Seeing Should Probably Not Be Believing: The Role of Deceptive Support in COVID-19 Misinformation on Twitter

CHAOYUAN ZUO and RITWIK BANERJEE, Stony Brook University, USA
FATEME HASHEMI CHALESHTORI, HOSSEIN SHIRAZI, and INDRAKSHI RAY,
Colorado State University, USA

With the spread of the SARS-CoV-2, enormous amounts of information about the pandemic are disseminated through social media platforms such as Twitter. Social media posts often leverage the trust readers have in prestigious news agencies and cite news articles as a way of gaining credibility. Nevertheless, it is not always the case that the cited article supports the claim made in the social media post. We present a cross-genre *ad hoc* pipeline to identify whether the information in a Twitter post (i.e., a "Tweet") is indeed supported by the cited news article. Our approach is empirically based on a corpus of over 46.86 million Tweets and is divided into two tasks: (i) development of models to detect Tweets containing claim and worth to be fact-checked and (ii) verifying whether the claims made in a Tweet are supported by the newswire article it cites. Unlike previous studies that detect unsubstantiated information by post hoc analysis of the patterns of propagation, we seek to identify reliable support (or the lack of it) *before* the misinformation begins to spread. We discover that nearly half of the Tweets (43.4%) are not factual and hence not worth checking—a significant filter, given the sheer volume of social media posts on a platform such as Twitter. Moreover, we find that among the Tweets that contain a seemingly factual claim while citing a news article as supporting evidence, at least 1% are not actually supported by the cited news and are hence misleading.

CCS Concepts: • **Computing methodologies → Natural language processing;**

Additional Key Words and Phrases: Misinformation detection, check-worthiness, COVID-19

ACM Reference format:
Chaoyuan Zuo, Ritwik Banerjee, Fateme Hashemi Chaleshtori, Hossein Shirazi, and Indrakshi Ray. 2022. Seeing Should Probably Not Be Believing: The Role of Deceptive Support in COVID-19 Misinformation on Twitter. *J. Data Information Quality* 15, 1, Article 9 (December 2022), 26 pages.
https://doi.org/10.1145/3546914

This work was supported in part by the U.S. National Science Foundation (NSF) under the awards IIS 2027750, CNS 1822118, and SES 1834597, and by NIST, ARL, Statnett, AMI, Cyber Risk Research, NewPush, and State of Colorado Cybersecurity Center.

Authors' addresses: C. Zuo, School of Journalism and Communication, Nankai University, Tianjin, Tianjin 300350, China; email: zuocy@nankai.edu.cn; R. Banerjee, Department of Computer Science, Stony Brook University, Stony Brook, New York 11794-2424, USA; email: rbanerjee@cs.stonybrook.edu; F. H. Chaleshtori, School of Computing, University of Utah, Salt Lake City, UT 84112-9249, USA; email: f.hashemichaleshtori@utah.edu; H. Shirazi, Management Information Systems Department, Fowler College of Business, San Diego State University, San Diego, California 92182-8230, USA; email: hshirazi@sdsu.edu; I. Ray, Department of Computer Science, Colorado State University, Fort Collins, Colorado 80523-1873, USA; email: iray@colostate.edu.

1936-1955/2022/12-ART9 $15.00
https://doi.org/10.1145/3546914

1 INTRODUCTION

The **World Health Organization (WHO)** defines a pandemic as "the worldwide spread of a new disease," and on March 11, 2020, it declared COVID-19 as one [51, 53]. Declaring a "pandemic" has the potential to trigger large-scale panic and fearmongering. Indeed, nearly a month before declaring COVID-19 a pandemic, the agency stated, "we're not just fighting an epidemic; we're fighting an infodemic," pointing to the deluge of misinformation and rumors particularly when trustworthy information was most needed [52, 76]. This can have devastating consequences: Individuals may take decisions based on falsehoods and social cohesion may be damaged by sowing distrust.

A recent study by Kouzy et al. [38] on more than 600 Tweets related to COVID-19 found that approximately 70% of the posts disseminated contained medical claims or public health information, but nearly 25% of them included misinformation, while another 107 (17.4%) propagated unverifiable information. The prevalence of misinformation abreast every major outbreak—Ebola [55], Zika [45], Yellow Fever [54], and now COVID-19—points to a pattern. Several studies have analyzed the dissemination of pandemic-related misinformation and rumor on social media (e.g., Reference [67]), but these analyses are *post hoc* and do not help with prevention. Once misinformation starts to spread, curtailing it is an uphill battle, especially since prior exposure to misinformation increases the chances that false information will be perceived as accurate [56]. This snowball effect leads to misinformation propagating much faster than accurate news [54, 64, 73]. There is, thus, a need for *timely* identification of misinformation on social media to stymie the spread of false claims. Early work in misinformation on social media often analyzed the dissemination patterns of false or unverifiable information in the network [34, 69], and some recent research has followed this approach for pandemic-related misinformation as well [67]. Others have focused on identifying topic-specific rumor-bearing posts [23]. In both approaches, the veracity of a specific nugget of information and its prior propagation in the network is requisite knowledge. Thus, they are not suitable for the *preemptive identification* of misinformation.

A social media post with original content is fundamentally different from one that includes a retransmission. Arif et al. [5] distinguish between them as "original" and "derivative" content. They report that when a claim enters the network with a large footprint, i.e., through a trusted account with a large number of followers, it spurs a greater volume of derivative content, which in turn creates a snowball effect. It is unlikely that ordinary users of social media deliberately believe and propagate misinformation. Instead, a claim gets propagated, because it is *perceived* as credible (as illustrated by Figure 1 and its propagation in Figure 2). When sharing information, users often cite trustworthy sources—including prestigious news agencies—to serve as markers of credibility [19]. While the accuracy and verification of information have long been held as a cornerstone of journalistic identity [65], there are no similar impositions on the commentary social media users may post while citing news articles. Such commentary may deviate from the claims made in the cited source, even to an extent that makes the source entirely irrelevant to the core of the commentary. Readers of such posts, however, often continue to rely on the credibility of the cited source and trust the claims in the commentary simply because the citation exists—the belief is born without a perusal of the original material. This may be due to homophily in social networks, where many are reading the commentary at least partly by reason of confirmation bias [16, 68]. Such posts are pernicious, especially because they spread misinformation by masquerading as trustworthy. This, of course, is what we would like to prevent. To this end, our work is geared toward (i) identifying posts that are presented as factual claims derived from trusted sources, carrying an information nugget worth verifying, and (ii) juxtaposing the information in the derived post against the cited source to check whether the propagated claim is supported or if the user has falsely imputed the information to that source.

Fig. 1. Original content entering Twitter through the "New York Post" institutional account. With 2.1M followers (Retrieved May 21, 2021), this has a large footprint.

Fig. 2. A corresponding derived content: remarks added along with source retransmission.

1.1 Problem Statement

For each COVID-19-related post that cites a news article, we pose two questions:

(1) Does the post include an objectively presented claim, i.e., a *factual claim*, and is that claim deemed important enough to check for veracity?
(2) Does the cited news article support the claim in the post?

We distinguish between *check-worthy* posts (which contain factual claims that are deemed important) and others—which are discarded from further analyses in this work. Next, we discriminate between derived content based on whether or not the post is faithful to the source. Posts that cite a news article but present claims unsupported by the source are candidates for misinformation.

1.2 Scope and Approach

Information propagated through social media can often be dissected along several dimensions. Imran et al. [32] categorize these dimensions in terms of time, location, topic, type of information, subjectivity (i.e., factual claims as opposed to opinions or other emotional content), information source, and credibility. Our work is unique, because we investigate "perceived credibility" in posts. We investigate whether or not the derived content is faithful to the original content, as it is retransmitted through the network. Further, we only consider Twitter posts (i.e., "Tweets") that

(A) pertain to the COVID-19 pandemic, thus restricting our dataset along the topic-dimension,
(B) contain factual claims, additionally controlling for the subjectivity-dimension,
(C) appear to provide support by citing a news article, which controls for the perception of credibility by providing an external information source, and
(D) are check-worthy, i.e., important enough (vis-à-vis their information content and their potential to snowball) to warrant an investigation into their veracity.

Tweets that voice opinions, share emotional content, or present factual claims without explicit external support to provide the perception of credibility are beyond the scope of this work.

(A) Controlling for the topic: We use a large dataset of COVID-19 Tweets, created by Banda et al. [7] to aid integrated research in epidemiology, misinformation, and related fields.

(B) Filtering subjectivity: A significant fraction of posts do not contain subjective information. For instance, Tweets often share personal anecdotes, contain emotional language, issue sarcastic

remarks, and so on. Our first step, therefore, is to distill Tweets that contain factual claims from the dataset.

(C) Controlling for perceived credibility: Not all posts that present a factual claim are readily credible. This perception is created by including a link to a news article in the Tweet, often along with statements made by the user who is creating the derived content from the original. Thus, we retain only those Tweets that contain a link to a news article. These links may be external to Twitter or introduced into Twitter through the institutional account of a news agency.

(D) Check-worthiness: Prior research in fake news detection has often ranked information nuggets in order of importance, especially in crises like natural disasters or epidemics (e.g., Reference [39]). This approach gave birth to a sizeable body of work on scoring information nuggets based on the check-worthiness [6, 28, 80]. Given the deluge of information available on the Internet, discriminating check-worthy information from the rest has become increasingly important in recent years. Consequently, we incorporate the identification of check-worthiness into this work as well and discard Tweets that are deemed unimportant.

The above steps form the first task of our entire pipeline. Its output—a dataset of factual check-worthy claims in the form of Tweets that link to news articles—becomes the input to our second task, where we identify whether or not a Tweet is, indeed, propagating a claim made in the cited news article. We use transformer-based models for the first task and then use the model that achieves the best performance to provide the input for the second.

We present the detailed architecture of our pipeline in Section 2 and the data preparation steps in Section 3. Then, in Sections 4 and 5, we present the two core steps of our pipeline where (i) check-worthy factual claims are identified and (ii) faithfully represented derived content is distinguished from potential misinformation and unverifiable claims. Subsequently, we discuss prior research in this field in Section 6 before concluding in Section 7.

2 ARCHITECTURE

We begin by conferring the basic requirements of a fake news detection algorithm, as discussed by Rubin et al. [62], and then present the primary components of the pipeline, responsible for (i) data collection, (ii) preprocessing, (iii) identifying check-worthy factual claims, and (iv) identifying verifiable claims.

2.1 Requirements

We take care to meet the nine fundamental criteria for fake news detection systems within the scope of **natural language processing (NLP)** research, originally proposed by Rubin et al. [62]:

(1) Our data satisfy the *availability of both truthful and deceptive instances*.
(2) It also satisfies *digital textual format accessibility*.
(3) It offers *verifiability of "ground truth"* by virtue of the manual annotation of two datasets with ground-truth labels. Our annotations offer high inter-annotator scores (details are discussed in the context of data preparation in Section 3 and experimental results in Section 4).
(4) Since we use Twitter posts, which are limited to 280 characters, our data adhere to *homogeneity in length*. Further, even though Twitter expanded its character count limit to 280 in November, 2017 [57], only 5% of the English language Tweets over the subsequent one year were longer than 190 characters, and only 9% used more than 140 [58], thus providing even more homogeneity in length than one would expect.
(5) Our work adheres to *homogeneity in writing matter*, in both topic (COVID-19 pandemic) and genre, and offers comparison across multiple news agencies and social media users.

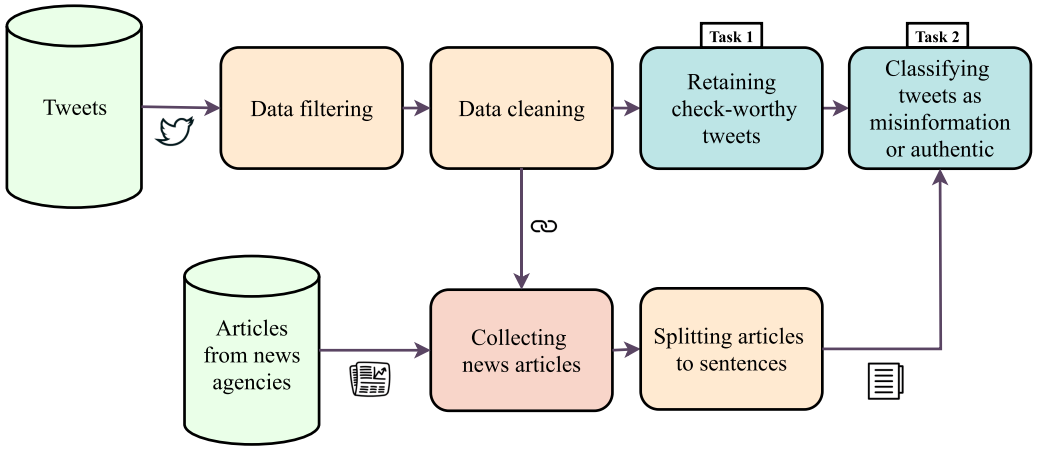


Fig. 3. System architecture. The pipeline comprises (i) the data collection from Twitter posts and news articles, (ii) data preprocessing (which includes the filtering, cleaning, and splitting into sentence-level chunks), (iii) the first task of identifying Tweets containing check-worthy factual claims, and (iv) the second task of distinguishing the information faithful to the original news content from the rest.

(6) The data were collected over a period of 3 months, during the COVID-19 pandemic, and therefore has a *predefined timeframe* of data collection, thereby reducing arbitrary variations that are typically present in corpora collected over shorter "snapshot" periods.
(7) We also control for *the manner of delivery* of the information, since we only consider posts that contain links to reputable news agencies and discard content derived from other kinds of user-generated content (e.g., blogs or other social media platforms).
(8) The corpus is created from publicly available data [7]. As such, it is not hindered by any of the *pragmatic concerns* cited by Rubin et al. [62].
(9) *Language and culture* are important factors affecting any NLP-based research, of course. Thus, we use only English-language Tweets in this work (although the approach can be applied to other languages, subject to availability of adequate volume of data in that language).

2.2 An Overview of the Components

Figure 3 shows the complete system architecture. Additionally, Table 1 illustrates (with examples) the correspondence between the data and the steps in the pipeline.

Data collection: We use an open dataset [7] as the starting point, to obtain the large collection of Tweets pertaining to the COVID-19 pandemic. In parallel, we also collect the complete news articles cited by the Tweets in this collection. The news articles are collected only for those Tweets that are retained after the data filtering step.

Data preprocessing: On the one hand, each Tweet is passed through multiple filters, token-level cleaning such as removal of function words, and non-linguistic features (discussed in greater detail in Section 3). On the other hand, the news articles cited by these Tweets are collected and processed as well, thereby removing spurious material around the article's content and then splitting the article's content and title into sentence-level chunks for subsequent use in our final task.

Task 1: Identification of check-worthy factual claims: This is designed as a supervised binary classification task, where each Tweet is designated as **check-worthy (CW)** or non-**check-worthy (NCW)**. We present the details of this component in Section 4.

Table 1. Sample Twitter Posts (Tweets) from Our Data

	Tweet (derived content)	Corresponding original content (cited news)
(1)	Africa deporting Europeans we love to see it [https://bit.ly/3vEtIyj] (Retrieved June 6, 2021)	*– no news cited –*
(2)	Coronavirus Map: How To Track Coronavirus Spread Across The Globe via @forbes [https://bit.ly/3upHDao] (Retrieved June 6, 2021)	**[H]** Coronavirus Map: How To Track Coronavirus Spread Across The Globe **[B]** As COVID-19 (coronavirus) spreads across the globe, it is helpful and interesting to track the transmission patterns through a coronavirus map
(3)	Native American Health Center Receives Body Bags Instead of Coronavirus Supplies. [https://bit.ly/39LBBJc] (Retrieved June 6, 2021)	**[H]** Native American health center receives body bags instead of coronavirus supplies **[B]** A community health center treating Native Americans in the Seattle area issued an urgent call for medical supplies …
(4)	Misinformation about Mr. Gates is now the most widespread of all coronavirus falsehoods –New York Times [https://nyti.ms/3fLCoO2] (Retrieved June 6, 2021)	**[H]** Bill Gates, at Odds With Trump on Virus, Becomes a Right-Wing Target **[B]** … Misinformation about Mr. Gates is now the most widespread of all coronavirus falsehoods …
(5)	Italy coronavirus: Italians who attempt to flee lockdown may face jail –CNN [https://cnn.it/3rVRZx8] (Retrieved June 6, 2021)	**[H]** All of Italy in lockdown as coronavirus cases rise **[B]** (CNN) Italy has been put under a dramatic total lockdown, as the coronavirus spreads in the country
(6)	Dow drops 200 points as unemployment claims surge once again via CNBC #news #CNBC [https://rb.gy/jxhy55] (Retrieved February 6, 2022)	**[H]** Stocks rise slightly, led by tech; Netflix hits record **[B]** Stocks rose slightly on Thursday, led by tech, as Wall Street grappled
(7)	Federal officials accuse two groups of selling fake coronavirus vaccines and treatment –CNN [https://cnn.it/3eewwck] (Retrieved February 6, 2022)	**[H]** Memorial Day weekend: Americans visit beaches and attractions with pandemic warnings in mind **[B]** The country has started a most unusual kind of Memorial Day weekend.

Tweets often cite news articles to lend credibility to the shared information: (1) a post not containing terms related to COVID-19 or a link to a news article; (2) a post without any specific check-worthy claim; (3) a statement worth checking vis-à-vis the headline (**[H]**) of the linked news article; (4) a statement worth checking vis-à-vis the body (**[B]**) of the linked news article; and ((5), (6), and (7)) a check-worthy claim that is not supported by the cited article, thus merely *appearing* trustworthy.

Task 2: Identifying whether the derived content in the Tweet is faithful to the original content in the cited news: Among the multiple models developed for the first task, we use the one with the best performance to feed Tweets with the CW label into the second task. This, too, is designed as binary classification. Multiple models and experimental setups are explored and discussed in Section 5.

3 DATA PREPARATION

In this section, we provide the details of the primary Twitter dataset used as the starting point of our pipeline, the data filtering steps to retain only relevant posts, the preprocessing done to clean the natural language data on which we conduct the classification experiments, and our own additional data collection of newswire articles.

Table 2. COVID-19 Keywords: The 52 Keywords Used to Filter Tweets

case, CDC, China, corona, covid, crisis, die, disease, distancing, drug, economy, emergency, Fauci, global, government, hands, health, hospital, immune, infected, kill, lab, lockdown, mask, medical, medicine, news, NHS, nursing, outbreak, pandemic, panic, patient, prevent, public, quarantine, recovery, restrictions, risk, safe, sick, social, spread, stock, symptoms, test, treatment, vaccine, virus, wash, watching, Wuhan

3.1 Data Filtering

Our pipeline begins by leveraging a large open dataset of Tweets related to COVID-19, developed and made available by Banda et al. [7]. This is a continually growing collection, and at the time of this work, it offered 383M Tweets collected from January through June 2020. Our work utilizes a subset (46.86M Tweets gathered from March to May) of this large collection. Even though this dataset is related to COVID-19, it is not immediately suitable for our tasks. Thus, we inject significant additional filtering and data cleaning steps:

Retweets: Re-posting of a Tweet is intended to facilitate quick sharing and re-transmission of information in the network. The original large dataset includes Retweets, which are often derived content, but with no additional information or commentary. While this may be useful for analyses of information propagation, it has no utility in our study. Thus, we remove all retweets.

Non-English Tweets: Controlling for language is an important requirement [62] (see Section 2). The dataset, however, includes Tweets from multiple languages. Therefore, we discard non-English posts.[1]

Tweets not containing topic-specific keywords: Compared to the original dataset, we impose a stricter condition to establish relevance of each post to the COVID-19 pandemic. We do this by using a set of 52 keywords and retain only those Tweets that contain at least one of these keywords. This set, shown in Table 2, was created by removing all function words[2] as provided by the English-language list of function words in the Python Natural Language Toolkit [10], sorting the remaining words by frequency, and then manually selecting from the most frequent entries. The Tweets collected by Banda et al. [7] include responses to other posts. Often, a response by itself has no content relevant to COVID-19, even if it were relevant in the context of the original Tweet. Most common examples include emotive expressions of sorrow, faith, hope, anger, or sarcasm.

Tweets without a link to a news agency of repute: Our work focuses on identifying instances where the original content (the cited news article) belies that claim made in the derived content (the Tweet). Thus, we further restrict our attention to Tweets that include a link to a news article. To this end, we check whether the external link from a Tweet is to a top English-language news website in the Alexa website ranking.[3] Table 3 shows the list of these news agency domains. Tweets with no external link to one of these domains are removed from our study.

3.2 Data Preprocessing

After applying the filters described above, we retain over 246k Tweets and prepare them for the subsequent NLP components of our pipeline by adding a few preprocessing steps. Some of these are standard domain-nonspecific practice in NLP research, while the others are particularly meant for the social media landscape.

[1]The Twitter API provides many properties based on a Tweet's ID (known as *hydration*), including the language used.
[2]Function words are words that play an important role in syntactic correctness of a sentence but offer little semantic content. They comprise determiners, pronouns, prepositions, and conjunctions (e.g., "the," "and," "his," "she," and "although").
[3]https://www.alexa.com/topsites/category/Top/News (this service was last available on September 17, 2020).

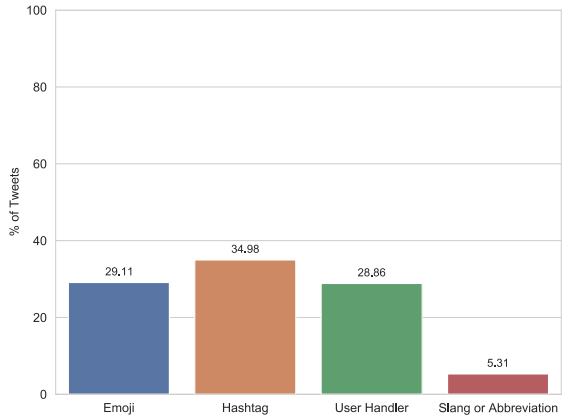


Fig. 4. Percentage (after filtering) of Tweets using various types of informal register.

Table 3. List of News Agencies Used as Original Content

List of new agencies we verified Tweets
Reuters, The Guardian, The Wall Street Journal, Washington Post, The New York Times, CNN, CNBC, CBS, New York Post, Fox News, USA Today, The Atlantic, SFGATE, Los Angeles Times, The Hollywood Reporter, BBC, The Hill, Chicago Tribune, U.S. News & World Report, The Daily Beast, Houston Chronicle, Time, NBC News, Deutsche Welle, Variety, Euronews

News agencies in the top-50 English-language news sources, as ranked by Alexa Website Ranking. In this work, we remove some domains from the original list due to paywall models, difficulty of data crawling, or topic/genre-specificity (e.g., weather news). The remaining 27 domains are shown here.

First, we remove non-linguistic tokens (i.e., non-words) in each Tweet. This comprises a removal of punctuation, URLs, and Twitter user handles. Links to the relevant news agencies (shown in Table 3) are decoupled from the post and maintained separately. Twitter extensively uses hashtags too. We remove the hash symbol, but retain the term. For example, "#quarantine" and "#staysafe" are converted to "quarantine" and "staysafe," respectively. Social media users frequently depart from dictionary-based lexicon and make ample use of informal register (see Figure 4). Most commonly, this includes emojis and colloquial non-standard abbreviations and misspellings that have become socially accepted. One may argue that emojis convey linguistic information (albeit not in the traditional sense) and thus, removing them alters the content in a post. We therefore use the `demoji` library[4] to replace each emoji with its corresponding text form. Abbreviations, especially if non-standard, are seldom handled well by readily available NLP tools (e.g., a syntactic parser), and may not even have a meaningful representation in language models unless the model was trained on large amounts of data containing these tokens. The same holds true for misspellings that have recently gained social acceptance on a platform. Therefore, we use a list of more than 5,700 such terms[5] and replace them with their formal register counterparts. This results in abbreviations like "wru" being converted to "where are you" and misspellings such as "wutevr" being replaced by "whatever." Finally, we observe that some Tweets are duplicated in the dataset, so we remove the spurious copies and retain only one.

[4]Available at pypi.org/project/demoji.
[5]Gathered from www.noslang.com/dictionary.

3.3 Newswire Data Collection

As mentioned earlier, this work investigates whether original claims found in news articles are faithfully reproduced in a Tweet. This is the reason behind discarding Tweets that do not contain a link to a news agency of repute (see Section 3.1). The data obtained from Banda et al. [7] do not contain this external information, however. Therefore, we collect the newswire articles linked from the Tweets. For this data collection, we use the `Newspaper3k` library.[6] Some articles could not be collected due to paywall restrictions, leading to a final corpus of 46,117 Tweets together with 23,841 unique newswire articles from the 27 news agency domains shown in Table 3. The number of unique articles is understandably lower, since multiple Tweets often propagate the same article published by widely known news agencies. For each newswire article, we retain its full text as well as the headline. Images, videos, and metadata are discarded. Subsequently, the articles are tokenized and split into individual sentences (using Reference [10]).

4 TASK 1: IDENTIFICATION OF CHECK-WORTHY TWEETS

After all the filtering and data cleaning steps have been taken, the first component of our pipeline is the identification and retention of check-worthy Tweets (see Figure 3). This is a precursor to the final objective, because social media posts do not always contain check-worthy factual claims. It thus behooves us to decouple this task from the final analysis of faithful representation and propagation of information. We design it as supervised binary classification, where each Tweet is given one of two possible labels: CW or NCW.

Classical supervised learning consists of training followed by evaluation on a test dataset. With the advent of Transformer-based deep learning models [72], however, supervised learning in NLP research is now often divided into (i) the use of embeddings that have been pretrained on a large corpus, thus yielding a pretrained **language model (LM)**, and (ii) tuning the embedded representations for a specific task. This is the approach we adopt as well. To this end, we experiment with multiple pretrained LMs, each with task-specific tuning. In the remainder of this section, we first present a short discussion of the pretrained LMs, followed by the datasets on which they are further tuned, before discussing the results.

4.1 Pretrained Language Models

We use 10 models pretrained on general data, plus two with domain-specific pretraining. They all use the Transformer architecture to learn contextual word representations, known as **Bidirectional Encoder Representations from Transformers (BERT)** [17]. BERT is pretrained on two NLP tasks, *viz.*, (i) **masked language modeling (MLM)** (where some input tokens are replaced with [MASK] and the model is trained to reconstruct the original tokens) and (ii) next sentence prediction (where the model is trained to understand whether one sentence can logically follow another). There are two variants (Base and Large), which differ in the size of the network used for training. BERT demonstrated state-of-the-art performance on multiple downstream language understanding tasks on benchmark datasets and inspired variations, including the following:

(1) DistilBERT [63], which pretrains a smaller general-purpose language model while providing comparable performance on the NLU benchmarks.
(2) RoBERTa [41], which discards the next sentence understanding task from pretraining but uses additional corpora. While the original BERT was pretrained on approximately 16 GB of unlabeled plain text data, RoBERTa used over 160 GB and achieved improved performance on several NLU benchmarks.

[6] github.com/codelucas/newspaper.

Table 4. Summary Statistics of the Three Collections Used for Supervised Learning in Task 1

	Dataset	Size			Description
		CW	NCW	Total	
DS1	Barrón-Cedeño et al. [2020]	231 (34.4%)	441 (65.6%)	672	COVID-19 Tweets
DS2	Hassan et al. [2017]	5,413 (24.06%)	17,088 (75.94%)	22,501	U.S. Presidential debates
DS3	*This article* [2021]	55 (55%)	45 (45%)	100	COVID-19 Tweets

(3) COVID-Twitter-BERT [48], two models pretrained on COVID-19 Tweets (CT-BERT-v1 and v2), the latter being pretrained on a much larger collection of 97 million Tweets.

A closely related model is ELECTRA [12], which is Transformer based, but instead of MLM uses a discriminative approach where some input tokens are intentionally replaced. The model is then trained to identify the replaced tokens. When pretrained using comparable amounts of data and similar model sizes, ELECTRA outperforms the original BERT models on various NLU benchmarks.

Yet another set of state-of-the-art NLU results were achieved by XLNet [75], which uses generalized autoregressive pretraining (in contrast to BERT's use of denoising autoencoder) to capture bidirectionality in a token's linguistic context. Moreover, it uses Transformer-XL [14] to overcome some restrictions of the original Transformer architecture (e.g., fixed-length context).

We use the multiple versions of BERT, DistilBERT, RoBERTa, CT-BERT, ELECTRA, and XLNet, giving us 12 pretrained models altogether. Next, we discuss their tuning for our first task.

4.2 Ground-truth Data for Model Tuning

Prior research on identification of fake news, while different from the investigation in this work, provides several noteworthy datasets that can be leveraged for supervised learning in this first task in our pipeline. In particular, we use three corpora under the monikers DS1, DS2, and DS3. Their basic statistics are shown in Table 4.

DS1: As the amount of information available on the Internet grew, so did the amount of false information. Realizing that human participation in fact-checking is likely to remain necessary in the foreseeable future, Barrón-Cedeño et al. [8] designed a shared task for fact-checking in social media, where the first step was to rank information nuggets based on their "check-worthiness." The dataset does, however, provide binary ground-truth labels for check-worthiness and can thus be directly used for supervision in our task.

DS2: The second dataset we use to supervise our classifiers is the well-known *ClaimBuster* corpus [28]. This collection provides three ground-truth labels for each datum: (i) check-worthy factual sentences, which present a factual claim whose authenticity is of interest to the general public; (ii) unimportant factual sentences, which contain factual claims but the claims are deemed to be not of interest to the general public; and (iii) non-factual sentences, which do not contain factual claims but instead consist of opinions, beliefs, questions, or other subjective content. We use the first category as CW and coalesce the remaining two into NCW.

DS3: We manually annotate 100 randomly selected Tweets from the corpus created based on the dataset available from Reference [7]. Three annotators carry out this task, and thus, each Tweet was assigned a CW or NCW label by each annotator independently. To measure the consensus on check-worthiness, we use Fleiss' kappa [18] (a measure of inter-rater reliability), but unlike the more commonly used Cohen's kappa, this can be applied in scenarios with more than two raters. We achieve $\kappa = 0.822$, indicating that the annotators are in near-perfect agreement [61]. There were disagreements only on 13 Tweets, where one of three annotators disagreed with the other two. In these cases, we used majority voting to assign the final label.

Table 5. Performance on Task 1: Identification of Check-worthy Tweets

Model	DS1				DS2				DS2 + DS1			
	P	**R**	**F_1**	**TP**	**P**	**R**	**F_1**	**TP**	**P**	**R**	**F_1**	**TP**
BERT-Base	57.6	89.1	70.0	49	86.5	58.2	69.6	32	86.8	60.0	71.0	33
BERT-Large	57.3	100	72.8	55	90.9	36.4	51.9	20	82.4	50.9	62.9	28
RoBERTa-Base	55.6	100	71.4	55	77.8	76.4	77.1‡	42	79.6	70.9	75.0‡	39
RoBERTa-Large	55.6	100	71.4	55	79.6	70.9	75.0‡	39	80.0	58.2	67.4	32
DistilBERT-Base	69.7	41.8	52.3	23	75.9	80.0	77.9‡	44	77.2	80.0	78.6‡	44
CT-BERT-v1	57.8	94.5	71.7	31	84.1	67.3	74.7	37	78.0	38.0	51.0	39
CT-BERT-v2	68.4	47.3	55.9	26	85.7	10.9	19.4	6	79.3	41.8	54.8	23
Electra-Base	56.4	96.4	71.1	53	88.5	41.8	56.8	23	85.7	43.6	57.8	24
Electra-Small	57.5	76.4	65.6	42	70.2	60.0	64.7	33	71.0	62.1	66.3	22
Electra-Large	62.2	92.7	74.5	51	80.0	43.6	56.5	24	81.6	56.4	66.7	31
XLNet-Base	87.8	65.5	75.0‡	19	88.0	64.5	74.4	36	84.4	69.1	76.0‡	38
XLNet-Large	58.1	65.5	61.5	36	84.4	49.1	62.1	27	78.4	72.7	75.5‡	40

The classification results on 12 models, each fine-tuned on DS1, DS2, and both. The evaluation is done on DS3, showing the **P**recision, **R**ecall, $\mathbf{F}_1$ score, and the number of true positives (**TP**) of the 55 check-worthy elements in DS3. Models considered as candidates for providing input to our second task are marked by ‡. XLNet-Base, shown in bold, is the pretrained model that achieves (upon fine-tuning) the highest precision among the candidates.

4.3 Experiments and Results

Our experiments for the first task are categorized based on the pretrained model and the corpus on which that model was tuned. Thus, each experiment can be represented as a ⟨model,dataset⟩ pair. We conduct three sets of experiments, where each model is tuned (i) on the COVID-19 Tweets corpus (DS1), (ii) on ClaimBuster (DS2), and (iii) on both corpora, tuning first on ClaimBuster and then on COVID-19 Tweets (DS2+DS1). We then evaluate each ⟨model,dataset⟩ pair on the manually annotated sample, DS3. The results are shown above in Table 5.

Since this first task in our pipeline is meant to feed check-worthy Tweets as input to the second task, the immediate and natural step is to select the "best" tuned model. Unfortunately, no single ⟨model,dataset⟩ pair achieves a clearly superior performance across the three standard metrics of precision, recall, and F_1 score. As lower precision means a greater number of falsely labeled cw Tweets will enter the second task, it is clear that we need to prioritize a high-precision model even at the expense of potentially lower recall. However, extremely low recall will quite likely cause the second task to receive inadequate amount of input data and therefore build a less robust model. We thus use a threshold F_1 score of 75 to remove some models from further consideration. Among the remaining (shown in Table 5 with ‡), ⟨XLNet-Base, DS1⟩ and ⟨XLNet-Base, DS2+DS1⟩ achieve the best precision. However, due to the extremely low recall of the former, we move forward to the second task with XLNet-Base tuned on DS2+DS1 as our choice.

5 TASK 2: NEWS VERIFICATION

Of the 46,117 Tweets retained after the filtering and preprocessing steps described in Section 3, the ⟨XLNet-Base, DS2+DS1⟩ model (described above in Section 4) feeds 39,458 Tweets into the second NLP component in our pipeline. Here, our goal is to identify whether or not the claim made in a Tweet containing a link to a news article is *actually* supported by the cited article.

5.1 Design and Setup of Experiments

The Tweets that reach this second task have already been labeled as check-worthy by the best-performing classifier in the previous step. We add another filter, however: removing Tweets that

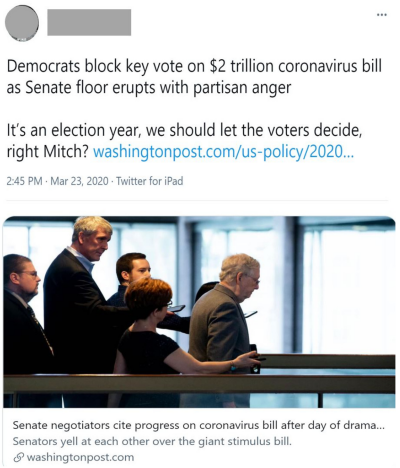


Fig. 5. A Tweet comprising multiple sentences: The first is objective and contains a check-worthy factual claim, while the second does not.

consist of multiple sentences. This is done to remove the noise of lengthy posts where one sentence may have a check-worthy factual claim, thus justifying the CW label, but the other sentences may be subjective opinions or expressions of sentiment, sarcasm, humor, and so on. Figure 5 presents such an example, where a check-worthy factual claim is followed by a possibly sarcastic question posed by the person sharing the piece of information. This filtration reduces the corpus size to 29,392 Tweets. We keep 11,800 Tweets for training, 12,335 for validation and hyperparameter tuning, and 5,257 for testing.

We observe that Tweets are often a near-verbatim reproduction of the news headline. Indeed, approximately 54% of all the Tweets provided as input to our second task fall into this category. The remaining cases, however, require a deeper understanding of the body of the news article to determine if the claim made in the Tweet is supported by the cited article. Thus, we further divide the second task into two steps where we consider (i) only the headline of the cited news article and (ii) the entire body of the article. The complete flowchart for this task is shown in Figure 6.

5.1.1 Distant Supervision. For both steps, the initial challenge is to obtain sufficient labeled data for training any supervised learning algorithm. We address this by *distant supervision*, an approach originally motivated by the use of *weakly labeled data* in bioinformatics [13]. In this approach, an assumption is made about the unlabeled data obtained or extracted from a corpus. Its success in learning relations from natural language, for instance, relied on a relation-triple $\langle \text{entity}_1, \text{entity}_2, \text{relation} \rangle$ being obtained from the Freebase corpus and *assuming* that any sentence mentioning the two entities express their relation in some way [46]. Similarly, the presence of specific emoticons and keywords has been used to obtain large amounts of distantly supervised Tweets for sentiment classification and topic identification [15, 43]. In our work, the assumption made for distant supervision is that if a news article is hyperlinked by a Tweet, then the article supports the claim made in the Tweet. In the absence of such a hyperlink, the ⟨Tweet, news⟩ pair is marked as unsupported. Our collection, by design, would yield only positive labels according to the above assumption of distant supervision. Thus, all ⟨Tweet, news⟩ pairs in the training set are given the weak label of "supported." We then create ⟨Tweet, news⟩ pairs by coupling each Tweet in the training set with an arbitrary but different headline from the collection of news articles. These pairs are given the weak label of "unsupported," thus forming the negative sample. This

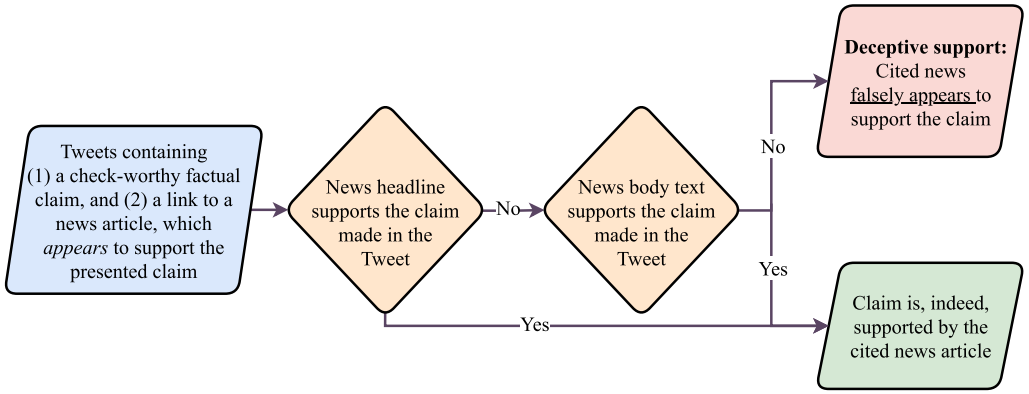


Fig. 6. Information verification in Task 2: The input comprises Tweets containing check-worthy factual claims that offer a news article as supporting evidence for that claim. The output is a binary decision about whether the support is deceptive.

strategy of creating negative samples by random pairing has shown promise in prior work on fact-checking [26, 50]. We use this same method to generate positive and negative weak labels for the validation set as well. This weakly labeled corpus of ⟨tweet, headline⟩ pairs is utilized in the first step (shown in Figure 6). For the second step, we build a weakly labeled corpus of ⟨Tweet, article⟩ pairs using the same method, where each Tweet is paired with the entirety (i.e., the headline plus the body) of a news article.

5.1.2 Step 1: Determining Support from the Cited Headline. For this first step, we use five pretrained language models (the base version when applicable): BERT [17], CT-BERT-v2 [48], XLNet [75], RoBERTa [41], and DistilRoBERTa [63]. We described the first four models earlier (Section 4). The last model, DistilRoBERTa, is a lighter version of RoBERTa, pretrained on a smaller general-purpose model. Additionally, we also use DistilRoBERTa trained on a large paraphrase dataset (henceforth denoted by DistilRoBERTa^p), which has been shown to achieve state-of-the-art performance on multiple tasks on semantic similarity. Our inclusion of this additional model is motivated by prior studies corroborating that a claim and its supporting evidence are bound to have relatively high semantic similarity [3, 47]. We tune all models on the ⟨tweet, headline⟩ weakly labeled collection.

5.1.3 Step 2: Determining Support from the Cited Article's Text. When a news article presents a factual claim, there may exist a single sentence in the article from which this claim follows. It is, however, also possible that the claim can only be gleaned from multiple sentences in the article. We thus follow a two-pronged strategy to determine support. On the one hand, we split the body of the article into a sequence of sentences and pair each sentence with the Tweet citing this article. Each such ⟨Tweet, sentence⟩ pair is then provided to the classifiers used in the first step (Section 5.1.2), since the data are structurally identical to that used in determining support from the cited headline. If any pair created from the article is labeled as "supported," then the ⟨Tweet, article⟩ pair is deemed "supported." Otherwise, it is deemed "unsupported." On the other hand, we also investigate ⟨Tweet, article⟩ pairs directly, without any sentence-splitting of the text. The same models are used again, except for DistilRoBERTa^p, which is not designed for long token sequences. To account for longer texts, we use Longformer [9] instead, which combines local windowed attention and global attention, thus allowing it to process sequences of thousands of tokens. Indeed, compared to RoBERTa, it has demonstrated superior performance on long-document tasks.

5.1.4 Technical Runtime Setup. All our experiments are conducted on NVIDIA Tesla V100 GPUs. We train every model for 1 and 2 epochs, with batch sizes of 16 and 24, and a learning rate set to 5×10^{-5}. For the first step, where only the news headline is paired with the Tweet, we set the maximum sequence length to be 128, and for the second step, we set it to 512. The only exception to this being Longformer, where the maximum sequence length is 4,096.

5.2 Evaluation, Results, and Discussion

On the validation set, all models achieve an $F1$ score of nearly 0.98, whether they classified ⟨Tweet, headline⟩ pairs or ⟨Tweet, article⟩ pairs. Given that our *weak labeling* builds the negative samples by combining a Tweet with a randomly selected different news article, the extremely high score is not unexpected, as discussed by Zuo et al. [79]. A more important point, arguably, concerns the false negatives of these models. In contrast to a standard supervised learning setup, these pairs are only *weakly false* negatives. That is, the Tweet does provide a link to a news article, but the model predicts the claim to be unsupported by the news article's headline. These pairs are the most likely candidates for deceptive hyperlinks, i.e., the cited news does not actually support the claim being made by the social media post. At the very least, these are the candidates for which the support is not obvious from the news headline alone. Thus, we collect these *weakly false* negative ⟨Tweet, headline⟩ pairs and feed them to the second step where the classifiers investigate entire articles.

5.2.1 Sample Annotation. Since this is a downstream task, some errors from the previous component are likely to pass through. Thus, before starting the second step, we analyze these weakly false negative pairs by performing another annotation task. The number of such pairs varies from one model to another, and the first step yields a total of 258 of them. Three annotators work independently on this collection, each answering the following:

(1) *Is the given Tweet check-worthy?* The annotators answer this question on the basis of the same guidelines provided to them during the first task.
(2) *If the Tweet is check-worthy, then does the cited article support the Tweet?* Each annotator peruses the entire article vis-à-vis the Tweet, and determines whether any information provided in the article supports the claim made in the Tweet. Accordingly, they assign one of two labels to the pair: *supported* or *unsupported*.

Of the 258 pairs, 51 were labeled as *not check-worthy* by at least two annotators. We discard these from the evaluation of the second step. Further, there were disagreements on 7 other Tweets, which we discard as well. Of the remaining 200 pairs, 55 were labeled as *unsupported* by at least two annotators. This annotation process showed substantial agreement among the three members, yielding a Fleiss' kappa score of $\kappa = 0.756$. Our inspection finds two main reasons for the disagreements. First, it is due to differing opinions on expressions of causality in human language. For instance, a Tweet announced "Dow drops 200 points as unemployment claims surge once again," while the corresponding news article mentioned the two events "Dow drops" and "unemployment claims surge" in separate paragraphs. For some readers, this is an indication of causality but no explicit mention of a causal relation between the two. A second reason is a difference among the annotators regarding the inclusion of metadata in the verification process, going beyond the purely linguistic expression of a claim. For example, a Tweet states "Yesterday more than 2K in the US died of coronavirus," where the dates of the post and the news article are, clearly, relevant.

Of the 200 manual annotations discussed above, 55 are labeled as deceptive (i.e., 27.5%). This, however, is sampled from the test of approximately 5,000 Tweets. Thus, our test data show that *at least* 55 of 5,000 Tweets (i.e., 1%) contain deceptive hyperlinks. In Figure 7, the number of ⟨Tweet, headline⟩ pairs predicted to be *unsupported* by the models are shown after the removal of erroneous

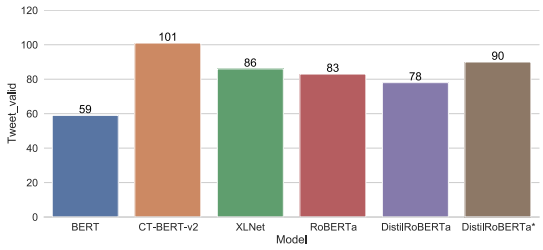


Fig. 7. *Weakly false* negative pairs for each model: check-worthy factual claims in tweets that cite a news article as external support, but the model labels them as *unsupported*, based on the ⟨tweet, headline⟩ pair.

Table 6. Experiment Results

	Step 1				*Step 2*										*Pipeline*					
					Sentence					*Full News*					*Sentence*			*Full News*		
Transformer	**P**	**R**	**F1**	**U†**	**P**	**R**	**F1**	**U$^{\#}$**	**TN$^{\#}$**	**P**	**R**	**F1**	**U$^{\#}$**	**TN$^{\#}$**	**U$^{\#}$**	**TN$^{\#}$**	**TN**	**U$^{\#}$**	**TN$^{\#}$**	**TN**
BERT	47.2	47.3	47.2	59	56.0	81.3	53.8	8	7	45.4	53.1	49.0	47	25	7	6	85.7	31	18	58.1
CT-BERT-v2	38.9	41.2	37.9	101	55.5	60.4	55.0	24	11	56.3	44.3	49.6	70	31	20	10	50	54	26	48.1
XLNet	**50.4**	**50.4**	**49.1**	86	**59.6**	**84.1**	**59.6**	12	11	45.4	**73.5**	56.1	34	25	9	8	88.9	26	18	69.2
RoBERTa	46.4	47.1	45.7	83	58.4	79.6	57.8	12	10	58.1	61.5	59.8	52	32	11	9	81.8	34	21	61.8
DistilRoBERTa	44.4	45.3	44.3	78	54.7	74.7	51.9	8	6	**67.8**	72.7	**69.4**	39	25	8	6	75	32	20	62.5
DistilRoBERTa*	49.1	49.2	47.5	90	53.6	86.9	49.3	4	4	—	—	—	—	—	4	4	100	—	—	—
Longformer	—	—	—	—	—	—	—	—	—	49.0	52.9	50.9	51	27	—	—	—	38	21	55.3

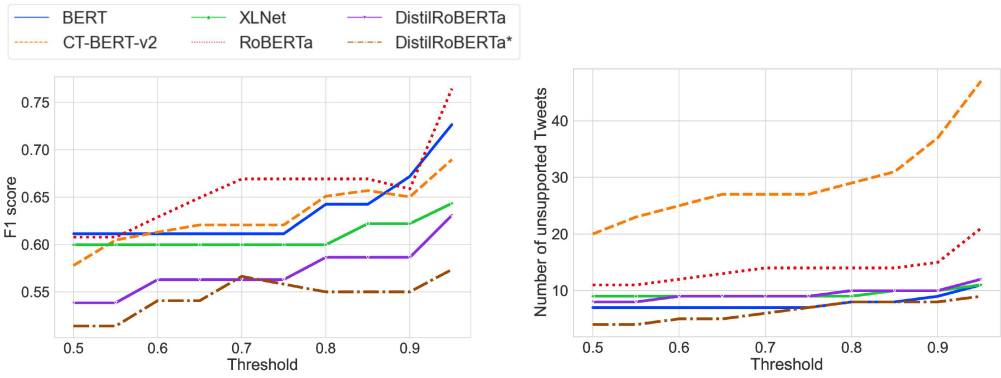
Model tuned on the paraphrase dataset marked with *. The number of check-worthy pairs labeled *unsupported* in step 1 are shown as U†. The numbers of *unsupported* are shown as U$^{\#}$. The number of pairs that are labeled *unsupported* by the model and indeed *unsupported* by annotation is shown as TN$^{\#}$. The ratio of truly unsupported claims to predicted unsupported claims is shown as TN. The best results for each step are shown in bold.

samples propagated by Task 1 (i.e., claims that are not check-worthy). Also, Table 1 includes three Tweets with deceptive hyperlinks, each citing a news article from a well-known news agency. However, the news article does not support the Tweet, as shown with examples (5) and (6) in Table 1 or is even irrelevant (see Table 1 example (7)).

5.2.2 Evaluation and Discussion. The performance of each model is evaluated on the 200 annotated pairs, with the annotation labels serving as the ground truth. For both steps of Task 2, we measure the performances using macro-average precision, recall, and F_1 score. Given the class imbalance, where only a minority of the samples offer deceptive support to the reader, macro-average associates more value to the minority class by disregarding the overwhelming effect of the majority class.

For step 2, we provide two ways of evaluating each model. First, we feed all 200 annotated samples into Step 2. That is, the entirety of the news articles are checked by the sentence-level models tuned on ⟨tweet, headline⟩ pairs, as well as the article-level models tuned on ⟨Tweet, article⟩ pairs. This evaluation is effectively an ablation study to understand how well our system can detect deceptive cues of support, in the absence of a separate first step in Task 2. Second, we follow the pipeline approach shown in Figure 6 and provide only the check-worthy *weakly false* negative samples from step 1 into step 2. For example, BERT labels 59 check-worthy ⟨tweet, headline⟩ pairs as *unsupported*, and we evaluate BERT in step 2 using only these 59 pairs. Since we Longformer only in step 2, for this evaluation we use the results of DistilRoBERTa^{p} from step 1.

Table 6 shows the comprehensive results of our evaluation of the second task. In the first step, where only the ⟨tweet, headline⟩ pairs are used, CT-BERT-v2 provides the worst performance. It labels the highest number of pairs as *unsupported*, which leads to low precision. But it achieves the lowest recall as well. This is perhaps not surprising, given that our task spans two genres: social media and newswire, while CT-BERT is armed with domain-specific pre-training only on Twitter. It may thus be ill equipped to understand the lexical context of words in newswire sentences.



(a) F1 score (b) Number of unsupported Tweets

Fig. 8. Varying threshold and results. The results under different thresholds in step 2 as a sentence-level pipeline. Model tuned on the paraphrase dataset marked with *.

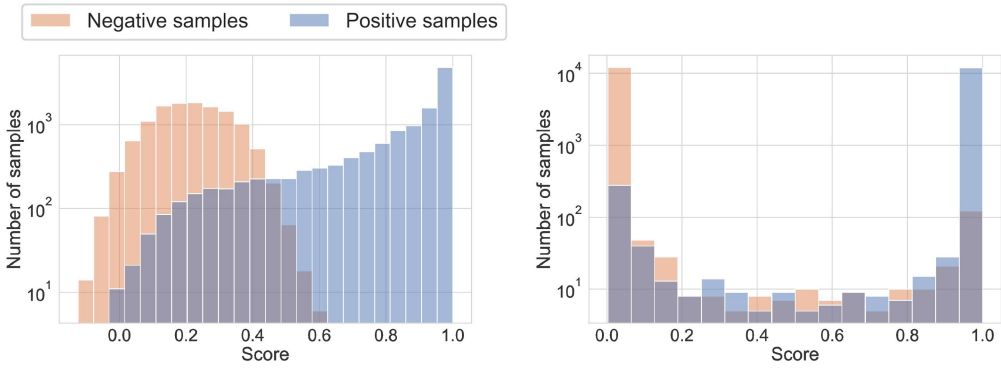
We also see that across all models, the second step, where the entire article is fed sentence by sentence, achieves significantly better performance when compared to only working with the headlines. A major difference between the two strategies used in step 2—using (i) ⟨Tweet, sentence⟩ pairs and (ii) ⟨Tweet, article⟩ pairs—is that the former tends to tag significantly fewer pairs as *unsupported*. This happens because the classifiers often find a sentence that is similar to the Tweet, and labels the pair as *supported*. Their true negative rate (also known as *specificity*) is thus significantly lower than the models using the latter strategy. It is worth noting, however, that for each model, the *negative predictive values* (i.e., the ratio of truly unsupported claims to predicted unsupported claims) are comparable across the two strategies. As such, if a model (except CT-BERT-v2) labels a pair as unsupported, then it is very likely that the citation is, indeed, deceptive.

There is no consistent improvement between DistilRoBERTa and DistilRoBERTa^p, even though the latter was expected to perform better due to its training on a large number of paraphrases. We believe it is the topic-specific nature of our work that removes the advantage. That is, if DistilRoBERTa^p were trained on a paraphrase corpus related to COVID-19, then its improvements would have been more significant. We also do not see Longformer exceeding the other models, in spite of it being designed for longer texts. This can be attributed to the "inverted pyramid" structure of newswire articles, which attempts to place all the essential information in the lead paragraph [59]. Thus, the other models can also capture the relevant information to a similar extent, eroding the relative advantage enjoyed by Longformer in many other tasks with long texts.

5.3 Additional Experiments and Discussion

Throughout our experiments, each ⟨Tweet, news⟩ pair—whether sentence-by-sentence or as the entire article—was put through a binary classifier, and the classification probability scores were used to determine the final label. A question may be raised at this point regarding the choice of the threshold probability score (0.5) that works as the decision boundary. In Figure 8, we show the results of varying the threshold for the second step in Task 2, where ⟨Tweet, news⟩ pairs were labeled on the basis of sentence-level analysis (discussed previously in Section 5.1.3).

Our approach has, in part, been motivated by indications from prior research that a claim and its supporting evidence are semantically similar [3, 47]. A pertinent question, thus, is whether measuring semantic similarity is enough to identify support. To investigate this, we design an additional experiment where the Tweet and the corresponding cited headline are converted to vectors, and



(a) Cosine similarity score (b) Probability score of XLNet

Fig. 9. Distribution of scores for Tweet-headline pairs on the development set. The y-axis is the number of Tweet-news pairs in log scale within the score range, with (a) showing the distribution of cosine similarity scores among the negative and positive samples respectively and (b) showing the classification probability score calculated by XLNet on those samples.

their cosine similarity is computed. This is in contrast to the experiments in the previous sections, where the ⟨Tweet, news⟩ pairs were put through a binary classifier, and the classification probability scores were used to determine the final label.

Now, we use the pre-trained DistilRoBERTa language model to obtain the vector representations of each Tweet and headline in the development set. The distribution of the cosine similarity scores are shown in Figure 9(a). For almost all the negative samples, the similarity is under 0.5, but this is true for a significant portion of the positive samples as well. Indeed, 12.2% of the positive samples have a cosine similarity score less than 0.5. A manual inspection of a random sample, however, reveals that only 5% of these are *unsupported*. In contrast, our investigation of the first step of Task 2 shows that 24–33% (varying between the various models) of the weakly false negative samples are, indeed, *unsupported*. Further, we juxtapose the cosine similarity scores obtained from DistilRoBERTa with the probability scores of XLNet, shown in Figure 9(b). It immediately becomes clear that the classification approach we took is significantly better at distinguishing the claims accompanied by genuinely supporting news articles from those with deceptive support. The cosine similarity scores obtained using the other pretrained language models provide very similar results and have not been included for the sake of brevity.

The results of this comparison decidedly indicate that our classifiers, which used the language models and further tuned them for this task, learn certain linguistic signals beyond just semantic similarity. This in turn leads to the system achieving significantly higher specificity (i.e., true negative rate). A higher specificity is a crucially important measure in a practical "real-world" scenario of misinformation detection. After all, higher specificity means that fewer genuine Tweets are mislabeled as containing deceptive support. A low-specificity detection system, however, is likely to annoy the typical user by labeling more of their social media posts as misinformation, and may gradually lead to consumers leaving the platform.

6 RELATED WORK

6.1 Data Collection

Systems designed for early detection of misinformation often rely on a combination of signals from the user, the dissemination pattern, and the content of the post [77, 78]. Jain et al. [33], for

instance, collected and clustered the Tweets, found similar content from credible news channels as ground-truth information, and then compared the semantics and sentiment of the Tweet to the reliable content. In case of a mismatch, the authors labeled the Tweet as misinformation. In this body of work, a fixed set of sources was assumed to be trustworthy—an approach that has been criticized by qualitative research for its potential implicit bias [29, 71]. There are very few exceptions to this approach, e.g., Al-Rakhami and Al-Amri [1], who instead rely on large-scale manual annotations—a particularly time-intensive approach to resolve a time-sensitive issue.

Large-scale high-quality data are critically important to misinformation detection using machine learning, and several efforts have sought to fill this need. Banda et al. [7] released a very large open-source dataset with more than 383 million Tweets. Their corpus includes only the Tweet IDs but is accompanied by the scripts needed to rehydrate the Tweets. The original dataset contains both Tweets and retweets, which allows tracking information dissemination. A cleaned version has also been released, however, without the retweets. This step removes around 75% of the Tweets. While this work does not directly attempt to detect misinformation, their dataset is valuable to others who intend to detect pandemic misinformation on social media.

The dataset released by Banda et al. [7] includes Tweets in other languages (French, German, Russian, and Spanish) but predominantly consists of English Tweets. Others have developed multilingual corpora. Notable among them are the contributions made by Gao et al. [21], providing English and Japanese posts on Twitter, and Chinese posts on *Weibo*, and Alqurashi et al. [4], who released an Arabic COVID-19 dataset of Tweets. A larger corpus of Arabic language Tweets related to COVID-19 was developed by Haouari et al. [27], which includes retweets and can thus be used to study pandemic information dissemination. In English language posts, propagation has been studied extensively. For instance, rumor propagation patterns have been studied for several years now, with applications in early detection, determining support, and veracity [25, 60]. Similar studies in other languages remain to be done.

6.2 Misinformation Detection

Memon and Carley [44] manually annotated more than 4.5K COVID-19-related tweets. The dataset having different types of information and misinformation was classified into 17 classes (*Irrelevant*, *Conspiracy*, *True Treatment*, *Fake Cure*, *Fake Treatment*, etc.) One cause for concern is that the data have been annotated by only one annotator. The authors looked at various attributes of two target groups: (i) misinformed users (who are actively posting misinformation) and (ii) informed users (who are actively spreading true information). Their methodology involves two steps. In the first step, the authors used a keyword-based Twitter search API for data collection. In the second step, the annotator categorized and labeled the Tweets into 17 classes, based on the types of information. The authors concluded that misinformed users' communities might be denser and more organized, while informed users use more narrative language. The authors observed that bots exist in both misinformed and informed communities, noticeably more among the misinformed users.

Hossain et al. [30] divided misinformation detection task into two sub-tasks of (i) retrieval of misconceptions relevant to posts being checked for veracity and (ii) stance detection to identify whether the posts *Agree*, *Disagree*, or express *No Stance* toward the retrieved misconceptions. Authors then collected and rephrased a set of COVID-19-related misconceptions from a Wikipedia entry, paired with 6.7K Tweets, and determined the stance of the Tweets against that misconception. Their goal was to determine whether NLP models can be adapted to the task of detecting misinformation without further training. The authors used relevant datasets to pre-train the models and to make the models domain specific. They have selected multiple NLP models, some that are suitable for misconception retrievals such as BM25 and Cosine Similarity with different embedding models like BERTSCORE and some that can be used for stance detection. The stance detection

sub-task can be considered to be equivalent to **Natural Language Inference (NLI)** problem, and thus, the authors used linear classifiers trained on NLI datasets combined with other models such as average GloVe embeddings as well as Sentence-BERT and Bidirectional LSTM encoding. Their results demonstrate that domain adoption, retraining language models on a corpus of COVID-19 tweets, increase the performance noticeably in both tasks of misconception retrieval and stance detection. Keeping the dataset updated is challenging as new rumors are being circulated and older ones may get obsolete as the pandemic continues. In addition, many Tweets in the dataset may not be available due to various reasons, e.g., deletion by users or removal by Twitter.

Kim and Walker [37] used a different strategy for defining misinformation. This study relied on the official recommendations of reputable health institutions to find the reply Tweets that make the same claim. They confirmed that this method is more effective at identifying Tweets with misinformation than searching based on keywords. The authors investigated the applicability of the proposed model with an example of advice from WHO related to *antibiotics* and *COVID-19 cure*. They collected more than 16K English reply Tweets over 3 months based on a specific combination of keywords closely related to the selected authentic advice, and parent Tweets were then obtained. These parent Tweets could potentially contain misinformation. Ignoring non-English and self-reply parent Tweets and filtering them based on another set of keywords, 573 pairs of the parent–reply pair Tweets were collected. Afterward, the sentence-BERT model converted reply Tweets and the advice to vectors, and the cosine similarity between each vector of reply Tweets and the vector of the advice is calculated. Two hundred reply Tweets with unique parent Tweets are selected where they have the highest cosine similarity scores calculated between the reply Tweet and the advice vectors. By manual inspection, authors detected parent Tweets with misinformation. Then they added metadata obtained from the users posting Tweets with misinformation, like timelines of friends and followers, to realize the extent of the spread of misinformation locally. This approach requires replies in response to a misinformation Tweet that has authentic information. Consequently, misinformation without replies containing authentic information will not be detected. In addition, this approach requires manual checking, which is laborious and error prone.

One example of studying non-English misinformation detection has been done by Kar et al. [36] on Indic languages (Bengali and Hindi) using **Multilingual BERT (mBERT)**.[7] Authors used the labeled English Tweets in the Infodemic COVID-19 dataset [2] as well as their translation into Bengali with Google Translate API, while retaining the same labels, as a part of their training dataset. They also used the Bengali dataset released in Reference [22] and manually annotated 100 randomly selected Tweets. The Hindi dataset was created in the same manner; they collected a set of Tweets by keyword searching and then added their Hindi translation. The authors used a zero-shot learning, which requires that the set of labels in the training data be different than the set of labels for the data that the model will be used to classify [74]. To perform zero-shot learning, they had experiments in which Tweets in one language were kept for testing and the rest of Tweets in other languages for training the model. They have further augmented the datasets by adding metadata of the Tweets, including the number of retweets and the number of likes, and 22 more features. The authors also defined three novel features: first, *Fact Verification Score*, which is obtained by searching the Tweet text in the Google search engine and taking the average Levenshtein distance between the Tweet text and the titles of search results only from reliable websites; second, *Bias Score*, which is defined using a Linear **Support Vector Machine (SVM)** Classifier for specifying the probability that a Tweet contains offensive language; and, third, *Source Tweet Embedding*, which is the vector representation of the Tweet text using BERT-based models. Four classifiers, i.e., **Multi-Layer Perceptron (MLP)**, Random Forest Classifier, SVM, and mBERT,

[7] https://github.com/google-research/bert/blob/master/multilingual.md.

were examined, and their results show that fine-tuned mBERT achieved the best F1-score of 89% in detecting Tweets with fake news. The disadvantage of this work is the need for manual annotation of a relatively large dataset.

Madani et al. [42] proposed a similar approach for the Moroccan language, using both Tweet and other metadata. For data collection, they got a dataset of fake news represented in Reference [66] that is based on ground-truth information from fact-checking websites. Based on that, the authors collected 10K Tweets with fake news related to COVID-19 by keyword searching, and they manually annotated the Tweets as fake or real. These English Tweets and the metadata that they extracted from them, such as Tweet length, Tweet sentiment, friends and followers number of Tweet's owner, and 10 more features, form their training and testing dataset. To gather the unlabeled Tweet dataset, they used the Tweepy library and translated the Tweets to Moroccan. For fake Tweet detection, six different machine learning models (Decision Tree, Random Forest, Naive Bayes, Gradient Boosting, Support Vector Machines, and MLP) were used. The Random Forest classifier outperformed all other models, including the MLP model, with respect to four evaluation metrics, accuracy, precision, recall, and F1-score. The authors observed that a positive correlation between the sentiment of a Tweet and its authenticity means that Tweets with positive sentiment are more likely to be authentic and Tweets with a negative sentiment most probably contain misinformation. The positive effect of metadata on performance is another observation. In our work, we do not use metadata as we focus on investigating the connection between the Tweet text and the cited news article.

Gupta et al. [24] implemented a semi-supervised ranking model that assesses the credibility of Tweets in real time. They have collected more than 10M Tweets about different events, and among them, they randomly selected 500 Tweets for annotation to build a training set for their model. They used crowdsourcing to classify the Tweets into four classes: *Definitely credible, Seems credible, Definitely incredible,* and *None of the above (skip Tweet).* The model extracts 45 content-related features from the Tweets and the users posting those Tweets, such as the number of characters, swear words, pronouns, positive and negative emoticons, number of retweets, and replies by the users. Based on these features, it gives credibility scores to the Tweets, ranging from 1 (low) to 7 (high). They tested four models that are commonly used for information retrieval, namely Coordinate AdaRank, RankBoost, Ascent, and SVM-rank. To compare these models, they used two evaluation metrics (**Normalized Discounted Cumulative Gain (NDCG)**) to obtain correctness and model running time. Finally, they chose the SVM-rank model, which is the second-best model in terms of NDCG@n[8] and is the best one in terms of training time. The model has been used in browser plugins and tested on 1,127 Twitter users over the course of 3 months, and 5.4 million Tweets' credibility scores were computed. They observed that features extracted from the Tweets content are more effective in credibility assessment than those extracted from the user accounts. We are also focusing on the content of the Tweets in our work to identify misinformation among the Tweets. The difference between this approach and ours is that we do not look at misinformation detection as a ranking problem but we offer a binary classification model.

Nguyen et al. [49] designed a shared task, WNUT-2020, to automatically identify informative COVID-19 Tweets, as manual annotation is a cost-intensive solution. This task, while not directly on COVID-related misinformation, can be viewed as a requisite step that can provide helpful data. Here the authors defined an "informative" Tweet as one that offers specific and clear information, and not rumor or prediction, about suspected, affirmed, healed, and deceased COVID-19 cases along with the travel history or location of the cases. From March 1 to June 30, about 23M non-repeating Tweets related to COVID-19 were gathered. Authors filtered this corpus by particular

[8]This means that to calculate the NDCG, the first n records in the ranked list are considered.

keywords like "positive," "discharge," "death," and so on, to separate candidates for informative Tweets. Among this dataset, a random sample set of 2K Tweets were manually annotated by three annotators with two labels, *informative* and *uninformative*. A classifier is trained on this subset to predict the probability of Tweets being informative for the rest of the Tweets in the dataset. Authors sampled 8K Tweets with different informative probabilities. These Tweets were also manually annotated; altogether, they formed a set of 10K Tweets as the final gold standard corpus used for training, validation, and testing the models for the shared task. Authors used fastText [35], a text classification task, as a baseline. The baseline classifier achieves the F1-score, harmonic mean of precision and recall, of 75%. Considering the F1-score, 48 of 55 participants outperform the baseline model; most of the teams are benefiting from pre-trained language models such as BERT, RoBERTa, XLNet, and so on. The top 6 teams used CT-BERT while more than half of the teams are leveraging ensemble techniques. The best participant's model reached the F1-score of 96.06% and accuracy of 91.50%. This work motivated our choice of using pre-trained transformers and fine-tuning them. While eliminating some of the Tweets is a similar task between our work and this study, we considered different definitions based on which we decide to ignore a Tweet; we keep a Tweet if it contains a factual claim that is of interest to the public, while in this work a Tweet is classified as *informative* if it provides direct and clear information about COVID-19 cases.

6.3 Misinformation Propagation

Huang and Carley [31] collected more than 67 million Tweets from 12 million users with metadata concerning geographical information, social identities, and the political orientation of users by tracking COVID-19 Twitter conversations. Their analysis found that misinformation has a higher likelihood of being spread within a single country by regular users, and not across nations.

Some recent work has looked at the spread of misinformation using epidemiological models as well. For example, Cinelli et al. [11] analyzed the spread of more than 8 million posts on social networks with epidemic models using reproduction number (R0), i.e., the average number of secondary cases an infected individual will create. They concluded that both questionable and reliable news spread with similar diffusion patterns, indicating that it may not be possible to accurately detect misinformation by means of metadata alone. Others, however, have reported that misinformation spreads significantly faster than the truth [64, 73]. Shahi et al. [64] conducted an exploratory study and relied on a list of 7,623 COVID-19-related fact-checked news articles and searched for news articles that are cited in Tweets, resulting in a set of 1,565 unique Tweets. Four classes of *False*, *Partially False*, *True*, and *Other* have been defined. Their analysis reveals that in 70% of the false and partially false categories of misinformation verified Twitter handles such as celebrities and organizations either create the content or help to spread it. Vosoughi et al. [73] investigated the publication of fake, verified, and mixed information on Twitter. Instead of focusing on a specific topic, they considered a longer duration (2006 to 2017). The diffusion of rumor cascades has been analyzed by considering the replies and retweets. It reported that false information on Twitter tends to be retweeted by many more users and spreads much faster than authentic information, especially when it is about a political issue.

7 CONCLUSION

In this work, we investigate a previously unexplored aspect of misinformation, *viz.*, where information is presented in social media with the *appearance* that it is supported by valid and reputable news agencies but the appearance is deceptive. That is, a claim is made on social media, and a news article is cited, but the article does not actually support the claim! It is often the case that users trust the existence of such support, without verifying any further. Our work uses Twitter posts about the COVID-19 pandemic. To this end, we provide a new dataset of COVID-19 Tweets,

where each Tweet cites a newswire article. We model this as an information retrieval task, where check-worthy claims are first separated from other social media posts and then put through classifiers to determine if the apparent support is deceptive. Our approach relies on distant supervision and shows that this is a viable option when annotated data are limited. Our findings reveal that a significant fraction of check-worthy claims—27.5% of the annotated sample (which corresponds to at least 1% of the test data)—contain deceptive support. Further, we provide experimental evidence that while semantic similarity plays an important role in finding support for a claim, there are deeper linguistic signals at play, captured by task-specific fine-tuning of language models.

We underscore that our technique is not specific to COVID-19 or other medical scenarios. The approach we have described can be applied to study misinformation and deception in any other topic, as long as the training corpus is domain specific (which, in our work, is health-related information). We selected COVID-19 due to its relevance in our current information landscape, the availability of data, and the existence of domain-specific language models like CT-BERT.

Our work here is a first step in the direction of identifying deceptive support across two genres: social media and newswire articles. There is significant scope for improvement, which we intend to pursue in the near future with larger datasets and seek collaborators to gain access to other social media platforms like Facebook or WhatsApp, where misinformation has been heavily discussed [20, 40, 70]. Our study indicates that to fight an infodemic, there is a need to look across genres instead of attending exclusively to social media posts. We hope that our findings can stimulate discussions aimed at making the Internet a more trustworthy landscape among its users, as well as making social media a more reliable source of information.

REFERENCES

[1] M. S. Al-Rakhami and A. M. Al-Amri. 2020. Lies kill, facts save: Detecting COVID-19 misinformation in Twitter. *IEEE Access* 8 (2020), 155961–155970. https://doi.org/10.1109/ACCESS.2020.3019600

[2] Firoj Alam, Shaden Shaar, Alex Nikolov, Hamdy Mubarak, Giovanni Da San Martino, Ahmed Abdelali, Fahim Dalvi, Nadir Durrani, Hassan Sajjad, Kareem Darwish, and Preslav Nakov. 2020. Fighting the COVID-19 infodemic: Modeling the perspective of journalists, fact-checkers, social media platforms, policy makers, and the society. arXiv:2005.00033. Retrieved from https://arxiv.org/abs/2005.00033.

[3] Aimée Alonso-Reina, Robiert Sepúlveda-Torres, Estela Saquete, and Manuel Palomar. 2019. Team GPLSI. Approach for automated fact checking. In *Proceedings of the 2nd Workshop on Fact Extraction and VERification (FEVER'19)*. Association for Computational Linguistics, 110–114. https://doi.org/10.18653/v1/D19-6617

[4] Sarah Alqurashi, Ahmad Alhindi, and Eisa Alanazi. 2020. Large Arabic Twitter dataset on COVID-19. arXiv:2004.04315. Retrieved from https://arxiv.org/abs/2004.04315.

[5] Ahmer Arif, Kelley Shanahan, Fang-Ju Chou, Yoanna Dosouto, Kate Starbird, and Emma S. Spiro. 2016. How information snowballs: Exploring the role of exposure in online rumor propagation. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW'16)*. Association for Computing Machinery, New York, NY, 466–477. https://doi.org/10.1145/2818048.2819964

[6] Fatma Arslan, Naeemul Hassan, Chengkai Li, and Mark Tremayne. 2020. A benchmark dataset of check-worthy factual claims. *Proceedings of the 14th International AAAI Conference on Web and Social Media (ICWSM'20)* 14, 1 (2020), 821–829.

[7] Juan M. Banda, Ramya Tekumalla, Guanyu Wang, Jingyuan Yu, Tuo Liu, Yuning Ding, and Gerardo Chowell. 2020. A Large-scale COVID-19 Twitter chatter Dataset for open scientific research—An international collaboration. arXiv:2004.03688. Retrieved from https://arxiv.org/abs/2004.03688.

[8] Alberto Barrón-Cedeño, Tamer Elsayed, Preslav Nakov, Giovanni Da San Martino, Maram Hasanain, Reem Suwaileh, and Fatima Haouari. 2020. CheckThat! at CLEF 2020: Enabling the automatic identification and verification of claims in social media. In *Advances in Information Retrieval*, Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva, and Flávio Martins (Eds.). Springer International Publishing, Cham, 499–507. https://doi.org/10.1007/978-3-030-45442-5_65

[9] Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. arXiv:2004.05150. Retrieved from https://arxiv.org/abs/2004.05150.

[10] Steven Bird, Edward Loper, and Ewan Klein. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc., Sebastopol, CA, USA.

[11] Matteo Cinelli, Walter Quattrociocchi, Alessandro Galeazzi, Carlo Michele Valensise, Emanuele Brugnoli, Ana Lucia Schmidt, Paola Zola, Fabiana Zollo, and Antonio Scala. 2020. The covid-19 social media infodemic. *Sci. Rep.* 10, 1 (2020), 1–10.

[12] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *Proceedings of the 8th International Conference on Learning Representations*. OpenReview.net.

[13] Mark Craven and Johan Kumlien. 1999. Constructing biological knowledge bases by extracting information from text sources. In *Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, 77–86.

[14] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2978–2988. https://doi.org/10.18653/v1/P19-1285

[15] Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Enhanced sentiment learning using Twitter hashtags and smileys. In *Proceedings of the International Conference on Computational Linguistics (COLING'10)*. COLING 2010 Organizing Committee, 241–249.

[16] Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H. Eugene Stanley, and Walter Quattrociocchi. 2016. The spreading of misinformation online. *Proc. Natl. Acad. Sci. U.S.A.* 113, 3 (2016), 554–559. https://doi.org/10.1073/pnas.1517441113

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, Minneapolis, 4171–4186. https://doi.org/10.18653/v1/n19-1423

[18] Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychol. Bull.* 76, 5 (1971), 378–382. https://doi.org/10.1037/h0031619

[19] B. J. Fogg, Gregory Cuellar, and David Danielson. 2007. Motivating, influencing, and persuading users: An introduction to captology. In *The Human Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications*, Julie A. Jacko, Julie A. Jacko, and Andrew Sears (Eds.). CRC Press, New York, NY, 159–172. https://doi.org/10.1201/9781410615862

[20] Sheera Frenkel. 2021. *White House Dispute Exposes Facebook Blind Spot on Misinformation*. The New York Times. Retrieved August 1, 2021 from https://www.nytimes.com/2021/07/19/technology/facebook-misinformation-blind-spot.html.

[21] Zhiwei Gao, Shuntaro Yada, Shoko Wakamiya, and Eiji Aramaki. 2020. NAIST COVID: Multilingual COVID-19 Twitter and Weibo dataset. arXiv:2004.08145. Retrieved from https://arxiv.org/abs/2004.08145.

[22] Avishek Garain. 2020. COVID-19 Tweets Dataset for Bengali Language. https://doi.org/10.21227/wdt0-ya78

[23] Amira Ghenai and Yelena Mejova. 2017. Catching zika fever: Application of crowdsourcing and machine learning for tracking health misinformation on Twitter. In *Proceedings of the IEEE International Conference on Healthcare Informatics (ICHI'17)*. IEEE, Park City, UT, 518–518. https://doi.org/10.1109/ICHI.2017.58

[24] Aditi Gupta, Ponnurangam Kumaraguru, Carlos Castillo, and Patrick Meier. 2014. TweetCred: Real-time credibility assessment of content on Twitter. In *Proceedings of the 6th International Conference on Social Informatics (Lecture Notes in Computer Science, Vol. 8851)*. Springer, 228–243. https://doi.org/10.1007/978-3-319-13734-6_16

[25] Sardar Hamidian and Mona Diab. 2016. Rumor identification and belief investigation on Twitter. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Association for Computational Linguistics, 3–8. https://doi.org/10.18653/v1/W16-0403

[26] Andreas Hanselowski, Hao Zhang, Zile Li, Daniil Sorokin, Benjamin Schiller, Claudia Schulz, and Iryna Gurevych. 2018. UKP-athene: Multi-sentence textual entailment for claim verification. In *Proceedings of the 1st Workshop on Fact Extraction and VERification (FEVER'18)*. ACL, 103–108. https://doi.org/10.18653/v1/W18-5516

[27] Fatima Haouari, Maram Hasanain, Reem Suwaileh, and Tamer Elsayed. 2021. ArCOV-19: The first Arabic COVID-19 Twitter dataset with propagation networks. In *Proceedings of the 6th Arabic Natural Language Processing Workshop*. Association for Computational Linguistics, 82–91.

[28] Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. 2017. Toward automated fact-checking: Detecting check-worthy factual claims by ClaimBuster. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'17)*. Association for Computing Machinery, New York, NY, 1803–1812. https://doi.org/10.1145/3097983.3098131

[29] Jennifer L. Hochschild and Katherine Levine Einstein. 2015. *Do Facts Matter?: Information and Misinformation in American Politics*. University of Oklahoma Press, Norman, OK.

[30] Tamanna Hossain, Robert L. Logan IV, Arjuna Ugarte, Yoshitomo Matsubara, Sean Young, and Sameer Singh. 2020. COVIDLies: Detecting COVID-19 misinformation on social media. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*. Association for Computational Linguistics, Online, 11 pages. https://doi.org/10.18653/v1/2020.nlpcovid19-2.11

[31] Binxuan Huang and Kathleen M. Carley. 2020. Disinformation and misinformation on Twitter during the novel coronavirus outbreak. arXiv:2006.04278. Retrieved from https://arxiv.org/abs/2006.04278.

[32] Muhammad Imran, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. 2015. Processing social media messages in mass emergency: A survey. *ACM Comput. Surv.* 47, 4 (2015), 1–38. https://doi.org/10.1145/2771588

[33] S. Jain, V. Sharma, and R. Kaushal. 2016. Towards automated real-time detection of misinformation on Twitter. In *Proceedings of the International Conference on Advances in Computing, Communications and Informatics (ICACCI'16)*. IEEE, 2015–2020. https://doi.org/10.1109/ICACCI.2016.7732347

[34] F. Jin, W. Wang, L. Zhao, E. Dougherty, Y. Cao, C. Lu, and N. Ramakrishnan. 2014. Misinformation propagation in the age of Twitter. *Computer* 47, 12 (2014), 90–94. https://doi.org/10.1109/MC.2014.361

[35] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Association for Computational Linguistics, 427–431.

[36] Debanjana Kar, Mohit Bhardwaj, Suranjana Samanta, and Amar Prakash Azad. 2020. No rumours please! A multi-Indic-lingual approach for COVID fake-tweet detection. arXiv:2010.06906. Retrieved from https://arxiv.org/abs/2010.06906.

[37] Hyunuk Kim and Dylan Walker. 2020. Leveraging volunteer fact checking to identify misinformation about COVID-19 in social media. *Harv. Kennedy School Misinf. Rev.* 1, 3 (2020), 10 pages. https://doi.org/10.37016/mr-2020-021

[38] 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 (2020), e7255.

[39] Rui Li, Kin Hou Lei, Ravi Khadiwala, and Kevin Chen-Chuan Chang. 2012. TEDAS: A Twitter-based event detection and analysis system. In *Proceedings of the IEEE 28th International Conference on Data Engineering*. IEEE, 1273–1276. https://doi.org/10.1109/ICDE.2012.125

[40] Rupali Jayant Limaye, Molly Sauer, Joseph Ali, Justin Bernstein, Brian Wahl, Anne Barnhill, and Alain Labrique. 2020. Building trust while influencing online COVID-19 content in the social media world. *Lancet Digit. Health* 2, 6 (2020), e277–e278. https://doi.org/10.1016/S2589-7500(20)30084-4

[41] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. arXiv:1907.11692. Retrieved from https://arxiv.org/abs/1907.11692.

[42] Youness Madani, Mohammed Erritali, and Belaid Bouikhalene. 2021. Using artificial intelligence techniques for detecting Covid-19 epidemic fake news in Moroccan tweets. *Results Phys.* 25 (2021), 104266. https://doi.org/10.1016/j.rinp.2021.104266

[43] Micol Marchetti-Bowick and Nathanael Chambers. 2012. Learning for microblogs with distant supervision: Political forecasting with Twitter. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 603–612.

[44] Shahan Ali Memon and Kathleen M. Carley. 2020. Characterizing covid-19 misinformation communities using a novel Twitter dataset. In *Proceedings of the CIKM 2020 Workshops*. CEUR-WS.org.

[45] Michele Miller, Tanvi Banerjee, Roopteja Muppalla, William Romine, and Amit Sheth. 2017. What are people tweeting about Zika? An exploratory study concerning its symptoms, treatment, transmission, and prevention. *JMIR Publ. Health Surveill.* 3, 2 (2017), e38. https://doi.org/10.2196/publichealth.7157

[46] Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. Association for Computational Linguistics, 1003–1011.

[47] Mitra Mohtarami, Ramy Baly, James Glass, Preslav Nakov, Lluís Màrquez, and Alessandro Moschitti. 2018. Automatic stance detection using end-to-end memory networks. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, 767–776. https://doi.org/10.18653/v1/N18-1070

[48] Martin Müller, Marcel Salathé, and Per Egil Kummervold. 2020. COVID-Twitter-BERT: A natural language processing model to analyse COVID-19 content on Twitter. arXiv:2005.07503. Retrieved from https://arxiv.org/abs/2005.07503.

[49] Dat Quoc Nguyen, Thanh Vu, Afshin Rahimi, Mai Hoang Dao, Linh The Nguyen, and Long Doan. 2020. WNUT-2020 task 2: Identification of informative COVID-19 English tweets. In *Proceedings of the 6th Workshop on Noisy User-generated Text (W-NUT'20)*. Association for Computational Linguistics, 314–318. https://doi.org/10.18653/v1/2020.wnut-1.41

[50] Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining fact extraction and verification with neural semantic matching networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. Association for the Advancement of Artificial Intelligence, 6859–6866. https://doi.org/10.1609/aaai.v33i01.33016859

[51] World Health Organization. 2010. What Is a Pandemic? Retrieved April 27, 2021 from https://www.who.int/csr/disease/swineflu/frequently_asked_questions/pandemic/en/.

[52] World Health Organization. 2020. Munich Security Conference. Retrieved April 27, 2021 from https://www.who.int/director-general/speeches/detail/munich-security-conference.

[53] World Health Organization. 2020. WHO Director-general's Opening Remarks at the Media Briefing on COVID-19—11 March 2020. Retrieved April 27, 2021 from https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020.

[54] Yeimer Ortiz-Martínez and Luisa F. Jiménez-Arcia. 2017. Yellow fever outbreaks and Twitter: Rumors and misinformation. *Am. J. Infect. Contr.* 45, 7 (2017), 816–817.

[55] Sunday Oluwafemi Oyeyemi, Elia Gabarron, and Rolf Wynn. 2014. Ebola, Twitter, and misinformation: A dangerous combination? *Br. Med. J.* 349 (2014), g6178. https://doi.org/10.1136/bmj.g6178

[56] Gordon Pennycook, Tyrone D. Cannon, and David G. Rand. 2018. Prior exposure increases perceived accuracy of fake news. *J. Exp. Psychol.: General* 147, 12 (2018), 1865. https://doi.org/10.2139/ssrn.2958246

[57] Sarah Perez. 2017. *Twitter Officially Expands Its Character Count to 280 Starting Today.* TechCrunch. Retrieved June 6, 2021 from https://techcrunch.com/2017/11/07/twitter-officially-expands-its-character-count-to-280-starting-today/.

[58] Sarah Perez. 2018. *Twitter's Doubling of Character Count from 140 to 280 Had Little Impact on Length of Tweets.* TechCrunch. Retrieved June 6, 2021 from https://techcrunch.com/2018/10/30/twitters-doubling-of-character-count-from-140-to-280-had-little-impact-on-length-of-tweets/.

[59] Horst Pöttker. 2003. News and its communicative quality: The inverted pyramid–when and why did it appear? *Journal. Stud.* 4, 4 (2003), 501–511.

[60] Vahed Qazvinian, Emily Rosengren, Dragomir R. Radev, and Qiaozhu Mei. 2011. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 1589–1599.

[61] J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33, 1 (1977), 159–174.

[62] Victoria L. Rubin, Yimin Chen, and Nadia K. Conroy. 2015. Deception detection for news: Three types of fakes. *Proc. Assoc. Inf. Sci. Technol.* 52, 1 (2015), 1–4. https://doi.org/10.1002/pra2.2015.145052010083

[63] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. arXiv:1910.01108. Retrieved from https://arxiv.org/abs/1910.01108.

[64] Gautam Kishore Shahi, Anne Dirkson, and Tim A. Majchrzak. 2021. An exploratory study of COVID-19 misinformation on Twitter. *Online Soc. Netw. Media* 22 (2021), 100104. https://doi.org/10.1016/j.osnem.2020.100104

[65] Ivor Shapiro, Colette Brin, Isabelle Bèdard-Brûlé, and Kasia Mychajlowycz. 2013. Verification as a strategic ritual. *Journal. Pract.* 7, 6 (2013), 657–673. https://doi.org/10.1080/17512786.2013.765638

[66] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big Data* 8, 3 (2020), 171–188.

[67] Lisa Singh, Shweta Bansal, Leticia Bode, Ceren Budak, Guangqing Chi, Kornraphop Kawintiranon, Colton Padden, Rebecca Vanarsdall, Emily Vraga, and Yanchen Wang. 2020. A first look at COVID-19 information and misinformation sharing on Twitter. arXiv:2003.13907 [cs.SI]. Retrieved from https://arxiv.org/abs/2003.13907.

[68] Beth St. Jean, Mega Subramaniam, Natalie Greene Taylor, Rebecca Follman, Christie Kodama, and Dana Casciott. 2015. The influence of positive hypothesis testing on youths' online health-related information seeking. *New Libr. World* 116, 3/4 (2015), 136–154. https://doi.org/10.1108/NLW-07-2014-0084

[69] Kate Starbird, Jim Maddock, Mania Orand, Peg Achterman, and Robert M. Mason. 2014. Rumors, false flags, and digital vigilantes: Misinformation on Twitter after the 2013 Boston Marathon Bombing. In *Proceedings of the iConference (iConference'14)*. iSchools, Grandville, MI, 654–662. https://doi.org/10.9776/14308

[70] Mayowa Tijani. 2020. *How to Spot COVID-19 Misinformation on WhatsApp.* Agence France-Presse. Retrieved August 1, 2021 from https://factcheck.afp.com/how-spot-covid-19-misinformation-whatsapp.

[71] Joseph E. Uscinski and Ryden W. Butler. 2013. The epistemology of fact checking. *Crit. Rev.* 25, 2 (2013), 162–180. https://doi.org/10.1080/08913811.2013.843872

[72] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*. Curran Associates, Inc., Red Hook, NY, 5998–6008.

[73] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (2018), 1146–1151. https://doi.org/10.1126/science.aap9559

[74] Wei Wang, Vincent W. Zheng, Han Yu, and Chunyan Miao. 2019. A survey of zero-shot learning: Settings, methods, and applications. *ACM Trans. Intell. Syst. Technol.* 10, 2 (2019), 1–37.

[75] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019*, Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (Eds.). Curran Associates, Inc., Red Hood, NY, 5754–5764.

[76] John Zarocostas. 2020. How to fight an infodemic. *Lancet* 395, 10225 (2020), 676. https://doi.org/10.1016/S0140-6736(20)30461-X

[77] Zhe Zhao, Paul Resnick, and Qiaozhu Mei. 2015. Enquiring minds: Early detection of rumors in social media from enquiry posts. In *Proceedings of the 24th International Conference on World Wide Web (WWW'15)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 1395–1405. https://doi.org/10.1145/2736277.2741637

[78] Xing Zhou, Juan Cao, Zhiwei Jin, Fei Xie, Yu Su, Dafeng Chu, Xuehui Cao, and Junqiang Zhang. 2015. Real-time news certification system on sina weibo. In *Proceedings of the 24th International Conference on World Wide Web (WWW'15 Companion)*. Association for Computing Machinery, New York, NY, 983–988. https://doi.org/10.1145/2740908.2742571

[79] Chaoyuan Zuo, Narayan Acharya, and Ritwik Banerjee. 2020. Querying across genres for medical claims in news. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 1783–1789. https://doi.org/10.18653/v1/2020.emnlp-main.139

[80] Chaoyuan Zuo, Ayla Karakas, and Ritwik Banerjee. 2019. To check or not to check: Syntax, semantics, and context in the language of check-worthy claims. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction – Proceedings of the 10th International Conference of the CLEF Association (Lecture Notes in Computer Science, Vol. 11696)*, Fabio Crestani, Martin Braschler, Jacques Savoy, Andreas Rauber, Henning Müller, David E. Losada, Gundula H. Bürki, Linda Cappellato, and Nicola Ferro (Eds.). Springer International Publishing, Lugano, Switzerland, 271–283. https://doi.org/10.1007/978-3-030-28577-7_23

Received 31 August 2021; revised 4 April 2022; accepted 25 May 2022