the executive branch such as the President, Vice President, and various cabinet members.

For each day of tweets from November 1, 2020 through November 17, 2020, we filter for retweets of tweets originally authored by the politicians of interest. Because we only collect a 10% sample of tweets, we expect that we only get a 10% sample of retweets for any given tweet. However, due to the high follower count and engagement that the majority of these accounts receive, we expect that we collect the majority of tweets by the accounts in the list that remained active during the period.

For each retweet of a tweet by the accounts in question, we retain the timestamp of the retweet, the engagement metrics (of likes or "favorites", retweets, quote tweets, and replies) for the original tweet at the time of retweet, and the original tweet ID and original tweet author. From November 1, 2020 until November 17, 2020, we collected 37,598 original tweets by the 3,482 individuals and organizations of interest; from those original tweets, we collected 3,692,702 retweets.

2 Data Labelling and Definitions

We examine the tweets 48 hours after their initial publication to see if they received a hard intervention, soft intervention, or no intervention, ensuring that there was time for Twitter to apply the intervention before we categorize it in our analysis. If Twitter applied the intervention after the tweet was included in our analysis, we do not know that the tweet had an intervention applied, as we do not revisit tweets after we have initially labeled them. However, we assume

this is an insignificant subset of the data.

Further, because we examine tweets after 48 hours but limit our analysis to the first 24 hours after tweet publication, we may label tweets as receiving an intervention even though the intervention would not have been applied during the period of our analysis. This is unfortunately the case because Twitter does not release data for the time the intervention was applied. However, from noticeable drops in engagement after 1-2 hours for most tweets that received interventions—as well as Twitter's own reporting that 74% of users viewed the tweet after the intervention was applied—we assume that most interventions were applied relatively quickly and so maintain that this would also be a small and insignificant subset of the data.

We consider a tweet to have "no intervention" if there is no indication that Twitter took steps to limit the spread of the tweet or add additional context to the information in the tweet. Most tweets would be considered "no intervention."

We define a soft intervention as a tweet that Twitter gives a context label to. These labels appear in conjunction with the tweet and provide a link to more information about the disputed topic. For example, in the graphic below, the tweet violates Twitter's policy on misinformation about a civic process, so it is labelled with a link containing more information about voter fraud in the United States. These tweets remain fully visible, but additional friction is added before a user can retweet it, asking them to provide a quote with additional context before sharing the tweet.

Figure 1: Example of a soft intervention tweet

We define a hard intervention as a tweet that receives Twitter's more severe form of intervention. This intervention includes being blocked from the timeline, meaning that a user must click on the Tweet to see its contents. Additionally, Twitter prevents users from retweeting, favoriting, or replying to the tweet. Users may only amplify the tweet by quote tweeting.

Figure 2: A hard intervention tweet from the timeline (top) and the same tweet as viewed from clicking on the tweet (bottom).

From November 1, 2020 until November 17, 2020, we collected 37,598 original tweets by the 3,482 individuals and organizations of interest; from those original tweets, we collected 3,692,702 retweets. Of the original tweets, 160 received the soft intervention, and 44 received the hard intervention. Of the tweets that received the soft intervention, 68 were by Donald Trump, and 16 of the 44 hard intervention Tweets were by Donald Trump as well.

Below are tables with the counts of hard interventions and soft interventions by the account that received them. Note that Tweets with hard interventions received both a hard intervention and a soft intervention.

Table 1: Top accounts with more than one soft intervention by number of soft interventions

Twitter Handle	Soft Intervention Count
realDonaldTrump	68
mtgreenee	20
LaurenWitzkeDE	17
AntonioSabatoJr	9
DrPaulGosar	3
GeorgePapa19	3
Jim_Jordan	3
GOP	2
RepThomasMassie	2
TTuberville	2
DrDenaGrayson	2
RandPaul	2
theangiestanton	2
Dude4Liberty	2

Table 2: Top accounts by number of hard interventions

Twitter Handle	Hard Intervention Count
mtgreenee	23
realDonaldTrump	16
mattgaetz	2
RepLaMalfa	1
BarnettforAZ	1
realannapaulina	1

Below is a table of intervention messages by the number of times they appeared. Again, note that hard intervention labels contained two messages, so there are more labels than tweets.

Table 3: Intervention labels and the number of times they appeared

Label	Count
This claim about election fraud is disputed	118
Some or all of the content shared in this Tweet is disputed and might be misleading about an election or other civic process	57
Learn about US 2020 election security efforts	54
Some votes may still need to be counted	19
Official sources may not have called the race when this was Tweeted	19
Official sources called this election differently	15
Learn how voting by mail is safe and secure	10
Multiple sources called this election differently	5
Manipulated media	1
This claim about election fraud is disputed	1

3 Data Aggregation

Tweets were split into two groups: tweets by Donald Trump and tweets by other users in the list. We do this for two reasons: (1) Donald Trump, on average, receives far more interventions than any other account, and (2) he generally has a higher level of engagement than the other accounts with labelled tweets.

For each retweet in our corpus, we calculate the amount of time that transpired between publication of the original tweet and the retweeting of that tweet, in minutes. Then, for each tweet group (no intervention, soft intervention, and hard intervention), we calculate the average number of retweets for any given minute after the tweet was posted. We then smooth this curve using a rolling average of 60 minutes to account for data sparsity. We calculate the standard error of the mean.