Social media polarization and echo chambers: A case study of COVID-19

Julie Jiang[1,2], **Xiang Ren**[1], **Emilio Ferrara**[1,2,3]
[1] USC Department of Computer Science
[2] USC Information Sciences Institute
[3] USC Annenberg School of Communication
juliej@isi.edu, xiangren@usc.edu, ferrarae@isi.edu

Abstract

During 2020, social media chatter has been largely dominated by the COVID-19 pandemic. In this paper, we study the extent of polarization of COVID-19 discourse on Twitter in the U.S. First, we propose Retweet-BERT, a scalable and highly accurate model for estimating user polarity by leveraging language features and network structures. Then, by analyzing the user polarity predicted by Retweet-BERT, we provide new insights into the characterization of partisan users. Right-leaning users, we find, are noticeably more vocal and active in the production and consumption of COVID-19 information. Our analysis also shows that most of the highly influential users are partisan, which may contribute to further polarization. Crucially, we provide empirical evidence that political echo chambers are prevalent, exacerbating the exposure to information in line with pre-existing users' views. Our findings have broader implications in developing effective public health campaigns and promoting the circulation of factual information online.

1 Introduction

As the unprecedented COVID-19 pandemic continues to put millions of people at home in isolation, online communication, especially on social media, is seeing a staggering uptick in engagement (Koeze and Popper 2020). Prior research has shown that COVID-19 has become a highly politicized subject matter, with political preferences linked to beliefs (or disbelief) about the virus (Calvillo et al. 2020; Uscinski et al. 2020) and support for safe practices (Jiang et al. 2020). As the United States was simultaneously undergoing one of the largest political events – the 2020 presidential election – public health policies may have been undermined by those who disagree politically with health officials and prominent government leaders. As it happens with topics that become politicized, people may fall into *echo chambers* – the idea that one is only presented with information they already agree with, thereby reinforcing one's confirmation bias (Garrett 2009; Barberá et al. 2015).

Social media platforms have been criticized for enhancing political echo chambers and driving political polarization (Conover et al. 2011b; Cinelli et al. 2020). The lack of diversity in multi-perspective and evidence-based information can present serious consequences on society by fueling the spread of misinformation (Del Vicario et al. 2016; Shu et al. 2017; Motta, Stecula, and Farhart 2020).

1.1 Research Questions

In this paper, we focus on the issue of COVID-19 and present a large-scale empirical analysis on the prevalence of echo chambers and the effect of polarization on social media. Our research is guided by the following research questions surrounding COVID-19 discussions on Twitter:

- **RQ1:** What are the roles of partisan users on social media in spreading information? How polarized are the most influential users? (See §5.)
- **RQ2:** Do echo chambers exist? And if so, what are the extents of the echo chambers? (See §6.)

The technical challenge for addressing these questions is posed by the need to build a scalable and reliable method to estimate user political leanings. To this end, we propose Retweet-BERT, an end-to-end model that estimates user polarity from their profiles and retweets on a spectrum from left- to right-leaning (§4). Retweet-BERT requires only a small initial set of labeled users that can be achieved with weak-supervision. We demonstrate that Retweet-BERT attains 96% accuracy as measured in cross-validation by AUC.

Using the estimated polarity scores for all 232,000 Twitter users in our data, we observe and compare the Twitter usage trends of partisan users. Our analyses show that right-leaning users are more vocal in creating original content, more active in broadcasting information (by retweeting), and more impactful through distributing information (by getting retweeted) than their left-leaning counterparts (§5.1). Moreover, influential users are usually highly partisan, a finding that holds irrespective of the influence measure used (§5.2).

Finally, we provide evidence that political echo chambers are apparent at both political extremes, though the degrees of cross-ideological interactions are highly asymmetrical (§6): While communication channels remain open between left-leaning and neutral users, right-leaning users are found in a densely-connected political bubble of their own. Information rarely travels in/out of the right-leaning echo chamber.

As our work offers unique insights into the polarization of COVID-19 discussions on Twitter, it carries broader implications for identifying and combating misinformation spread, as well as strengthening the online promotion of public health campaigns. Further, since communication across the two echo chambers functions very differently, we stress that communication effectiveness must be evaluated separately for people in each echo chamber.

2 Related Work

Representation learning on Twitter. Analysis of Twitter data takes in the form of two, often combined, approaches, namely content-based and network-based. In content-based approaches, users are characterized by the account metadata, hashtags, tweet content and other language-related features extracted from their profiles (Conover et al. 2011a; Badawy, Ferrara, and Lerman 2018; Addawood et al. 2019); in network-based approaches, users are represented in the retweet network or the mention network, both are directed networks where edges indicate the flow of communication (Conover et al. 2011b; Garimella et al. 2018a) – the use of user-follower networks is rare due to the time-consuming nature of its data collection (Martha, Zhao, and Xu 2013).

Both approaches can benefit from recent advances in representation learning, and specifically embedding methods. Techniques like word embedding (Mikolov et al. 2013) or more recently transformers (Devlin et al. 2019), have shown to improve sentiment analysis on tweets (Naseem et al. 2020) and tweet topic classification (Lilleberg, Zhu, and Zhang 2015). Network embedding (Goyal and Ferrara 2018) can aid user type detection. For instance, Ribeiro et al. (2018) used representation learning on both the retweet network structure and the tweet content to detect hateful users. Xiao et al. (2020) used network representations to classify users in a politically-centered network. In this work, we propose a new strategy based on combining content and network embedding for user polarity detection.

Ideology detection. The ability to detect user ideology is of interest to many researchers, e.g., to enable studies of political preference. Most methods are rooted in the observation that people sharing similar political beliefs are often situated in tightly-knit communities (Conover et al. 2011b). Earlier methods such as Conover et al. (2011b) classify users' political leanings from the hashtag they used. The same challenge has been tackled with label propagation, with users who have linked left-winged or right-winged media outlets in their tweets as seed users (Badawy, Ferrara, and Lerman 2018; Addawood et al. 2019). Barberá et al. (2015) proposed a latent space model to estimate the polarity of users, assuming that users tend to follow politicians who share similar ideological stances. Darwish et al. (2020) developed an unsupervised approach to cluster users who share similar political stances based on their hashtags, retweet texts, and retweet accounts. Word embeddings have also been applied to user tweets to generate clusters of topics, which helps inform the political leaning of users (Preoţiuc-Pietro et al. 2017). Recently, Xiao et al. (2020) formulated a multi-relational network to detect binary ideological labels. Our proposed method stands out because it (i) combines both language and network features for a more comprehensive estimation of ideology, and (ii) is scalable and can be trained in a limited time with limited labeled data.

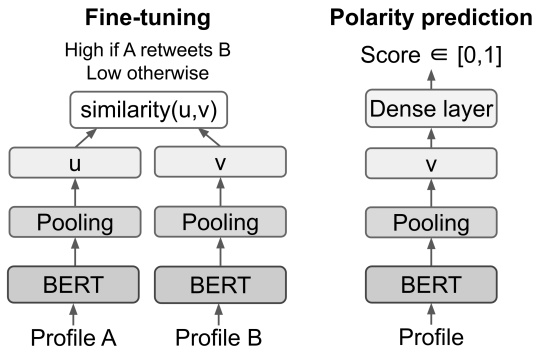

Figure 1: Illustration of the proposed Retweet-BERT. We first fine-tune it on the retweet network (left) using a Siamese network structure, where the two BERT networks share weights. We then train a denser layer on top to predict polarity (right).

Echo chambers. Echo chambers have been found on numerous social media platforms (Schmidt et al. 2017; Cinelli et al. 2020). In part, this is due to a conscious decision made by users when choosing who or what to follow, selectively exposing themselves to contents they already agree with (Garrett 2009); but this may also be a consequence of the algorithms social media platforms use to attract users (Schmidt et al. 2017). Numerous studies have shown that echo chambers are prevalent on Twitter (Conover et al. 2011b; Colleoni, Rozza, and Arvidsson 2014; Barberá et al. 2015; An et al. 2014; Cossard et al. 2020), and that those who attempt to bridge the gap between two opposite echo chambers have to pay a "price of bipartisanship" with their influenceGarimella et al. (2018a). In some cases, the internal structure of the echo chambers may be distinctive, e.g., regarding vaccination, Cossard et al. (2020) highlighted that vaccine advocates ignore the skeptics while the skeptics criticize the advocates.

3 Data

We use a large COVID-19 Twitter dataset collected by Chen, Lerman, and Ferrara (2020), containing data from January 21 to July 31, 2020 (v2.7). All tweets collected contain keywords relevant to COVID-19. The tweets can be an original tweet, retweets, quoted tweets (retweets with comments), or replies. Each tweet also contains the user's profile description, the number of followers they have, and the user-provided location. Some users are verified, meaning they are authenticated by Twitter in the interest of the public, reducing the chance that they are fake or bot accounts (Hentschel et al. 2014). All users can optionally fill in their profile descriptions, which can include personal descriptors (e.g.,

Table 1: 5-fold CV results for political leaning classification for various models. The best AUC score for each model type is shown in bold; the best overall score is indicated with *.

Model	Accuracy	AUC
Average word embeddings		
GloVe-wiki-gigaword-300	0.856	0.875
Word2Vec-google-news-300	0.852	**0.877**
Average transformer output		
BERT-base-uncased	0.859	0.882
BERT-large-uncased	0.862	0.885
DistilBERT-uncased	0.863	0.888
RoBERTa-base	0.870	0.898
RoBERTa-large	0.882	**0.914**
Fine-tuned transformers		
BERT-base-uncased	0.900	**0.932**
DistilBERT-uncased	0.899	0.931
RoBERTa-base	0.893	0.916
S-BERT		
S-BERT-large-uncased	0.869	0.890
S-DistilBERT-uncased	0.864	0.885
S-RoBERTa-large	0.879	**0.903**
Retweet-BERT (our model)		
Retweet-DistilBERT-one-neg	0.900	0.933
Retweet-DistilBERT-mult-neg	0.935	0.965
Retweet-BERT-based-mult-neg	0.934	**0.966***

"*Dog-lover*", "*Senator*", "*Best-selling author*") and the political party or activism they support (e.g., "*Republican*", "*#BLM*").

Interaction networks. The retweet network $G_R = (V, E)$ is modeled as a weighted, directed graph. Each user $u \in V$ is a node in the graph, each edge $(u, v) \in E$ indicates that user u has retweeted from user v, and the weight of an edge $w(u, v)$ represents the number of retweets. We use the terms retweet interaction and edges of the retweet network interchangeably. Similarly, we construct the mention network G_M, where the edges are mentions instead of retweets. A user can be mentioned through retweets, quoted tweets, replies, or otherwise directly mentioned in any tweet.

Data pre-processing. We restrict our attention to users who are likely in the United States, as determined by their self-provided location (Jiang et al. 2020). Following Garimella et al. (2018b), we only retain edges in the retweet network with weights of at least 2. Since retweets often imply endorsement (Boyd, Golder, and Lotan 2010), a user retweeting another user more than once would imply stronger endorsement and produce more reliable results. As our analyses depend on user profiles, we remove users with no profile data. We also remove users with degrees less than 10 (in- or out-degrees) in the retweet network, as these are mostly inactive Twitter users. To remove biases from potential bots infiltrating the dataset (Ferrara 2020), we calculate bot scores using Davis et al. (2016), which assigns a score from 0 (likely human) to 1 (likely bots), and remove the top 10% of users by bot scores as suggested by Ferrara (2020).

Our final dataset contains 232,000 users with 1.4 million retweet interactions among them. The average degree of the retweet network is 6.15. For the same set of users in the mention network, there are 10 million mention interactions, with an average degree of 46.19. Around 18,000, or approximately 8% of all, users are verified.

4 Polarity estimation

This section describes our proposed method to estimate the polarity of users in a spectrum from left to right. We first use weak-supervision to detect two polarized groups of users, which we treat as seed users (§4.1). Then we explore various models to predict the political leaning of users (§4.2). Finally, these models are evaluated on labeled data using 5-fold cross-validation and the best model is applied to the remaining users to obtain their polarity scores (§4.3).

4.1 Political leaning of seed users

We use two weakly-supervised strategies to find the "ground truth" labeling of political leanings for a subset of users (i.e., *seed users*). For the first method, we gather the top 50 most used hashtags in user profiles and annotate them as left- or right-leaning depending on what political party or candidate they support (or oppose). Of these hashtags (uncased), 17 are classified as left-leaning (e.g., `#TheResistance`, `#VoteBlue`) and 12 as right-leaning (e.g., `#MAGA`, `#KAG`). Users are labeled as left-leaning (right-leaning) if their profile contains more left-leaning (right-leaning) hashtags. We do not consider hashtags used in tweets, for the reason that hashtags in tweets can be used to inject opposing content into the feed of other users (Conover et al. 2011b). Instead, in line with Badawy, Ferrara, and Lerman; Addawood et al. (2018; 2019), we assume that hashtags appearing in user profiles more accurately capture true political affiliation.

An alternative method makes use of the media outlets mentioned in users' tweets through mentions or retweets (Badawy, Lerman, and Ferrara 2019; Bovet and Makse 2019; Ferrara et al. 2020). Similar to Ferrara et al. (2020), we identify 29 prominent media outlets on Twitter. Each media outlet has their media bias scored by the non-partisan media watch-dog *AllSides.com* on a scale of 1 to 5 (*left*, *center-left*, *neutral*, *center-right*, *right*). An endorsement from a user is either an explicit retweet from a media's official Twitter account or a mention of a link from the media's Website. Given a user who has given at least two endorsements, we calculate their media bias score from the average of the scores of their media outlets. A user is considered left-leaning (right-leaning) if their media bias score is equal to or below 2 (above 4).

Using a combination of the profile hashtag method and the media outlet method, we categorized 79,370 (34% of all) users as either left- or right-leaning. In case of any disagreements between the two detection methods, we defer to the first, hashtag-based method. We refer to these users as seed users for political leaning estimation. A total of 59,832, or 75% of all, seed users are left-leaning, compared

to 19,538 who are right-leaning, consistent with previous research which revealed that there are more liberal users on Twitter (Wojcik and Hughes 2019).

4.2 Methods

To predict user political leanings, we explore several representation learning methods based on the users' profile description and their retweet interactions. We first examine several language models that produce sentence-level embeddings based on words, which we use to produce profile embeddings. We then propose a new model that includes retweet interactions to supplement the profile representations. All profiles are pre-processed and tokenized according to the instructions for each language model.

Average word embeddings. As baselines, we use pre-trained Word2Vec (Mikolov et al. 2013) and GloVe (Pennington, Socher, and Manning 2014) word embeddings from Gensim (Řehůřek and Sojka 2010). The sentence (i.e., profile) embeddings are formed from the average embeddings of each word embedding. We fit a logistic regression model on the sentence embeddings for the classification task.

Transformers. Transformers such as BERT (Devlin et al. 2019), RoBERTa (Liu et al. 2019), and DistilBERT (Sanh et al. 2019) are pre-trained language models that lead to significant performance gains across many NLP tasks. Unlike word embeddings, transformers can disambiguate words with different meanings under different contexts. They are also designed to easily adapt to various downstream tasks by fine-tuning the output layers.

There are a few ways to adapt transformers for sequence classification. One way is to *average* the output embeddings of each word token in the sentence. We fit a logistic regression model on the averaged transformer output embeddings for classification. The other, more time-consuming method is to *fine-tune* the transformer through the initial token embedding of the sentence (e.g., `[CLS]` for BERT, `<s>` for RoBERTa) with a sequence classification head. We use the sequence classification head published with *HuggingFace*'s open-sourced transformers library (Wolf et al. 2019), which adds a linear dense layer on top of the pooled output of the initial token embedding of the transformers.

S-BERT. Reimers and Gurevych (2019) proposed Sentence Transformers (S-BERT), which consists of a Siamese and triplet model on top of a transformer to produce sentence-level embeddings. S-BERT outperform naive transformer-based methods for semantic textual similarity tasks, while massively reducing the time complexity. A basic S-BERT model consists of pooling the output embeddings of each token and a loss function that tailors to pre-defined sentence pair objectives, such as finding the most similar pairs of sentences. Using pre-trained S-BERT models, we retrieve embeddings for every profile. The profile embeddings are fit with a dense layer with sigmoid activation for classification. We select S-BERT models that have been pre-trained for semantic textual similarity.

Our model: Retweet-BERT. Inspired by S-BERT (Reimers and Gurevych 2019), we propose Retweet-BERT (visualized in Fig. 1), a sentence embedding model that incorporates the retweet network. We base our model on the assumption that users who retweet each other are more likely to share similar ideologies. As such, the intuition of our model is to make profile embeddings more similar for users who retweet each other. Specifically, using any of the aforementioned models that can produce sentence-level embeddings, let s_i denote the profile embedding for user i. For every positive retweet interaction from user i to j (i.e., $(i, j) \in E$), we optimize the objective

$$\sum_{k \in V, (i,k) \notin E} \max(||s_i - s_j|| - ||s_i - s_k|| + \epsilon, 0), \qquad (1)$$

where $||\cdot||$ is a distance metric and ϵ is a margin hyperparameter. We follow the default configuration of S-BERT, which uses the Euclidean distance and $\epsilon = 1$.

To optimize the training procedure, we use two negative sampling strategies. The first is negative sampling (`one-neg`), in which we randomly sample one other node k for every anchor node in each iteration (Mikolov et al. 2013). For simplicity, we assume all nodes are uniformly distributed. The second is multiple negative sampling (`mult-neg`), in which the negative examples are drawn all other examples in the same batch (Henderson et al. 2017). For instance, if the batch of positive examples are $[(s_{i1}, s_{j1}), (s_{i2}, s_{j2}), ..., (s_{in}, s_{jn})]$, then the negative examples for pair at index k are (s_{ik}, s_{jk}) are all the $\{s_{jk'}\}$ for $k' \in [1, n]$ and $k' \neq k$.

It is worth noting that Retweet-BERT disregards the directionality of the network and only considers the immediate neighbors of all nodes. In practice, however, we find that this model balances the trade-off between training complexity and testing performance. Building on the convenience of S-BERT for sentence embeddings, we use the aforementioned S-BERT models pre-trained for semantic textual similarity as the basis for fine-tuning.

4.3 Polarity estimation results

Table 1 shows the cross-validated results for political leaning classification on the seed users, using different prediction models. Of all models that do not consider the retweet network, fine-tuned transformers are demonstrably better. Averaging transformer outputs and fine-tuning S-BERTs lead to similar results. For transformers that have a *base* and *large* variant, where the *large* version has roughly twice the number of tunable parameters than the *base*, we see very little added improvement with the *large* version. The lack of improved performance for the *large* models may be attributed to having to vastly reduce the batch size due to memory issues, which could hurt performance[1]. DistilBERT, a smaller and faster version of BERT, produces results comparable with or even better than the other models.

Our proposed model, Retweet-BERT, delivers the best results on BERT-base using the multiple negatives training

[1] https://github.com/google-research/bert#out-of-memory-issues

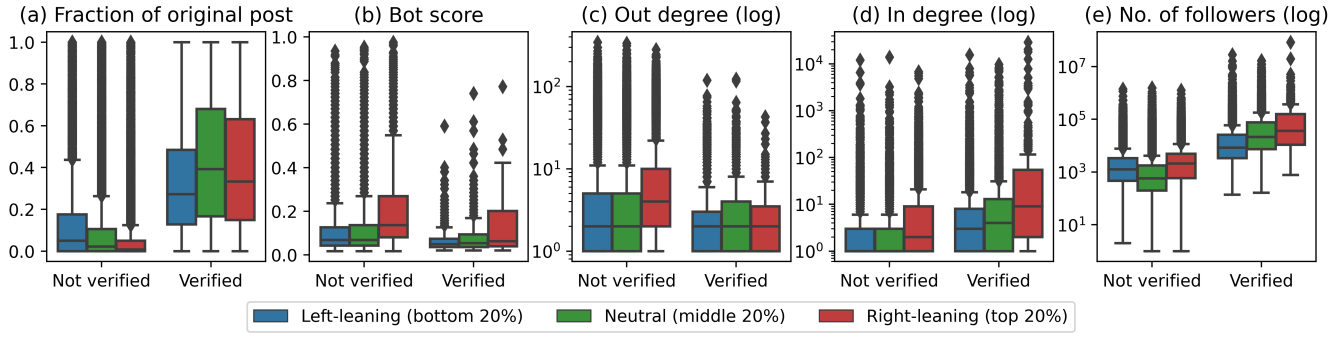

Figure 2: Dataset statistics of left-leaning (bottom 20%), neutral (middle 20%), and right-leaning (top 20%) users, partitioned by their verification status. The degree distributions are taken from the retweet network.

strategy. We train this model on all the seed users with political leaning labels and infer polarity scores for the rest of the users, ranging from 0 (far-left) to 1 (far-right). These scores will be referred to as the *polarity scores*. Since there are more left-leaning seed users, the polarity scores are naturally skewed towards 0 (left). Therefore, we bin users by evenly distributed deciles of the polarity scores, with each decile containing exactly 10% of all users.

5 Characterizing partisan users

5.1 The roles of partisan users

We first examine the characteristics of extremely polarized users, defined as the users in the bottom (left-leaning/far-left) or top (right-leaning/far-right) 20% of the polarity scores. As a point of comparison, we also include neutral users who are in the middle 20% of the polarity scores. Considering various aspects of user tweeting behaviors, we characterize the Twitter user roles as follows:

1. *Information creators*: those who create original content, and are usually the source of new information.
2. *Information braodcasters*: those who foster the distribution of existing content, such as through retweeting other people and promoting the visibility of other's content.
3. *Information distributors*: those whose contents are likely to be seen by many people, either through passive consumption by their followers or through broadcasting (retweeting) by others.

According to these definitions, a user can be all of these or none of these at the same time. In Fig. 2, we plot several Twitter statistics regarding the polarized and neutral users, disaggregated by their verification status.

Compared to unverified users, verified users are more likely information creators. This is unsurprising, given that verified users can only be verified if they demonstrate they are of public interest and noteworthy. Comparatively, left-leaning verified have the smallest fraction of original post. However, this is reversed for unverified users, with unverified left-leaning users having the highest fraction of original content and unverified right-leaning users having little to no original content. We note that this may be related to the distribution of bot scores. Fig. 2(b) reveals that right-leaning users score significantly higher on the bot scale. Since bots retweet significantly more than normal users (Ferrara et al. 2016), we cannot rule out the possibility that right-leaning bots are confounding the analysis. However, users scoring the highest on the bot scale have already been removed from the data (§3).

Unverified right-leaning users, in comparison with their left-leaning counterparts, are more likely information broadcasters as they have the highest out-degree distribution (Fig. 2(c)). As out-degree measures the number of people a user retweets from, a user with a high out-degree function critically in information broadcasting. The fact that they also have very little original content (Fig. 2(a)) further suggests that unverified right-leaning users primarily retweets from others.

Finally, all right-leaning users function as information distributors regardless of their verification status. Their tweets are much more likely to be shared and consumed by others. Their high in-degree distribution indicates they get retweeted more often (Fig. 2(d)), and the higher number of followers they have indicates that their posts are likely seen by more people (Fig. 2(e)).

As right-leaning users play larger roles in both the broadcasting and distributing of information, we question if these users form a political echo chamber, wherein right-leaning users retweet frequently from, but only from, users who are also right-leaning. As we will see in §6, we indeed find evidence that right-leaning users form a strong echo chamber.

5.2 The polarity of influencers

The above characterizes the Twitter activities of users who are extremely left or right-biased. However, the majority of the social influence is controlled by a few key individuals (Wu et al. 2011; Lou and Tang 2013; Zhang et al. 2015). In this section, we consider five measures of social influence: verification status, number of followers, number of retweets, number of mentions, and PageRank in the retweet network (Page et al. 1999). A user is considered influential if they are in the top 5% of all people according to the measure of influence. Fig. 3 reveals the proportion of users in each

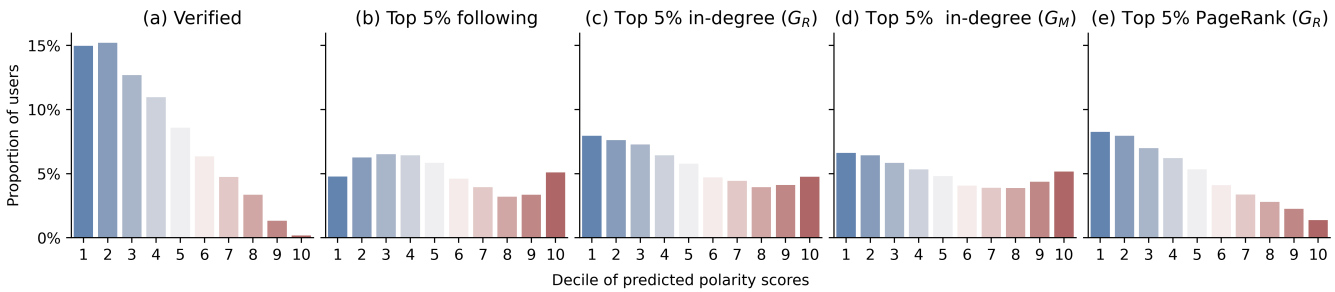

Figure 3: Proportion of users in each decile of predicted political bias scores that are (a) verified, (b) top 10% in the number of followers, (c) top 5% of in-degrees in the retweet network (most retweeted by others), (c) top 5% of in-degrees in the mention network (most mentioned by others), and (e) top 5% in PageRank in the retweet network.

.

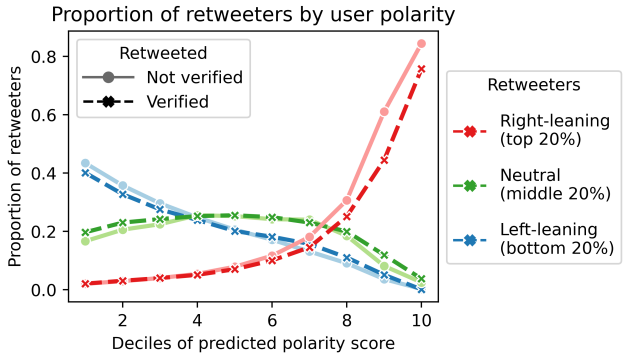

Figure 4: The distribution of left-leaning (bottom 20% of the polarity scores), center (middle 20%), and right-leaning (top 20%) retweeters (y-axis) for users across the polarity score deciles (x-axis). The retweeted users are either verified or not verified.

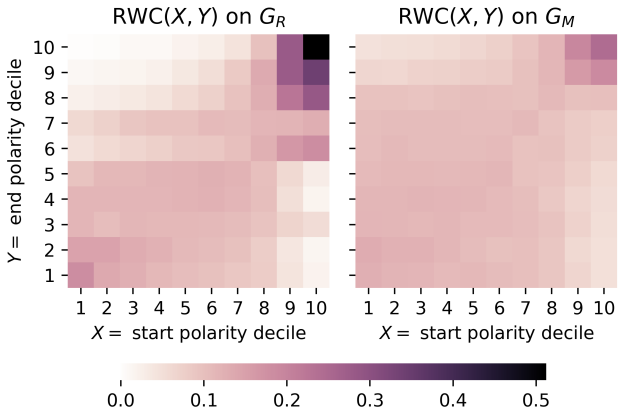

Figure 5: The RWC(X, Y) for every pair of polarity deciles X and Y on the retweet (left) and mention (right) networks using Eq. 2.

decile of polarity score that are influential. We show that, consistent with all of the influence measures above, partisan users are more likely to be found influential.

The verification status is correlated with partisan bias, with the proportion of verified users decreasing linearly as we move from the most left- to the most right-leaning deciles of users (Fig. 3(a)). 15% of users in the 1st and 2nd deciles, which are most liberal, are verified, compared to less than 1% of users in the extremely conservative 10th decile. As verified accounts generally mark the legitimacy and authenticity of the user, the lack of far-right verified accounts opens up the question of whether there is a greater degree of unverified information spreading in the right-leaning community. We stress, however, that our result is cautionary. A closer investigation is needed to establish if there are other politically driven biases, such as a liberal bias from Twitter as a moderating platform, that may contribute to the underrepresentation of conservative verified users.

While being verified certainly aids visibility and authenticity, users do not need to be verified to be influential. We observe bimodal distributions (U-shaped) in the proportion of users who are influential with respect to their polarity according to three measures of influence: top most followed, retweeted, and mentioned (Fig. 3(b)-(d)), indicating that partisan users have more influence in these regards. In particular, far-right users having some of the highest proportion of most-followed users. Far-left users are more likely to be highly retweeted and mentioned, but the far-right also holds considerable influence in those regards.

Lastly, we look at PageRank, a well-known algorithm for measuring node centrality in directed networks (Page et al. 1999). A node with a high PageRank is indicative of high influence and importance. Much like the distribution of verified users, the proportion of users with high PageRank in each polarity decile is correlated with how left-leaning the polarity decile (Fig. 3(b)), which suggests that left-leaning users hold higher importance and influence.

6 Echo chambers

As most influential users are partisan, we question if echo chambers exist how prevalent they are. We first provide evidence of echo chambers (§6.1), then examine how cross-

ideological information flows between the far-left and far-right (§6.2). Finally, we consider the influence of users who are popular among the left and the right to provide further context on the extent of echo chambers (§6.3).

6.1 User polarity vs. audience polarity

We begin by exploring the partisan relationship between the retweeted and the retweeter, where the latter is considered as the (immediate) audience of the former. Fig. 4 plots the proportion of left-leaning, neutral, or right-leaning retweeters for users in each of the 10 deciles of polarity scores, revealing that users on both ends of the political spectrum reach an audience that primarily agrees with their political stance. In fact, the far-left and far-right users have virtually no retweeters from supporters of the opposite party. However, the echo chamber effect is much more prominent on the far-right. About 80% of the audience reached by far-right users are also right. In comparison, only 40% of the audience reached by far-left users are also left. There is little difference in the distribution of retweeters between verified and unverified users.

Since the polarized users are mostly preoccupied in their echo chambers, the politically neutral users (Fig. 4, green) would serve the important function of bridging the echo chambers and allowing for cross-ideological interactions. Most of them (30-40%) retweet from sources that are also neutral, and around 20% of them retweet from very liberal sources. When it comes to broadcasting tweets from the far-right, they behave similarly to the far-left retweeters: Almost no neutral users retweet from the far-right. Such observations would imply a much stronger flow of communication between the far-left users and neutral users, whereas the far-right users remain in a political bubble.

6.2 Random walk controversy

Previously, we explored the partisan relationship between users and their immediate audience. To quantify how information is disseminated throughout the Twitter-sphere and its relationship with user polarity, we conduct random walks on the graphs to measure the degree of controversy between any two polarity deciles of users. Our method extends the Random Walk Controversy (RWC) score for two partitions (Garimella et al. 2018b), which uses random walks to measure the empirical probability of any node from one polarity decile being exposed to information from another.

A walk begins with a given node and recursively visits a random out-neighbor of the node. It terminates when the maximum walk length is reached or if a node previously seen on the walk is revisited. Following Garimella et al. (2018b), we also halt the walk if we reach an authoritative node, which we define as the top 1000 nodes ($\approx 4\%$) with the highest in-degree in any polarity decile. By stopping at nodes with high in-degrees, we can capture how likely a node from one polarity decile receives highly endorsed and well-established information from another polarity decile. To quantify the controversy, we measure the RWC from polarity decile A to B by estimating the empirical probability

$$\text{RWC}(A, B) = Pr(\text{start in } A | \text{end in } B). \quad (2)$$

The probability is conditional on the walks ending in any partition to control for varying distribution of high-degree vertices in each polarity decile. RWC yields a probability, with a high RWC(A, B) implying that random walks landing in B started from A. Compared to the original work (Garimella et al. 2018b), we simplify the definition of RWC as we do not need to consider the varying number of users in each echo chamber.

We initiate the random walks 10,000 times randomly in each polarity decile for a maximum walk length of 10. The RWC between any two polarity deciles for the retweet and mention networks are visualized in Fig. 5. For both networks, the RWC scores are higher along the diagonal, indicating that random walks most likely terminate close to where they originated. Moreover, the intensities of the heatmap visualizations confirm that there are two separate echo chambers. The right-leaning echo chamber (top right corner) is much denser and smaller than the left-leaning echo chamber (bottom left corner). Any walk in the retweet network that originates in polarity deciles 9 and 10 will terminate in polarity deciles 8 to 10 about 80% of the time. In contrast, walks that start in deciles 1–7 have a near equal, but overall much smaller, probability of landing in deciles 1–7. In essence, users who are right-leaning form a smaller but stronger echo chamber, while other users form a larger and more distributed echo chamber.

The RWC scores on the mention network confirm the presence of the two echo chambers, but the intensities are reduced. Compared to random walks on the retweet network, those on the mention network are much more likely to end faraway. As a result, while there is rarely any cross-ideological retweet interactions, there exists a greater degree of direct communication through mentions, likely done to speak to or criticize against the opposing side (Conover et al. 2011b). We note that, because the RWC scores appear highly symmetrical about the diagonals, there is little difference in the cross-ideological interaction between opposite directions of communication flow.

6.3 Popular users among the left and right

Retweeting is the best indication of active endorsement (Boyd, Golder, and Lotan 2010), and is commonly used as the best proxy for gauging popularity and virality on Twitter (Cha et al. 2010). Fig. 6 shows the users who are the most popular users among the left and the right according to the number of left- or right-leaning retweeters they have.

Analyzing the identities of the top-most retweeted users by partisans gives us the first hint at the presence of political echo chambers. There is no overlap between the most retweeted users by the left- and by the right-leaning audience, and they tend to be politically aligned with the polarization of their audience. Almost all users who are most retweeted by left-leaning users are Democratic politicians, liberal-leaning pundits, or journalists working for left-leaning media. Notably, @ProjectLincoln is a political action committee formed by Republicans to prevent the re-election of the Republican incumbent President Trump. Similarly, almost all users who are most retweeted by right-leaning users are Republican politicians or right-leaning

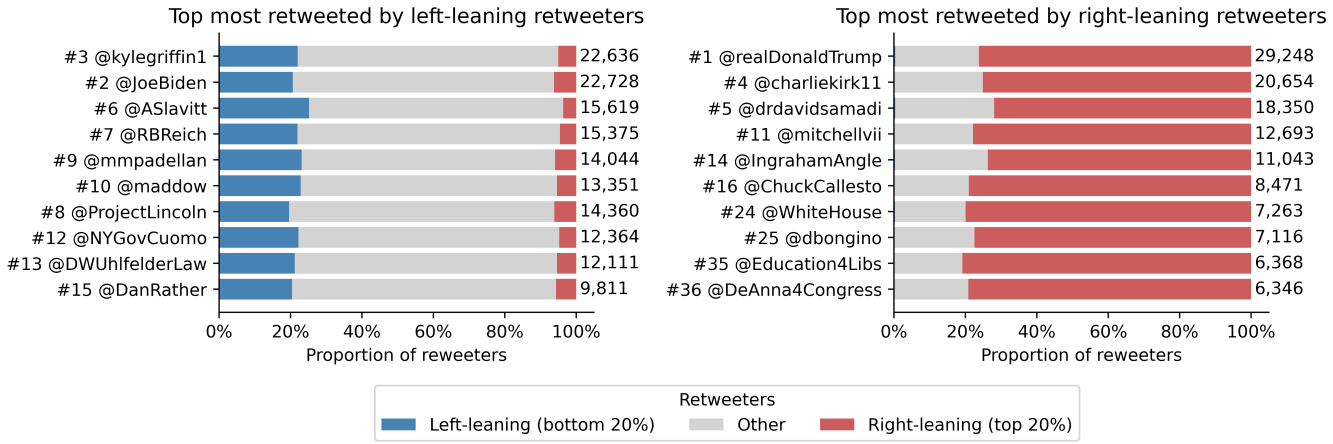

Figure 6: Users with the highest number of retweeters from left- and right-leaning users. The bar plots show the distribution of their unique retweeters by political leaning. Users are also ranked by their total number of retweeters (i.e. #1 @realDonaldTrump means that @realDonaldTrump has the most retweeters). Numbers appended to the end of the bars show their total number of retweeters.

pundits, or journalists working for right-leaning media. Despite its username, @Education4Libs is a far-right account promoting QAnon, a far-right conspiracy group[2].

These popular users are not only popular among the partisan users, but are considerably popular overall, as indicated by the high overall rankings by the number of total retweeters. With a few exceptions, users who are popular among the left are more popular among the general public than do users who are popular among the right.

The distribution of the polarity of retweeters of these most popular users reveals another striking observation: The most popular users among the far-right rarely reach an audience that is not also right, whereas those of the far-left reach a much wider audience in terms of polarity. Users who are popular among the far-left hails the majority of their audience from non-partisan users (around 75%) and, importantly, draw a sizable proportion of far-right audience (around 5%). In contrast, users who are popular among the far-right has an audience made up almost exclusively of the far-right (around 80%) and amass only a negligible amount of far-left audience.

7 Discussion

In this paper, we study the extent of echo chambers and political polarization in COVID-19 conversations on Twitter in the US. We propose Retweet-BERT, a model that leverages user profile descriptions and retweet interactions to effectively and accurately measure the degree and direction of polarization (§4). Applying Retweet-BERT, we provide insightful characterizations of partisan users and the echo chambers in the Twitter-sphere to address our research questions.

[2]At the time of writing, @Education4Libs has been banned by Twitter.

RQ1. *What are the roles of partisan users on social media in spreading information? How polarized are the most influential users?* From characterizing partisan users, we find that right-leaning users stand out as being more vocal, more active, and more impactful than their left-leaning counterparts (§5.1).

Our finding that many influential users are partisan suggests that online prominence is linked with partisanship (§5.2). This result is in line with previous literature on the "price of bipartisanship", which is that bipartisan users must forgo their online influence if they expose information from both sides (Garimella et al. 2018a). In another simulated study, Garibay et al. (2019) show that polarization can allow influential users to maintain their influence. Consequently, an important implication is that users may be incentivized to capitalize on their partisanship to maintain or increase their online popularity, thereby further driving polarization. Information distributed by highly polarized yet influential users can reinforce political predispositions that already exist, and any polarized misinformation spread by influencers risks being amplified.

RQ2. *Do echo chambers exist? And if so, what are the extents of the echo chambers?* Though COVID-19 is a matter of public health, we discover strong evidence of political echo chambers on this topic on both ends of the political spectrum, but particularly within the right-leaning community. Right-leaning users are almost exclusively retweeted by users who are also right-leaning, whereas the left-leaning and neutral users have a more proportionate distribution of retweeter polarity (§6.1). From random walk simulations (§6.2), we find that information rarely travels in or out of the right-leaning echo chamber, forming a small yet intense political bubble. In contrast, far-left and non-partisan users are much more receptive to information from each other. Comparing users who are popular among the far-left and the

far-right, we reveal that users who are popular among the right are *only* popular among the right, whereas users who are popular among the left are also popular among all users (§6.3).

Despite Twitter's laudable recent efforts in fighting misinformation and promoting fact-checking (Fowler 2020), we shed light on the fact that communication is not just falsely manipulated, but also hindered, by communication bubbles segregated by partisanship. It is imperative that we not only dispute misinformation but also relay true information to all users. As we have shown, outside information is extremely difficult to get through to the right-leaning echo chamber, which could present unique challenges for public figures and health officials outside this echo chamber to effectively communicate information.

Future direction. Though the question of whether social media platforms *should* moderate polarization is debated, we note that *how* they can do so remains an open problem. It is unclear how much of the current polarization is attributed to users' selective exposure versus the platform's recommendation algorithm. Moreover, whether users are even aware that they are in an echo chamber, and how much conscious decision is being made by the users to combat that, remains to be studied in future work.

Another future avenue of research could focus on studying how misinformation travels in different echo chambers. Since our study highlights that there is an alarmingly small number of far-right verified users, and given that verified users are typically believed to share legitimate and authentic information, further research is required to establish if the right-leaning echo chamber is at greater risk of being exposed to false information from unverified users. Detailed content analysis on the tweets can reveal if there are significant disparities in the narratives shared by left- and right-leaning users. Crucially, our work provides a basis for more in-depth analyses on how and what kind of misinformation is spread in both echo chambers.

Limitations. There are several limitations regarding this work. First, we cannot exclude any data bias. The list of keywords was manually constructed, and the tweets collected are only a sample of all possible Tweets containing these keywords. Since the data was collected based on keywords strictly related to COVID-19, we only gathered data that are relevant to the virus and not tainted by political commentary. Therefore, the data provides us a natural setting to study the polarization of COVID-19 discourse on Twitter.

Second, our study hinges on the fact that retweets imply endorsement, which may be an over-simplification. To reduce noisy, isolated retweet interactions, we consider only retweets that have occurred at least twice between any two users.

Finally, our political detection model is built on a weakly-supervised labeling of users using politically-relevant hashtags and the polarization of news media as the sources of ground-truth. We took a conservative approach and only seeded users who explicitly use politicized hashtags in their profile or have repeatedly interacted with polarized new sources.

References

Addawood, A.; Badawy, A.; Lerman, K.; and Ferrara, E. 2019. Linguistic cues to deception: Identifying political trolls on social media. In *ICWSM '19*, 15–25. AAAI.

An, J.; Quercia, D.; Cha, M.; Gummadi, K.; and Crowcroft, J. 2014. Sharing political news: the balancing act of intimacy and socialization in selective exposure. *EPJ Data Sci.* 3(1):12.

Badawy, A.; Ferrara, E.; and Lerman, K. 2018. Analyzing the digital traces of political manipulation: The 2016 russian interference twitter campaign. In *ASONAM '18*, 258–265. IEEE.

Badawy, A.; Lerman, K.; and Ferrara, E. 2019. Who falls for online political manipulation? In *WWW '19*, 162–168. ACM.

Barberá, P.; Jost, J. T.; Nagler, J.; Tucker, J. A.; and Bonneau, R. 2015. Tweeting from left to right: Is online political communication more than an echo chamber? *Psychol. Sci.* 26(10):1531–1542.

Bovet, A., and Makse, H. A. 2019. Influence of fake news in Twitter during the 2016 US presidential election. *Nature Commun.* 10(1):1–14.

Boyd, D.; Golder, S.; and Lotan, G. 2010. Tweet, tweet, retweet: Conversational aspects of retweeting on Twitter. In *HICSS '10*, 1–10. IEEE.

Calvillo, D. P.; Ross, B. J.; Garcia, R. J. B.; Smelter, T. J.; and Rutchick, A. M. 2020. Political ideology predicts perceptions of the threat of COVID-19 (and susceptibility to fake news about it). *Soc. Psychol. Personal Sci.* 11(8):1119–1128.

Cha, M.; Haddadi, H.; Benevenuto, F.; Gummadi, P. K.; et al. 2010. Measuring user influence in Twitter: The million follower fallacy. In *ICWSM '10*, 10–17. AAAI.

Chen, E.; Lerman, K.; and Ferrara, E. 2020. Tracking social media discourse about the COVID-19 pandemic: Development of a public Coronavirus Twitter data set. *JMIR Public Health Surveill* 6(2):e19273.

Cinelli, M.; Morales, G. D. F.; Galeazzi, A.; Quattrociocchi, W.; and Starnini, M. 2020. Echo chambers on social media: A comparative analysis. *arXiv preprint arXiv:2004.09603*.

Colleoni, E.; Rozza, A.; and Arvidsson, A. 2014. Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data. *J. Commun.* 64(2):317–332.

Conover, M. D.; Goncalves, B.; Ratkiewicz, J.; Flammini, A.; and Menczer, F. 2011a. Predicting the political alignment of Twitter users. In *PASSAT/SocialCom '11*, 192–199. IEEE.

Conover, M. D.; Ratkiewicz, J.; Francisco, M. R.; Gonçalves, B.; Menczer, F.; and Flammini, A. 2011b. Political polarization on Twitter. In *ICWSM '11*, 89–96. AAAI.

Cossard, A.; Morales, G. D. F.; Kalimeri, K.; Mejova, Y.; Paolotti, D.; and Starnini, M. 2020. Falling into the echo chamber: the italian vaccination debate on twitter. In *ICWSM '20*, 130–140. AAAI.

Darwish, K.; Stefanov, P.; Aupetit, M.; and Nakov, P. 2020. Unsupervised user stance detection on Twitter. In *ICWSM '20*, 141–152. AAAI.

Davis, C. A.; Varol, O.; Ferrara, E.; Flammini, A.; and Menczer, F. 2016. Botornot: A system to evaluate social bots. In *WWW '16*, 273–274.

Del Vicario, M.; Bessi, A.; Zollo, F.; Petroni, F.; Scala, A.; Caldarelli, G.; Stanley, H. E.; and Quattrociocchi, W. 2016. The spreading of misinformation online. *PNAS* 113(3):554–559.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT '19*, volume 1, 4171–4186. ACL.

Ferrara, E.; Varol, O.; Davis, C.; Menczer, F.; and Flammini, A. 2016. The rise of social bots. *Commun. ACM* 59(7):96–104.

Ferrara, E.; Chang, H.; Chen, E.; Muric, G.; and Patel, J. 2020. Characterizing social media manipulation in the 2020 US presidential election. *First Monday* 25(11).

Ferrara, E. 2020. What types of COVID-19 conspiracies are populated by Twitter bots? *First Monday* 25(6).

Fowler, G. A. 2020. Twitter and Facebook warning labels aren't enough to save democracy. *The Washington Post*. https://www.washingtonpost.com/technology/2020/11/09/facebook-twitter-election-misinformation-labels/ Accessed: 2020-12-14.

Garibay, I.; Mantzaris, A. V.; Rajabi, A.; and Taylor, C. E. 2019. Polarization in social media assists influencers to become more influential: analysis and two inoculation strategies. *Sci. Rep.* 9(1):1–9.

Garimella, K.; De Francisci Morales, G.; Gionis, A.; and Mathioudakis, M. 2018a. Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship. In *WWW '18*, 913–922. ACM.

Garimella, K.; Morales, G. D. F.; Gionis, A.; and Mathioudakis, M. 2018b. Quantifying controversy on social media. *ACM TCS* 1(1):1–27.

Garrett, R. K. 2009. Echo chambers online?: Politically motivated selective exposure among internet news users. *J. Comput.-Mediat. Commun.* 14(2):265–285.

Goyal, P., and Ferrara, E. 2018. Graph embedding techniques, applications, and performance: A survey. *Knowl. Based Syst.* 151:78–94.

Henderson, M.; Al-Rfou, R.; Strope, B.; Sung, Y.-H.; Lukács, L.; Guo, R.; Kumar, S.; Miklos, B.; and Kurzweil, R. 2017. Efficient natural language response suggestion for smart reply. *arXiv preprint arXiv:1705.00652*.

Hentschel, M.; Alonso, O.; Counts, S.; and Kandylas, V. 2014. Finding users we trust: Scaling up verified Twitter users using their communication patterns. In *ICWSM '14*, 591–594. AAAI.

Jiang, J.; Chen, E.; Yan, S.; Lerman, K.; and Ferrara, E. 2020. Political polarization drives online conversations about COVID-19 in the united states. *Hum. Behav. Emerg. Technol.* 2(3):200–211.

Koeze, E., and Popper, N. 2020. The virus changed the way we internet. *The New York Times*. https://www.nytimes.com/interactive/2020/04/07/technology/coronavirus-internet-use.html Accessed: 2020-12-14.

Lilleberg, J.; Zhu, Y.; and Zhang, Y. 2015. Support vector machines and word2vec for text classification with semantic features. In *ICCI*CC '15*, 136–140. IEEE.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lou, T., and Tang, J. 2013. Mining structural hole spanners through information diffusion in social networks. In *WWW '13*, 825–836. ACM.

Martha, V.; Zhao, W.; and Xu, X. 2013. A study on Twitter user-follower network: a network based analysis. In *ASONAM '13*, 1405–1409. ACM.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS '13*, 3111–3119. Curran Associates Inc.

Motta, M.; Stecula, D.; and Farhart, C. 2020. How right-leaning media coverage of COVID-19 facilitated the spread of misinformation in the early stages of the pandemic in the US. *Canadian J. Polit. Sci.* 1–8.

Naseem, U.; Razzak, I.; Musial, K.; and Imran, M. 2020. Transformer based deep intelligent contextual embedding for twitter sentiment analysis. *Future Gener. Comput. Syst.* 113:58 – 69.

Page, L.; Brin, S.; Motwani, R.; and Winograd, T. 1999. The PageRank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. GloVe: Global vectors for word representation. In *EMNLP '14*, 1532–1543. ACL.

Preoţiuc-Pietro, D.; Liu, Y.; Hopkins, D.; and Ungar, L. 2017. Beyond binary labels: Political ideology prediction of Twitter users. In *ACL '17*, 729–740. ACL.

Řehůřek, R., and Sojka, P. 2010. Software Framework for Topic Modelling with Large Corpora. In *LREC '10 Workshop*, 45–50. ELRA.

Reimers, N., and Gurevych, I. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *EMNLP-IJCNLP '19*, 3982–3992. ACL.

Ribeiro, M. H.; Calais, P. H.; Santos, Y. A.; Almeida, V. A.; and Meira Jr, W. 2018. Characterizing and detecting hateful users on Twitter. In *ICWSM '18*, 676–679. AAAI.

Sanh, V.; Debut, L.; Chaumond, J.; and Wolf, T. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Schmidt, A. L.; Zollo, F.; Del Vicario, M.; Bessi, A.; Scala, A.; Caldarelli, G.; Stanley, H. E.; and Quattrociocchi, W. 2017. Anatomy of news consumption on facebook. *PNAS* 114(12):3035–3039.

Shu, K.; Sliva, A.; Wang, S.; Tang, J.; and Liu, H. 2017. Fake news detection on social media: A data mining perspective. *SIGKDD Explor. Newsl.* 19(1):22–36.

Uscinski, J. E.; Enders, A. M.; Klofstad, C.; Seelig, M.; Funchion, J.; Everett, C.; Wuchty, S.; Premaratne, K.; and Murthi, M. 2020. Why do people believe COVID-19 conspiracy theories? *HKS Misinformation Rev.* 1(3).

Wojcik, S., and Hughes, A. 2019. Sizing up Twitter users. *Pew Res. Cent.*

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; Davison, J.; Shleifer, S.; von Platen, P.; Ma, C.; Jernite, Y.; Plu, J.; Xu, C.; Scao, T. L.; Gugger, S.; Drame, M.; Lhoest, Q.; and Rush, A. M. 2019. HuggingFace's transformers: State-of-the-art natural language processing. *ArXiv preprint arXiv:1910.03771*.

Wu, S.; Hofman, J. M.; Mason, W. A.; and Watts, D. J. 2011. Who says what to whom on Twitter. In *WWW '11*, 705–714. ACM.

Xiao, Z.; Song, W.; Xu, H.; Ren, Z.; and Sun, Y. 2020. TIMME: Twitter ideology-detection via multi-task multi-relational embedding. In *SIGKDD '20*, 2258–2268. ACM.

Zhang, J.; Tang, J.; Li, J.; Liu, Y.; and Xing, C. 2015. Who influenced you? Predicting retweet via social influence locality. *ACM TKDD* 9(3).