A Dashboard for Mitigating the COVID-19 Misinfodemic

Zhengyuan Zhu[1], Kevin Meng[2], Josue Caraballo[1], Israa Jaradat[1],
Xiao Shi[1], Zeyu Zhang[1], Farahnaz Akrami[1], Haojin Liao[1], Fatma Arslan[1],
Damian Jimenez[1], Mohammed Samiul Saeef[1], Paras Pathak[1], and Chengkai Li[1]

[1]The University of Texas at Arlington
[2]Massachusetts Institute of Technology

Abstract

This paper describes the current milestones achieved in our ongoing project that aims to understand the surveillance of, impact of, and effective interventions against the COVID-19 misinfodemic on Twitter. Specifically, it introduces a public dashboard which, in addition to displaying case counts in an interactive map and a navigational panel, also provides some unique features not found in other places. Particularly, the dashboard uses a curated catalog of COVID-19 related facts and debunks of misinformation, and it displays the most prevalent information from the catalog among Twitter users in user-selected U.S. geographic regions. The paper explains how to use BERT-based models to match tweets with the facts and misinformation and to detect their stance towards such information. The paper also discusses the results of preliminary experiments on analyzing the spatio-temporal spread of misinformation.

1 Introduction

Alongside the COVID-19 pandemic, there is a raging global misinfodemic (Mian and Khan, 2020; Roozenbeek et al., 2020) just as deadly. As fear grows, false information related to the pandemic goes viral on social media and threatens to affect an overwhelmed population. Such misinformation misleads the public on how the virus is transmitted, how authorities and people are responding to the pandemic, as well as its symptoms, treatments, and so on. This onslaught exacerbates the vicious impact of the virus, as the misinformation drowns out credible information, interferes with measures to contain the outbreak, depletes resources needed by those at risk, and overloads the health care system. Although health misinformation is not new (Oyeyemi et al., 2014), such a dangerous interplay between a pandemic and a misinfodemic is unprecedented. It calls for studying not only the outbreak but also its related misinformation; the fight on these two fronts must go hand-in-hand.

This demo paper describes the current milestones achieved in our ongoing project that aims to understand the surveillance of, impact of, and effective interventions against the COVID-19 misinfodemic. 1) For *surveillance*, we seek to discover the patterns by which different types of COVID-19 misinformation spread. 2) To understand the *impact* of misinformation, we aim to compare the spreading of the SARS-CoV-2 virus and misinformation and derive their correlations. 3) To understand what types of *interventions* are effective in containing misinformation, we will contrast the spreading of misinformation before and after debunking efforts. 4) To understand whether the outcomes related to 1), 2) and 3) differ by geographical locations and demographic groups, we will study the variability of misinformation and debunking efforts across geographical and demographic groups.

While we continue to pursue these directions, we have built an online dashboard at https://idir.uta.edu/covid-19/ to directly benefit the public. A screencast video of the dashboard is at bit.ly/3c6v5xf. The dashboard provides a map, a navigation panel, and timeline charts for looking up numbers of cases, deaths, and recoveries, similar to a number of COVID-19 tracking dashboards. [1][2][3] However, our dashboard also provides several features not found in other places.

[1]https://www.covid19-trials.com/
[2]https://coronavirus.jhu.edu/map.html
[3]https://www.cdc.gov/covid-data-tracker/index.html

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, pages 99–105
April 19 - 23, 2021.

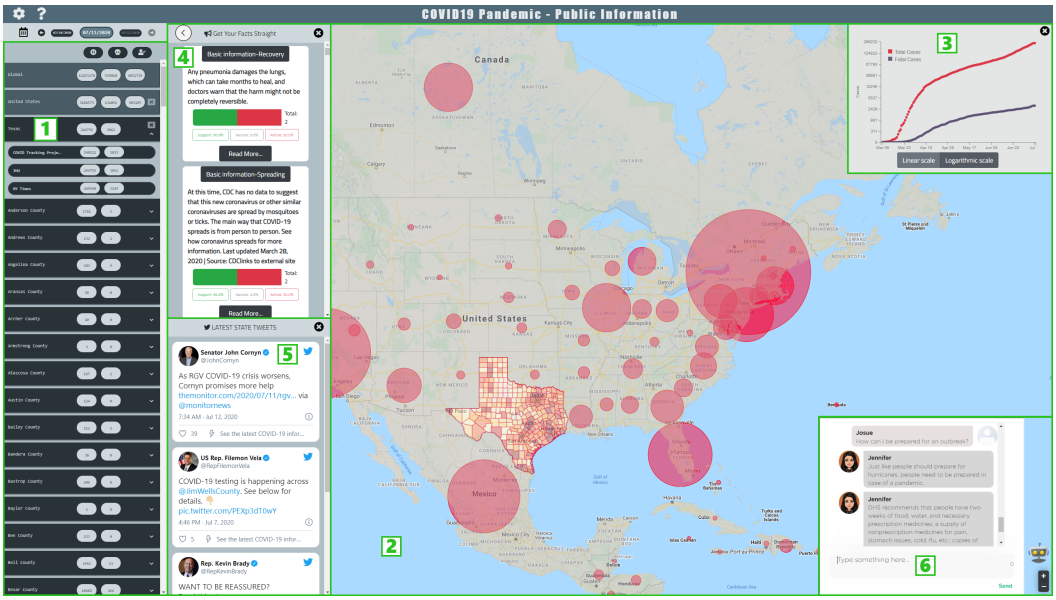

Figure 1: The user interface of the dashboard for mitigating the COVID-19 misinfodemic

1) It displays the most prevalent factual information among Twitter users in any user-selected U.S. geographic region. 2) The "factual information" comes from a catalog that we manually curated. It includes statements from authoritative organizations, verdicts, debunks, and explanations of (potentially false) factual claims from fact-checking websites, and FAQs from credible sources. The catalog's entries are further organized into a taxonomy. For simplicity, we refer to it as the *catalog and taxonomy of COVID-19 facts* or just *facts* in ensuing discussion. 3) The dashboard displays COVID-19 related tweets from local authorities of user-selected geographic regions. 4) It embeds a chatbot built specifically for COVID-19 related questions. 5) It shows case-statistics from several popular sources which sometimes differ.

The codebase of the dashboard's frontend, backend, and data collection tools are open-sourced at https://github.com/idirlab/covid19. All collected data are at https://github.com/idirlab/covid19data. Particularly, the catalog and taxonomy of facts are also available through a SPARQL endpoint at https://cokn.org/deliverables/7-covid19-kg/ and the corresponding RDF dataset can be requested there.

What is particularly worth noting about the underlying implementation of the dashboard is the adaptation of state-of-the-art textual semantic similarity and stance detection models. Tweets are first passed through a *claim-matching* model, which selects the tweets that semantically match the facts in our catalog. Then, the *stance detection* model determines whether the tweets agree with, disagree with, or merely discuss these facts. This enables us to pinpoint pieces of misinformation (i.e., tweets that disagree with known facts) and analyze their spread.

A few studies analyzed and quantified the spread of COVID-19 misinformation on Twitter (Kouzy et al., 2020; Memon and Carley, 2020; Al-Rakhami and Al-Amri, 2020) and other social media platforms (Brennen et al., 2020). However, these studies conducted mostly manual inspection of small datasets, while our system automatically sifts through millions of tweets and matches tweets with our catalog of facts.

2 The Dashboard

Figure 1 shows the dashboard's user interface, with its components highlighted.

Geographic region selection panel (Component 1). A user can select a specific country, a U.S. state, or a U.S. county by using this panel or the interactive map (Component 2). Once a region is selected, the panel shows the counts of confirmed cases, deaths and recovered cases for the region in collapsed or expanded modes. When a region is expanded by the user, counts from all available sources are displayed; on the other hand, if it is collapsed, only counts from

the default (which the user can customize) data source are displayed. These sources do not provide identical numbers.

Interactive map (Component 2). On each country and each U.S. state, a red circle is displayed, with an area size proportional to its number of confirmed cases. When a state is selected, the circle is replaced with its counties' polygons in different shades of red, proportional to the counties' confirmed cases.

Timeline chart (Component 3). It plots the counts of the selected region over time and can be viewed in linear or logarithmic scale.

Panel of facts (Component 4). For the selected region, this panel displays facts from our catalog, and the distribution of people discussing, agreeing, or disagreeing with them on Twitter. A large number of people refuting these facts would indicate wide spread of misinformation. To avoid repeating misconceptions, the dashboard displays facts from authoritative sources only.

Government tweets (Component 5). It displays COVID-19 related tweets in the past seven days from officials of the user-selected geographic region. These tweets are from a curated list of 3,744 Twitter handles that belong to governments, officials, and public health authorities at U.S. federal and state levels.

Chatbot (Component 6). This component embeds the *Jennifer Chatbot* built by the New Voices project of the National Academies of Sciences, Engineering and Medicine (Li et al., 2020), which was built specifically for answering COVID-19 related questions. As part of the collaborative team behind this chatbot, we are expanding it using the aforementioned catalog.

3 The Datasets

The dashboard uses the following three datasets.

1) Counts of confirmed cases, deaths, and recoveries. We collected these counts daily from Johns Hopkins University, [4] the New York Times (NYT) [5] and the COVID Tracking Project. [6] These sources provide statistics at various geographic granularities (country, state, county).

2) Tweets. We are using a collection of approximately 250 million COVID-19 related tweets from January 1st, 2020 to May 16th, 2020, obtained from (Banda et al., 2020) (version 10.0). We removed tweets and Twitter handles (and their tweets) that do not have location information, resulting in 34.6 million remaining tweets. We then randomly selected 10.4% of each month's tweets, leading to 3.6 million remaining tweets. We used the OpenStreetMap (Quinion et al., 2020) API to map the locations of Twitter accounts from user-entered free text to U.S. county names. We used the ArcGIS API [7] to map the locations of tweets from longitude/latitude to counties.

3) A catalog and a taxonomy of COVID-19 related facts.

The manually curated catalog currently has 9,512 entries from 21 credible websites, including statements from authoritative organizations (e.g., WHO, CDC), verdicts, debunks, and explanations of factual claims (of which the truthfulness varies) from fact-checking websites (e.g., the IFCN CoronaVirusFacts Alliance Database, [8] PolitiFact), and FAQs both from credible sources (e.g., FDA, NYT) and a dataset curated by (Wei et al., 2020).

We organized the entries in this catalog into a taxonomy of categories, by integrating and consolidating the available categories from a number of source websites, placing entries from other websites into these categories or creating new categories, and organizing the categories into a hierarchical structure based on their inclusion relationship. The taxonomy is as follows, in the format of {level-1 categories [*level-2 categories* (`level-3 categories`)]}: [9]
{Animals, Basic Information [*Causes, Definition, Disease Alongside, Recovery, Spreading, Symptoms, Testing*], Cases, Contribution, Diplomacy, Economics/Finance [*Crisis, Grants/Stimulus, Tax, Unemployment*], Family Preparation, Funeral, Government Control [*Administration* (`Lockdown, Reopen, Staff`), *Law, Medical Support, Military*], Mental Health, Prevention [*Actions to Prevent* (`Hand Hygiene, Isolation, Masks, Social Distancing`), *Medication, Vaccines*], Religion, Schools/Universities, Travel, Treatment [*Medication, Minor Symptom, Severe Symptom*], Violence/Crime}.

We also stored the catalog and the taxonomy

[4] https://github.com/CSSEGISandData/COVID-19
[5] https://github.com/nytimes/covid-19-data
[6] https://covidtracking.com/
[7] https://developers.arcgis.com/python/guide/reverse-geocoding/
[8] https://www.poynter.org/ifcn-covid-19-misinformation/
[9] Not every level-1 or level-2 category has subcategories.

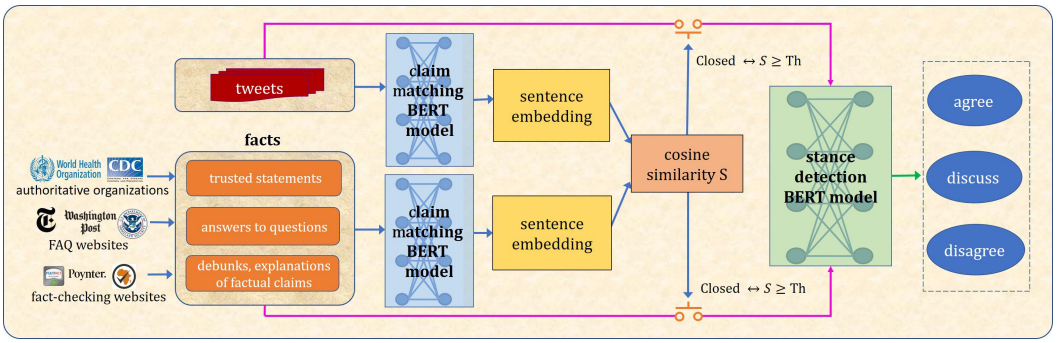

Figure 2: Matching tweets with facts and stance detection

Tweet	Fact	Taxonomy Categories	Similarity	Stance
Coronavirus cannot be passed by dogs or cats but they can test positive.	There has been no evidence that pets such as dogs or cats can spread the coronavirus.	Animals, Spreading	0.817	agree
More people die from the flu in the U.S. in 1 day than have died of the Coronavirus across the world ever.	Right now, it appears that COVID-19, the disease caused by the new coronavirus, causes more cases of severe disease and more deaths than the seasonal flu.	Cases	0.816	disagree

Table 1: Example results of matching tweets with facts and stance detection

as an RDF dataset, in which each entry of the catalog is identified by a unique resource identifier (URI). It is connected to a mediator node that represents the multiary relation associated with the entry. For example, Figure 3 shows a question about COVID-19, its answer and source, and the lowest-level taxonomy nodes that the entry belongs to, all connected to a mediator node. This RDF dataset, with 12 relations and 78,495 triples, is published in four popular RDF formats—N-Triples, Turtle, N3, and RDF/XML. Furthermore, we have set up a SPARQL query endpoint at https://cokn.org/deliverables/7-covid19-kg/ using OpenLink Virtuoso.[10]

4 Matching Tweets with Facts and Stance Detection

Given the catalog of COVID-19 related facts F and the tweets T, we first employ *claim-matching* to locate a set of tweets $\mathbf{t}^f \in T$ that discuss each fact $f \in F$. Next, we apply *stance detection* on pairs $\mathbf{p}^f = \{(t, f) \mid t \in \mathbf{t}^f\}$ to determine whether each t is agreeing with, disagreeing with, or neutrally discussing f. Finally, aggregate results are displayed on Component 4 of the dashboard to summarize the public's view on each fact. Figure 2 depicts the overall claim-matching and stance detection pipeline. For both tasks, we employed Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019). Table 1 shows some example results of claim matching and stance detection.

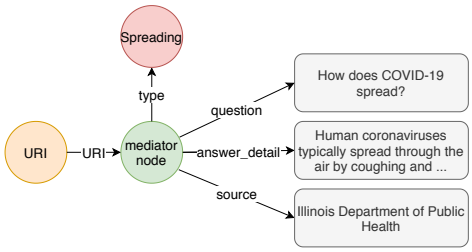

Figure 3: An entry of the catalog stored in RDF

Claim matching.

We generate sentence embeddings $\mathbf{s}^t$ and $\mathbf{s}^f$, for t and f respectively, using the mean-tokens pooling strategy in Sentence-BERT (Reimers and Gurevych, 2019). The relevance between t and f is then calculated as:

$$R^{t,f} = \frac{\mathbf{s}^t \cdot \mathbf{s}^f}{\|\mathbf{s}^t\| \times \|\mathbf{s}^f\|} \quad (1)$$

Given $R^{t,f}$, we model claim-matching as a ranking task on the relevance between facts and tweets. Thus, the output of this stage is $\mathbf{t}^f = \{t \in T \mid R^{t,f} \geq \theta\}$ for each fact $f \in F$, where the threshold θ is 0.8 in our implementation.

[10] https://virtuoso.openlinksw.com/

Stance detection. Given $\mathbf{t}^f$, we detect the stance that each tweet t takes toward fact f. There could be 3 classes of stance: agree (t supports f), discuss (t neutrally discusses f), and disagree (t refutes f). For this task, we obtained a pre-trained $\text{BERT}_{\text{Base}}$ model [11] and trained it on the Fake-News Challenge Stage 1 (FNC-1) dataset. [12] We denote this model Stance-BERT.

We first pre-process $\mathbf{p}^f$ to conform with BERT input conventions by 1) applying $W(\cdot)$, the WordPiece tokenizer (Wu et al., 2016), 2) applying $C(a_1, a_2, \ldots, a_n)$, a function that concatenates arguments in appearance order, and 3) inserting specialized BERT tokens [CLS] and [SEP]. Since BERT has a maximum input length of $M = 512$ and some facts can exceed this limit, we propose a sliding-window approach inspired by (Devlin et al., 2019) to form input $\mathbf{x}^f$:

$$\mathbf{x}^f = \Big\{ \{C([\text{CLS}], W(t), [\text{SEP}], W(f)_{[i*S, i*S+L]}, [\text{SEP}]) \mid 0 \le i < \left\lceil \frac{|W(f)|}{S} \right\rceil \} \mid (t, f) \in \mathbf{p}^f \Big\} \quad (2)$$

where S defines the distance between successive windows and $L = M - (|W(t)| + 3)$ is the sequence length available for each fact. If $i * S + L$ is an out-of-bounds index for $W(f)$, the extra space is padded using null tokens.

Each element $\mathbf{w} \in \mathbf{x}^f$ contains a set of windows representing a tweet-fact pair. Each window $w_i \in \mathbf{w}$ is passed into Stance-BERT, which returns probability distributions (each containing 3 entries, 1 for each class) $\hat{\mathbf{y}}^f_{w_i}$ for each window.

Stance aggregation. For each fact f, the stance detection results are accumulated to generate scores S^f_C, where $C \in \{\text{agree}, \text{discuss}, \text{disagree}\}$ that denote the percentage of tweets that agree, discuss, and disagree with f: [13]

$$S^f_C = \frac{\sum_{\mathbf{w} \in \mathbf{x}^f} \left[\text{argmax}\, \sigma(\{\hat{\mathbf{y}}^f_{w_i} \mid w_i \in \mathbf{w}\}) = C \right]}{|\mathbf{x}^f|} \quad (3)$$

where $\sigma(\cdot)$ is a function that averages the model's output scores for each class across all windows of tweet-fact pair. The 3 final stance scores are passed to the dashboard's panel of facts (Component 4) for display.

[11] https://github.com/google-research/bert

[12] http://www.fakenewschallenge.org/

[13] We use the Iverson bracket: $[P] = 1$ if P is true, else 0

5 Evaluation and Results

5.1 Performance of Claim Matching

To evaluate the performance of the claim matching component, we first created a Cartesian product of the 3.6 million tweets with 500 "facts" from the catalog (see Section 3 for description of datasets), followed by randomly selecting 800 tweet-fact pairs from the Cartesian product. To retain a balanced dataset, 400 pairs were drawn from those pairs scored over 0.8 by the claim matching component, and another 400 pairs were drawn from the rest. To obtain the ground-truth labels on these 800 pairs, we used three human annotators. 183 pairs were labeled "matched" (i.e., the tweet and the fact have matching topics) and 617 pairs "unmatched". Table 2 shows the claim matching component's performance on these 800 pairs, measured by precision@k and nDCG@k(normalized Discounted Cumulative Gain at k). Both precision@k and nDCG@k are metrics of ranking widely used in classification problem, the order of top k prediction is considered in nDCG@k but not in precision@k.

Metric	@5	@10	@20	@50	@100
Precision	0.80	0.80	0.70	0.56	0.52
nDCG	0.62	0.72	0.78	0.81	0.83

Table 2: Performance of claim matching on the 800 tweet-fact pairs

5.2 Performance of Stance-BERT

Model	F1 score			
	agree	discuss	disagree	macro
Stance-$\text{BERT}_{\text{window}}$(FNC-1)	0.65	**0.45**	**0.84**	**0.65**
Stance-$\text{BERT}_{\text{trunc}}$(FNC-1)	**0.66**	0.41	0.82	0.63
(Xu et al., 2018)(FNC-1)	0.55	0.15	0.73	0.48
Stance-$\text{BERT}_{\text{window}}$(COVID-19)	0.75	0.03	0.58	0.45

Table 3: Performance of Stance-BERT on the FNC-1 test dataset and 200 matched tweet-fact pairs

Table 3 shows Stance-BERT's performance on the FNC-1 competition test dataset and our tweet-fact pairs, using F1 scores for all 3 classes as well as macro-F1. On FNC-1, we tested 2 variations of the same model: $\text{Stance-BERT}_{\text{window}}$, which uses the sliding-window approach (Section 4), and $\text{Stance-BERT}_{\text{trunc}}$, a model that truncates/discards all inputs after M tokens but is otherwise identical to $\text{Stance-BERT}_{\text{window}}$. Both variants significantly outperformed the method

used in (Xu et al., 2018), one of the recent competitive methods on FNC-1.

Note that FNC-1 also includes a fourth "unrelated" class that we discarded, since we already have a claim-matching component. Because other recent stance detection methods (Mohtarami et al., 2018; Fang et al., 2019) only reported macro-F1 scores calculated using all four classes including "unrelated", we cannot report a direct comparison with their methods. However, we argue that our macro-F1 of 0.65 remains highly competitive. The model of (Xu et al., 2018) achieved a 0.98 F1 score on "unrelated", which suggests that "unrelated" (i.e., separating related and unrelated pairs) is far easier than the other 3 classes (i.e., discerning between different classes of related pairs). Given that Stance-BERT significantly outperformed (Xu et al., 2018) on all other 3 classes, it is plausible that Stance-BERT will remain a top performer under all four classes.

To evaluate Stance-BERT's performance on our tweet-fact pairs, the three human annotators produced ground-truth labels on another set of 481 randomly selected tweet-fact pairs. 200 pairs are labeled as "matched". These 200 pairs are further labeled as "agree"/"discuss"/"disagree", in a distribution of 110/73/17 tweet-fact pairs. Ultimately, we discovered that Stance-BERT performs remarkably well on "agree" and "disagree" classes but falters on "discuss".

5.3 Misinformation Analysis

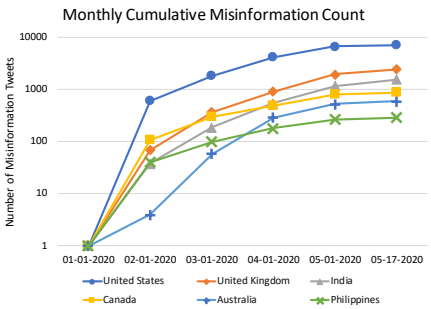

Figure 4: 6 countries with the most misinformation tweets

Figure 4 is the cumulative timeline for the top-6 countries with the most COVID-19 misinformation tweets in the dataset. "Misinformation tweets" refer to tweets that go against known facts as judged by our stance detection model.

We also conducted a study on the correlation between misinformation tweet counts and COVID-19 case counts. We looked at the percentage of cases relative to a country's population size, and the percentage of misinformation tweets relative to the total number of tweets from a country. The Pearson correlation coefficients between them are in Table 4. We find that the number of misinformation tweets most positively correlates with the number of confirmed cases. In contrast, its correlation with the number of recovered cases is weaker.

Country	Confirm	Death	Recover
United States	0.763	0.738	0.712
United Kingdom	0.862	0.833	-
India	0.794	0.798	0.755
Canada	0.706	0.667	0.663
Australia	0.954	0.922	0.887
Philippines	0.720	0.696	0.618

Table 4: Correlation between the percentage of confirmed/deceased/recovered cases and the percentage of misinformation tweets. The number of recovered cases in U.K. after April 13th is missing from the data source.

Finally, we manually categorized the misinformation tweets based on the taxonomy (Section 3). Table 5 lists the five most frequent categories of misinformation tweets. These five categories make up 49.9% of all misinformation tweets, with the other 50.1% being spread out over the other 33 categories.

Category	Count	Percentage
Definition	2503	15.1
Spreading	2118	12.7
Other	1450	8.7
Testing	1301	7.8
Disease Alongside	936	5.6
Total	8308	49.9

Table 5: Most frequent categories of misinformation tweets

6 Conclusion

This paper introduces an information dashboard constructed in the context of our ongoing project regarding the COVID-19 misinfodemic. Going forward, we will focus on developing the dashboard at scale, including more comprehensive tweet collection and catalog discovery and collection. We will also introduce more functions into the dashboard that are aligned with our project goal of studying the surveillance of, impact of, and intervention on COVID-19 misinfodemic.

References

Mabrook S Al-Rakhami and Atif M Al-Amri. 2020. Lies kill, facts save: Detecting covid-19 misinformation in twitter. *IEEE Access*, 8:155961–155970.

Juan M. Banda, Ramya Tekumalla, Guanyu Wang, Jingyuan Yu, Tuo Liu, Yuning Ding, Katya Artemova, Elena Tutubalin, and Gerardo Chowell. 2020. A large-scale COVID-19 twitter chatter dataset for open scientific research - an international collaboration.

J Scott Brennen, Felix M Simon, Philip N Howard, and Rasmus Kleis Nielsen. 2020. Types, sources, and claims of COVID-19 misinformation. *Reuters Institute*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186.

Wei Fang, Moin Nadeem, Mitra Mohtarami, and James Glass. 2019. Neural multi-task learning for stance prediction. In *EMNLP Workshop on Fact Extraction and Verification*, pages 13–19.

Ramez Kouzy, Joseph Abi Jaoude, Afif Kraitem, Molly B El Alam, Basil Karam, Elio Adib, Jabra Zarka, Cindy Traboulsi, Elie W Akl, and Khalil Baddour. 2020. Coronavirus goes viral: quantifying the COVID-19 misinformation epidemic on twitter. *Cureus*, 12(3).

Yunyao Li, Tyrone Grandison, Patricia Silveyra, Ali Douraghy, Xinyu Guan, Thomas Kieselbach, Chengkai Li, and Haiqi Zhang. 2020. Jennifer for COVID-19: An nlp-powered chatbot built for the people and by the people to combat misinformation. In *ACL Workshop on Natural Language Processing for COVID-19*, pages 1–9.

Shahan Ali Memon and Kathleen M Carley. 2020. Characterizing covid-19 misinformation communities using a novel twitter dataset. *arXiv preprint arXiv:2008.00791*.

Areeb Mian and Shujhat Khan. 2020. Coronavirus: the spread of misinformation. *BMC medicine*, 18(1):1–2.

Mitra Mohtarami, Ramy Baly, James Glass, Preslav Nakov, Lluís Màrquez, and Alessandro Moschitti. 2018. Automatic stance detection using end-to-end memory networks. In *NAACL*, pages 767–776.

Sunday Oluwafemi Oyeyemi, Elia Gabarron, and Rolf Wynn. 2014. Ebola, twitter, and misinformation: a dangerous combination?. *BMJ*, 349:g6178.

Brian Quinion, Sarah Hoffmann, and Marc T. Metten. 2020. Nominatim: A search engine for openstreetmap data.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP-IJCNLP*, pages 3973–3983.

Jon Roozenbeek, Claudia R Schneider, Sarah Dryhurst, John Kerr, Alexandra LJ Freeman, Gabriel Recchia, Anne Marthe Van Der Bles, and Sander Van Der Linden. 2020. Susceptibility to misinformation about covid-19 around the world. *Royal Society open science*, 7(10):201199.

Jerry Wei, Chengyu Huang, Soroush Vosoughi, and Jason Wei. 2020. What are people asking about covid-19? a question classification dataset. *arXiv preprint arXiv:2005.12522*.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Brian Xu, Mitra Mohtarami, and James Glass. 2018. Adversarial domain adaptation for stance detection. In *NeurIPS*.