Retweet-BERT: Political Leaning Detection Using Language Features and Information Diffusion on Social Networks

Julie Jiang,[1, 2]**, Xiang Ren,**[1, 2] **Emilio Ferrara**[1,2, 3]

[1] Information Sciences Institute, University of Southern California, Marina Del Rey, CA, USA
[2] Department of Computer Science, Viterbi School of Engineering, University of Southern California, Los Angeles, CA, USA
[3] Annenberg School of Communication, University of Southern California, Los Angeles, CA, USA
juliej@isi.edu, xiangren@usc.edu, ferrarae@isi.edu

Abstract

Estimating the political leanings of social media users is a challenging and ever more pressing problem given the increase in social media consumption. We introduce Retweet-BERT, a simple and scalable model to estimate the political leanings of Twitter users. Retweet-BERT leverages the retweet network structure and the language used in users' profile descriptions. Our assumptions stem from patterns of networks and linguistics homophily among people who share similar ideologies. Retweet-BERT demonstrates competitive performance against other state-of-the-art baselines, achieving 96%-97% macro-F1 on two recent Twitter datasets (a COVID-19 dataset and a 2020 United States presidential elections dataset). We also perform manual validation to validate the performance of Retweet-BERT on users not in the training data. Finally, in a case study of COVID-19, we illustrate the presence of political echo chambers on Twitter and show that it exists primarily among right-leaning users. Our code is open-sourced and our data is publicly available.

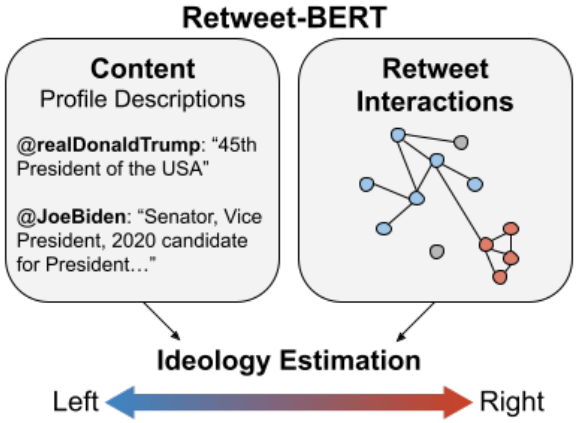

Figure 1: The two key motivating components of Retweet-BERT.

Introduction

Online communities play a central role as the glue of the very fabric of our digital society. This has become even more obvious during the unprecedented times of physical isolation brought by the COVID-19 pandemic, during which social media have seen a significant uptick in engagement (Koeze and Popper 2020). Recent work revealed that COVID-19 quickly became a highly politicized and divisive topic of discussion online (Calvillo et al. 2020; Jiang et al. 2020). The latest literature suggests that political affiliations may have an impact on people's favorability of public health preventive measures (e.g., social distancing, wearing masks) (Jiang et al. 2020), vaccine hesitancy (Peretti-Watel et al. 2020; Hornsey et al. 2020), and conspiracy theories (Uscinski et al. 2020). Though polarization on social media has been a long-standing phenomenon (Conover et al. 2011b; Colleoni, Rozza, and Arvidsson 2014; An et al. 2014; Cinelli et al. 2020), it is particularly imperative we study how polarization affects the consumption of COVID-19 information. Divisive politicized discourse can be fueled by the presence of *echo chambers*, where users are mostly exposed to information that well aligns with ideas they already agree with, further reinforcing one's positions due to confirmation bias (Garrett 2009; Barberá et al. 2015). Political polarization can contribute to the emergence of echo chambers (Conover et al. 2011b; Cinelli et al. 2020), which may accelerate the spread of misinformation and conspiracies (Del Vicario et al. 2016; Shu et al. 2017; Motta, Stecula, and Farhart 2020; Rao et al. 2020; **?**). To facilitate research in online polarization, such as the COVID-19 infodemic, we present *Retweet-BERT*, a lightweight tool to accurately detect user ideology in large Twitter datasets (illustrated in Fig. 1). Our method simultaneously captures *(i)* semantic features about the user's textual content in their profile descriptions (e.g., affiliations, ideologies, sentiment, and linguistics) and *(ii)* the patterns of diffusion of information – i.e., the spread of a given message on the social network – and how they can contribute to the formation of particular network structures (e.g., echo chambers). Prior works on polarization primarily focus on only one of these aspects (Conover et al. 2011b,a; Barberá et al. 2015; Preoţiuc-Pietro et al. 2017; Wong et al. 2016).

There are two important assumptions behind Retweet-BERT. One is that the act of retweets implies endorsement (Boyd, Golder, and Lotan 2010), which further implies support for another's ideology (Wong et al. 2016). The other is that people who share similar ideologies also share sim-

ilar textual content in their profile descriptions, including not only similar keywords (e.g. *Vote Blue!*) and sentiment, but also linguistics. The idea of *linguistic homophily* among similar groups of people has been documented and explored in the past (Yang and Eisenstein 2017; Kovacs and Kleinbaum 2020). People who adopt similar language styles have a higher likelihood of friendship formation (Kovacs and Kleinbaum 2020).

Retweet-BERT leverages both network structure and language cues to predict user ideology. Our method is simple, intuitive, and scalable. The two steps to Retweet-BERT are

1. **Training** in an unsupervised manner on the full dataset by learning representations based on users' profile descriptions and retweet interactions
2. **Fine-tuning** the model for polarity estimation on a smaller labeled subset

An illustration of Retweet-BERT is shown in Fig. 2. Crucially, our method does not require human annotations. Instead, we label a small set of users heuristically based on hashtags and mentions of biased new media outlets, as was done in prior works (Conover et al. 2011a; Badawy, Ferrara, and Lerman 2018; Addawood et al. 2019). In addition, since we only use profile descriptions instead of all of the users' tweets, Retweet-BERT can be easily deployed.

The datasets we use are two large-scale Twitter datasets collected in recent years. The COVID-19 Twitter dataset was collected from January to July of 2020 for 232,000 active users. We demonstrate that Retweet-BERT attains 96% cross-validated macro-F1 on this dataset and outperforms other state-of-the-art methods based on transformers, graph embedding, etc. We also perform extensive evaluations of our model on a second Twitter dataset on the 2020 presidential elections to showcase the reliability of Retweet-BERT (97% macro-F1).

Using Retweet-BERT, we estimate polarity scores for all users in the COVID-19 dataset and characterize patterns of information distribution in a case study COVID-19 on Twitter. Left- and right-leaning users exhibit distinct and asymmetrical patterns of communication. Moreover, we observe a significant presence of echo chambers in the right-leaning population. Our results underscore the urgency and importance of further research in this area.

In sum, the contributions of this work are:

- We present Retweet-BERT, a simple and elegant approach to estimate user ideology based on linguistic homophily and social network interactions.
- We conduct experiments and manual validations to highlight the effectiveness of Retweet-BERT on two public recent Twitter datasets compared to baselines: COVID-19 and the 2020 US presidential elections.
- We illustrate the presence of polarization and political echo chambers on Twitter by applying Retweet-BERT to the COVID-19 dataset.

Our code is open-sourced and our data is publicly available through the original dataset papers (see Appendix).

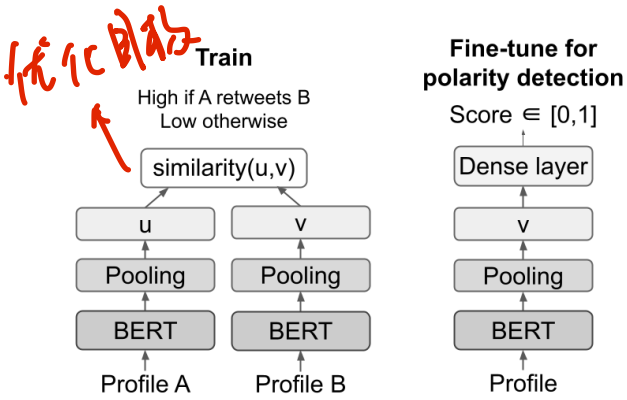

Figure 2: Illustration of the proposed Retweet-BERT. We first train it in an unsupervised manner on the retweet network (left) using a Siamese network structure, where the two BERT networks share weights. We then train a new dense layer on top to predict polarity on a labeled dataset (right).

Related Work

Ideology Detection

There is growing interest in estimating expressed ideologies. Many works focused on opinion mining and stance detection (Somasundaran and Wiebe 2009; Walker et al. 2012; Abu-Jbara et al. 2013; Hasan and Ng 2014; Sridhar et al. 2015; Darwish et al. 2020). Of particular interest are political ideology detection of textual data (Sim et al. 2013; Iyyer et al. 2014; Bamman and Smith 2015) as well as of Twitter users (Conover et al. 2011a,b; Barberá et al. 2015; **?**; Wong et al. 2016; Preoţiuc-Pietro et al. 2017; Badawy, Ferrara, and Lerman 2018; Badawy, Lerman, and Ferrara 2019; Xiao et al. 2020). There are two general strategies for identifying Twitter user ideologies: content-based and network-based. Content-based strategies are concerned with the user's tweets and other textual data. An earlier study used hashtags in tweets to classify users' political ideologies (Conover et al. 2011a). Preoţiuc-Pietro et al. (2017) applied word embedding on tweets to detect tweets of similar topics. Network-based strategies leverage cues from information diffusion to inform ideological differences. These models observe that users interact more with people they share similar ideologies with (**?**). Interactions can be retweets (Wong et al. 2016) or followings (Barberá et al. 2015). Xiao et al. (2020) formulated a multi-relational network using retweets, mentions, likes, and follows to detect binary ideological labels. Other works used a blend of both content- and network-based approaches (Badawy, Lerman, and Ferrara 2019). Hashtag-based methods were combined with label propagation to infer the leanings of users from the retweet network (Conover et al. 2011a,b; Badawy, Ferrara, and Lerman 2018). Closely related to our work, Darwish et al. (2020) clustered users by projecting them on a space jointly characterized by their tweets, hashtags, and retweeted accounts; however, this algorithm comes at a high computational cost.

Socially-infused Text Mining

More related to our work is a recent line of work that learns from socially-infused text data. Li and Goldwasser (2019) combined user interactions and user sharing of news media to predict the bias of new articles. Pan et al. (2016) used node structure, node content, and node labels to learn node representations to classify categories of scientific publications. Yang and Eisenstein (2017) used social interactions to improve sentiment detection by leveraging the idea of linguistics homophily. Johnson, Jin, and Goldwasser (2017) used lexical, behavioral, and social information to categorize tweets from politicians into various topics of political issues. These works provide promising results for combining social network data and textual data.

Our Work: Retweet-BERT is unique from the approaches described above in two substantial ways: (i) it combines both language features, in particular the state-of-the-art transformers (BERT (Devlin et al. 2019)) for natural language processing, and social network features for a more comprehensive estimation of user ideology, and (ii) it is scalable to large datasets without supervision.

Data

We use two recent large-scale Twitter datasets. The primary dataset is on COVID-19 (`COVID`) from January 21 to July 31, 2020 (v2.7) (Chen, Lerman, and Ferrara 2020). All tweets collected contain COVID-related keywords. We also use a secondary dataset on the 2020 presidential elections (`Elections`) collected from March 1 to May 31, 2020 (Chen, Deb, and Ferrara 2021). Both datasets are publicly available. Each tweet contains user metadata, including their profile description, the number of followers, the user-provided location, etc. Users can be verified, which means they are authenticated by Twitter in the interest of the public.

Although a number of Twitter accounts have since been banned by Twitter (notably, @realDonaldTrump was suspended in January 2021 (Twitter Inc. 2021)), our data collection was done in real-time and so all tweets by banned accounts are still in our dataset.

Content Cues: Profiles

For the purposes of this work, we do not use tweet contents but rather user profile descriptions. In addition to different users posting various numbers of tweets, our main assumption behind this work is that profile descriptions are more descriptive of a user's ideology than tweets. The profile description is a short biography that is displayed prominently when clicking into a user. It usually includes personal descriptors (e.g., "*Father*", "*Governor*", "*Best-selling author*") and, when appropriate, the political ideology or activism they support (e.g., "*Democratic*", "*#BLM*"). Capped at 160 characters, these descriptions have to be short, which motivates users to convey essential information about themselves clearly, succinctly, and attractively. Previous work established a positive link between the number of followers and the character length of the user (Mention 2018), which would suggest that more influential users will have a more meaningful profile.

Interaction Cues: Retweet Network

In this work, we use *retweets* to build the interaction network. Retweets refer only to tweets that were shared verbatim. Retweets are distinct from *quoted tweets*, which are essentially retweets with additional comments. We do not use the *following* network as it is rarely used due to the time-consuming nature of its data collection (Martha, Zhao, and Xu 2013). The retweet network G_R is a weighted, directed graph where vertices V are users and edges E are retweet connections. An edge $(u, v) \in E$ indicates that user u retweeted from user v and the weight $w(u, v)$ represents the number of retweets.

Data Pre-processing

We removed inactive users and users who are likely not in the U.S. (see Appendix for details). Users in our dataset must have posted more than one tweet. 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). The `COVID` dataset contains 232,000 users with 1.4 million retweet interactions. The average degree of the retweet network is 6.15. Around 18k users ($\approx 8\%$) are verified. The `Elections` dataset contains are 115,000 users and 3.6 million retweet interactions.

Method

This section describes our proposed method to estimate the polarity of users as a binary classification problem. We first use heuristics-based methods to generate "pseudo"-labels for two polarized groups of users, which are used as seed users for training and evaluating polarity estimation models. We then introduce several baseline models followed by Retweet-BERT.

Pseudo-label Generation

We consider two reliable measures to estimate political leanings for some users, which can be used for model training and automatic, large-scale evaluation. These measures will be used to generate "pseudo" political leaning labels for a subset of users (i.e., *seed users*). These seed users will be used as the set of training users.

Hashtag-based method. The first method involves annotating the 50 most popular hashtags used in user profiles as left- or right-leaning depending on what political party or candidate they support (or oppose). 17 of these hashtags are classified as left-leaning (e.g. *#Resist*) and 12 as right-leaning (e.g. *#MAGA*). The list of hashtags can be found in the Appendix. Users are labeled left-leaning if their profiles contain more left-leaning than right-leaning hashtags and vice versa. We do not consider hashtags appearing in tweets because hashtags in tweets can be used to reply to opposing ideology content (Conover et al. 2011b). Instead, following prior work (Badawy, Ferrara, and Lerman 2018;

Addawood et al. 2019), we assume that hashtags appearing in users' self-reported profile descriptions are better indicators of their true ideological affiliations.

News media-based method. The second method utilizes media outlets mentioned in users' tweets through mentions or retweets (Badawy, Lerman, and Ferrara 2019; Bovet and Makse 2019; Ferrara et al. 2020). Following Ferrara et al. (2020), we determined 29 prominent media outlets on Twitter. Each media outlet's political bias is evaluated by the non-partisan media watchdog *AllSides.com* on a scale of 1 to 5 (*left*, *center-left*, *neutral*, *center-right*, *right*). If a user mentions any of these media outlets, either by retweeting the media outlet's Twitter account or by link sharing, the user is considered to have endorsed that media outlet. Given a user who has given at least two endorsements to any of these media (to avoid those who are not extremely active in news sharing), we calculate their media bias score from the average of the scores of their media outlets. A user is considered left-leaning if their media bias score is equal to or below 2 and right-leaning if their score is equal or above 4.

Pseudo-labeling seed users. Using a combination of the profile hashtag method and the media outlet method, we categorized 79,370 ($\approx$ 34% of all) users as either left- or right-leaning. The first, hashtag-based, method alone was only able to label around 16,000 users, while the second, media-based, method labeled around 49,000 users. The two methods overlapped in labeling around 10,000 users. In case of any disagreements between the two methods, which were exceedingly rare at only 200 instances, we defer to the first, hashtag-based method. These users are considered *seed users* for political leaning estimation. 75% of these seed users are left-leaning, a finding consistent with previous research which revealed that there are more liberal users on Twitter (Wojcik and Hughes 2019). In our secondary `Elections` dataset, we tagged 75,301 seed users.

Pseudo-labeling validation. This pseudo-labeling method is limited in its capacity for labeling *all* users (*i.e.*, low coverage ratio, covering only 34% of all users), but it serves as a good starting point for its simplicity. We validated this labeling strategy by annotating 100 randomly sampled users from the main `COVID` dataset. Two authors independently annotated the data by considering both the tweets and the profile descriptions to determine the users' political leaning, keeping political neutrality to the extent possible. We then discussed and resolved any annotation differences until reaching a consensus. We attained a substantial inter-annotator agreement (Cohen's Kappa) of 0.85. 96 users' annotated labels agree with the pseudo-labels and 4 users' labels cannot be conclusively determined manually. The high agreement with the pseudo-labels makes us highly confident in the precision of our pseudo-label approach.

Methods for Polarity Estimation

While the pseudo-labels can assign confident political leaning labels for a third of all users, they cannot determine the political leaning of the rest. To predict political leanings for all users, we explore several representation learning methods based on users' profile description and/or their retweet interactions. In all of our methods in this section and the one that follows (our proposed method), We do not consider users' tweets. This is because the datasets contain sampled tweets based on keywords and do not encompass any user's full tweets histories. Considering tweets in isolation can bias an algorithm for political leaning detection.

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 profile embeddings are formed by averaging the embeddings of the profile tokens.

Transformers. Transformers (Devlin et al. 2019; Liu et al. 2019; Sanh et al. 2019) are state-of-the-art pre-trained language models that have led to significant performance gains across many NLP tasks. We experiment with two different ways to apply transformers for our task: (1) *averaging* the output embeddings of all words in the profile to form profile embeddings, and (2) *fine-tuning* a 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 by Wolf et al. (2020), which adds a dense layer on top of the pooled output of the transformer's initial token embedding.

S-BERT. Reimers and Gurevych (2019) proposed Sentence Transformers (S-BERT), which is a Siamese network optimized for sentence-level embeddings. S-BERT outperforms naive transformer-based methods for sentence-based tasks, while massively reducing the time complexity. We directly retrieve profile embeddings for each user using S-BERT's pre-trained model for semantic textual similarity.

Network-based models. We explore network-based models such as node2vec (Grover and Leskovec 2016), which learns node embeddings based on structural similarity and homophily, and label propagation, which deterministically propagates labels using the network. Neither of these models can classify isolated nodes in the network. We also experiment with GraphSAGE (Hamilton, Ying, and Leskovec 2017), an inductive graph neural network method that utilizes node attributes to enable predictions for isolated nodes. We use the aforementioned profile embeddings as node attributes. All profile or network embeddings are subsequently fit with a logistic regression model for the classification task. Hyperparameter-tuning details can be found in the Appendix. The profiles are pre-processed and tokenized according to the instructions for each language model.

With the exception of GraphSAGE, all of these aforementioned methods use either the textual features of the profile description or the network content, but not both. Purely network-based models will do poorly for nodes with only a few connections and may only be suitable for non-isolated nodes. Purely text-based models will do poorly when there are insufficient textual features to inform the models.

Proposed Method: Retweet-BERT

Combining textual and social content. To overcome the aforementioned issues, we propose Retweet-BERT (Fig. 2),

Model Type	Model	Profile	Network	COVID Acc.	COVID AUC	COVID F1	Elections Acc.	Elections AUC	Elections F1
Random and Majority	Random	✗	✗	0.585	0.501	0.706	0.499	0.499	0.506
	Majority	✗	✗	0.706	0.500	0.828	0.508	0.500	0.674
Average word embeddings	Word2Vec-google-news-300	✓	✗	0.852	0.877	0.907	0.831	0.906	0.839
	GloVe-wiki-gigaword-300	✓	✗	0.856	0.875	**0.909**	0.835	0.908	**0.844**
Average transformer output	BERT-base-uncased	✓	✗	0.859	0.882	0.910	0.837	0.912	0.844
	BERT-large-uncased	✓	✗	0.862	0.887	0.911	0.842	0.913	0.848
	DistilBERT-uncased	✓	✗	0.863	0.888	0.912	0.845	0.919	0.851
	RoBERTa-base	✓	✗	0.870	0.898	0.917	0.853	0.925	**0.859**
	RoBERTa-large	✓	✗	0.882	0.914	**0.924**	-	-	-
Fine-tuned transformers	BERT-base-uncased	✓	✗	0.900	0.932	**0.934**	0.902	0.963	**0.906**
	DistilBERT-uncased	✓	✗	0.899	0.931	**0.934**	0.899	0.962	0.904
	RoBERTa-base	✓	✗	0.893	0.916	0.930	0.888	0.953	0.895
S-BERT	S-BERT-large-uncased	✓	✗	0.869	0.890	0.916	0.849	0.924	0.855
	S-DistilBERT-uncased	✓	✗	0.864	0.885	0.913	0.843	0.917	0.849
	S-RoBERTa-large	✓	✗	0.879	0.903	**0.922**	0.874	0.944	**0.878**
Graph embedding	node2vec*	✗	✓	0.928	0.955	**0.949**	0.882	0.944	**0.883**
	GraphSAGE + RoBERTa-base	✓	✓	0.789	0.725	0.873	-	-	-
Retweet-BERT (our model)	Retweet-DistilBERT-one-neg	✓	✓	0.900	0.933	0.935	-	-	-
	Retweet-DistilBERT-mult-neg	✓	✓	0.935	0.965	**0.957**	0.973	0.984	**0.973**
	Retweet-BERT-base-mult-neg	✓	✓	0.934	0.966	**0.957**	0.971	0.984	0.971

*node2vec, a transductive-only model, can only be applied to non-isolated users in the retweet network.

Table 1: 5-fold CV results for political leaning classification on seed users for various models that are tuned via grid-search on the main COVID dataset ($N = 79,000$) and the secondary Elections dataset ($N = 75,000$). The best F1 (macro) scores for each model type are shown in bold and the best overall scores are underlined. Retweet-BERT outperforms all other models on both datasets.

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 encourage the profile embeddings to be more similar for users who retweet each other. Retweet-BERT is trained in two steps. The first step involves training in an unsupervised manner on the retweet network, and the second step involves supervised fine-tuning on the labeled dataset for classification. Similar to the training of S-BERT (Reimers and Gurevych 2019), the unsupervised training step of Retweet-BERT uses a Siamese network structure. Specifically, using any of the aforementioned models that can produce sentence-level embeddings, we apply it to a profile description to obtain the profile embedding s_i for user i. For every positive retweet interaction from user i to j (i.e., $(i, j) \in E$), we optimize the objective

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

where $||\cdot||$ is a distance metric and ϵ is a margin hyperparameter. We follow the default configuration as in S-BERT (Reimers and Gurevych 2019), which uses the Euclidean distance and $\epsilon = 1$. We then freeze the learned weights and add a new layer on top to fine-tune on a labeled dataset for classification.

Negative sampling. To optimize the training procedure during the unsupervised training step, we employ negative sampling. We explore two types of negative sampling strategies. The first is a simple 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 from 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 (s_{ik}, s_{jk}), the pair at index k, are $\{s_{jk'}\}$ 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, we find that doing so 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 base model for fine-tuning.

Results

We conduct two sets of evaluation to compare the methods: 1) cross-validation over the pseudo-labeled seed users, as an automatic, large-scale evaluation; 2) in-house human evaluation on a set of held-out users, as a complementary evaluation to the first one. We use the macro-averaged F1 score as

the primary metric due to data imbalance. We note that due to our setup, many of the aforementioned related work are not directly comparable. We do not use the following network (Barberá et al. 2015; Xiao et al. 2020). We also do not use manual labeling (Wong et al. 2016) or additional external sources to determine user ideology (Wong et al. 2016; Preoţiuc-Pietro et al. 2017). We do include a comparison with the label propagation method used in Conover et al. (2011a,b); Badawy, Ferrara, and Lerman (2018) on the held-out users.

Finally, the best model (ours) is selected to classify all the remaining users (non-seed users) to obtain their polarity leaning labels in the COVID dataset. These labels are used to conduct a case study of polarization COVID-19 on Twitter.

Automatic Evaluation on Seed Users

Baselines. We conduct a 5-fold cross-validation on the seed users (i.e., full set of training users) comparing Retweet-BERT with baselines. In addition, we also use a random label predictor (based on the distribution of the labels) and a majority label predictor model as additional baselines. Table 1 shows the cross-validated results for political leaning classification on the seed users, Overall, the models perform comparatively similarly between the two datasets. 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 as the *base*, we see very little added improvement with the *large* version, which 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 comparable or even better results than BERT or RoBERTa. Though the network-based model, node2vec, achieves good performance, it can only be applied on nodes that are not disconnected in the retweet network. While GraphSAGE can be applied to all nodes, it vastly underperforms compared to other models due to its training complexity and time efficiency (Wu et al. 2020).

Our proposed model, Retweet-BERT, delivers the best results using the DistilBERT base model and the multiple negatives training strategy on both datasets. Other Retweet-BERT variants also achieve good results, which shows our methodology can work robustly with any base language model.

Human Evaluation on Held-out Users

For further validation, the authors manually annotated the political leanings of 100 randomly sampled users *without* seed labels. We annotated these users as either left- or right-leaning based on their tweets and their profile descriptions. We were unable to determine the political leanings of 15 people. We take the best model from each category in Table 1 and evaluate them on this labeled set. In this experiment, we also include label-propagation, a simple but efficient method to propagate pseudo-labels through the network commonly used in past work (Conover et al. 2011a,b; Badawy, Ferrara, and Lerman 2018). The results are reported in Table 2 for the 85 labeled users. With a macro-F1 of 0.932, Retweet-BERT outperforms all baselines, further strengthening our confidence in our model.

Model	Profile	Network	F1
RoBERTa-large (*average*)	✓	✗	0.892
BERT-base-uncased (*fine-tuned*)	✓	✗	0.908
S-RoBERTA-large (*S-BERT*)	✓	✗	0.909
Label Propagation*	✗	✓	0.910
node2vec*	✗	✓	0.922
Retweet-BERT-base-mult-neg	✓	✓	0.932

*Label propagation and node2vec only predicts labels for nodes connected to the training network (*transductive*), but 10 nodes were not connected and thus were excluded from this evaluation.

Table 2: Results on 85 users with human-annotated political-leaning labels from a random sample of 100 users without seed labels. Retweet-BERT outperforms all models.

Case Study of COVID-19

To demonstrate the applicability of Retweet-BERT, we apply it to our entire COVID dataset to obtain polarity scores for each user to characterize the extent of polarization online. We reproduce results from our follow-up work, which analyzes the characteristics of partisan users in the COVID dataset. Here, we use the output sigmoid logits of the Retweet-BERT model, which we interpret as the probability of users being labeled as right-leaning (vs. left-leaning). Since the dataset is imbalanced, we consider only the most likely left-leaning and the most likely right-leaning users of all, which is defined as the top 20% and bottom 20% of all users in terms of their predicted polarity scores. We visualize the most retweeted users by the right- and left-leaning user-based and their subsequent audience (retweeters) distribution in Fig. 3, which we break down in detail below.

The most popular users among the left the right. The users depicted in Fig. 3 are the users most retweeted by the left- or the right-leaning retweeters. We use retweet amount as it is a reliable indication of active endorsement (Boyd, Golder, and Lotan 2010) and is also commonly used as a proxy for gauging popularity and virality on Twitter (Cha et al. 2010).

The identities of the top-most retweeted users by partisanship highlight the extent of political polarization. Partisan users mainly retweet from users of their political party. 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 pundits, or journalists working for right-leaning media. @Education4Libs is a far-right far-

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

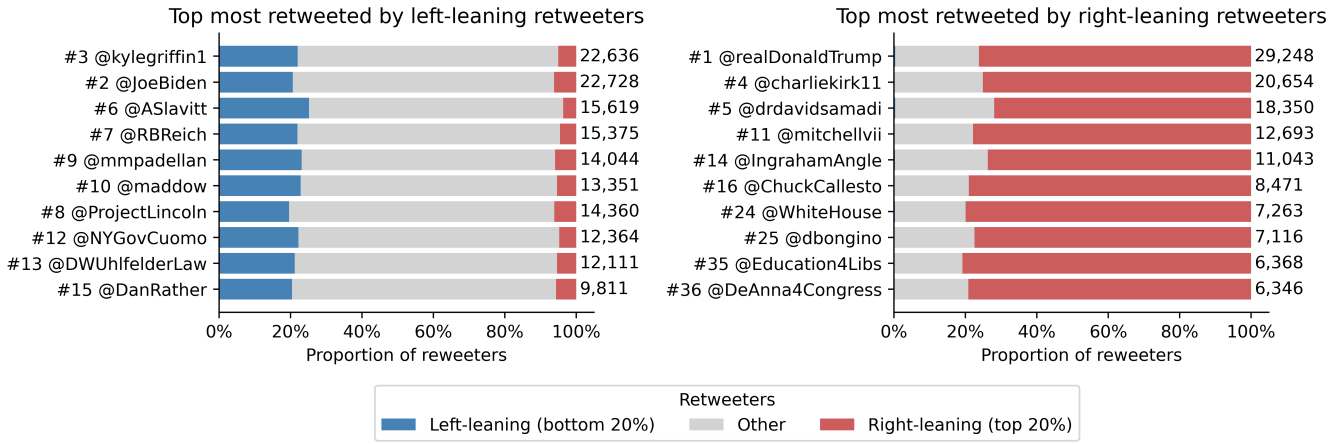

Figure 3: The most retweeted users by the left- and right-leaning user base (reprinted from Anonymous Authors (2021)). The bar plots show the distribution of their unique retweeters by political leaning. Users are also ranked by their total number of retweeters (i.e. #1 @realDonaldTrump means that @realDonaldTrump has the most retweeters overall). Numbers appended to the end of the bars show their total number of retweeters. Accounts most retweeted by left-leaning retweeters are made of 20% left-leaning retweeters and 5% right-leaning retweeters, whereas accounts most retweeted by right-leaning retweeters are made of 80% right-leaning retweeters and virtually no left-leaning retweeters.

right conspiracy group promoting QAnon.[2]

Overall popularity of popular users among the left and the right. 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 (notably @realDonaldTrump), users who are popular among the left are more popular than users who are popular among the right.

Audience of the popular users. Furthermore, we observe a striking discrepancy in the distribution of the audience. 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, hailing 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 have an audience made up almost exclusively of the far-right (around 80%) and amass only a negligible amount of far-left audience.

Summary: Our results highlight that the popular users (i.e., most retweeted) by the left-leaning users are also left-leaning, and vice versa for right-leaning. Additionally, we see that the audience (retweeters) of popular right-leaning users are tweeted almost exclusively by right-leaning users. These results suggest the presence of political echo chambers and asymmetrical information flows between and within the two echo chambers. Additional evaluations of this case study can be found in our follow-up work Anonymous Authors (2021).

[2] @Education4Libs is banned by Twitter as of January 2021.

Discussion

Implications

The effectiveness of Retweet-BERT is mainly attributed to the use of both social and textual data. Using both modalities led to significant improvement gains over using only one. This finding has also been validated in other contexts (Li and Goldwasser 2019; Pan et al. 2016; Yang and Eisenstein 2017; Johnson, Jin, and Goldwasser 2017), but ours is the first to apply this line of thought to detecting user ideology on social media.

Our work can be utilized by researchers to understand the political and ideological landscapes of social media users. For instance, we used it to understand the polarization and the partisan divide of COVID-19 discourse on Twitter. Our results suggest the existence of echo chambers, which warrants further investigation into how political echo chambers may contribute to the spread of politically biased, distorted, or non-factual information.

Though we apply Retweet-BERT specifically to the retweet network on Twitter, we note that it can be extended to *any* data with a social network structure and textual content, which is essentially any social media. Though we use hashtags as the method to initiate weak labels in place of manual supervision, other methods can be used depending on the social network platform such as user-declared interests in community groups (e.g., Facebook groups, Reddit Subreddits, Youtube channels). We leave investigations of using Retweet-BERT on other social network structures such as following networks and commenting networks for future work.

Limitations

Since our method relied on mining both user profile descriptions and the retweet network, it was necessary to remove

users that did not have profile descriptions or have sufficient retweet interactions (see Appendix). As such, our dataset only contains some of the most active and vocal users. The practical use of our model, consequently, should only be limited to active and vocal users of Twitter.

Additionally, we acknowledge that Retweet-BERT is most accurate on datasets of polarizing topics where users can be distinguished almost explicitly through verbal cues. This is driven by two reasons. First, polarizing datasets makes it clearer to evaluate detection performance. Second, and more importantly, the applications of Retweet-BERT are realistically more useful when applied to controversial or polarizing topics. Since our detection method relies on users revealing explicit cues for their political preference in their profile descriptions or their retweet activities, we focus on the top 20% (most likely right-leaning) and the bottom 20% (most likely left-leaning) when conducting the case study on the polarization of COVID-19 discussions. The decision to leave most users out is *intentional*: we only want to compare users for which Retweet-BERT is most confident in predicting political bias. Detecting user ideology is a difficult and largely ambiguous problem, even for humans (Elfardy and Diab 2016). Cohen and Ruths (2013) raised concerns that it is much more difficult to predict the political leanings of the general Twitter public, who are much more "modest" in vocalizing their political opinions. Thus, we focus our efforts on detecting the more extreme cases of political bias in an effort to reduce false positives (predicting users as politically biased when in fact they are neutral) over false negatives (predicting users as politically neutral when in fact they are biased).

Conclusion

We propose Retweet-BERT, a simple and elegant method to estimate user political leanings based on social network interactions (the social) and linguistic homophily (the textual). We evaluate our model on two recent Twitter datasets and compare it with other state-of-the-art baselines to show that Retweet-BERT achieves highly competitive performance (96%-97% macro-F1 scores). Our experiments demonstrate the importance of including both the textual and the social components. Additionally, we propose a modeling pipeline that does not require manual annotation, but only a training set of users labeled heuristically through hashtags and news media mentions. Applying Retweet-BERT to users involved in COVID-19 discussions on Twitter in the US, we find strong evidence of echo chambers and political polarization, particularly among the right-leaning population. Importantly, our work has the potential to advance future research in studying political leanings and ideology differences on social media.

Ethical Statement

We believe our work has the potential to be used in combating misinformation and conspiracy spread, as well as identifying communication patterns between and within polarized communities. However, we are aware of the ways our work can be misused. For instance, malicious actors can use our work to politically target certain groups of users and propagate the spread of misinformation. As such, we encourage researchers to use these tools in a way that is beneficial for society. Further, to protect the privacy of the users and also in accordance with Twitter's data sharing policy, we will not be sharing our actual dataset, nor the partisan labels, but only the Tweet IDs used in this paper through the original dataset release papers (Chen, Lerman, and Ferrara 2020; Chen, Deb, and Ferrara 2021). Please see the Appendix for more details. All data used in this paper are public and registered as IRB exempt by University Southern California IRB (approved protocol UP-17-00610).

Acknowledgements.

The authors are grateful to DARPA (award number HR001121C0169) for its support.

Appendix

Reproducibility

Code and Data Availability. We uploaded the code of Retweet-BERT to https://github.com/julie-jiang/retweet-bert. Upon acceptance, we will also publicly release the Tweet IDs of the preprocessed data used in our analyses. In accordance with Twitter data-sharing policies, we cannot release the actual tweets. To reproduce our work, the tweets need to be hydrated (see https://github.com/echen102/COVID-19-TweetIDs) to obtain the profile descriptions of users and to build the retweet network.

Heuristics-based Pseudo-Labeling Details. We show the exact hashtags in Table 3 and media bias ratings in Table 4 used in the heuristics-based pseudo-labeling of user political leanings. In the labeling process, all hashtags are treated as case insensitive.

Left	Right
Resist	MAGA
FBR	KAG
TheResistance	Trump2020
Resistance	WWG1WGA
Biden2020	QAnon
VoteBlue	Trump
VoteBlueNoMatterWho	KAG2020
Bernie2020	Conservative
BlueWave	BuildTheWall
BackTheBlue	AmericaFirst
NotMyPresident	TheGreatAwakening
NeverTrump	TrumpTrain
Resister	
VoteBlue2020	
ImpeachTrump	
BlueWave2020	
YangGang	

Table 3: Hashtags that are categorized as either left-leaning or right-leaning from the top 50 most popular hashtags used in user profile descriptions in the `COVID` dataset.

Media (Twitter)	URL	Rating
@ABC	abcnews.go.com	2
@BBCWorld	bbc.com	3
@BreitbartNews	breitbart.com	5
@BostonGlobe	bostonglobe.com	2
@businessinsider	businessinsider.com	3
@BuzzFeedNews	buzzfeednews.com	1
@CBSNews	cbsnews.com	2
@chicagotribune	chicagotribune.com	3
@CNBC	cnbc.com	3
@ CNN	cnn.com	2
@DailyCaller	dailycaller.com	5
@DailyMail	dailymail.co.uk	5
@FoxNews	foxnews.com	4
@HuffPost	huffpost.com	1
InfoWars*	infowars.com	5
@latimes	latimes.com	2
@MSNBC	msnbc.om	1
@NBCNews	nbcnews.com	2
@nytimes	nytimes.com	2
@NPR	npr.org	3
@OANN	oann.com	4
@PBS	pbs.org	3
@Reuters	reuters.com	3
@guardian	theguardian.com	2
@USATODAY	usatoday.com	3
@YahooNews	yahoo.com	2
@VICE	vice.com	1
@washingtonpost	washingtonpost.com	2
@WSJ	wsj.com	3

*The official Twitter account of InfoWars was banned in 2018.

Table 4: The Twitter handles, media URL, and bias ratings from AllSides.com for the popular media on Twitter.

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. (2018), 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 a 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.

Hyperparameter Tuning All models producing user (profile and/or network) embeddings are fit with a logistic regression model for classification. We search over parameter {`C`: [1, 10, 100, 1000]} to find the best 5-fold CV value. We also use randomized grid search to tune the base models. For node2vec, the search grid is {`d`: [128, 256, 512, 768], `l`: [5, 10, 20, 80], `r`: [2, 5, 10], `k`: [10, 5], `p`: [0.25, 0.5, 1, 2, 4], `q`: [0.25, 0.5, 1, 2, 4]}. For GraphSAGE, the search grid is {activation: [`relu`, `sigmoid`], S_1: [10, 25, 50], S_2: [5, 10, 20], negative samples: [5, 10, 20]}. Both node2vec and GraphSAGE are trained for 10 epochs with hidden dimensions fixed to 128. Retweet-BERT is trained for 1 epoch.

References

Abu-Jbara, A.; King, B.; Diab, M.; and Radev, D. 2013. Identifying opinion subgroups in Arabic online discussions. In *ACL '13*, 829–835. ACL.

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.

Anonymous Authors. 2021. Anonymous Published Paper. *Anonymous Journal*.

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.

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

Bamman, D.; and Smith, N. A. 2015. Open extraction of fine-grained political statements. In *EMNLP '15*, 76–85. ACL.

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.

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.; Deb, A.; and Ferrara, E. 2021. #Election2020: The first public Twitter dataset on the 2020 US Presidential election. *J. Comput. Soc. Sci.*, 1–18.

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*.

Cohen, R.; and Ruths, D. 2013. Classifying political orientation on Twitter: It's not easy! In *ICWSM '13*, volume 7. AAAI.

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.

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.

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.

Elfardy, H.; and Diab, M. 2016. Addressing annotation complexity: The case of annotating ideological perspective in Egyptian social media. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, 79–88.

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

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).

Garimella, K.; Morales, G. D. F.; Gionis, A.; and Mathioudakis, M. 2018. 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.

Grover, A.; and Leskovec, J. 2016. Node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 855–864. ISBN 9781450342322.

Hamilton, W.; Ying, Z.; and Leskovec, J. 2017. Inductive representation learning on large graphs. In *NIPS '17*, 1025–1035. Curran Associates, Inc.

Hasan, K. S.; and Ng, V. 2014. Why are you taking this stance? Identifying and classifying reasons in ideological debates. In *EMNLP '14*, 751–762. ACL.

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*.

Hornsey, M. J.; Finlayson, M.; Chatwood, G.; and Begeny, C. T. 2020. Donald Trump and vaccination: The effect of political identity, conspiracist ideation and presidential tweets on vaccine hesitancy. *J. Exp. Soc. Psychol.*, 88: 103947.

Iyyer, M.; Enns, P.; Boyd-Graber, J.; and Resnik, P. 2014. Political ideology detection using recursive neural networks. In *ACL '14*, 1113–1122. ACL.

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

Johnson, K.; Jin, D.; and Goldwasser, D. 2017. Leveraging behavioral and social information for weakly supervised collective classification of political discourse on Twitter. In *ACL'17*, 741–752. ACL.

Koeze, E.; and Popper, N. 2020. The virus changed the way we internet. *The New York Times*. Accessed: 2020-12-14.

Kovacs, B.; and Kleinbaum, A. M. 2020. Language-style similarity and social networks. *Psychol. Sci.*, 31(2): 202–213.

Li, C.; and Goldwasser, D. 2019. Encoding social information with graph convolutional networks for political perspective detection in news media. In *ACL '19*, 2594–2604. ACL.

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*.

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

Mention. 2018. Mention's Twitter engagement report 2018. Accessed: 2021-08-07.

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.

Pan, S.; Wu, J.; Zhu, X.; Zhang, C.; and Wang, Y. 2016. Tri-party deep network representation. In *IJCAI '16*. ACM.

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

Peretti-Watel, P.; Seror, V.; Cortaredona, S.; Launay, O.; Raude, J.; Verger, P.; Fressard, L.; Beck, F.; Legleye, S.; L'Haridon, O.; Léger, D.; and Ward, J. K. 2020. A future vaccination campaign against COVID-19 at risk of vaccine hesitancy and politicisation. *The Lancet Infectious Diseases*, 20(7): 769–770.

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.

Rao, A.; Morstatter, F.; Hu, M.; Chen, E.; Burghardt, K.; Ferrara, E.; and Lerman, K. 2020. Political partisanship and anti-science attitudes in online discussions about COVID-19. *arXiv preprint arXiv:2011.08498*.

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.

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*.

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.

Sim, Y.; Acree, B. D. L.; Gross, J. H.; and Smith, N. A. 2013. Measuring ideological proportions in political speeches. In *EMNLP '13*, 91–101. ACL.

Somasundaran, S.; and Wiebe, J. 2009. Recognizing Stances in Online Debates. In *ACL-IJCNLP '09*, 226–234. ACL.

Sridhar, D.; Foulds, J.; Huang, B.; Getoor, L.; and Walker, M. 2015. Joint models of disagreement and stance in online debate. In *ACL-IJCNLP '15*, 116–125. ACL.

Twitter Inc. 2021. Permanent suspension of @realDonaldTrump. Accessed: 2021-10-17.

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).

Walker, M.; Anand, P.; Abbott, R.; and Grant, R. 2012. Stance Classification using Dialogic Properties of Persuasion. In *NAACL '12*, 592–596. ACL.

Wojcik, S.; and Hughes, A. 2019. Sizing up Twitter users. https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/. Accessed: 2020-12-14.

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.; Le Scao, T.; Gugger, S.; Drame, M.; Lhoest, Q.; and Rush, A. 2020. Transformers: State-of-the-Art Natural Language Processing. In *EMNLP'20: System Demonstrations*, 38–45. Online: ACL.

Wong, F. M. F.; Tan, C. W.; Sen, S.; and Chiang, M. 2016. Quantifying political leaning from tweets, retweets, and retweeters. *TKDE*, 28(8): 2158–2172.

Wu, Z.; Pan, S.; Chen, F.; Long, G.; Zhang, C.; and Philip, S. Y. 2020. A comprehensive survey on graph neural networks. *IEEE Trans. Neural Netw. Learn. Syst.*, 32(1): 4–24.

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. ISBN 9781450379984.

Yang, Y.; and Eisenstein, J. 2017. Overcoming language variation in sentiment analysis with social attention. *Trans. Assoc. for Comput. Linguist.*, 5: 295–307.