Combating the COVID-19 Infodemic: Untrustworthy Tweet Classification using Heterogeneous Graph Transformer

Lin Ai, Zizhou Liu, Julia Hirschberg
Columbia University, New York, NY, USA
lin.ai@cs.columbia.edu, zl2889@columbia.edu, julia@cs.columbia.edu

Abstract

While COVID-19 has affected most of the world, attempts to control it have been difficult due to the lack of trustworthy information about the virus's origin, severity, effective treatments, and prevention measures. To address this, we have collected **RTCas-COVID-19**, a large corpus of 35M COVID-19 tweets from 2020, and weak-labeled 2M with a semi-supervised approach. We have also developed an inductive framework, **RTCS-HGT** (**Ret**weet **C**ascade **S**ubgraph Sampling **H**eterogeneous **G**raph **T**ransformer), which achieves 0.918 test accuracy on tweet trustworthiness classification on our dataset and improves training time by 93%.[1]

1 Introduction

The spread of misinformation has become a major issue in modern society, aided by the increasing popularity of social media (Pazzanese 2020). Misinformation online has degraded trust in many mainstream media outlets and influenced the way governments, political parties and public individuals are perceived (Ognyanova et al. 2020), leading to increased suspicion and division in society. Recently, misinformation has played a large role in the persistence of the COVID-19 pandemic, as much false information about it has been spread: how serious it is, what cures are effective, how dangerous vaccination is and how to avoid infection. As a result, the public's ability to respond to COVID-19 is seriously affected (Barua et al. 2020). This brings to light the many challenges in distinguishing between true and false information, and the negative consequences of failing to do so. While automated fact-checkers and misinformation identification systems are widely used on social media platforms (Facebook 2020), they do not always achieve their purpose. Thus it is important to continue to develop more robust systems to identify misinformation.

To improve the effectiveness of fact-checking models, one possible strategy is to employ a robust initial screening process that can identify low-credibilty information and determine whether it requires additional fact-checking. To address this challenge, we have constructed a novel corpus of 35M COVID-19 tweets, namely **RTCas-COVID-19** (**Ret**weet **Cas**cade **COVID-19**), including source tweets (original tweets that initiate retweet cascades) and retweet cascades, with 2M source tweets weak-labeled as trustworthy or untrustworthy and a small subset of human-annotated source tweets. Using these corpora, we propose an inductive framework, **RTCS-HGT** (**Ret**weet **C**ascade **S**ubgraph Sampling **H**eterogeneous **G**raph **T**ransformer), that effectively captures social context and tweet propagation patterns, and improves the performance of tweet trustworthiness classification. The contributions of this paper are twofold:

- We have constructed RTCas-COVID-19, a large Twitter corpus of COVID-19 tweets and their corresponding retweet cascades. The corpus is weak-labeled for untrustworthy information detection, and provides richer and higher quality social context information compared to other currently existing rumor detection corpora.
- We propose RTCS-HGT, an inductive tweet trustworthiness classification framework that utilizes textual information, tweet propagation patterns, and social context information. The model outperforms all the baseline models on our corpus, and is more scalable to larger real-world data.

2 Related Work

2.1 Fake News Detection on Social Media

As Allcott, Gentzkow, and Yu (2019) note, the spread of misinformation has declined sharply on Facebook but has continued to rise on Twitter since 2016; thus, much recent work on social media misinformation detection focuses on Twitter. While early work on fake news detection relied primarily on linguistic features (Pérez-Rosas et al. 2018; Rashkin et al. 2017; Ajao, Bhowmik, and Zargari 2018), information propagation patterns provide richer contexts for detecting misinformation, as fake news propagates differently from true news (Vosoughi, Roy, and Aral 2018; Kwon, Cha, and Jung 2017). Propagation-based approaches (Ma et al. 2016; Ma, Gao, and Wong 2017; Yu et al. 2017; Ma, Gao, and Wong 2018) make use of tree-structured propagation patterns of microblog posts and learn contextual representations using Recurrent Neural Network (RNN) or Convolutional Neural Network (CNN) models.

[1] Our corpus and source code are available in the Github repository https://github.com/lynneeai/RTCS-HGT.git.

Since users on social media tend to follow like-minded people (Shu, Bernard, and Liu 2019), false information such as conspiracy theory generates homogeneous and polarized communities having similar information consumption patterns (Del Vicario et al. 2016), resulting in an *echo chamber* effect. Therefore, many studies (Ruchansky, Seo, and Liu 2017; Monti et al. 2019) add social context components to the process, using user profiles in addition to textual features and information propagation patterns. Shu, Wang, and Liu (2019) make use of tri-relationships among users, news, and news publishers to detect fake information, creating multiple components which are however less flexible and scalable in adapting to new datasets or adding new features and entities.

Most graph neural networks (GNNs) models, such as GCN (Kipf and Welling 2016), GAT (Veličković et al. 2017) and GraphSage (Hamilton, Ying, and Leskovec 2017), are natively designed only for homogeneous graphs. The Heterogeneous Graph Transformer (HGT) (Hu et al. 2020) is proposed to tackle this problem, with node-type and edge-type parameters built with an attention mechanism applied over each edge type during target-specific message aggregation. Recently, a heterogeneous graph to represent user and contents has been utilized in fake contents detection (Huang et al. (2020), Agarwal et al. (2022), He et al. (2022), Min et al. (2022)). However, Huang et al. (2020) and He et al. (2022) fail to capture users' interaction in their models. Min et al. (2022) and Agarwal et al. (2022) do not utilize sampling methods and thus are potentially slow in training time with millions of nodes and edges in their proposed data.

2.2 COVID-19 False Information

Since the COVID-19 outbreak, a large amount of false information has been spread over social media platforms. The World Health Organization has also labeled the spread of fake news on COVID-19 as an "infodemic" (Thomas 2020). To address this challenge, many studies have attempted to identify false information on COVID-19 in social media. Chen, Lerman, and Ferrara (2020) and Banda et al. (2020) have collected large-scale COVID-19 Twitter datasets that are publicly available. Sharma et al. (2020) maintain a dashboard tracking unreliable information on Twitter between March and May 2020. Shaar et al. (2020) and Cheema, Hakimov, and Ewerth (2020) detect COVID-19 tweets worth fact-checking using linguistic features and language models. Zhao et al. (2022) analyze the user profile and method utilized to spread COVID-19 related misinformation in social media. While most existing studies on COVID-19 misinformation only conduct separate analyses on either linguistic or social context characteristics, we address this challenge by incorporating both, along with tweet propagation patterns.

3 Data Collection

3.1 RTCas-COVID-19 Corpus

We have collected and cleaned a new COVID-19 corpus based on 2 publicly available datasets (Chen, Lerman, and Ferrara 2020; Banda et al. 2020), namely RTCas-COVID-19. So far, we have retrieved 480M tweets from January to December 2020. To filter out less informative tweets, we select a set of source tweets, defined as original tweets posted by users, in contrast to retweets, quote tweets, and replies, from the full corpus using keyword search. The keywords are extracted from the debunked COVID-19 rumor statements provided by CMU IDeaS[2]. This process selects source tweets on specific topics that potentially contain balanced portions of rumor tweets and debunking tweets. In addition, we keep only the source tweets that have embedded URLs and are in English only. For each source tweet, its associated retweet cascade is also collected. The cleaned corpus includes 35M tweets (10M source tweets, 25M retweets).

Given the expense and time-consuming nature of human annotation, particularly for a new and time-sensitive topic like the COVID-19 pandemic, we employ a semi-supervised weak-labeling approach to label each tweet as ”trustworthy” (i.e. ”reliable”) or ”untrustworthy” (i.e. ”unreliable”) using the source credibility of the URLs shared in the tweet. To elaborate further, we categorize tweets that have a low credibility and necessitate further fact-checking. Therefore, the process of ”trustworthiness classification” serves as a precursor for identifying the veracity of the information tweeted. We collect a set of trustworthy sources from Media Bias/Fact Check (MBFC)[3]. We also add the list of mainstream news media (Wikipedia contributors 2021b) to our trustworthy sources set. In total, we identify 1357 trustworthy sources. Similarly, we collect a set of untrustworthy sources from MBFC, NewsGuard[4], and the Zimdars' list of fake and misleading news websites (Zimdars 2016). We also add the list of satirical news websites (Wikipedia contributors 2021a) to our untrustworthy sources set to include tweets that are intended to be amusing but whose content is not intended to be believed as true. In total, we identify 1518 untrustworthy sources. A source tweet will be automatically labeled as trustworthy or untrustworthy if it shares a URL from our trustworthy or untrustworthy sources sets. Overall, 2M tweets are weak-labeled (1.64M trustworthy tweets, 360K untrustworthy tweets).

3.2 Human Annotation

To verify the effectiveness of our weak-labeling approach, we (lab members) have manually annotated 380 tweets, sampled randomly from our full corpus, where each is annotated by three annotators. Both Cohen's kappa and Fleiss' kappa inter-annotator agreement scores on these annotated tweets are 0.810. Evaluated with the human-annotated set, our weak-labeling approach achieves an accuracy of 0.71, with an F1 score of 0.64 for the trustworthy class and 0.76 for the untrustworthy class. Table 1 shows the overall statistics of our corpus.

4 Model Design

4.1 Tweet-User Heterogeneous Graph

Since users tend to interact more with like-minded people, false information is more likely to be spread and more eas-

[2] https://www.cmu.edu/ideas-social-cybersecurity/

[3] https://mediabiasfactcheck.com/

[4] https://www.newsguardtech.com/

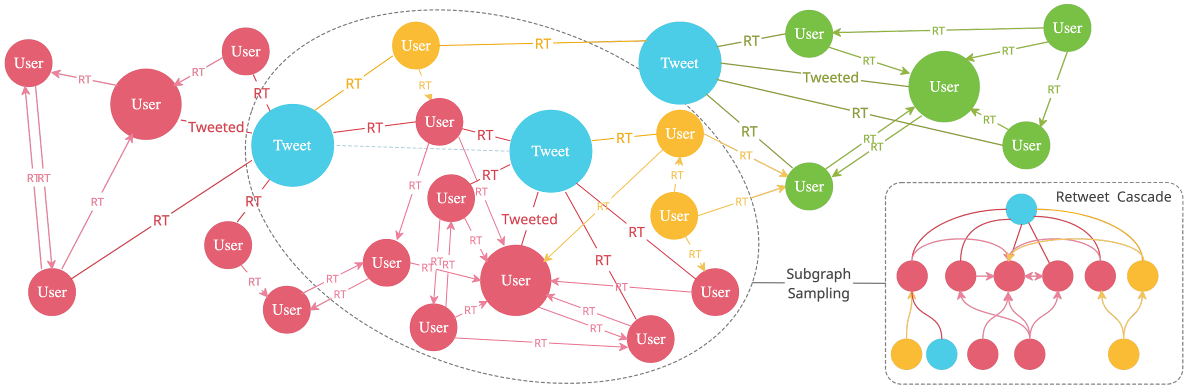

Figure 1: Tweet-user heterogeneous graph: Red user nodes are users who often tweet or retweet untrustworthy tweets; green user nodes are users who often tweet or retweet trustworthy tweets; yellow nodes are users with mixed behavior. The bottom right section shows a 2-hop retweet cascade subgraph sampling.

	Total	Source	Retweets
Full Corpus	35M	10M	25M
	Total	**Trust**	**Untrust**
Weak-Labeled	2M	1.64M	360K
Human-Annotated	380	215	165

Table 1: RTCas-COVID-19 Statistics

ily diffused within certain communities (Del Vicario et al. 2016; Shu, Bernard, and Liu 2019); thus social context may be very useful for false information detection. For example, on Twitter, retweeting can be understood as a form of information diffusion, by which users amplify tweets to new audiences and publicly agree or validate these tweets (Boyd, Golder, and Lotan 2010). Therefore, we represent social context as a tweet-user heterogeneous graph, where the nodes are connected by ”tweeting” and ”retweeting” interactions, as illustrated in Figure 1. This heterogeneous graph captures tweet propagation patterns and user relationships. The two node types in this graph are **Tweet** nodes and **User** nodes, and the types of edges are summarized in Table 2. The **Tweet** nodes are embedded using the BERTweet (Nguyen, Vu, and Tuan Nguyen 2020) model. With these nodes, we are able to take linguistic features into account in addition to social context features. For the **User** nodes, we extract Twitter profile features using the Twitter Developer API[5]. In addition, Ferrara (2020) suggest that high *bot score* accounts are used to promote political conspiracies alongside with COVID-19 content. Thus we also extract users' bot score features using a Botometer API (Sayyadiharikandeh et al. 2020). The user nodes are therefore embedded using the concatenation of Twitter profile features and bot score features.

4.2 RTCS-HGT Framework

Heterogeneous Graph Transformer We utilize an inductive HGT for node representation learning. Our model also optimizes 2 loss types simultaneously during training:

Edge	Type	Weight
user-tweeted-tweet	undirected	$\frac{1}{1+T} = 1$
user-retweeted-tweet	undirected	$\frac{1}{1+T}$
user-retweeted-user	directed	retweet count

Table 2: Tweet-User Heterogenoues Graph Edges: T, time difference in minutes between tweet's posting and user's retweeting; for the tweeters, $T = 0$.

- **Supervised Tweet Classification Loss:** The negative log likelihood classification loss. Our tweet trustworthiness classifier stacks 2 layers of HGT and concatenates the output tweet nodes' representations with a feed forward layer and a log softmax output layer to perform classification.
- **Unsupervised User Proximity Loss:** Based on the hypothesis that users who interact with each other frequently often share similar behaviors or characteristics, we want to encourage closely-connected user nodes to learn similar representations, while enforcing distanced ones to learn distinct representations, specifically:

$$\begin{aligned} proxloss = \mathbb{E}[&- \mathbb{E}\log\sigma(z_u^\top z_{v_p}) \\ &- Q \cdot \mathbb{E}\log\sigma(-z_u^\top z_{v_n})] \end{aligned} \qquad (1)$$

  for all users $u \in G$, $v_p \in P_u$, and $v_n \in N_u$. P_u denotes a set of randomly sampled neighboring nodes of node u, N_u denotes a set of randomly sampled non-neighboring nodes of node u, and z_v denotes the representation of a node v.

Retweet Cascade Subgraph Sampling To make the model scalable to large graphs and to reduce training time, we propose a tweet-centered subgraph sampling approach, visualized in Figure 1. For each tweet, we perform a 2-hop neighbor sampling, where the first hop samples the tweeter and the retweet user cascade of the tweet, and the second hop samples entities that are closely related to the tweet, including users in the same communities and some other tweets spread within these communities. With this sampling

[5] https://developer.twitter.com/en

approach, the concise sampled subgraph captures the diffusion patterns and enough social context features with respect to the tweets. This approach optimizes the training process and provides rich information for newly-seen tweets during inference time.

5 Experiments

5.1 Dataset

We train and test the RTCS-HGT model on our weak-labeled subset of the RTCas-COVID-19 corpus. To build a tweet-user heterogeneous graph with significant density, we choose equal sets of trustworthy and untrustworthy tweets that have been retweeted at least 100 times and sample the top 1% users by their number of interactions (either tweet or retweet) with the selected tweets. Table 3 summaries the data statistics.

Nodes		**Edges**	
# Source Tweets	5,714	# u-tweeted-t	5,714
# Users	39,822	# u-rt-t	562,284
# Trust	2,857	# u-rt-u	482,544
# Untrust	2,857		

Table 3: RTCS-HGT Data Statistics

5.2 Experimental Settings and Results

We compare our model with the following text classification baseline models:

- **RCNN** (Lai et al. 2015): A model with a recurrent structure that captures the contextual information and a max-pooling layer that captures the influential word in the given class of labels. We encode the tweets using a BERTweet model as inputs to the RCNN model and concatenate the last layer hidden states output of RCNN with a feed forward layer and a log softmax output layer.
- **BERTweet** (Nguyen, Vu, and Tuan Nguyen 2020): A pre-trained BERT$_{\text{base}}$ (Devlin et al. 2019) language model for English tweets, which outperforms a RoBERTa$_{\text{base}}$ (Liu et al. 2019) model on text classification. We concatenate the last layer hidden states output of BERTweet with a feed forward layer and a log softmax output layer.
- **CT-BERT** (Müller, Salathé, and Kummervold 2020): A pre-trained BERT$_{\text{LARGE}}$ model trained on 160M COVID-19 tweets. We concatenate the last layer hidden states output of CT-BERT with a feed forward layer and a log softmax output layer.
- **HGATRD** (Huang et al. 2020): A heterogeneous graph attention network framework that captures global semantic relations of text content and source tweet propagation patterns.
- **HGT** (Hu et al. 2020): This model is equivalent to our framework without retweet cascade subgraph sampling.

Our model has 55142 trainable parameters. Hyperparameters are tuned with a held-out validation set for all models, with a train-validation-test split ratio of $7/1/2$. For a fair comparison, 5-fold cross validation is utilized, and all numbers reported in Table 4 are the average results of the 5-fold test sets. As shown in Table 4, our RTCS-HGT model outperforms all baseline models on our weak-labeled subset of RTCas-COVID-19 corpus. Specifically, the RTCS-HGT model trained without user proximity loss achieves an average test accuracy of 0.918.

The CT-BERT model also achieves comparable performance, but since this model is specifically pre-trained on COVID-19 tweets, it cannot be easily applied to data on other topics without re-training on a large amount of data. HGATRD is also a strong baseline. However, it is transductive, which means it cannot make inference on unseen data, nor is it scalable to other larger datasets. In addition, HGT shows close performance to our model but requires significantly longer runtime compared to our model. We illustrate this gap in Section 5.3.

Model	**Test Acc.**	**Macro $\mathbf{F_1}$**	**Trust $\mathbf{F_1}$**	**Untrust $\mathbf{F_1}$**
RCNN-LSTM	0.844	0.844	0.846	0.842
RCNN-GRU	0.844	0.843	0.848	0.839
BERTweet	0.847	0.847	0.850	0.843
CT-BERT	0.893	0.893	0.894	0.893
HGATRD	0.894	0.894	0.894	0.895
HGT	0.908	0.908	0.910	0.906
RTCS-HGT	0.913	0.913	0.913	0.912
RTCS-HGT (no *proxloss*)	**0.918**	**0.918**	**0.918**	**0.918**

Table 4: RTCS-HGT vs. Baselines on Weak-Labeled RTCas-COVID-19 Test Sets

5.3 Ablation Study

We evaluate the effectiveness of the tweet-centered retweet cascade subgraph sampling approach and the user proximity loss by comparing different model variants with a baseline HGT model. For all model variants, we report test accuracy, macro-average F_1, and epoch elapsed time. All model variants are trained for 30 epochs with 1 Nvidia Quadro RTX 8000 GPU.

Retweet Cascade Subgraph Sampling In order to study whether the retweet cascade subgraph sampling approach improves model performance and how the sampled subgraph size makes a difference, we set a base sampler configuration and increase the sampling size by N times. For each batch of tweets, the base sampler samples the tweeter user and 10 retweet users in the first hop. In the second hop, each sampled user samples: **(a)** 5 other unsampled users that this user has retweeted, **(b)** 5 other unsampled users that has retweeted this user, **(c)** 5 tweets that this user has posted, and **(d)** 12 tweets that this user has retweeted. These numbers are the results of grid searches from 0 to the average node degrees listed in Table 7. We then multiply these numbers by a sampler multiplier N ranging from 2 to 5. This set of model variants is trained without user proximity loss.

From Table 5, we see that, in general, adding retweet cascade subgraph sampling improves the model's performance.

The best performing model is trained with the base sampler, achieving 0.918 test accuracy, outperforming the HGT baseline model by 1%. Larger sampler multiplier N increases training time, and does not necessarily improve model accuracy. On our dataset, the base sampler provides a very good balance between time and accuracy.

We further increase the sampler multiplier independently for hop 1 and hop 2 sampling. Table 5 shows that increasing the hop 1 sampler multiplier boosts performance, meaning that sampling more retweet users helps with the trustworthiness classification; with more retweet users sampled, richer social context information is obtained in the second hop as well. Increasing the hop 2 sampler alone however does not improve model performance.

The most significant improvement produced by adding the subgraph sampling is the training time. From Table 5, we see that the average time elapsed per epoch for training a HGT model is 3 minutes and 49.32 seconds, whereas the average epoch elapsed time for training RTCS-HGT models without user proximity loss ranges from 15.06 seconds to 40.13 seconds, which is 83% to 93% faster than training a HGT model. Therefore, we conclude that the retweet cascade subgraph sampling approach improves model accuracy and also significantly decreases the training time by utilizing a small subgraph with rich social context information.

Unsupervised User Proximity Loss We also evaluate the effectiveness of unsupervised user proximity loss. In the base $proxloss$ configuration, we randomly sample 10% of the user nodes from the sampled subgraph in each iteration; then each sampled user node samples 5 neighboring user nodes and 5 non-neighboring user nodes to calculate $proxloss$. These numbers are empirical results. We then train different RTCS-HGT variants by multiplying these numbers by a $proxloss$ multiplier N ($N \in [2, 5]$). We also train a HGT+$proxloss$ model with the base $proxloss$ configuration.

From Table 6, we see that, in general, adding user proximity loss makes model performance worse. Our assumption is that, when calculating $proxloss$, the randomly sampled non-neighboring user nodes add noise into the training, and therefore, worsen performance. Sampling a good quality of negative samples has always been a challenging problem in the area of self-supervised contrastive learning. However, we believe that learning good representations for users simultaneously with training the tweet trustworthiness model could potentially boost model performance, and thus benefit other downstream tasks such as communities identification and unreliable accounts detection. Therefore, improving the sampling approach for calculating $proxloss$ is one of the future directions of our study.

6 Corpus Social Context Information Analysis

In order to utilize tweet propagation patterns in misinformation detection, we construct a COVID-19 corpus of 10M source tweets along with their retweet cascades, our RTCas-COVID-19 corpus. Previously, Liu et al. (2015) and Ma et al. (2016) have collected Twitter 15 and Twitter 16, with source tweets and their corresponding propagation threads, which have been used as 2 benchmark corpora for rumor detection in social media. In this section, we compare the user interaction density and community distinction between our corpus and Twitter 15 and 16. We argue that our corpus provides richer and higher quality social context information, which better mimics the real-world data we see in social media.

6.1 User Interaction Density

For a graph $G(V, E)$, where V is the set of nodes and E is the set of edges, and a subgraph $S(V', E')$, where $V' \in V$ and $E' \in E$, we define the *density* $d(S)$ of the subgraph to be $d(S) = \frac{|E'|}{|V'|}$, and the *density* $d(G)$ of the graph G to be

$$d(G) = max_{S \subseteq G}\{d(S)\} \qquad (2)$$

We construct tweet-user heterogeneous graphs using the method described in Section 4.1 with Twitter 15 and 16 data: these have graph densities of 2.41 and 2.34, respectively. For comparison, the RTCas-COVID-19 graph has a density of 23.07, which is more than 9.5 times higher than that of Twitter 15 and 16. To better study the effect of graph density on model performance, we select denser subgraphs from the full Twitter 15 and 16 graphs using Charikar's greedy approximation algorithm (Charikar 2000) and train and test RTCS-HGT models on them. We convert the annotations of Twitter 15 and 16 to binary labels for a more direct comparison, where ”non-rumor” and ”true” tweets are considered trustworthy, and ”unverified” and ”false” tweets are considered untrustworthy. As listed in Table 8, the model achieves the highest accuracy on the Twitter 15 and 16 subsets with graph densities of 5.36 and 4.85, which are both denser than the full datasets, achieving test accuracies of 0.781 and 0.744, respectively. Thus, denser graphs help model performance. However, we observe that performance drops when the model is tested on subsets with densities higher than these. This is reasonable, as a denser subgraph might not contain enough nodes for the models to perform well, given the size of the corpora — Twitter 15 and 16 contain 1490 and 818 source tweets respectively, and our sampled corpus contains 5714 source tweets, as summarized in Table 3. Therefore, Twitter 15 and 16 do not provide enough graph density as RTCas-COVID-19 does for our model to achieve a SOTA performance. Without further specification, all subsequent experiments of user interaction density are conducted on the Twitter15/5.35 and Twitter16/4.85 variants, with 5.36 and 4.85 graph densities, respectively.

We further examine the density of user retweeting interactions by examining the node degrees of user nodes in tweet-user heterogeneous graphs. As summarized in Table 7, on average, each user in RTCas-COVID-19 posts 0.14 source tweets. However, in Twitter 15 and 16, the numbers drop 79% to 86%, where each user posts only 0.03 to 0.02 source tweets on average. Similarly, in RTCas-COVID-19, each user retweets 14.12 times on average, whereas in Twitter 15 and 16, each user retweets 2.96 and 2.67 times on average, respectively — 79% and 81% less frequently. In RTCas-COVID-19, each user is retweeted by 12.12 other distinct users on average, whereas in Twitter 15 and 16, the

Model		Test Acc.	Macro $\mathbf{F_1}$	Epoch Elapsed Time (%H:%M:%S)
HGT		0.908	0.908	0:03:49.32
RTCS-HGT Sampler N				
1		**0.918**	**0.918**	**0:00:15.06**
2		0.912	0.912	0:00:23.24
3		0.905	0.906	0:00:29.06
4		0.914	0.914	0:00:33.32
5		0.91	0.91	0:00:40.13
Hop 1 Sampler N	**Hop 2 Sampler N**			
1	2	0.913	0.912	0:00:18.53
1	3	0.913	0.912	0:00:22.45
1	4	0.913	0.913	0:00:25.87
1	5	0.912	0.912	0:00:30.26
2	1	0.902	0.902	0:00:18.69
3	1	0.913	0.913	0:00:20.83
4	1	0.914	0.914	0:00:21.84
5	1	0.917	0.916	0:00:23.02

Table 5: Accuracy and elapsed time comparison between HGT and RTCS-HGT with different subgraph sampler multipliers N. All models in the table are trained without $proxloss$.

Model	Test Acc.	Macro $\mathbf{F_1}$	Epoch Elapsed Time
HGT (no *proxloss*)	0.908	0.908	0:03:49.32
HGT (*proxloss* N=1)	0.892	0.892	0:09:07.11
RTCS-HGT (no *proxloss*)	**0.918**	**0.918**	**0:00:15.06**
***proxloss* N**			
1	0.913	0.913	0:01:47.73
2	0.913	0.912	0:03:14.58
5	0.913	0.913	0:07:02.16

Table 6: Accuracy and elapsed time comparison between HGT and RTCS-HGT with different $proxloss$ multipliers N.

numbers are 2.51 and 2.28 on average, 79% and 81% fewer than that of RTCas-COVID-19.

Dataset	User Nodes		
	#tweets	**#retweets**	**rt #users**
RTCas-COVID-19	0.14	14.12	12.12
Twitter15/5.36	0.03	2.96	2.51
Twitter16/4.85	0.02	2.67	2.28

Table 7: Average user nodes' degrees of RTCas-COVID-19, Twitter 15, and Twitter 16.

We also specifically examine the number of users each user has retweeted and the number of times each user has been retweeted by other users in these 3 corpora. Figure 2 demonstrates the distribution of these retweet counts. We exclude numbers larger than 99% percentile to avoid outliers when plotting the figures. In RTCas-COVID-19, the majority of the users have retweeted at least 5 to 30 other users and have been retweeted 0 to 300 times themselves by other users. These numbers are significantly smaller in Twitter 15 and 16, where the majority of users have retweeted at most 8 other users and have been retweeted less than 40 times by other users.

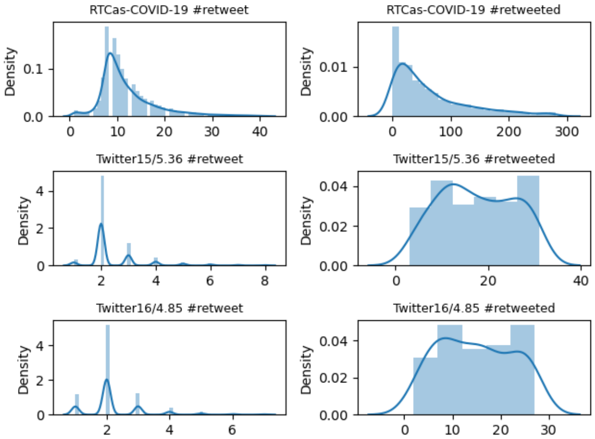

Figure 2: Retweet counts distribution. Figures in the left column illustrate the distribution of number of users each user has retweeted in RTCas-COVID-19, Twitter 15, and Twitter 16 respectively. Figures in the right column illustrate the distribution of number of times each users has been retweeted in these corpora.

We select popular tweets that have been retweeted at least 100 times and sample the top 1% of users with the most number of interactions to construct the graph we build in Section 5. This makes our graph significantly denser than those built with Twitter 15 and 16. Although Ma, Gao, and Wong (2017) state that Twitter 15 and 16 are also constructed with popular source tweets that are highly retweeted or replied, the analysis above shows that our RTCas-COVID-19 corpus not only contains more popular tweets with a larger numbers of retweets, but also captures a denser social context graph, which enables the model to utilize more community-based knowledge to perform untrust-

Dataset	Graph Density	# Users	Test Acc.	Macro F_1
Twitter 15	2.41 (full)	477,293	0.765	0.761
	5.36	**53,951**	**0.781**	**0.779**
	7.49	20,216	0.729	0.726
	9.84	7,787	0.727	0.724
Twitter 16	2.34 (full)	287,119	0.734	0.728
	4.85	**33,216**	**0.744**	**0.737**
	7.00	10,159	0.666	0.627
	9.49	2,137	0.595	0.477

Table 8: RTCS-HGT on Twitter 15 and 16 with different graph densities. The RTCS-HGT models are trained without $proxloss$ and with $SamplerN = 1$.

worthy information identification. More importantly, our corpus is significantly larger in size compared to Twitter 15 and 16, even with our sampling standard. We would be able to sample a even larger subset if we are satisfied with tweets that have comparable numbers of retweets to those in Twitter 15 and 16.

6.2 Community Distinction

In order to further analyze the quality of social context information captured by RTCas-COVID-19 — specifically, whether we can identify communities where false information is spread more easily and frequently, we visualize the tweet-user heterogeneous graph of RTCas-COVID-19, Twitter 15, and Twitter 16, as shown in Figures 3, 4, and 5. For RTCas-COVID-19, we randomly sample 2000 nodes to plot the graph as the full corpus is too large. For Twitter 15 and 16, we convert the labeling into binary. In these graphs, green nodes are trustworthy tweet nodes, red nodes are untrustworthy tweet nodes, and black nodes are user nodes.

As illustrated in Figure 3, the nodes naturally form 2 clusters where, in one cluster, the tweet nodes are mostly green, meaning that they are trustworthy, whereas in the other cluster, the tweet nodes are mostly red, implying that they are untrustworthy. We also observe that users within each cluster tend to interact more frequently with other users in the same cluster rather than with users from the other cluster, creating a distinct boundary between the two clusters. However, we do not observe such clear boundaries and clusters in the Twitter 15 and 16 graphs. In Figures 4 and 5, green nodes and red nodes are mostly mixed together, meaning that the user interactions are not dense enough to form distinguishable communities. Twitter 16's graph is slightly better than that of Twitter 15, in which we see a relatively denser cluster of red nodes, but it is still not as clear as what we see in RTCas-COVID-19's graph. This observation indicates that our RTCas-COVID-19 corpus contains popular source tweets with more complex retweet cascades; in addition, the denser user interactions also make it possible to distinguish ”red” communities from ”green” communities, where much more untrustworthy tweets are being propagated within ”red” communities. These features would also benefit other challenging tasks such as unreliable accounts detection.

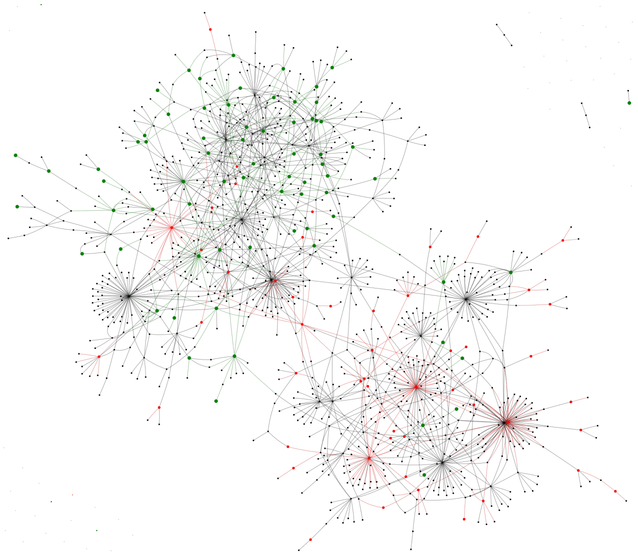

Figure 3: Sampled tweet-user heterogeneous graph of RTCas-COVID-19. Green and red nodes are trustworthy and untrustworthy tweet nodes. Black nodes are user nodes.

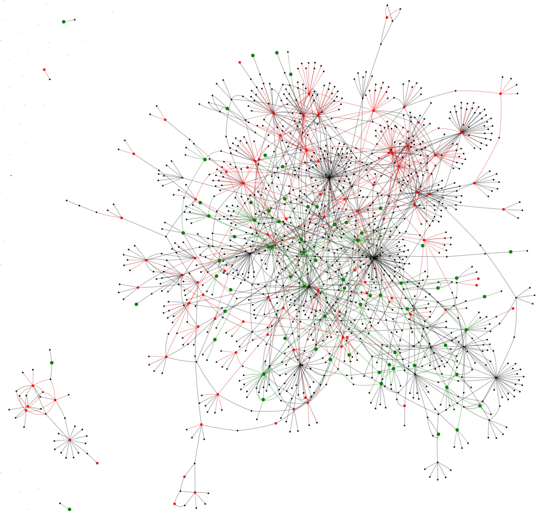

Figure 4: Tweet-user heterogeneous graph of Twitter 15. Green and red nodes are trustworthy and untrustworthy tweet nodes. Black nodes are user nodes.

7 Conclusions and Future Work

In this paper, we present **RTCas-COVID-19**, a novel corpus of 35M COVID-19 tweets, including source tweets and retweet cascades, along with 2M weak-labeled source tweets labeled with their trustworthiness and a small subset of human-annotated source tweets. We demonstrate that RTCas-COVID-19 provides richer and higher quality social context information compared with other currently existing rumor detection corpora. There are significantly more user interactions in this corpus, making the social context graph much denser than those of other corpora and forming clearly

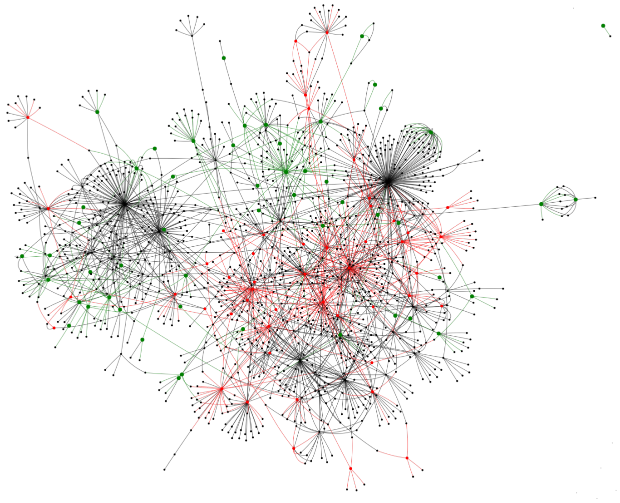

Figure 5: Tweet-user heterogeneous graph of Twitter 16. Green and red nodes are trustworthy and untrustworthy tweet nodes. Black nodes are user nodes.

distinguishable communities where different information is being spread. With these characteristics, this corpus can be used for studying not only untrustworthy information detection tasks, but also other computational social media tasks, such as early rumor detection, communities identification, and unreliable accounts detection.

In addition, we propose **RTCS-HGT**, an inductive heterogeneous graph framework, and present its results on classifying tweet trustworthiness. RTCS-HGT outperforms all the baseline models, demonstrating the effectiveness of our tweet-user heterogeneous graph and retweet cascade subgraph sampling approach in capturing social context features and tweet propagation patterns on tweet trustworthiness classification. Specifically, the retweet cascade subgraph sampling approach improves model performance, both in accuracy and runtime, by utilizing a concentrated subgraph with rich social context information. As an inductive learning model, it is also more flexible and scalable when adapting to new datasets.

In future work, we will incorporate more Twitter user interactions into our corpus and heterogeneous graph, such as replies and follows, to interpret more complex social context features. Moreover, we are investigating fact-checking approaches, which can be used as an addition to source credibility for weak-labeling the tweets. We will also explore semi-supervised or unsupervised training approaches, such as Xie et al. (2020), to train the model efficiently with a small amount of gold data, avoiding the cost of high quality human annotation. Furthermore, we plan to investigate different neighbor and non-neighbor sampling approaches when calculating the unsupervised user proximity loss in order to learn better user representations. This would potentially benefit many downstream tasks, such as unreliable accounts detection.

8 Broader Perspective and Ethics Statement

We discuss the ethical considerations of our collected RTCas-COVID-19 corpus and the further usage of RTCS-HGT as follows:

Data Collection: We present a novel COVID-19 Twitter corpus, namely RTCas-COVID-19, which is collected based on the two publicly available datasets: (Chen, Lerman, and Ferrara 2020; Banda et al. 2020). Since both of these datasets provide only Tweet IDs, we hydrate the full tweets' content using Twarc[6]. All content retrieved from Twarc is public information, and is provided by the official Twitter Developer API.

Feature Extraction: Twitter users' public profile information is extracted using the official Twitter Developer API, and users' bot scores are extracted using the Botometer API (Sayyadiharikandeh et al. 2020). Neither of these APIs provides personal information or can be used to identify individuals, and the fully hydrated data will not be released publicly.

Data Annotation: All the annotations are done by our lab members, who participate voluntarily and are fully aware of any risks of harm associated with their participation.

Data Release: To comply with Twitter's Terms of Service[7], we will only publicly release the Tweet IDs and User IDs of the collected Tweets for non-commercial research purposes.

Potential Biases: Tweets in our corpus are English-only, which potentially causes the analysis and models to be biased towards English-speaking populations. In addition, to train our model, we select tweets with large retweet counts and users with more interactions, which might lead the analysis and results biased towards users with higher activity or stronger attractiveness on Twitter.

Mis-classification: Note that the current RTCS-HGT model is trained on weakly-labeled data, and misclassification is inevitable in real-world applications. Any further usage of the model and interpretations of the results should be made with caution and under expert judgment to avoid misinterpretation and altering of credibility.

9 Acknowledgments

This research was done with funding from the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001120C0123. The views, opinions and/or findings expressed are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

References

Agarwal, P.; Srivastava, M.; Singh, V.; and Rosenberg, C. 2022. Modeling User Behavior With Interaction Networks for Spam Detection. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, 2437–2442.

[6] https://github.com/DocNow/twarc

[7] https://developer.twitter.com/en/developer-terms/agreement-and-policy

New York, NY, USA: Association for Computing Machinery. ISBN 9781450387323.

Ajao, O.; Bhowmik, D.; and Zargari, S. 2018. Fake news identification on twitter with hybrid cnn and rnn models. In *Proceedings of the 9th international conference on social media and society*, 226–230.

Allcott, H.; Gentzkow, M.; and Yu, C. 2019. Trends in the diffusion of misinformation on social media. *Research & Politics*, 6(2): 2053168019848554.

Banda, J. M.; Tekumalla, R.; Wang, G.; Yu, J.; Liu, T.; Ding, Y.; and Chowell, G. 2020. A large-scale COVID-19 Twitter chatter dataset for open scientific research–an international collaboration. *arXiv preprint arXiv:2004.03688*.

Barua, Z.; Barua, S.; Aktar, S.; Kabir, N.; and Li, M. 2020. Effects of misinformation on COVID-19 individual responses and recommendations for resilience of disastrous consequences of misinformation. *Progress in Disaster Science*, 8: 100119.

Boyd, D.; Golder, S.; and Lotan, G. 2010. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *2010 43rd Hawaii international conference on system sciences*, 1–10. IEEE.

Charikar, M. 2000. Greedy approximation algorithms for finding dense components in a graph. In *International workshop on approximation algorithms for combinatorial optimization*, 84–95. Springer.

Cheema, G. S.; Hakimov, S.; and Ewerth, R. 2020. Check_square at CheckThat! 2020: Claim Detection in Social Media via Fusion of Transformer and Syntactic Features. *arXiv preprint arXiv:2007.10534*.

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 and Surveillance*, 6(2): e19273.

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. *Proceedings of the National Academy of Sciences*, 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 *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Facebook. 2020. Here's how we're using AI to help detect misinformation.

Ferrara, E. 2020. What types of covid-19 conspiracies are populated by twitter bots? *First Monday*.

Hamilton, W. L.; Ying, R.; and Leskovec, J. 2017. Inductive representation learning on large graphs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 1025–1035.

He, L.; Xu, G.; Jameel, S.; Wang, X.; and Chen, H. 2022. Graph-Aware Deep Fusion Networks for Online Spam Review Detection. *IEEE Transactions on Computational Social Systems*, 1–9.

Hu, Z.; Dong, Y.; Wang, K.; and Sun, Y. 2020. Heterogeneous graph transformer. In *Proceedings of The Web Conference 2020*, 2704–2710.

Huang, Q.; Yu, J.; Wu, J.; and Wang, B. 2020. Heterogeneous Graph Attention Networks for Early Detection of Rumors on Twitter. In *2020 International Joint Conference on Neural Networks (IJCNN)*, 1–8. IEEE.

Kipf, T. N.; and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Kwon, S.; Cha, M.; and Jung, K. 2017. Rumor detection over varying time windows. *PloS one*, 12(1): e0168344.

Lai, S.; Xu, L.; Liu, K.; and Zhao, J. 2015. Recurrent convolutional neural networks for text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29.

Liu, X.; Nourbakhsh, A.; Li, Q.; Fang, R.; and Shah, S. 2015. Real-time rumor debunking on twitter. In *Proceedings of the 24th ACM international on conference on information and knowledge management*, 1867–1870.

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*.

Ma, J.; Gao, W.; Mitra, P.; Kwon, S.; Jansen, B. J.; Wong, K.-F.; and Cha, M. 2016. Detecting rumors from microblogs with recurrent neural networks. In *25th International Joint Conference on Artificial Intelligence, IJCAI 2016*, 3818–3824. International Joint Conferences on Artificial Intelligence.

Ma, J.; Gao, W.; and Wong, K.-F. 2017. Detect Rumors in Microblog Posts Using Propagation Structure via Kernel Learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 708–717. Vancouver, Canada: Association for Computational Linguistics.

Ma, J.; Gao, W.; and Wong, K.-F. 2018. Rumor Detection on Twitter with Tree-structured Recursive Neural Networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1980–1989. Melbourne, Australia: Association for Computational Linguistics.

Min, E.; Rong, Y.; Bian, Y.; Xu, T.; Zhao, P.; Huang, J.; and Ananiadou, S. 2022. Divide-and-Conquer: Post-User Interaction Network for Fake News Detection on Social Media. In *Proceedings of the ACM Web Conference 2022*, WWW '22, 1148–1158. New York, NY, USA: Association for Computing Machinery. ISBN 9781450390965.

Monti, F.; Frasca, F.; Eynard, D.; Mannion, D.; and Bronstein, M. M. 2019. Fake news detection on social media using geometric deep learning. *arXiv preprint arXiv:1902.06673*.

Müller, M.; Salathé, M.; and Kummervold, P. E. 2020. COVID-Twitter-BERT: A Natural Language Processing Model to Analyse COVID-19 Content on Twitter. *arXiv preprint arXiv:2005.07503*.

Nguyen, D. Q.; Vu, T.; and Tuan Nguyen, A. 2020. BERTweet: A pre-trained language model for English Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 9–14. Online: Association for Computational Linguistics.

Ognyanova, K.; Lazer, D.; Robertson, R. E.; and Wilson, C. 2020. Misinformation in action: Fake news exposure is linked to lower trust in media, higher trust in government when your side is in power. *Harvard Kennedy School Misinformation Review*, 1.

Pazzanese, C. 2020. Social media used to spread, create COVID-19 falsehoods – Harvard Gazette. https://news.harvard.edu/gazette/story/2020/05/social-media-used-to-spread-create-covid-19-falsehoods/.

Pérez-Rosas, V.; Kleinberg, B.; Lefevre, A.; and Mihalcea, R. 2018. Automatic Detection of Fake News. In *Proceedings of the 27th International Conference on Computational Linguistics*, 3391–3401. Santa Fe, New Mexico, USA: Association for Computational Linguistics.

Rashkin, H.; Choi, E.; Jang, J. Y.; Volkova, S.; and Choi, Y. 2017. Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2931–2937. Copenhagen, Denmark: Association for Computational Linguistics.

Ruchansky, N.; Seo, S.; and Liu, Y. 2017. Csi: A hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 797–806.

Sayyadiharikandeh, M.; Varol, O.; Yang, K.-C.; Flammini, A.; and Menczer, F. 2020. Detection of novel social bots by ensembles of specialized classifiers. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2725–2732.

Shaar, S.; Nikolov, A.; Babulkov, N.; Alam, F.; Barrón-Cedeno, A.; Elsayed, T.; Hasanain, M.; Suwaileh, R.; Haouari, F.; Da San Martino, G.; et al. 2020. Overview of CheckThat! 2020 English: Automatic identification and verification of claims in social media. *Conference and Labs of the Evaluation Forum*.

Sharma, K.; Seo, S.; Meng, C.; Rambhatla, S.; Dua, A.; and Liu, Y. 2020. Coronavirus on social media: Analyzing misinformation in Twitter conversations. *arXiv preprint arXiv:2003.12309*.

Shu, K.; Bernard, H. R.; and Liu, H. 2019. Studying fake news via network analysis: detection and mitigation. In *Emerging Research Challenges and Opportunities in Computational Social Network Analysis and Mining*, 43–65. Springer.

Shu, K.; Wang, S.; and Liu, H. 2019. Beyond news contents: The role of social context for fake news detection. In *Proceedings of the twelfth ACM international conference on web search and data mining*, 312–320.

Thomas, Z. 2020. WHO says fake coronavirus claims causing'infodemic'. *BBC. Available at: https://www. bbc. com/news/technology-51497800 (accessed 22 March 2020)*.

Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; and Bengio, Y. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Vosoughi, S.; Roy, D.; and Aral, S. 2018. The spread of true and false news online. *Science*, 359(6380): 1146–1151.

Wikipedia contributors. 2021a. List of satirical news websites — Wikipedia, The Free Encyclopedia. [Online; accessed 15-May-2021].

Wikipedia contributors. 2021b. News media in the United States — Wikipedia, The Free Encyclopedia. [Online; accessed 15-May-2021].

Xie, Q.; Dai, Z.; Hovy, E.; Luong, T.; and Le, Q. 2020. Unsupervised Data Augmentation for Consistency Training. *Advances in Neural Information Processing Systems*, 33.

Yu, F.; Liu, Q.; Wu, S.; Wang, L.; Tan, T.; et al. 2017. A Convolutional Approach for Misinformation Identification. In *IJCAI*, 3901–3907.

Zhao, Y.; Zhu, S.; Wan, Q.; Li, T.; Zou, C.; Wang, H.; and Deng, S. 2022. Understanding how and by whom covid-19 misinformation is spread on social media: Coding and network analyses.

Zimdars, M. 2016. False, misleading, clickbait-y, and satirical “news” sources. *Google Docs*.