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

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, (ii) verify 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.

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

ACM Reference Format:
. 2018. Seeing Should Probably not be Believing: The Role of Deceptive Support in COVID-19 Misinformation on Twitter. *Proc. ACM Meas. Anal. Comput. Syst.* 37, 4, Article 111 (August 2018), 25 pages. https://doi.org/10.1145/1122445.1122456

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 [58, 60]. It is not a word the agency uses lightly, especially after facing criticism for declaring the H1N1 outbreak in 2009 a pandemic, even as it infected nearly a quarter of the global population. The gravity of using the word "pandemic" lies in its potential to trigger large-scale panic and fear-mongering. 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 [59, 76]. On one end, this can have devastating consequences for individuals, who may take critical decisions based on false information, and on the other end, rip our social fabric by sowing distrust.

The ubiquity of such misinformation is particularly worrisome. 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 disease outbreak – Ebola [52], Zika [45], Yellow Fever [51], 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.*, [67]), but these analyses

Author's address:

2476-1249/2018/8-ART111 $15.00
https://doi.org/10.1145/1122445.1122456

Fig. 1. Original content entering the social media information space through the "New York Post" institutional Twitter account. With 2.1M followers (accessed: May 21, 2021), this is an entry with a large footprint.

Fig. 2. A corresponding derived content: re-transmission of the source with added remarks.

are *post-hoc* and do not help with prevention. Once a nugget of 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 [53]. This leads to a snowball effect, whereupon misinformation propagates many times faster than accurate news [51, 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 the rumor-bearing posts within a specific topic [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 *prior to its propagation.*

In social media, a post with original content is fundamentally different from one that includes a re-transmission. Arif et al. [5] have distinguished between them as "original content" and "derivative content". They report that when a claim enters the network with a large footprint, *e.g.*, 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 Fig. 1 and its propagation in Fig. 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.

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 not supported by the article, are candidates for misinformation or rumor propagation.

1.2 Scope and approach

Information propagated through social media can often be dissected along several dimensions. In their survey, 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 among the existing research on misinformation in social media because we investigate "perceived credibility" in posts. In other words, we investigate whether or not the derived content is faithful to the original content, as it is re-transmitted 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 Tweets related to COVID-19. This open dataset was created by Banda et al. [7] for the express purpose of integrated research in epidemiology, misinformation, and related fields. This is a continually growing dataset, and at the time of this work, it comprised 383 million Tweets.

B. Filtering subjectivity. A significant fraction of posts do not contain subjective information. For instance, Tweets often share personal anecdotes, contain emotional language, issue ironic or sarcastic remarks, etc. Our first step, therefore, is to distill Tweets that contain factual claims from the rest of the data.

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. In addition to the above controls, 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.*, [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.

In the remainder of this paper, 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 our work in the greater context of prior research in this field in Section 6 before concluding along with a brief discussion of future research directions in Section 7.

2 ARCHITECTURE

In our discussion of the overall architecture of the proposed system, 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, which are responsible for (i) data collection, (ii) data preprocessing, (iii) identifying check-worthy factual claims and (iv) discriminating verifiable claims from others.

2.1 Requirements

In their analysis of fake news detection systems within the scope of natural language processing (NLP) research, Rubin et al. [62] present nine fundamental requirements. In this work, we take care to ensure that our approach meets these criteria:

(1) Our data satisfies 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 adheres to *homogeneity in length*. Further, even though Twitter expanded its character count limit to 280 in November, 2017 [54], only 5% of the English language Tweets over the subsequent one year were longer than 190 characters, and only 9% used more than 140 [55], thus providing even more homogeneity in length than one would expect.
(5) Our work also adheres to *homogeneity in writing matter*, in terms of both topic (the COVID-19 pandemic) and genre, and offers comparison across multiple news agencies and social media users.
(6) The data used in this work was collected over a period of three months, during the prevalence of 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).

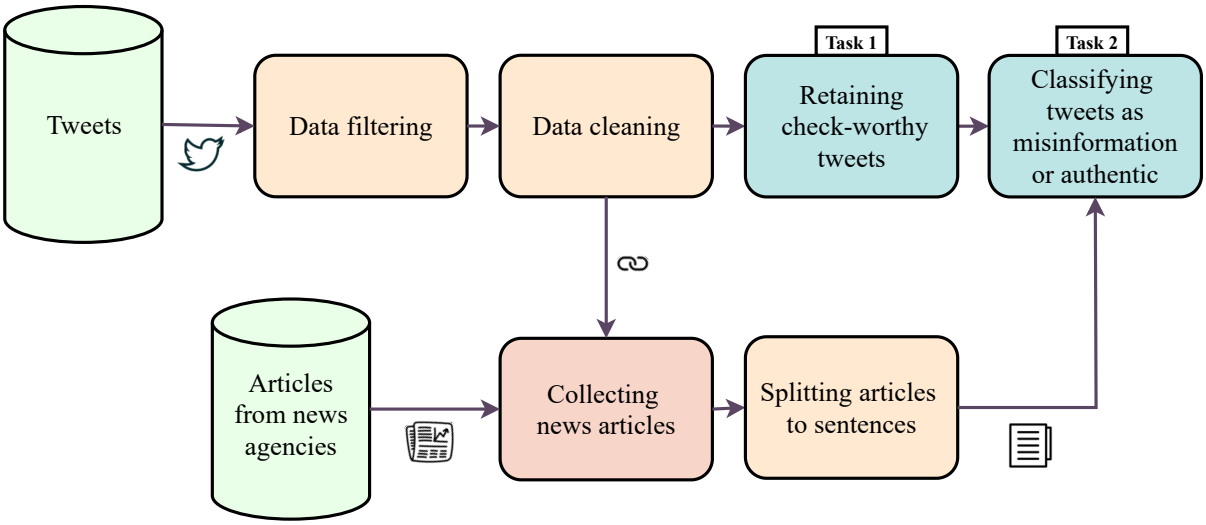


Fig. 3. **System architecture**. The pipeline comprises (i) the data collection from Twitter posts and news articles, (ii) data preprocessing – which includes the filtering, cleaning, and splitting into sentence-level chunks, (iii) the first task of identifying Tweets containing check-worthy factual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

(8) The corpus is created from publicly available data (in particular, based on the open dataset created for research by Banda et al. [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 overall system architecture, with the complete pipeline and its components. To provide a correspondence between the steps in our pipeline and the data, we also present examples of Tweets in Table 2.2.

Data collection. We use the open dataset created by Banda et al. [7] as the starting point, where we 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 dataset. The news articles are collected only for those Tweets that are retained after the data filtering step.

Data preprocessing. On 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 (along with its 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.** 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 of the linked news article; (4) a statement worth checking vis-à-vis the body of the linked news article; and (5) a check-worthy claim that is not supported by the cited article, thus merely *appearing* trustworthy.

	Tweet (derived content)	**Corresponding original content (cited news article)**
(1)	Africa deporting Europeans we love to see it [https://bit.ly/3vEtIyj] *Last accessed*: June 6, 2021	*– no news cited –*
(2)	Coronavirus Map: How To Track Coronavirus Spread Across The Globe via @forbes [https://bit.ly/3upHDao] *Last accessed*: June 6, 2021	**Headline:** Coronavirus Map: How To Track Coronavirus Spread Across The Globe **Body:** 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] *Last accessed*: June 6, 2021	**Headline:** Native American health center receives body bags instead of coronavirus supplies **Body:** 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] *Last accessed*: June 6, 2021	**Headline:** Bill Gates, at Odds With Trump on Virus, Becomes a Right-Wing Target **Body:** … 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] *Last accessed*: June 6, 2021	**Headline:** All of Italy is in lockdown as coronavirus cases rise **Body:** (CNN)Italy has been put under a dramatic total lockdown, as the coronavirus spreads in the country …

Task 2: Identifying whether or not 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 a binary classification task. 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.

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 46.86 million Tweets collected from March through May 2020. We inject additional filtering and data cleaning steps to it, however, which are discussed next.

3.1 Data filtering

Even though this Twitter dataset is related to COVID-19, it is not immediately suitable for the natural language processing tasks in our work. We have the following conditions to filter a significant part of this dataset:

Table 2. **COVID-19 keywords.** The 52 keywords used to filter out Tweets.

Keywords related to the COVID-19 pandemic
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

Table 3. **List of news agencies used as original content.** 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.

List of new agencies we verified Tweets
reuters.com, theguardian.com, wsj.com, washingtonpost.com, nytimes.com, cnn.com, cnbc.com, cbsnews.com, nypost.com, foxnews.com, usatoday.com, theatlantic.com, sfgate.com, latimes.com, hollywoodreporter.com, bbc.com, thehill.com, chicagotribune.com, usnews.com, thedailybeast.com, chron.com, time.com, nbcnews.com, bbc.co.uk, dw.com, variety.com, euronews.com

Retweets. A Retweet is a re-posting of a Tweet, 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 in a network, it is not useful for our study. Thus, we remove all Retweets.

Non-English Tweets. As we discussed earlier in Section 2, controlling for language is an important requirement [62]. The dataset, however, includes Tweets from five different languages. We therefore insert a step to filter out non-English entries.[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 (NLTK) [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.

[1]Given the ID of a Tweet, the Twitter API allows for the retrieval of many of its properties, a process known as *hydration*. A hydrated Tweet has several attributes, including one that specifies its language. We use the value of this attribute to determine if it is in English.
[2]Function words are words that play an important role in syntactic correctness of a sentence, but offer little semantic content. They consist mainly of determiners, pronouns, prepositions, and conjunctions. For example, "the", "and", "his", "she", "although".
[3]https://www.alexa.com/topsites/category/Top/News (this service was last available on Sep 17, 2020)

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.

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.

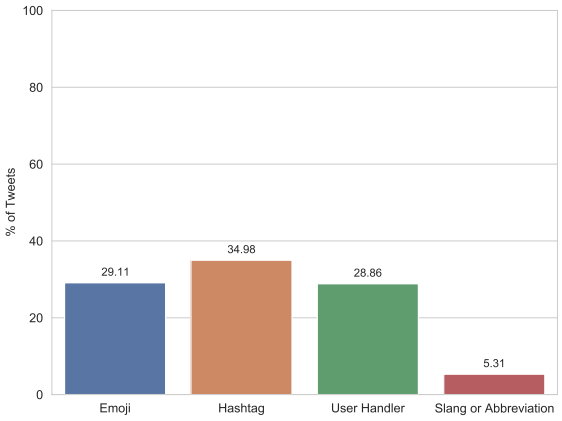


Fig. 4. **Extent of informal register usage.** Percentage of different parameters to total Tweets remaining after all filtering steps.

Social media users frequently depart from dictionary-based lexicon and make ample use of informal register. Most commonly, this includes emojis and colloquial non-standard abbreviations and misspellings that have become socially accepted. One may argue that emojis convey information (albeit not in the traditional linguistic sense) and thus, removing them alters the information 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.

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.

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

For each newswire article, we retain full text of the article, as well as the headline. Any images, videos, and metadata information (*e.g.*, authors, date of publication) are discarded. Subsequently, the articles are tokenized and split into individual sentences using the Python Natural Language Toolkit (NLTK) [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 (as shown earlier in 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 in social media. The task itself is designed as a supervised binary classification, where each Tweet is given one of two possible labels: *check-worthy* (CW), or *not check-worthy* (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, and (ii) tuning the embedded representations for a specific task. This is the approach we adopt in our work as well. To this end, we experiment with multiple pretrained language models, tuning each model in task-specific ways. In the remainder of this section, we first present a short discussion of the pretrained language models, followed by the datasets on which they are further tuned, before discussing the results.

4.1 Pretrained language models

We use ten language models pretrained on general data, plus two domain-specific pretrained models. These models all rely on the Transformer-based learning of contextual word representations, known as *Bidirectional Encoder Representations from Transformers* (BERT) [17]. BERT is pretrained on two NLP tasks, *viz.*, masked language modeling – where some input tokens are replaced with [MASK] and the model is trained to reconstruct the original tokens, and next sentence prediction – where the model is trained to understand whether or not one sentence can logically come after another. There are two variants of this model, BERT-Base and BERT-Large, which differ in the size of the network used for training (see Devlin et al. [17] for details). BERT demonstrated state-of-the-art performance on multiple downstream natural language understanding (NLU) tasks on benchmark datasets, and inspired variations of the original model. These include

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

A closely related model is ELECTRA [12], which is Transformer-based, but instead of the generative approach of BERT's masked language modeling, 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 a generalized autoregressive pretraining to capture bidirectionality in a token's linguistic context (in contrast to BERT, which uses denoising autoencoder to capture bidirectionality). Like BERT, it is a Transformer-based model, but it uses Transformer-XL [14] to overcome the restrictions of the basic Transformer models (*e.g.*, fixed-length context).

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 paper* [2021]	55 (55%)	45 (45%)	100	COVID-19 Tweets

As pretrained models, we use the multiple versions of BERT, DistilBERT, RoBERTa, CT-BERT, ELECTRA, and XLNet, giving us 12 models altogether. These are tuned on datasets specific to our first task, as discussed next.

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. In this work, 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 [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.

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

Table 5. **Performance on Task 1: Identification of check-worthy Tweets.** 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**) out 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.

Model	DS1				DS2				DS2 + DS1			
	P	**R**	$\mathbf{F}_1$	**TP**	**P**	**R**	$\mathbf{F}_1$	**TP**	**P**	**R**	$\mathbf{F}_1$	**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

lower precision means a greater number of falsely labeled check-worthy (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.

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 consist of multiple sentences. This is done in order 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, etc. 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.

5.1 Design and setup of experiments

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.

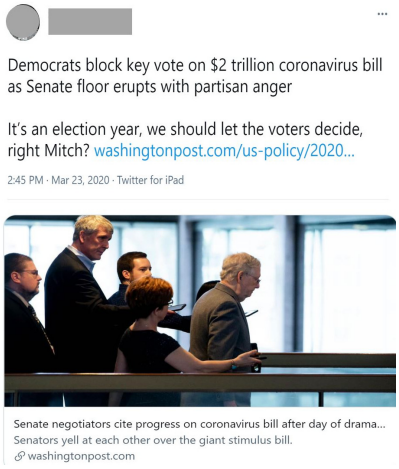


Fig. 5. A Tweet comprising multiple sentences. The first sentence is objective, and contains a check-worthy factual claim, while the second sentence does not.

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 employing *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 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 in Section 4. The last model, DistilRoBERTa, is a lighter version of RoBERTa, pretrained on a smaller general-purpose language 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]. All the models are tuned 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 can be distilled. It is, however, also possible that

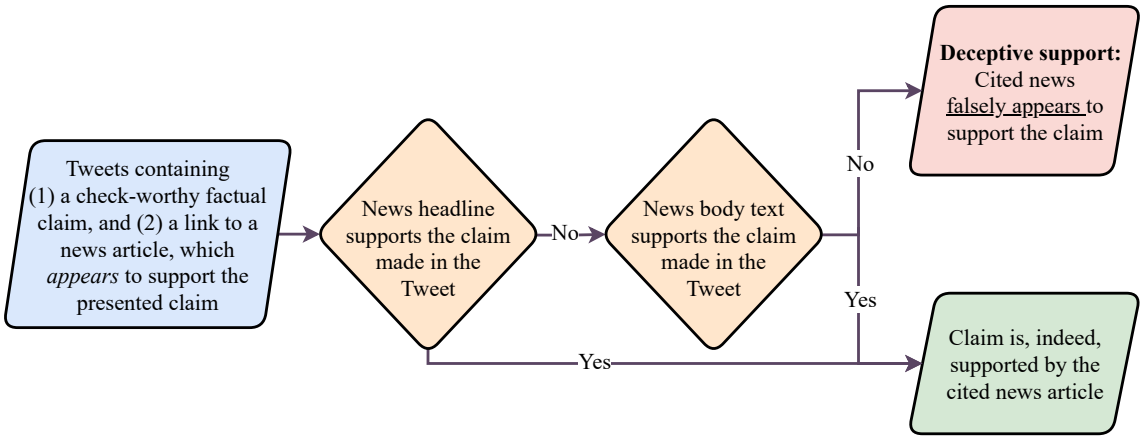


Fig. 6. **Information verification in Task 2:** The input comprises Tweets containing check-worthy factual claims that offer a news article as supporting evidence for that claim. The output is a binary decision about whether the support is deceptive.

the claim can only be gleaned from multiple sentences in the article. We thus follow a two-pronged strategy to determine support. On 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 described above (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", the ⟨Tweet, article⟩ pair is deemed "supported". Otherwise, it is deemed "unsupported". On the other hand, we also conduct experiments on the ⟨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 instead [9], which combines local windowed attention and global attention, thus allowing it to process sequences that are 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 where the hyperlink is deceptive, and the 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 entire article is investigated by the classifiers.

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, 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. Out 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. One such example is a Tweet that states "Yesterday more than 2K in the US died of coronavirus", where the dates of the post and the news article are, clearly, relevant.

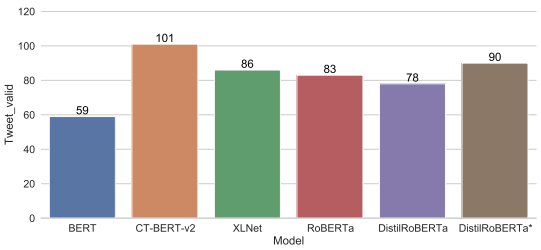


Fig. 7. **Number of *weakly false* negative pairs for each model.** These are check-worthy factual claims made in Tweets that link to a news article as a cue of external support, but the model labels them as *unsupported*, based on the ⟨Tweet, headline⟩ pair.

In Figure 7, the number of ⟨Tweet, headline⟩ pairs predicted to be *unsupported* by the models are shown after the removal of erroneous samples propagated by Task 1 (*i.e.*, claims that are not check-worthy).

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:

(1) First, we feed all the samples from Task 1 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.
(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⟩

				Phase 2										*Pipeline*					
	Phase1			*Sentence*					*Full News*					*Sentence*			*Full News*		
Transformer	**P**	**R**	**F1**	**P**	**R**	**F1**	**U#**	**TN#**	**P**	**R**	**F1**	**U#**	**TN#**	**U#**	**TN#**	**TN**	**U#**	**TN#**	**TN**
BERT	47.2	47.3	47.2	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	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**	**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	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	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	53.6	86.9	49.3	4	4	-	-	-	-	-	4	4	100	-	-	-
Longformer	-	-	-	-	-	-	-	-	49.0	52.9	50.9	51	27	-	-	-	38	21	55.3

Table 6. **Experiment results.** Model tuned on the paraphrase dataset marked with *. 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.

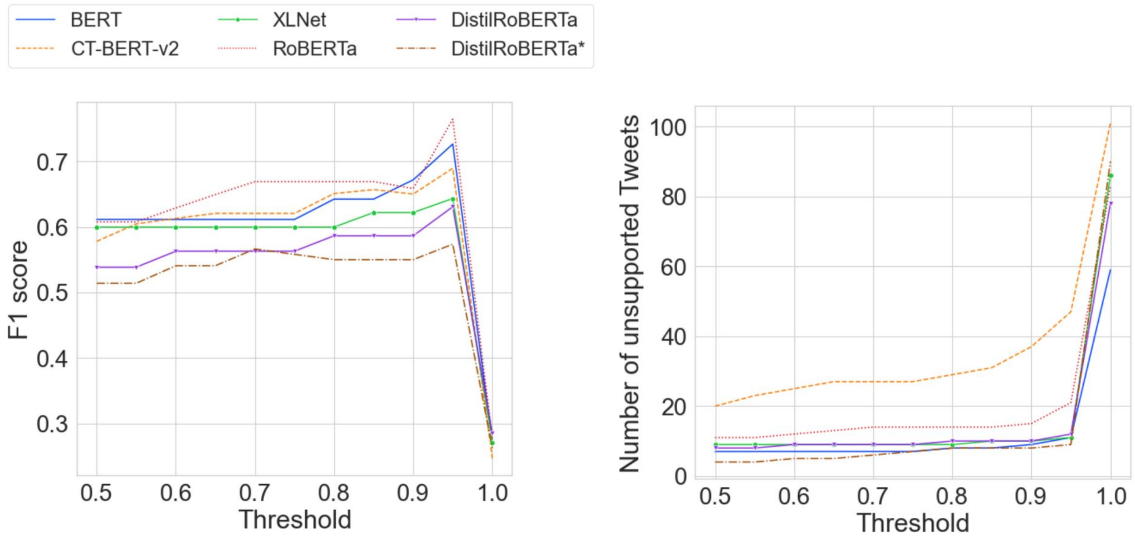
pairs as *unsupported*, and we evaluate BERT in step 2 using only these 59 pairs. Since we Longformer only in step 2, for this evaluation we use the results of DistilRoBERTa^{p} from step 1.

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: Twitter and newswire text, while CT-BERT is a language model with domain-specific pre-training. Thus, it may not be able to properly account for the lexical context of words found in newswire sentences.

We can 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. With the exception of CT-BERT-v2, we can see that if a model labels a pair as unsupported, it is highly 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 which removes the advantage. That is, if DistilRoBERTa^{p} were trained on a paraphrase corpus related to COVID-19, 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 [56]. 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.

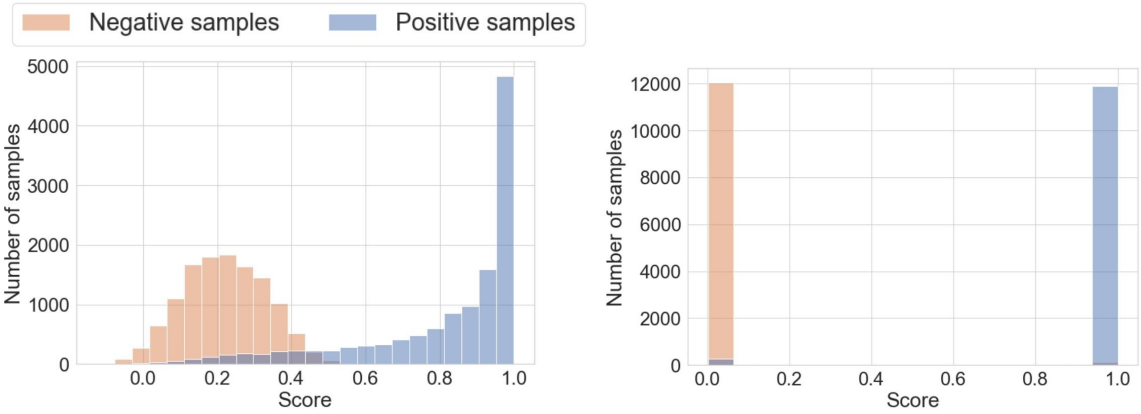
Throughout our experiments, each ⟨Tweet, news⟩ pair – whether sentence-by-sentence or as the entire article – was put through a binary classifier, and the classification probability scores were used to determine the final label. A question may be raised at this point, regarding the choice of the threshold probability score (0.5) that works as the decision boundary. Here, 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).



(a) F1 score

(b) Number of unsupported Tweets

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



(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 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.

A higher threshold correlates with a slightly higher specificity, thus capturing more samples with deceptive hyperlinks. This is shown in Figure 8. There is, however, also a corresponding mild increase in incorrectly labeling samples as *unsupported*. Thus, the corresponding increase in the F_1 score is not statistically significant.

5.3 Additional Experiments and Discussion

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. In order to investigate this question, we design an additional experiment where the Tweet and the corresponding cited headline are converted to vectors, and 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 Fig 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, on the other hand, 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

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]. As an example, Jain et al. [33] collected and clustered the Tweets, found similar content from credible news channels as ground-truth information, and then, they compared the Tweet's content to the reliable content by sentiment and semantic analysis. In case of a mismatch, the authors labeled the Tweet as misinformation. In this body of work, a fixed set of sources were 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], that rely on large-scale manual annotations – a particularly time-intensive approach to resolve a time-sensitive issue.

Accessing high-quality data is crucial in detecting misinformation in social networks by machine learning techniques. Various research attempted to address this challenge. Banda et al. [7] released a very large open-source dataset with more than 383 million Tweets. The dataset includes only the Tweet IDs but is accompanied by the required scripts to rehydrate the Tweets, *i.e.*, retrieve the contents of a Tweet through the use of the Twitter API. The original dataset contains both Tweets and retweets, which allows tracking the dissemination of Tweets; but, a cleaned version has been released as well that has no retweets, which is suitable for analysis of the context of the Tweets. On average, this cleaning step removes 75% of the Tweets. This work does not detect misinformation, but the dataset they published is invaluable to others who intend to research misinformation and examine ML models for this purpose.

Multilingual Datasets. While most released COVID-19 Twitter datasets are in English, the dataset released by Banda et al. [7] includes Tweets in other languages, such as French, German, Russian, and Spanish. Gao et al. [21] released another multilingual dataset of English and Japanese posts on Twitter, and Chinese posts on *Weibo*, while Alqurashi et al. [4] released an Arabic COVID-19 dataset of Tweets. Haouari et al. [27] presented a large Arabic language dataset of Tweets related to COVID-19, along with the propagation networks. In English language datasets, propagation has been studied extensively. Rumor propagation patterns have been studied for several years now, with application in early detection, determining support, and determining their veracity [25, 57], while for other languages it is not well-studied. In this study, we limit the scope of our work to only English Tweets without losing generality.

6.1 Misinformation Detection

Memon and Carley [44] manually annotated more than 4.5K COVID-19-related Tweets. The dataset has a diverse set of categories for 17 types of information and misinformation; *i.e. Irrelevant*, *Conspiracy*, *True Treatment*, *Fake Cure*, *Fake Treatment*, *etc.* One cause for concern is that the data has been annotated by only one annotator. In this work, they 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 may 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* towards 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, increases 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.* due to 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 confirm 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 during three 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. 200 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 and then they added meta-data obtained from the users posting Tweets with misinformation, like timelines of friends and followers, to realize the extent of the spread of misinformation locally. In this approach, there should be replies in response to a misinformation Tweet with 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 [22], and manually annotated 100 randomly selected Tweets. The Hindi dataset has been 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 approach; in general, meaning that the set of labels in the training data and the set of labels for the data that the model will be used to classify are disjoint [74]. To perform zero-shot learning in this work, 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 Multi-Layer Perceptron (MLP), Random Forest Classifier (RFC), SVM, and mBERT 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 [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, 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, and Support Vector Machines, and Multilayer perceptron (MLP)) have been used. In this study, the authors made three important observations. First, the Random Forest classifier outperformed all other models, including the MLP model, with respect to four evaluation metrics, accuracy, precision, recall, and F1-score. Positive correlation between the sentiment of a Tweet and its authenticity, meaning that Tweets with positive sentiment are more likely to be authentic and Tweets with negative sentiment most probably contain misinformation, and the positive effect of metadata on performance are two other observations. In our work, we do not use metadata as we focus on investigating the connection between the Tweet text and the cited news article.

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

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 number of characters, swear words, pronouns, positive and negative emoticons, number of retweets and replies by the users, and 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 a course of three months, and 5.4 million Tweets credibility scores computed. They observed that features extracted from the Tweets content are more effective in credibility assessment compared to the features 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 work is not focused on misinformation detection but can be considered as a data filtering step needed for fake news detection. The authors defined an informative Tweet as it 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 1st to June 30th, about 23M non-repeating Tweets related to COVID-19 have been gathered. Authors filtered this corpus by particular keywords like “positive”, “discharge”, “death”, *etc.* to separate candidates for informative Tweets. Among this dataset, a random sample set of 2K Tweets are 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 are 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 out of 55 participants outperform the baseline model; most of the teams are benefiting from pre-trained language models such as BERT, RoBERTa, XLNet, *etc.* 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 the accuracy of 91.50%. This work confirms 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 which 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.

6.2 Misinformation spread analysis

Huang and Carley [31] collected more than 67 million Tweets from 12 million users with metadata related to geographical information, social identities, and the political orientation of users by tracking COVID-19 Twitter conversations. The data includes metadata related to geographical information, social identities, and the political orientation of users. By analyzing the information about these 12 million users, they reported that misinformation

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

is more likely to be spread by regular users and within the source country, not internationally. In addition, they reported that many of the Tweets speaking of disinformation storylines and referring to unreliable news sites, are posted by regular users, some of them are bots. Similarly, others 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 are involved either by helping to spread or creating the content. The authors have not proposed a ready-to-use model that can be applied for misinformation detection tasks but their approach and the parameters they used for analysis can be considered in future works.

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 and reported that false information on Twitter tends to be retweeted by many more users and gets spread much faster compared to true information, especially when it is about a political issue.

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 infectious individual will create. They concluded that both questionable and reliable news spread with similar diffusion patterns, which indicates that it is impossible to detect fake news solely using meta-data, and analyzing the language and the content is crucially important.

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 focus here has been on Twitter posts pertaining to 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 whether or not the apparent support is deceptive. Our approach relies on distant supervision, and shows that this is a viable option in the face of a dearth of annotated data. Our findings reveal that a significant fraction of check-worthy claims – 27.5% of the annotated sample – 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.

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 data sets and seek collaborators to gain access to other social media platforms like Facebook or WhatsApp, where misinformation has been a highly discussed issue [20, 40, 70]. Our study indicates that in order to fight such 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. Beyond the claims, our work will also be extended to study counterclaims and counter-beliefs expressed in social media in the form or replies to posts or comments. Analyzing the stance, emotive content, and argumentation in such

responses, will offer methodological and epistemic breadth to our understanding of misinformation. By offering a holistic view of the issues pertaining to misinformation, we hope that this work, along with our future endeavors, will help us all to discover the truth in a timely fashion.

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 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 Second Workshop on Fact Extraction and VERification (FEVER). Association for Computational Linguistics, Hong Kong, China, 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

[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, USA, 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 2020) 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 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

[10] Steven Bird, Edward Loper, and Ewan Klein. 2009. Natural Language Processing with Python. O'Reilly Media Inc., Sebastopol, CA, USA. https://www.nltk.org/

[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. Scientific Reports 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 8th International Conference on Learning Representations. OpenReview.net, Addis Ababa, Ethiopia, 18 pages.

[13] Mark Craven and Johan Kumlien. 1999. Constructing Biological Knowledge Bases by Extracting Information from Text Sources. In Proceedings of the Seventh 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, Florence, Italy, 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 Coling 2010: Posters. COLING 2010 Organizing Committee, Beijing, China, 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. Proceedings of the National Academy of Sciences 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 2019 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, MN, USA, 4171–4186. https://doi.org/10.18653/v1/n19-1423

[18] Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. Psychological Bulletin 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, USA, 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
[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 2017 IEEE International Conference on Healthcare Informatics (ICHI). IEEE, Park City, UT, USA, 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 Social Informatics - 6th International Conference (Lecture Notes in Computer Science, Vol. 8851). Springer, Barcelona,Spain, 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, San Diego, California, 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 Proc. First Workshop on Fact Extraction and VERification (FEVER). ACL, Brussels, Belgium, 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 Sixth Arabic Natural Language Processing Workshop. Association for Computational Linguistics, Kyiv, Ukraine (Virtual), 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, USA, 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. 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
[32] Muhammad Imran, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. 2015. Processing social media messages in mass emergency: A survey. ACM Computing Surveys (CSUR) 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 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI). IEEE, Jaipur, India, 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. 2016. Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759 (2016).
[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
[37] Hyunuk Kim and Dylan Walker. 2020. Leveraging volunteer fact checking to identify misinformation about COVID-19 in social media. Harvard Kennedy School Misinformation Review 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 2012 IEEE 28th International Conference on Data Engineering. IEEE, Washington D.C., USA, 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. The Lancet Digital 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
[42] Youness Madani, Mohammed Erritali, and Belaid Bouikhalene. 2021. Using artificial intelligence techniques for detecting Covid-19 epidemic fake news in Moroccan tweets. Results in Physics 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, Avignon, France, 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, Galway, Ireland, 9 pages.
[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 Public Health and Surveillance 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, Suntec, Singapore, 1003–1011. https://www.aclweb.org/anthology/P09-1113
[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 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). Association for Computational Linguistics, New Orleans, Louisiana, 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
[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 Sixth Workshop on Noisy User-generated Text (W-NUT 2020). Association for Computational Linguistics, Online, 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 Proc. AAAI Conference on Artificial Intelligence, Vol. 33. 6859–6866. https://doi.org/10.1609/aaai.v33i01.33016859
[51] Yeimer Ortiz-Martínez and Luisa F Jiménez-Arcia. 2017. Yellow fever outbreaks and Twitter: Rumors and misinformation. American Journal of Infection Control 45, 7 (2017), 816–817.
[52] Sunday Oluwafemi Oyeyemi, Elia Gabarron, and Rolf Wynn. 2014. Ebola, Twitter, and misinformation: a dangerous combination? BMJ 349 (2014), g6178. https://doi.org/10.1136/bmj.g6178
[53] Gordon Pennycook, Tyrone D Cannon, and David G Rand. 2018. Prior exposure increases perceived accuracy of fake news. Journal of experimental psychology: general 147, 12 (2018), 1865. https://doi.org/10.2139/ssrn.2958246
[54] 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/
[55] 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/
[56] Horst Po¨ ttker. 2003. News and its communicative quality: The inverted pyramid—when and why did it appear? Journalism Studies 4, 4 (2003), 501–511.
[57] Vahed Qazvinian, Emily Rosengren, Dragomir R. Radev, and Qiaozhu Mei. 2011. Rumor has it: Identifying Misinformation in Microblogs. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Edinburgh, Scotland, UK, 1589–1599.
[58] 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/
[59] World Health Organization. 2020. Munich Security Conference. Retrieved April 27, 2021 from https://www.who.int/director-general/speeches/detail/munich-security-conference
[60] 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
[61] Landis. J. Richard 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. Proceedings of the Association for Information Science and Technology 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
[64] Gautam Kishore Shahi, Anne Dirkson, and Tim A. Majchrzak. 2021. An exploratory study of COVID-19 misinformation on Twitter. Online Social Networks and 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. Journalism Practice 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]

[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 Library 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 iConference 2014 Proceedings. iSchools, Urbana-Champaign, Illinois, 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. Critical Review 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., Long Beach, CA, USA, 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 Transactions on Intelligent Systems and Technology (TIST) 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., Vancouver, BC, Canada, 5754–5764.

[76] John Zarocostas. 2020. How to fight an infodemic. The 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, USA, 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 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, Online, 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