A Dashboard for Mitigating the COVID-19 Misinfodemic

Zhengyuan Zhu, Kevin Meng, Josue Caraballo, Israa Jaradat
Xiao Shi, Zeyu Zhang, Farahnaz Akrami, Haojin Liao, Fatma Arslan
Damian Jimenez, Mohammed Samiul Saeef, Paras Pathak, Chengkai Li
University of Texas at Arlington

Abstract

This paper describes the current milestones achieved in our ongoing project that aims to understand the surveillance of, impact of and intervention on 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 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 developing pandemic of COVID-19, there is a raging global misinfodemic just as deadly. As fear grows, false information related to the virus goes viral on social media and threatens to affect an overwhelmed population. Such misinformation misleads the public on how the virus is transmitted, its symptoms, self-treatments and cures, how authorities and people are responding to the pandemic, and so on. This onslaught of misinformation exacerbates the vicious impact of the virus, as the misinformation drowns 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), now might be the first time with such a dangerous interplay between a pandemic and a misinfodemic. The severe urgency of this unprecedented combination of pandemic and misinfodemic calls for rapid response in studying not only the outbreak but also its related misinformation together. In other words, 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 intervention on COVID-19 misinfodemic on social media, particularly Twitter. 1) For *surveillance*, we seek to discover the patterns by which different types of COVID-19 misinformation spread, i.e., by whom, when, and in what context misinformation gets posted and re-posted. 2) To understand the *impact* of misinformation, we compare the spreading of the SARS-CoV-2 virus and misinformation and derive correlations between misinformation, infections, and people's social behaviors. 3) To understand what types of *interventions* are effective in containing misinformation, we will contrast the spreading of misinformation before and after debunking efforts are made by various governments and organizations. 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 to directly benefit the public using what we have accomplished so far. The dashboard is at https://idir.uta.edu/covid-19/ and is being constantly updated with more features toward the project's goals. A screencast video of the dashboard

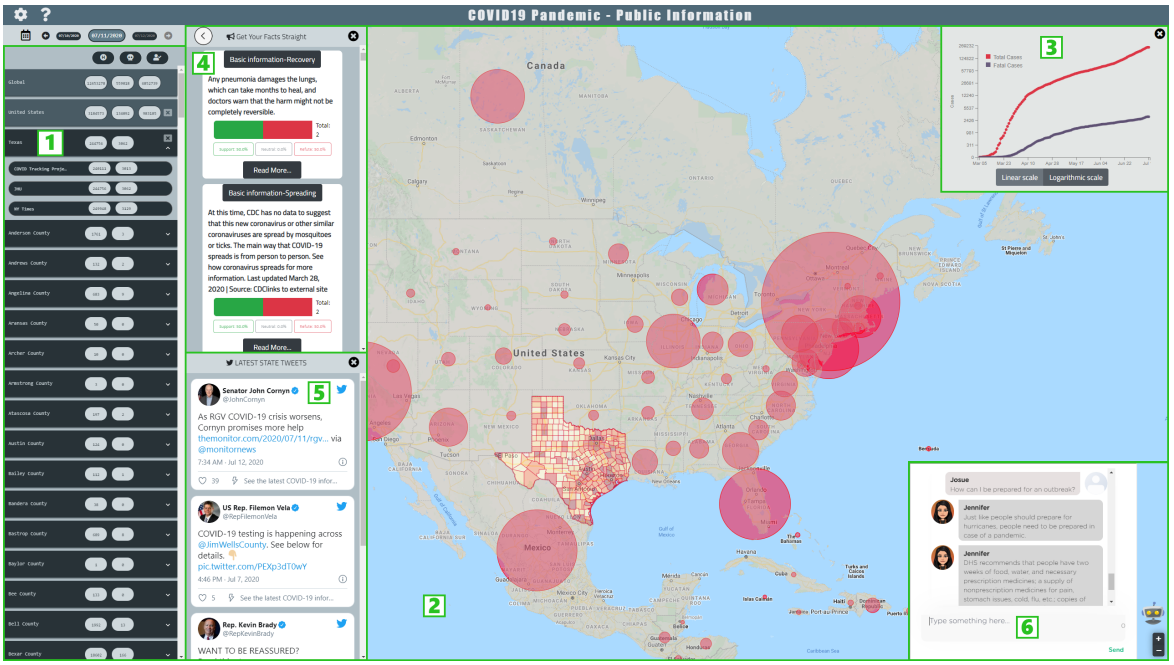

Figure 1: The user interface of the dashboard for mitigating the COVID-19 misinfodemic

is at https://vimeo.com/438102761. The codebase of the 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.

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 currently available. [1][2][3] However, our dashboard also provides several significant and unique features that are not found in other places. 1) The dashboard displays the most prevalent facts and debunks of misinformation among Twitter users in any user-selected U.S. geographic region. 2) The facts and debunks come from a catalog of COVID-19 information that we curated. 3) It shows case-statistics from several popular sources that do not always agree with each other. 4) It displays COVID-19 related tweets from local governments and authorities of user-selected geographic regions. 5) It also embeds a chatbot built specifically for COVID-19 related questions.

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

2 The Dashboard

Figure 1 shows the dashboard's user interface, with its components highlighted. A user can select a specific country, a U.S. state, or a U.S. county by using the geographic region selection panel (Component 1) or the interactive map (Component 2).

Geographic region selection panel (Component 1). 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,

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

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 and misinformation debunks (Component 4). For the selected region, this panel displays a few pieces of the most prevalent facts and misinformation among Twitter users in the region, as well as the distribution of supporting, neutral and refuting tweets for such facts and misinformation. Misinformation is displayed as debunks, to avoid repeating misconceptions.

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 posted by 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 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 on a daily basis from Johns Hopkins University, [4] the New York Times [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. We manually curated this catalog, which 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 from credible sources (e.g., FDA, NYT). The sources and numbers of facts are [(IFCN CoronaVirusFacts Alliance Database, 7480), (PolitiFact, 454), (Coronavirus.gov, 435), (WHO,307), (CDC, 295), (NYT, 119), (CNN, 65), (FDA, 60), (Global Health Now, 59), (ECDC, 47), (Washington Post, 35), (United Nations, 25), (COVID-19 FactCheck, 23), (World Vision, 20), (IDPH, 20), (FEMA, 17), (WJLA, 16), (News Guard, 13), (Johns Hopkins, 12), (Doctors Without Border, 6), (Johns Hopkins Medicine, 4)].

We categorized the entries in this catalog and organized them into a taxonomy of categories. This is done 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, Vac-*

[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.

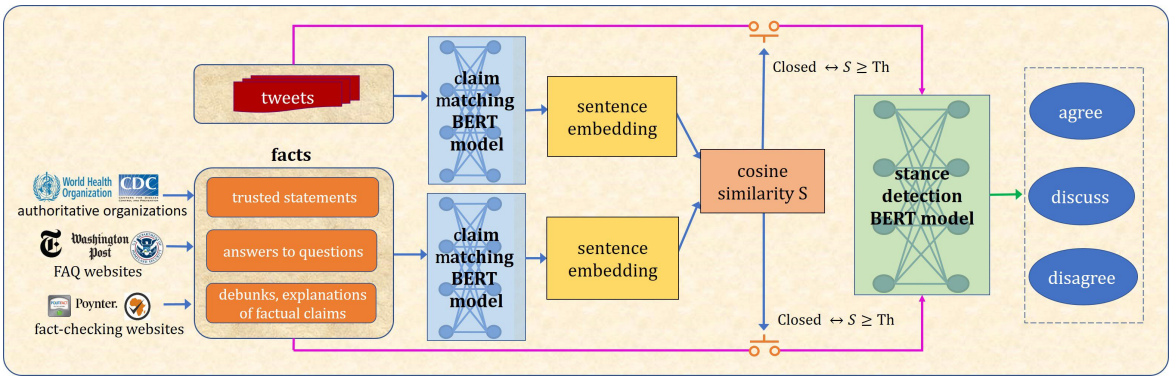

Figure 2: Matching tweets with facts and stance detection

Tweet	Fact	Taxonomy Category	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.	Spreading & Animals	0.817	agree
I think detergent would kill coronavirus but I wouldn't advise ingesting it.	Hand dryers are not effective in killing the coronavirus. You should frequently clean your hands with soap and water or an alcohol-based hand rub.	Prevention	0.848	disagree

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

cines], Religion, Schools/Universities, Travel, Treatment [*Medication, Minor Symptom, Severe Symptom*], Violence/Crime}.

We also stored the catalog and the taxonomy as an RDF (Cyganiak et al., 2014) dataset, to facilitate data sharing and querying. Each entry in 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, all connected to a mediator node. Each mediator node is connected through edge "type" to one or more of the lowest-level taxonomy nodes that the entry belongs to. The RDF of the catalog/taxonomy, 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 using OpenLink Virtuoso.[10] The endpoint is available at https://cokn.org/deliverables/7-covid19-kg/, together with a few sample SPARQL queries.

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

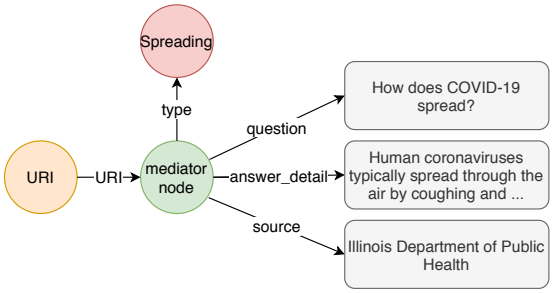

Figure 3: An example of a mediator node

4 Matching Tweets with Facts and Stance Detection

Given the catalog of COVID-related facts F and the dataset of 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 $\mathbf{p}^f = \{(t, f), \forall t \in \mathbf{t}^f\}$ to determine whether each t is supporting, refuting, or neutrally discussing f. Finally, aggregate results are displayed on Component 4 of the dashboard to summarize public view on each fact. Figure 2 depicts the overall claim-matching and stance detection pipeline. For both claim-matching and stance detection, we

employed Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019). Table 1 shows some example results of claim matching and stance detection.

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:

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

where t and f are considered related if $R^{t,f} \geq Th = 0.8$. Thus, the final output of this stage is $\mathbf{t}^f = \{t \in T \mid R^{t,f} \geq Th\}$ for each fact $f \in F$.

Stance detection. Given $\mathbf{t}^f$, we detect the stance that each tweet t takes toward each fact f. In stance detection, the relationship between t and f is categorized into one of 3 classes: 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] Denote this model Stance-BERT.

We first pre-process $\mathbf{p}^f$ to conform with BERT input conventions. Denote $W(\cdot)$ as the WordPiece tokenizer (Wu et al., 2016), $C(a_1, a_2, \ldots, a_n)$ as the concatenation of all arguments in appearance order, [CLS] as a BERT token denoting the start of an input, and [SEP] as a BERT token denoting the end of an input segment. 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 $\mathbf{x}^f$:

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

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

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

$$S^f_K = \frac{\sum_{\mathbf{w} \in \mathbf{x}^f} [\text{argmax}\, \sigma(\{\hat{\mathbf{y}}^f_{w_i}, \forall w_i \in \mathbf{w}\}) = K]}{|\mathbf{x}^f|} \quad (3)$$

where $\sigma(\cdot)$ is a function that returns an array in which each index i's value is the average of the value of all arguments' values at i. The 3 scores are finally passed to the dashboard's misinformation panel (Component 4) for display.

5 Evaluation and Results

5.1 Performance of Stance-BERT

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

Table 2: Performance of Stance-BERT

Table 2 shows Stance-BERT's performance on the FNC-1 competition test dataset. Evaluations are conducted using F1 scores for all 3 classes, as well as a macro-F1 score (the arithmetic mean of F1 scores across all classes). 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 all inputs after M tokens but is otherwise identical. Both versions of Stance-BERT significantly outperformed the method used in (Xu et al., 2018), one of the recent competitive methods on FNC-1. Because the sliding-window approach avoids discarding information from long inputs, it outperforms the truncation method.

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

[11] https://github.com/google-research/bert
[12] http://www.fakenewschallenge.org/

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.

5.2 Misinformation Analysis

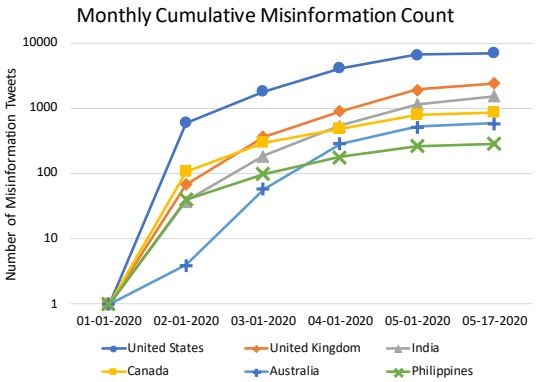

Figure 4: Top 6 countries with the most misinformation tweets.

Figure 4 is the cumulative timeline for the top six countries with the most COVID-19 misinformation tweets from January to May, 2020. In this context, "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 confirmed/deceased/recovered 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 product-moment correlation coefficients between them are shown in Table 3. 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.

Furthermore, we analyzed the categories of the misinformation tweets based on the taxonomy explained in Section 3. Table 4 lists the five most frequent categories of misinformation tweets: *Definition* - discussing the definition or properties of COVID-19; *Spreading* - describing the transmission of the virus; *Other* - a catch-all for the catalog entries that do not fit into other categories; *Testing* - describing methods, reliability, or other aspects of testing for COVID-19; and *Disease Alongside* - discussing possible diseases caused by or similar to COVID-19. These top five categories make up 49.9% of all misinformation tweets, with the other 50.1% being spread out over the other 33 categories.

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 3: 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.

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 4: The 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 collect more labeled data for improving and evaluating our Stance-BERT models. 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

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*.

Richard Cyganiak, David Wood, Markus Lanthaler, Graham Klyne, Jeremy J Carroll, and Brian McBride. 2014. Rdf 1.1 concepts and abstract syntax. *W3C recommendation*, 25(02).

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.

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.

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*.