Human-in-the-loop Evaluation for Early Misinformation Detection: A Case Study of COVID-19 Treatments

Ethan Mendes, Yang Chen, Wei Xu, Alan Ritter
Georgia Institute of Technology
{emendes3, yangc}@gatech.edu {wei.xu, alan.ritter}@cc.gatech.edu

Abstract

We present a human-in-the-loop evaluation framework for fact-checking novel misinformation claims and identifying social media messages that support them. Our approach extracts check-worthy claims, which are aggregated and ranked for review. Stance classifiers are then used to identify tweets supporting novel misinformation claims, which are further reviewed to determine whether they violate relevant policies. To demonstrate the feasibility of our approach, we develop a baseline system based on modern NLP methods for human-in-the-loop fact-checking in the domain of COVID-19 treatments. We make our data[1] and detailed annotation guidelines available to support the evaluation of human-in-the-loop systems that identify novel misinformation directly from raw user-generated content.

1 Introduction

As many people now get information from social networking websites such as Facebook and Twitter, misinformation has become a serious societal problem. To address this, social media companies have spent billions on content moderation.[2] Prior work on developing natural language processing systems to combat misinformation has mainly focused on various sub-tasks (Lee et al., 2021; Guo et al., 2022), including claim detection (Eger et al., 2017; Li et al., 2022), evidence retrieval (Jiang et al., 2020; Samarinas et al., 2021; Wan et al., 2021; Aly and Vlachos, 2022), fact verification (Aly et al., 2021; Wu et al., 2022; Chen et al., 2022; Gu et al., 2022), stance classification (Thorne et al., 2017; Conforti et al., 2018; Li et al., 2019), and fallacy recognition (Alhindi et al., 2022). Researchers have also attempted to perform early detection of novel misinformation claims (Yue et al., 2022), as it is crucial for supporting early interventions such as pre-bunking (Lewandowsky and Van Der Linden, 2021). However, evaluations are often set up automatically using datasets that were retrospectively constructed based on a predefined set of debunked claims.

Recent work by Glockner et al. (2022) presented convincing evidence that existing NLP fact-checking pipelines are unsuitable for detecting novel real-world misinformation. They show these systems rely on leaked counter-evidence from news sources that have already fact-checked the claim. In general, it is unrealistic to assume this type of evidence will be available for new claims that have not yet been widely spread.

In this paper, we address this challenge by presenting a more realistic human-in-the-loop detection and evaluation framework that can measure a system's capabilities for detecting novel check-worthy claims *in the wild* (see Figure 1). We focus on discovering new, domain-specific claims from raw tweets which are then verified by humans, rather than relying on a pre-defined list of claims that have already been fact-checked for evaluation. More importantly, we consider not only the *accuracy* but also the *volume*, *relevance*, and *timeliness* of misinformation claims automatically identified by a system, given a collection of raw tweets. We argue this approach provides more realistic experimental conditions because **(1)** it does not rely on leaked counter-evidence from claims that have already been fact-checked, **(2)** human expertise is vital in verifying the truthfulness of claims (Nakov et al., 2021; Karduni et al., 2018) and **(3)** it is more effective for humans to check aggregated claims within a specific domain (e.g., claims about the efficacy of COVID-19 treatments), before proceeding to individual social media messages to determine if they violate specific misinformation policies.

We validate our methodology for end-to-end misinformation detection in the domain of COVID-19

[1] https://github.com/ethanm88/hitl-evaluation-early-misinformation-detection
[2] https://www.cnbc.com/2021/02/27/content-moderation-on-social-media.html

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 15817–15835

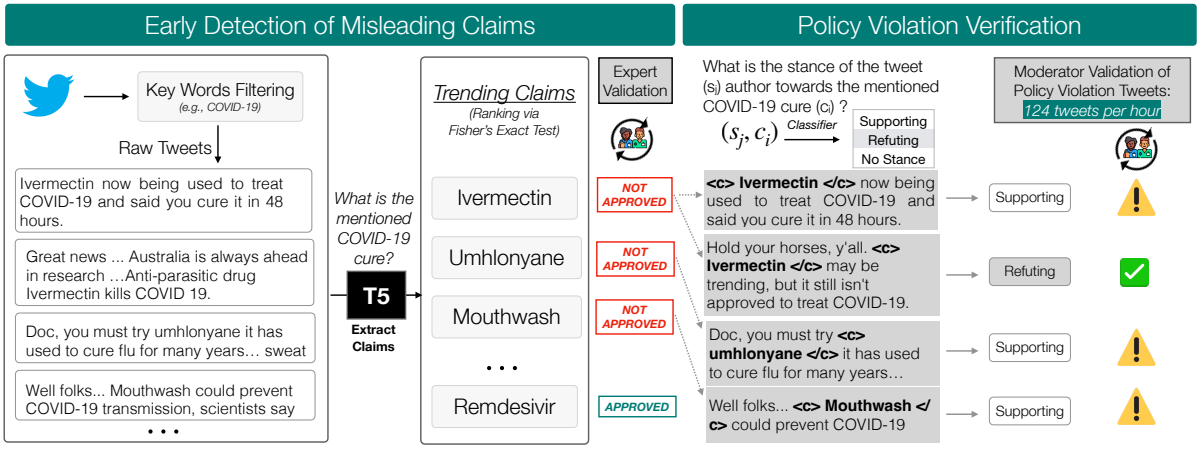

Figure 1: Overview of our human-in-the-loop evaluation framework for early misinformation detection. In stage one (left), a system extracts check-worthy claims directly from raw tweets *in the wild* (rather than retrieving relevant tweets based on provided claims), then aggregates trending claims to be validated by human experts. In stage two (right), the system classifies authors' stances toward false claims and flags tweets for further manual inspection.

treatments. COVID-19 treatments make an ideal testbed for human-in-the-loop misinformation extraction because Twitter has provided clearly defined policies in this area, which we use as guidelines in a realistic human evaluation of a system's output.[3] We evaluate our baseline system with our four defined metrics and find that 18% of the top-50 trending claims were actually misleading (*relevance*), 50% of new misleading claims (unapproved COVID-19 treatments) are detected before they are debunked by journalists in a news article (*timeliness*), 65% of tweets flagged constitute policy-violations (*accuracy*), and an average of 124 policy violations can be confirmed by a human-annotator per hour (*volume*) when using our system.

Our work fills an important gap in the literature, by showing that it is possible to construct a realistic end-to-end evaluation that supports the early detection of novel rumors directly from raw data. Instead of classifying individual tweets as rumorous or not, we extract phrase-level claims that can be aggregated and ranked across a large amount of data and thus can be reviewed more time-efficiently by fact-checkers for human evaluation and for real-world applications. Tweets that are automatically classified as supporting misinformation claims can then be reviewed to determine whether they violate relevant policies.

[3] https://tinyurl.com/CovidMisinformationPolicy

2 Related Work

There is a large body of misinformation-related research. Due to space limitations, we only highlight the most relevant work. See also the excellent surveys by Nakov et al. (2021) and Guo et al. (2022).

2.1 Detecting Check-worthy Claims

One of the most related works to ours is the CLEF-2022 CheckThat shared-task (Nakov et al., 2022), which evaluates three sub-tasks automatically and separately: **(1)** determine whether a tweet is worth fact-checking; **(2)** given a check-worthy claim in the form of a tweet, and a set of previously fact-checked claims, rank the tweets in order of their usefulness to fact-check; and **(3)** given the text and the title of a news article, determine whether the main claim it makes is true, partially true, false, or other. In contrast, our experimental setup is more realistic as it operationalizes over a large amount (e.g., millions) of raw tweets and requires span-level extraction to identify the exact claims (e.g., claims about the efficacy of COVID-19 treatments) rather than just "claims in the form of a tweet" (e.g., tweets that talk about COVID-19 treatments). We also present an end-to-end human-in-the-loop evaluation of the entire misinformation detection pipeline based on the accuracy, volume, and timeliness of all extracted claims, other than just the automatic intrinsic evaluation of each component separately.

Similar to CLEF CheckThat, there exist many other prior works that treat claim detection (or ru-

mor detection) as a text classification problem by predicting check-worthiness (or rumourousness) given a tweet or sentence. One representative work is ClaimBuster (Hassan et al., 2017) which classifies 20,617 sentences from the U.S. general election debate transcripts as non-factual, unimportant factual, and check-worthy factual. Researchers have also developed other datasets (Diggelmann et al., 2020; Konstantinovskiy et al., 2021; Redi et al., 2019; Thorne et al., 2018) and automatic models (Hansen et al., 2019; Jaradat et al., 2018; Wright and Augenstein, 2020). Another relevant work is by Sundriyal et al. (2022) which identifies claims as text fragments, such as "our wine keeps you from getting #COVID19" and "Better alternative to #DisinfectantInjection". The evaluations are mostly done automatically over a small fixed set (normally at the scale of 1k∼50k) of annotated tweets or sentences.

2.2 Early Rumor Detection

As briefly mentioned in §2.1, rumor detection is also commonly framed as a text classification task. The standard rumor detection setup (Zubiaga et al., 2016; Derczynski et al., 2017; Vosoughi et al., 2017; Gorrell et al., 2019; Shu et al., 2020) considers only accuracy without temporal information in the evaluation. More related to our work is a task called early rumor detection (Liu et al., 2015; Ma et al., 2017; Yu et al., 2017; Ruchansky et al., 2017; Zhou et al., 2019; Xia et al., 2020; Bian et al., 2020), which compares classification model's accuracy at different time points and has been extensively surveyed and discussed by Zeng and Gao (2022). However, as they pointed out, most existing methods were "designed with oversimplification" and evaluated automatically on datasets, such as TWITTER-WEIBO (Ma et al., 2016), PHEME (Zubiaga et al., 2016), and BEARD (Zeng and Gao, 2022), that were constructed retrospectively by collecting social media posts using manually curated search keywords (e.g., names of false treatment) based on a given set of debunked claims (e.g., from `snopes.com`). This setup does not measure systems' capability to discover unseen rumors in the wild as our human-in-the-loop evaluation does. In real-world scenarios, what exactly is needed from a misinformation detection system is to automatically figure out what keywords (e.g., names of potential false treatments) to search for – which we focus on and evaluate in this paper.

2.3 COVID-19 Misinformation Detection

Given the severity and pervasiveness of the issue, there exists a lot of research (not limited to NLP) about COVID-19 misinformation (Hossain et al., 2020; Glandt et al., 2021; Dimitrov et al., 2020; Shahi et al., 2021; Agley and Xiao, 2021; Chen and Hasan, 2021; Biamby et al., 2022). The most related work to ours is the CONSTRAINT shared-task (Patwa et al., 2021) at the AAAI-2021, which considers a binary text classification problem of 10,700 COVID-related tweets about real and fake news, and in particular, the work by Kou et al. (2022) that experimented on this dataset with a human-in-the-loop approach. For each input tweet (e.g., "Ethylene oxide used in COVID-19 testing swabs changes the structure of the DNA of human body"), Kou et al. (2022) asked crowd workers to write out the main message (e.g., "Ethylene oxide somehow damages human DNA"), which is then compared with information extracted from COVID-related fack-checking articles and medical papers to help automatic system to predict the tweet's truthfulness. While they prototyped the interesting idea of human-in-the-loop misinformation detection (Shabani et al., 2021), their design is unrealistic to require humans to manually write one sentence per tweet.

3 Human-in-the-Loop Evaluation Framework

One of the most important functions of a misinformation detection system is to identify new misinformation claims *in the wild*, and in a timely manner. We thus design our evaluation framework to measure not only the *accuracy* but also the *volume*, *relevance*, and *timeliness* of misinformation claims identified by a system, given a large collection of raw tweets (*not* collected based on already debunked claims). See Figure 1 for an overview of our framework.

3.1 Early Detection of Misleading Claims

Problem Definition. Given a large set of tweets $\mathbf{T}$, the goal is to automatically discover novel check-worthy claims, each denoted c_i, and aggregate a ranked list of claims, $C = [c_1, c_2, ..., c_n]$. In this task, we use *trendiness* (e.g. defined with Fisher's Exact Test in §4.1.3) as a factor in ranking claims, where a more popular or widely discussed claim will have a greater trendiness. More formally, a novel check-worthy claim c_i is charac-

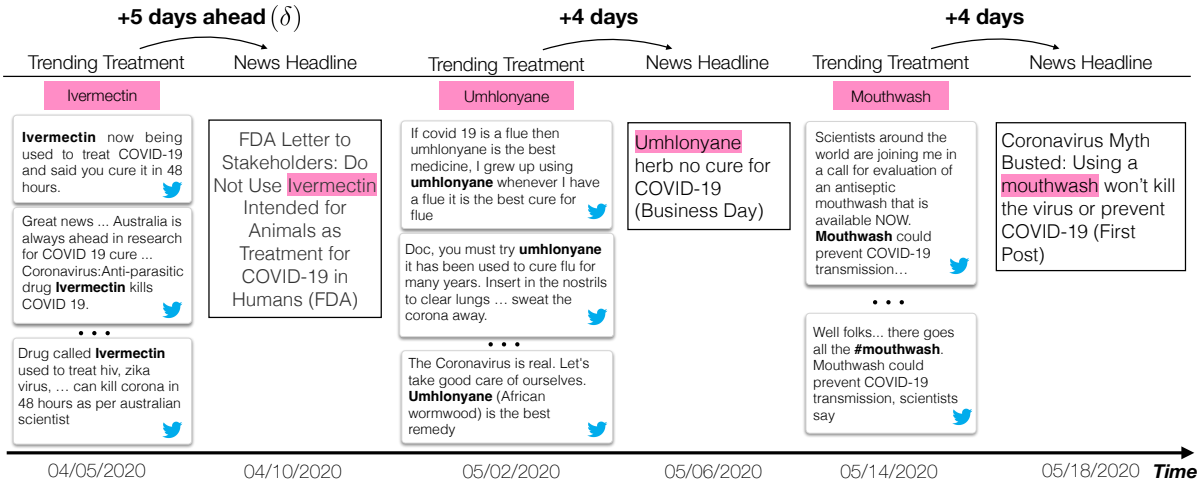

Figure 2: Examples of trending claims regarding notable unapproved treatments detected by our system. We also show the headline of the news reference identified by our annotators in the mock content moderation experiment. For each claim, two/three tweets are displayed that were classified as SUPPORTING from the date that the trend was detected.

terized by $(t_i, z_i, \mathbf{S}_i)$, where t_i is the first time the system identified the claim as trending i.e. when its trendiness first broke a set threshold, z_i is the claim's trendiness score at time t_i, and $\mathbf{S}_i \subset \mathbf{T}$ is the set of tweets supporting the claim. A filtering heuristic can also be applied to remove obvious non-misleading claims from consideration (e.g. filtering out claims supporting approved COVID-19 treatments in §4.1.3).

Evaluation Metrics. A human-in-the-loop evaluation is performed over the top-K trending claims in which annotators verify whether a claim is misleading and, if so, find the earliest news article debunking the claim. The evaluation is based on two metrics: **(1)** the percentage of misleading claims in the top-K trending claims (*relevance*) and **(2)** the number of days (or hours), denoted as δ, between t_i and the publication date of the earliest news article (*timeliness*). Figure 2 visually depicts the application of the δ metric for COVID-19 treatment misinformation. See §4.2 for details about this case study evaluation.

By annotating at the claim level before the tweet level, we reduce human annotator workload by limiting the total number of tweets evaluated. This approach also makes our framework more realistic and efficient, allowing for a more thorough and accurate evaluation of misinformation detection systems. In order to facilitate future research in this area, we provide a collection of raw tweets to evaluate and a baseline system to compare against. As most existing systems are not available as open-source, the release of our evaluation platform will enable fair and comparable evaluations of these systems.

3.2 Policy Violation Verification

Problem Definition. The objective of this task is to identify tweets within the set associated with a claim c_i, $\mathbf{S}_i = \{s_1, s_2, ..., s_{|\mathbf{S}_i|}\}$, that violate a misleading information policy. In general, a tweet s_j, is likely to violate a policy if it expresses a strong supportive stance towards a claim c_i, which was identified as misleading by the human-in-the-loop evaluation process from the prior stage (§3.1).

Evaluation Metrics. To evaluate a system's performance and effectiveness, a human-in-the-loop evaluation is performed on a random sample of N tweets that express a supportive stance toward misleading claims. The evaluation is based on two metrics: **(1)** the *accuracy* of the system in identifying policy-violation tweets and **(2)** the *volume* in terms of the number of policy violations found per hour by analysts using the system.

To measure the *accuracy* of the system, human annotators assign a score to each tweet in the sample, based on a five-point Likert scale, with 5 corresponding to a clear violation of the policy and 1 representing a clear non-violation. We set a threshold of score ≥ 4 to make a binary policy violation determination. This scoring scheme allows us to measure the system's accuracy based on the distribution of annotator scores for all tweets in the sample.

To quantify the *volume* of policy violations identified by analysts using the system, we define the metric *policy-violations per hour* as the number of tweets identified by the annotator containing policy violations, divided by the total number of hours spent by the annotator during the two-stage annotation process (§3.1 and 3.2):

$$\textit{policy violations / hr} = \frac{V}{C \times r_c + T \times r_t}$$

where V is the number of policy violations found, C and T are the numbers of claims and tweets checked respectively, and r_c and r_t are the average annotation rates for claims and tweets respectively.

This metric allows us to assess the efficiency of the system in identifying policy-violation tweets and the potential benefits for content moderators using the system.

4 A Case Study: COVID-19 Treatment Misinformation

To illustrate the usage of our human-in-the-loop evaluation framework outlined in §3, we present a case study for COVID-19 misinformation. Specifically, we target Twitter's COVID-19 policy on unapproved treatments, which states that:

> "*False or misleading information suggesting that unapproved treatments can be curative of COVID-19*"

are grounds for labeling tweets with corrective information (Twitter, 2021).

4.1 Our System

In this subsection, we describe the three components of our COVID-19 treatment misinformation detection system: claim extraction, stance classification, and claim ranking with their task-specific *intrinsic* performance. Later, in §4.2, we present an *extrinsic* human-in-the-loop evaluation of the entire system using our defined framework (§3).

4.1.1 Extracting Check-Worthy Claims

Data. We train and evaluate our claim extraction models on the human-annotated Twitter COVID-19 event extraction dataset created by Zong et al. (2022), which is collected between 2020/01/15 and 2020/04/26. In this work, we focus on claims of the form "*X is an effective COVID-19 treatment*", where *X* is an extracted span. We split the provided 1,271 training tweets in the CURE & PREVENTION category into 60% for training and 15% for development, and report token-level F1 scores on the 500 tweets used for evaluation in the 2020 W-NUT shared task.[4]

Models. We develop three approaches, outlined below, to extract claims as a text span from a sentence with a sequence tagging model and a question-answering model (Rajpurkar et al., 2018; Du et al., 2021). Details of training hyperparameters can be found in Appendix D.1.

(1) Sequence Tagging: a standard sequence-labeling task with a BIO tagging scheme, where 'B' and 'I' tags are used to identify treatment tokens. We follow a similar approach as the named entity recognition method used by Devlin et al. (2019) and experiment with two pre-trained models, including RoBERTa$_{large}$ (Liu et al., 2019) and a domain-specific COVID-Twitter-BERT$_{large}$(CT-BERT) (Müller et al., 2020).

(2) Question-answering (QA): we treat the claim extraction as a question-answering task and apply the SQuADv2.0 (Rajpurkar et al., 2018) formulation as some tweets may not include relevant claims (similar to unanswerable questions). We experiment with two approaches: a span-prediction model that predicts start and end positions for answer spans in context using RoBERTa/CT-BERT as the encoder, in addition to a text-to-text model that generates answers using T5$_{large}$ (Raffel et al., 2020; Wang and Lillis, 2020). The question template for extracting treatments discussed in tweets is "What is the mentioned COVID-19 cure?".

(3) QA-Pretraining: it has been shown that intermediate task pre-training can yield further gains for low-resource target tasks (Pruksachatkun et al., 2020; Poth et al., 2021). We thus experiment with pre-training QA models on the SQuADv2.0 dataset before fine-tuning on the claim extraction dataset.

Intrinsic Evaluation. Table 1 shows the claim extraction results on the COVID-19 treatment dataset. We observe QA models outperform tagging models across encoders. RoBERTa outperforms the domain-specific encoder (CT-BERT) for QA extraction. However, after QA pre-training, CT-BERT improves from 53.1 to 63.8 F_1 and outperforms RoBERTa by 2.3 points. Finally, as the T5 model achieves the best F_1 regardless of QA pre-training, we use T5$_{\text{SQuADv2 Pre-train}}$ as our final

[4] https://noisy-text.github.io/2020/extract_covid19_event-shared_task.html

Approach	Model	F_1
Tagging	RoBERTa	50.3
	CT-BERT	51.2
QA	RoBERTa	61.5
	CT-BERT	53.1
	T5	63.7
$QA_{\text{Pre-train}}$	$\text{RoBERTa}_{\text{SQuADv2 Pre-train}}$	59.9
	$\text{CT-BERT}_{\text{SQuADv2 Pre-train}}$	63.8
	$\text{T5}_{\text{SQuADv2 Pre-train}}$	**63.9**

Table 1: Token-level F_1 scores for claim extraction experiments on COVID-19 treatment dataset.

Majority Annotation	#Tweets
SUPPORTING	743
REFUTING	631
NO STANCE	400
No Consensus (NO STANCE)	226
Total	2000

Table 2: Distribution of annotated COVID-19 treatment stance dataset.

claim extraction model.

4.1.2 Task-Specific Stance Classification

Data. Due to the lack of datasets for COVID-19 treatment stance, we annotate a new dataset for our evaluation. To collect relevant tweets we tracked the keywords "cure", "prevention", "virus", and "COVID-19" from November 2020 to December 2020 using the Twitter API. We collected $1,055,559$ tweets and claims extracted using our model (§4.1.1). Out of $97,016$ tweets for which a treatment was able to be extracted, we randomly sample $2,000$ tweets to annotate the author's stance on the effectiveness of the treatment. We paid crowd workers on Amazon MTurk to annotate our data. Each task consists of a tweet with a highlighted treatment. We asked workers to determine the author's stance towards the treatment and select among three options (SUPPORTING, REFUTING, or NO STANCE). We decided not to include additional options for irrelevant and sarcastic tweets due to poor annotator agreement in pilot experiments. Each tweet is labeled by 5 independent crowd workers. Workers were paid $\$0.20$/HIT, which roughly equates to $\$8.50$/hour. A screenshot of the annotation interface is provided in Figure 6 and dataset statistics are summarized in Table 2.

Quality Control. During the annotation process, we monitored the quality of workers' annotations using their agreement with each other and split the data into 10 batches (200 tweets each) to detect poor annotations in the early stages. We calculate the annotation agreement of each worker against the majority of 5 workers. If the worker's agreement is less than 0.75 for a SUPPORTING annotation based on a majority vote, we do not allow them to participate in the subsequent annotation batch. Across all annotations, we find a 0.65 value of Cohen's κ (Artstein and Poesio, 2008) for the inter-annotator agreement between workers. The distribution of the dataset based on the majority vote of workers is shown in Table 2. In the case that there is no majority annotation for a given tweet i.e. the 5 individual annotators are split $2/2/1$ among the three annotation types, we assign the tweet a default annotation of NO STANCE. We randomly split our annotated dataset of 2,000 tweets into a training set of 1,200 tweets, a development set of 400, and a test set of 400.

Models. Using the annotated corpus, we develop classifiers to detect the author's stance toward a treatment. Specifically, given a claim and a tweet (c_i, s_j), our goal is to predict the author's stance $m_i \in \{$SUPPORTING, REFUTING, or NO STANCE$\}$. We experiment with three models including a baseline neural bag-of-words (NBOW) model, RoBERTa_{large} (Liu et al., 2019) and a $\text{COVID-Twitter-BERT}_{large}$ (CT-BERT) (Müller et al., 2020). To indicate the position of the claim in the input, we use relative position encoding (RPE) (Shaw et al., 2018) for the NBOW model. For the pre-trained language models, we add special markers around the claim following the best-performing model from Soares et al. (2019), [ENTITY MARKERS - ENTITY START]. Details of training hyperparameters are in Appendix D.2.

Intrinsic Evaluation. Table 3 presents the results for NBOW, RoBERTa, and CT-BERT. We observe that CT-BERT outperforms all other models, with an F_1 score of 66.7. Generally, we find that these models performed best in classifying tweets with SUPPORTING stance and worst on tweets with a NO STANCE label. This latter result is possibly due to annotators classifying those tweets that might be irrelevant to the task as NO STANCE.

4.1.3 Ranking of Trending Claims

Following Twitter's COVID-19 misinformation policy violation guidelines, we focus on tweets that advocate the efficacy of an unapproved treatment.

Stance	NBOW	RoBERTa	CT-BERT
SUPPORTING	59.8	70.0	**74.9**
REFUTING	32.8	61.9	**70.6**
NO STANCE	**56.0**	45.9	54.7
Aggregate F_1	49.5	59.3	**66.7**

Table 3: F_1 scores on stance classification dataset for classifying author's stance towards the extracted claim.

Thus, we filter out tweets that mention common approved treatments listed in Table 8 in Appendix C, which are prepared according to authorities and news agencies including the CDC and NYT. Upon determining stance on the filtered set, we only consider tweets with a SUPPORTING stance towards the effectiveness of the extracted treatment. Finally, we remove near-duplicates and cluster the remaining extracted treatments based on word overlap (Jaccard similarity) to enable treatment-level decision-making similar to Basu et al. (2013).

Ranking Claims. For the claims (treatments) in each cluster, we count the number of tweets mentioning the claim both daily and cumulatively. Based on these counts, we compute the claim's p-value on a given date using the one-tailed Fisher's Exact Test (Fisher, 1922) which has been shown to be effective in rare event detection (Moore, 2004; Johnson et al., 2007; Ritter et al., 2011). A claim's p-value is a measure of its trendiness denoted z_i (§3.1) by which it is ranked relative to other claims.

Detecting Novel Claims. Based on the results of Fisher's Exact Test, our system automatically detects novel trending claims and flags them for manual inspection by content moderators. A claim is considered as newly trending if it's p-value is less than a preset significance threshold (α-level) and it has never broken this threshold previously (further details in Appendix E). Using the notation from §3.1, if the claim (c_i) found to be newly trending on date t_i is judged to be misleading by a human moderator, our system then provides a list of individual tweets ($\mathbf{S}_i$) that SUPPORT the misleading claim for manual inspection.

4.2 Human-in-the-Loop Evaluation for Detecting COVID-19 Misinformation

In this section, we evaluate the system outlined in §4.1 using the human-in-the-loop evaluation methodology we define in §3. We follow the same procedure described in §4.1.2 to prepare a new dataset containing $14,741,171$ tweets for large-scale evaluation. We then extract treatments using our QA-based claim extractor and apply our stance detection model to classify authors' stance towards each treatment. After removing tweets without an extracted treatment, the resultant evaluation corpus consists of $1,905,424$ tweets.

4.2.1 Early Detection of Misleading COVID-19 Treatments

We first evaluate the ability of our system to detect newly trending misleading COVID-19 treatment claims and report metrics measuring *relevance* and *timeliness* as defined in §3.1.

Data Preparation and Human Evaluation. We set aside tweets collected from 2020/03/01 to 2020/03/31 (1-month time-frame) to serve as an initial base of historical data to compute cumulative counts for detecting novel trending claims using the Fisher's Exact Test. Newly trending treatments are then identified during the time period of 2020/04/01 to 2022/05/05 (2-year time-frame) using the methodology described in §4.1.3. The top 300 treatments are selected based on p-values from Fisher's Exact Test which equates to a significance level of $\alpha = 1.15\text{e}{-6}$.

We employ two in-house annotators, who act as mock content-moderators, to evaluate these 300 treatments and determine **(1)** if the extraction is a treatment **(2)** if the treatment is unapproved and **(3)** the earliest publication date of a news article debunking the treatment as effective for COVID-19 using the Google News engine. Appendix A contains further details about the annotation process.

Out of the 300 treatments, 100 were annotated by both annotators to determine inter-rater agreement on task **(2)**, which was 0.87 as given by Cohen's κ. On average, it took 89.7 seconds to complete each treatment annotation.

Results on Early Detection of Misleading Claims. In terms of *relevance*, Table 4 shows the percentage of the top 5/50/100 trending treatments ranked by p-value that were determined to be unapproved, and Figure 3 shows the cumulative number of potential unapproved trends identified over time along with the total number of trending treatments. To evaluate the *timeliness*, we calculate δ, which measures the number of days our system detects misleading claims before a debunking news article is published (§3.1). We find that our system is able to detect 50% of rumors before the publication date of the relevant news article ($\delta \geq 0$), with the median δ

being 21 days. Figure 2 shows three notable unapproved treatment examples from early in the pandemic with their relevant news article and δ values.

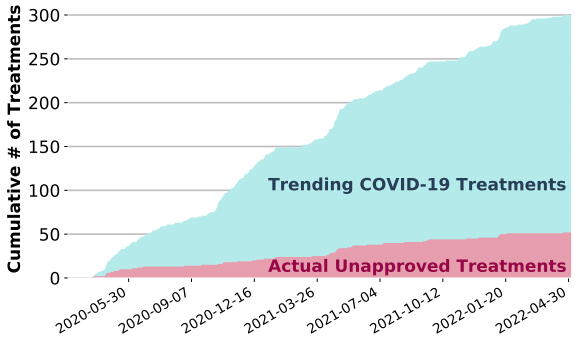

Figure 3: Cumulative number of potential and actual unapproved treatments detected.

	Top 5	Top 50	Top 100
% unapproved	60.0	18.0	14.0

Table 4: Percentage of top 5/50/100 trending treatments based on p-value that were classified as unapproved in the annotation process.

4.2.2 Identifying COVID-19 Policy Violations.

In addition to detecting novel rumors online, we also evaluate the ability of our system to identify tweets that violate Twitter's misleading information policy and report metrics measuring *accuracy* and *volume* as defined in §3.2.

Data Preparation and Human Evaluation. For each of the 40 treatments identified as unapproved in the previous experiment, we randomly sample 10 tweets that have SUPPORTING stance towards the claim. Near duplicate tweets were identified and removed leaving 361 unique tweets. The two in-house annotators then assign a score to each tweet based on a five-point Likert scale, with 5 corresponding to a clear violation of the policy and 1 representing a clear non-violation (See score descriptions in Table 7 in Appendix A).

To investigate the quality of annotations, we compute agreement on 206 tweets using ordinal Krippendorff's α ($0 \leq \alpha \leq 1$) (Krippendorff, 2011).[5] We find that annotators agreed moderately with Krippendorff's $\alpha = 0.54$, indicating fair agreement. On average, it took the annotators 16.1 seconds to annotate each tweet.

[5]We use Krippendorff's α instead of Cohen's κ to measure inter-rater agreement here because it is applicable to ordinal annotation data.

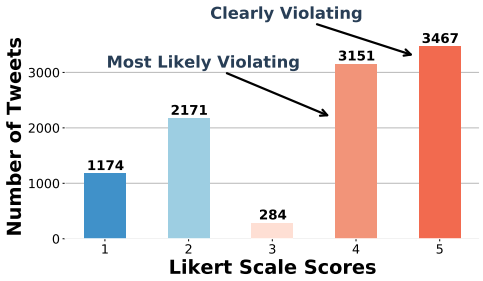

Figure 4: Expected distribution of Likert score annotations on full set of tweets mentioning one of the 40 treatments investigated.

Results on Identifying Policy Violations. In Figure 4, we find that 65% (*accuracy*) of tweets had scores indicating that it was either likely or clearly violating the policy with an average score of 3.54 out of 5. Figure 4 presents the estimated distribution of Likert scores over 10,246 tweets using the 10 annotated tweets sampled for each treatment and extrapolated to all tweets mentioning the treatment. In terms of *volume*, we estimate that an annotator can identify approximately 124.2 policy violations per hour with our system, based on the average annotation rate of the 300 treatments and the same extended set of tweets, where a Likert score of 4 or 5 constitutes a policy violation. See Appendix F for a full calculation of this statistic.

5 Conclusion

In this work, we present a novel end-to-end human-in-the-loop evaluation framework for the early identification of novel misinformation on social media from raw tweets. Unlike previous evaluation frameworks, our methodology captures the interplay between the system and human content moderators while also providing realistic metrics for early misinformation detection. We validate our misinformation detection framework for claims in the domain of COVID-19 treatments. By aggregating and ranking structured representations of claims, and relying on human fact-checkers to review trending claims, our system is able to detect 50% of misleading claims earlier than the news.

6 Limitations

While our approach does require domain-specific information extraction models to extract structured representations of novel misinformation claims for easy aggregation and review, there is significant prior work on event extraction that can be adapted to extract check-worthy claims (Ritter et al., 2012; Luan et al., 2019; Du and Cardie, 2020). Furthermore, we argue content moderators or fact-checkers are likely to be more effective when focusing on one claim type at a time (e.g. COVID-19 treatments, election integrity, vaccine effectiveness, etc.), rather than reviewing a mixture of claims on multiple topics.

Our COVID-19 case study also makes use of "mock" content moderators, rather than employees or contractors working for social media companies or fact-checking websites. However, we believe this methodology still provides valuable insight that would not be publicly available otherwise, as social media companies do not currently publish extensive details about their content moderation processes[6] and fact-checking websites vary widely in policy and have been shown to provide inconsistent claim classification (Marietta et al., 2015). Some prior user studies (Nguyen et al., 2018; Pennycook and Rand, 2019; Shabani et al., 2021) have also shown laypeople (e.g., Amazon Mechanical Turk workers) can be good at judging the veracity of claims or reliability of news articles.

As of late November 2022, Twitter has suspended enforcement of its COVID-19 misleading information policies such as the one we target in this paper.[7] However, per the Associated Press article, one of the possible reasons for the suspension was that Twitter has "struggled to respond to a torrent of misinformation about the virus" with many "bogus claims about home remedies" still on the site despite the previous enforcement of policies. While we do not have details about the internal automated systems Twitter has in place to assist with content moderation, an end-to-end early detection system might have helped stem the spread of misinformation on the platform. Additionally, despite the lack of official policy enforcement, our system can still be used by third-party fact-checking websites or researchers to measure and report misinformation on Twitter. Finally, the main goal of our work is not to create a system for COVID-19 misinformation detection but rather to propose a framework that allows for a fair and realistic evaluation of early misinformation detection systems in any domain.

[6] https://www.nytimes.com/2022/05/19/business/twitter-content-moderation.html

[7] https://apnews.com/article/twitter-ends-covid-misinformation-policy-cc232c9ce0f193c505bbc63bf57ecad6

7 Broader Impact and Ethical Considerations

We release our corpus of tweets annotated with stance, and our dataset of trending misinformation claims under Twitter's Developer Agreement,[8] which grants permissions for academic researchers to share Tweet IDs and User IDs (less than 1,500,000 Tweet IDs within 30 days) for non-commercial purposes, as of October 10th. 2022.

Our system is designed for research purposes and may contain unknown biases towards demographic groups or individuals (Sap et al., 2019). Further investigation into systematic biases should be conducted before our models are deployed in a production environment.

We believe this study helps shed light on how NLP tools developed to help combat online misinformation might be used in a real content moderation workflow. We hope this will encourage future research on human-in-the-loop systems and help shape the design of new tasks and datasets in this area. We believe it is beneficial for some research on combating misinformation to take place outside of social media companies to provide an unbiased view of the challenges involved in fighting online misinformation.

Acknowledgments

We thank anonymous reviewers for their helpful feedback on this work. We also thank Chase Perry and Andrew Duffy for their help with annotations and human evaluation. This research is supported in part by the NSF (IIS-2052498), ODNI and IARPA via the BETTER and HIATUS programs (2019-19051600004, 2022-22072200004). The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of NSF, ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

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

References

Jon Agley and Yunyu Xiao. 2021. Misinformation about COVID-19: evidence for differential latent profiles and a strong association with trust in science. *BMC Public Health*.

Tariq Alhindi, Tuhin Chakrabarty, Elena Musi, and Smaranda Muresan. 2022. Multitask instruction-based prompting for fallacy recognition. In *Proceedings of Empirical Methods in Natural Language Processing*.

Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. The fact extraction and VERification over unstructured and structured information (FEVEROUS) shared task. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*.

Rami Aly and Andreas Vlachos. 2022. Natural logic-guided autoregressive multi-hop document retrieval for fact verification. In *Proceedings of Empirical Methods in Natural Language Processing*.

Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*.

Sumit Basu, Chuck Jacobs, and Lucy Vanderwende. 2013. Powergrading: a clustering approach to amplify human effort for short answer grading.

Giscard Biamby, Grace Luo, Trevor Darrell, and Anna Rohrbach. 2022. Twitter-COMMs: Detecting climate, COVID, and military multimodal misinformation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Tian Bian, Xi Xiao, Tingyang Xu, Peilin Zhao, Wenbing Huang, Yu Rong, and Junzhou Huang. 2020. Rumor detection on social media with bi-directional graph convolutional networks. In *AAAI Conference on Artificial Intelligence*.

CDC. 2021. How to protect yourself and others.

Jifan Chen, Aniruddh Sriram, Eunsol Choi, and Greg Durrett. 2022. Generating literal and implied subquestions to fact-check complex claims. In *Proceedings of Empirical Methods in Natural Language Processing*.

Yuanzhi Chen and Mohammad Hasan. 2021. Navigating the kaleidoscope of COVID-19 misinformation using deep learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Costanza Conforti, Mohammad Taher Pilehvar, and Nigel Collier. 2018. Towards automatic fake news detection: Cross-level stance detection in news articles. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, Brussels, Belgium. Association for Computational Linguistics.

Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. 2017. SemEval-2017 task 8: RumourEval: Determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. Climate-FEVER: A dataset for verification of real-world climate claims. In *Proceedings of the NeurIPS 2020 Workshop: Tackling Climate Change with Machine Learning*.

Dimitar Dimitrov, Erdal Baran, Pavlos Fafalios, Ran Yu, Xiaofei Zhu, Matthäus Zloch, and Stefan Dietze. 2020. Tweetscov19 - a knowledge base of semantically annotated tweets about the covid-19 pandemic. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*.

Xinya Du and Claire Cardie. 2020. Event extraction by answering (almost) natural questions. In *Proceedings of Empirical Methods in Natural Language Processing*.

Xinya Du, Luheng He, Qi Li, Dian Yu, Panupong Pasupat, and Yuan Zhang. 2021. QA-driven zero-shot slot filling with weak supervision pretraining. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*.

Steffen Eger, Johannes Daxenberger, and Iryna Gurevych. 2017. Neural end-to-end learning for computational argumentation mining. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Ronald A. Fisher. 1922. On the interpretation of $\chi 2$ from contingency tables, and the calculation of P.

Kyle Glandt, Sarthak Khanal, Yingjie Li, Doina Caragea, and Cornelia Caragea. 2021. Stance Detection in COVID-19 Tweets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*.

Max Glockner, Yufang Hou, and Iryna Gurevych. 2022. Missing counter-evidence renders nlp fact-checking unrealistic for misinformation. *Proceedings of Empirical Methods in Natural Language Processing*.

Genevieve Gorrell, Elena Kochkina, Maria Liakata, Ahmet Aker, Arkaitz Zubiaga, Kalina Bontcheva, and Leon Derczynski. 2019. SemEval-2019 task 7: RumourEval, determining rumour veracity and support for rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation*.

Zihui Gu, Nan Fan, Ju andTang, Preslav Nakov, Xiaoman Zhao, and Xiaoyong Du. 2022. PASTA: Table-operations aware fact verification via sentence-table cloze pre-training. In *Proceedings of Empirical Methods in Natural Language Processing*.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A Survey on Automated Fact-Checking. In *Transactions of the Association for Computational Linguistics*.

Casper Hansen, Christian Hansen, Stephen Alstrup, Jakob Grue Simonsen, and Christina Lioma. 2019. Neural check-worthiness ranking with weak supervision: Finding sentences for fact-checking. Association for Computing Machinery.

Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. 2017. Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster.

Tamanna Hossain, Robert L. Logan IV, Arjuna Ugarte, Yoshitomo Matsubara, Sean Young, and Sameer Singh. 2020. COVIDLies: Detecting COVID-19 misinformation on social media. In *Proceedings of the 1st Workshop on NLP for COVID-19 at EMNLP*.

Israa Jaradat, Pepa Gencheva, Alberto Barrón-Cedeño, Lluís Màrquez, and Preslav Nakov. 2018. ClaimRank: Detecting check-worthy claims in Arabic and English. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. 2020. HoVer: A dataset for many-hop fact extraction and claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, Online. Association for Computational Linguistics.

Howard Johnson, Joel Martin, George Foster, and Roland Kuhn. 2007. Improving translation quality by discarding most of the phrasetable. In *Proceedings of Empirical Methods in Natural Language Processing*.

Alireza Karduni, Ryan Wesslen, Sashank Santhanam, Isaac Cho, Svitlana Volkova, Dustin Arendt, Samira Shaikh, and Wenwen Dou. 2018. Can you verifi this? studying uncertainty and decision-making about misinformation using visual analytics. In *Twelfth international AAAI conference on web and social media*.

Lev Konstantinovskiy, Oliver Price, Mevan Babakar, and Arkaitz Zubiaga. 2021. Toward automated factchecking: Developing an annotation schema and benchmark for consistent automated claim detection. *Digital threats: research and practice*.

Ziyi Kou, Lanyu Shang, Yang Zhang, Zhenrui Yue, Huimin Zeng, and Dong Wang. 2022. Crowd, expert & ai: A human-ai interactive approach towards natural language explanation based covid-19 misinformation detection. In *International Joint Conferences on Artificial Intelligence Organization*.

Klaus Krippendorff. 2011. Computing Krippendorff's alpha-reliability.

Nayeon Lee, Belinda Z. Li, Sinong Wang, Pascale Fung, Hao Ma, Wen-tau Yih, and Madian Khabsa. 2021. On unifying misinformation detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Stephan Lewandowsky and Sander Van Der Linden. 2021. Countering misinformation and fake news through inoculation and prebunking. *European Review of Social Psychology*.

Manling Li, Revanth Gangi Reddy, Ziqi Wang, Yi-Shyuan Chiang, Tuan Lai, Pengfei Yu, Zixuan Zhang, and Heng Ji. 2022. COVID-19 claim radar: A structured claim extraction and tracking system. In *Proceedings of the Association for Computational Linguistics*.

Quanzhi Li, Qiong Zhang, and Luo Si. 2019. Rumor detection by exploiting user credibility information, attention and multi-task learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy. Association for Computational Linguistics.

Xiaomo Liu, Armineh Nourbakhsh, Quanzhi Li, Rui Fang, and Sameena Shah. 2015. Real-time rumor debunking on twitter. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, New York, NY, USA. Association for Computing Machinery.

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 preprint arXiv:1907.11692*.

Yi Luan, Dave Wadden, Luheng He, Amy Shah, Mari Ostendorf, and Hannaneh Hajishirzi. 2019. A general framework for information extraction using dynamic span graphs. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J. Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*.

Jing Ma, Wei Gao, and Kam-Fai Wong. 2017. Detect rumors in microblog posts using propagation structure via kernel learning. In *Proceedings of the Association for Computational Linguistics*. Association for Computational Linguistics.

Morgan Marietta, David C. Barker, and Todd Bowser. 2015. Fact-checking polarized politics: Does the fact-check industry provide consistent guidance on disputed realities? *The Forum*, 13(4):577–596.

Robert C. Moore. 2004. On log-likelihood-ratios and the significance of rare events. In *Proceedings of Empirical Methods in Natural Language Processing*.

Martin Müller, Marcel Salathé, and Per E Kummervold. 2020. COVID-Twitter-BERT: A natural language processing model to analyse COVID-19 content on Twitter. In *arXiv preprint arXiv:2005.07503*.

Preslav Nakov, Alberto Barrón-Cedeño, Giovanni da San Martino, Firoj Alam, Julia Maria Struß, Thomas Mandl, Rubén Míguez, Tommaso Caselli, Mucahid Kutlu, Wajdi Zaghouani, Chengkai Li, Shaden Shaar, Gautam Kishore Shahi, Hamdy Mubarak, Alex Nikolov, Nikolay Babulkov, Yavuz Selim Kartal, Michael Wiegand, Melanie Siegel, and Juliane Köhler. 2022. Overview of the clef–2022 checkthat! lab on fighting the covid-19 infodemic and fake news detection. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, Cham. Springer International Publishing.

Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barr'on-Cedeno, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021. Automated fact-checking for assisting human fact-checkers. In *International Joint Conference on Artificial Intelligence*.

An T. Nguyen, Aditya Kharosekar, Saumyaa Krishnan, Siddhesh Krishnan, Elizabeth Tate, Byron C. Wallace, and Matthew Lease. 2018. Believe it or not: Designing a human-ai partnership for mixed-initiative fact-checking. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*. Association for Computing Machinery.

Parth Patwa, Mohit Bhardwaj, Vineeth Guptha, Gitanjali Kumari, Shivam Sharma, Srinivas PYKL, Amitava Das, Asif Ekbal, Md Shad Akhtar, and Tanmoy Chakraborty. 2021. Overview of constraint 2021 shared tasks: Detecting english covid-19 fake news and hindi hostile posts. In *Combating Online Hostile Posts in Regional Languages during Emergency Situation*.

Gordon Pennycook and David G. Rand. 2019. Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences*.

Clifton Poth, Jonas Pfeiffer, Andreas Rücklé, and Iryna Gurevych. 2021. What to pre-train on? Efficient intermediate task selection. In *Proceedings of Empirical Methods in Natural Language Processing*.

Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel R. Bowman. 2020. Intermediate-task transfer learning with pretrained language models: When and why does it work? In *Proceedings of the Association for Computational Linguistics*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the Association for Computational Linguistics*.

Miriam Redi, Besnik Fetahu, Jonathan Morgan, and Dario Taraborelli. 2019. Citation needed: A taxonomy and algorithmic assessment of wikipedia's verifiability.

Alan Ritter, Colin Cherry, and William B Dolan. 2011. Data-driven response generation in social media. In *Proceedings of Empirical Methods in Natural Language Processing*.

Alan Ritter, Oren Etzioni, and Sam Clark. 2012. Open domain event extraction from twitter. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*.

Natali Ruchansky, Sungyong Seo, and Yan Liu. 2017. CSI: A hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. Association for Computing Machinery.

Chris Samarinas, Wynne Hsu, and Mong Li Lee. 2021. Improving evidence retrieval for automated explainable fact-checking. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, Online. Association for Computational Linguistics.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the Association for Computational Linguistics*.

Shaban Shabani, Zarina Charlesworth, Maria Sokhn, and Heiko Schuldt. 2021. SAMS: Human-in-the-loop approach to combat the sharing of digital misinformation. In *Proceedings of the AAAI 2021 Spring Symposium on Combining Machine Learning and Knowledge Engineering*.

Gautam Kishore Shahi, Anne Dirkson, and Tim A. Majchrzak. 2021. An exploratory study of covid-19 misinformation on twitter. *Online Social Networks and Media*.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-attention with relative position representations. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

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*.

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the Association for Computational Linguistics*.

Megha Sundriyal, Atharva Kulkarni, Vaibhav Pulastya, Md Shad Akhtar, and Tanmoy Chakraborty. 2022. Empowering the fact-checkers! automatic identification of claim spans on twitter. In *Proceedings of Empirical Methods in Natural Language Processing*.

James Thorne, Mingjie Chen, Giorgos Myrianthous, Jiashu Pu, Xiaoxuan Wang, and Andreas Vlachos. 2017. Fake news stance detection using stacked ensemble of classifiers. In *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*.

Twitter. 2021. Covid-19 misleading information policy.

Soroush Vosoughi, Mostafa 'Neo' Mohsenvand, and Deb Roy. 2017. Rumor gauge: Predicting the veracity of rumors on twitter. *ACM Transactions on Knowledge Discovery from Data*.

Hai Wan, Haicheng Chen, Jianfeng Du, Weilin Luo, and Rongzhen Ye. 2021. A DQN-based approach to finding precise evidences for fact verification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Online. Association for Computational Linguistics.

Congcong Wang and David Lillis. 2020. UCD-CS at W-NUT 2020 shared task-3: A text to text approach for COVID-19 event extraction on social media. In *Proceedings of the Sixth Workshop on Noisy User-generated Text*.

Dustin Wright and Isabelle Augenstein. 2020. Claim check-worthiness detection as positive unlabelled learning. In *Findings of the Association for Computational Linguistics: EMNLP*, Online. Association for Computational Linguistics.

Xueqing Wu, Kung-Hsiang Huang, Yi Fung, and Heng Ji. 2022. Cross-document misinformation detection based on event graph reasoning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Rui Xia, Kaizhou Xuan, and Jianfei Yu. 2020. A state-independent and time-evolving network for early rumor detection in social media. In *Proceedings of Empirical Methods in Natural Language Processing*.

Feng Yu, Qiang Liu, Shu Wu, Liang Wang, and Tieniu Tan. 2017. A convolutional approach for misinformation identification. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*.

Zhenrui Yue, Huimin Zeng, Ziyi Kou, Lanyu Shang, and Dong Wang. 2022. Contrastive domain adaptation for early misinformation detection: A case study on covid-19. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*.

Fengzhu Zeng and Wei Gao. 2022. Early rumor detection using neural Hawkes process with a new benchmark dataset. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Kaimin Zhou, Chang Shu, Binyang Li, and Jey Han Lau. 2019. Early rumour detection. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota. Association for Computational Linguistics.

Carl Zimmer, Katherine J. Wu, Jonathan Corum, and Matthew Kristoffersen. 2020. Coronavirus drug and treatment tracker.

Shi Zong, Ashutosh Baheti, Wei Xu, and Alan Ritter. 2022. Extracting a knowledge base of COVID-19 events from social media. *Proceedings of the 29th International Conference on Computational Linguistics*.

Arkaitz Zubiaga, Maria Liakata, and Rob Procter. 2016. Learning reporting dynamics during breaking news for rumour detection in social media. *ArXiv*, abs/1610.07363.

A Annotation Guidelines for the Human-in-the-Loop Evaluation

Early Detection of Misleading COVID-19 Treatments. Given a list of trending claims (e.g., COVID-19 treatments), annotators are required to determine **(1)** if the extraction is a treatment **(2)** if the treatment is unapproved and **(3)** the earliest publication date of a news article they can find that debunks the treatment as effective. Annotators query the Google News engine with a query in the form of *"[treatment] cures COVID-19"* and sort by date to find the earliest published news article starting from 2020/04/01 that debunks the treatment as effective against COVID-19. Treatments are only considered to be unapproved if the annotators can identify a reputable news source as a reference. Table 5 shows the annotation questions and guidelines as they appeared to annotators during the human-in-the-loop evaluation for this task. Note that March 1st, 2020 was used as the starting date for the article search because it was the earliest date for which we had tweet data.

Identifying COVID-19 Policy Violation Tweets. Given a tweet with SUPPORTING stance towards the claim that "*treatment is effective in treating COVID-19*", annotators assign a score based on a five-point Likert scale, with 5 corresponding to a clear violation of the policy and 1 representing a clear non-violation. Table 7 shows the Likert score descriptions and Table 6 shows the annotation questions and guidelines as they appeared to annotators during the human-in-the-loop evaluation for this task.

B Annotation Interface for Stance Classification

Our stance data collection procedures (§4.1.2) on Amazon MTurk were approved by an ethics board. Before individuals were allowed to annotate data for our task, they were required to give consent by electronically signing off on the ethics statement found in Figure 5 (some portions have been redacted for anonymity purposes). Note that all annotators were MTurk workers in the United States who had previously annotated 1000 HITs with a pass-rate $\geq$ 95%. Figure 6 shows the interface used for collecting these stance annotations.

C Approved Treatments

Table 8 shows the approved treatments that were used for filtering in §4.1.3.

D Implementation Details

All experiments are performed with NVIDIA A40 GPUs. All hyperparameters are selected using a held-out development set.

D.1 Claim Extraction Models

Hyperparameters can be found in Table 9.

Sequence Tagging Models. We apply sequence-labeling models with a standard BIO tagging scheme ('B' and 'I' tags are used to identify treatment tokens), similar to the named entity recognition method used by Devlin et al. (2019). We experiment with RoBERTa$_{large}$(354M) (Liu et al., 2019) and a domain-specific COVID-Twitter-BERT$_{large}$(345M) (Müller et al., 2020).

Question Answering Models. We experiment with QA-based slot filling models (Du et al., 2021), which model claim extraction as a SQuADv2.0 question-answering task (Rajpurkar et al., 2018). We experiment with two approaches: a span-prediction model that predicts start and end positions for answer spans in context using RoBERTa or CT-BERT as the encoder, in addition to a text-to-text model that generates answers using T5$_{large}$(770M) (Raffel et al., 2020; Wang and Lillis, 2020).

Pre-training QA Models. We pre-train QA models on the SQuADv2.0 dataset for 2 epochs with a learning rate of 2e-5 and batch size of 16, followed by fine-tuning for claim extraction.

D.2 Stance Classification Models

Hyperparameters can be found in Table 10.

NBOW. We use an NBOW model and a relative position encoding (RPE) (Shaw et al., 2018) to indicate the position of the claim in the input sentence. Specifically, the 1D RPE encodes the relative distance of each token to the extracted treatment. We then concatenate NBOW with the RPE embedding and pass the concatenation through one layer of a feed-forward neural network.

RoBERTa/CT-BERT. We finetune a RoBERTa$_{large}$ (Devlin et al., 2019) and COVID-Twitter-BERT (CT-BERT) (Müller et al., 2020)

Question	Annotation Guidance
Should the extraction be considered for consideration of a new treatment?	• Answer “Repeat” if the same trend has been seen previously. • Answer“Approved” if treatment should have been marked as approved based on the approved trending list or is otherwise an obvious valid treatment. • Answer “Unsure” if the treatment was in clinical trials at the time or you are otherwise unsure if the treatment is a valid cure. • Answer “Not a Treatment” for extraction errors, preventative measures, etc. • Answer “General health advice" for any general strategies for staying healthy such as “hand washing”, “exercise”, “hygiene”, etc. • Answer “Unapproved” otherwise.
What is the date of publication of the earliest article stating that it is misinformation that this treatment cures COVID-19?	• Answer “NA” if you answered “Approved”, “Unsure”, “Repeat” in the previous question or if no such article can be found in the time frame of [03/01/2020 - Present]. Otherwise, provide the date of the earliest article found. • **How to search for articles:** 1. If X is the treatment, search: “X cures COVID-19” on the News tab in Google - correct any obvious misspellings - try some obvious variations if do not find appropriate results (e.g. “hcq” and “hydroxychloroquine”) 2. Set date range to [03/01/2020 - Present] (or narrow the date range i.e. month by month range if too many search results) 3. Select the option to order results by date 4. Go to the last page in search results (earliest) 5. Find the earliest article that debunks the claim - answer “NA” if no appropriate article is found • Make sure to verify the correct date of the article publication from the article webpage as the date on google news is not always updated and reliable
What is a link to the article?	• Article URL or "NA" is the answer to the previous question was "NA"

Table 5: Human evaluation question and guidelines provided to annotators during the early detection of misleading COVID-19 claims task

Question	Annotation Guidance
Is the tweet a duplicate (already seen)?	• “Yes” or [BLANK]
Does this tweet violate Twitter’s COVID-19 unapproved treatment policy?	• Answer “NA” if previous answer was [BLANK], otherwise answer “1” - “5” based on the attached table (Table 7)

Table 6: Human evaluation question and guidelines provided to annotators during the COVID-19 policy violation verification task

using a linear classification layer. We use the best-performing model from Soares et al. (2019), [ENTITY MARKERS - ENTITY START], which uses special tokens $[\mathtt{C}_{\mathtt{start}}], [\mathtt{C}_{\mathtt{end}}]$ to mark treatment span in a sentence. The modified sentence is then fed into BERT/CT-BERT and the representation of the starting marker $[\mathtt{C}_{\mathtt{start}}]$ is sent into a linear classification layer.

E Ranking Extracted Claims

To generate a list of treatments mentioned on a specific day sorted by trendiness or significance we cannot simply use the daily frequency counts of treatments mentioned, as more popular treatments will consistently be mentioned at a higher volume. In our evaluation dataset used in §4.2, for example, chloroquine and its variants are mentioned in a tweet with SUPPORTING stance approximately

"You are being asked to be a volunteer in a research study. The purpose of this study is to advance research on Computational Linguistics. The annotation form will take approximately 3 minutes to complete. You must be 18 years of age or older to participate. Your judgments will be used by researchers worldwide to help advance research on Computational Linguistics. They will enable machine learning techniques to be applied to problems in natural language understanding. We will keep your personal information (Mechanical Turk ID, etc.) confidential. The risks involved are no greater than those involved in daily activities. We will comply with any applicable laws and regulations regarding confidentiality. To make sure that this research is being carried out in the proper way, the [REDACTED] may review study records. The [REDACTED] may also look at study records. If you have any questions about the study, you may contact [REDACTED]. If you have any questions about your rights as a research subject, you may contact [REDACTED] at [REDACTED]. Thank you for participating in this study. By completing the online survey, you indicate your consent to be in the study. Subjects located in the EU are not allowed to join this study."

Figure 5: Amazon MTurk ethics statement which was shown to annotators before stance labeling task

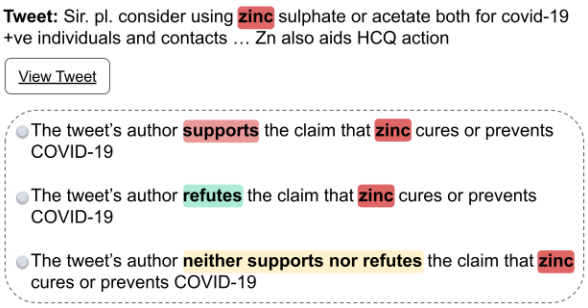

Figure 6: Amazon MTurk interface for stance annotation towards extracted claims in tweets.

Score	Description
1	Clearly not in violation of Twitter's policy.
2	Probably not violating the policy, but does seem to suggest a questionable treatment may be effective. For example, the treatment is in clinical trials at the time the tweet was written, or the tweet does not make a strong claim about effectiveness.
3	Unclear whether or not this violates the policy.
4	Most likely violating Twitter's policy. Seems like the treatment is not effective based on official sources or reputable news organizations, and the tweet is making a relatively strong claim that the treatment is effective.
5	Clearly in violation of Twitter's policy.

Table 7: Likert score descriptions that are presented to annotators. These are used to evaluate whether tweets supporting a misinformation claim are in violation of Twitter's policies.

Source	Treatments
CDC(*)	Mask, Face Mask, Social Distancing, Stay Home, Wash Hands, Hand Washing, Cover Coughs, Cover Sneezes
New York Times(**)	Remdesivir, REGEN-COV, Bamlanivimab, Etesevimab, Sotrovimab, Dexamethasone, Prone positioning, Ventilators, Evusheld, Paxlovid, Molnupiravir, Lagevrio, Baricitinib, Olumiant, Tocilizumab, Actemra

Table 8: Approved COVID-19 treatments used in evaluation based on lists from the New York Times (Zimmer et al., 2020) and the Centers for Disease Control (CDC, 2021).

	Tagging/QA (non-T5)	QA (T5)
learning rate	1e-5,2e-5,3e-5	1e-4,2e-4,3e-4
batch size	8,16	8,16
epoch	50	10

Table 9: Hyperparameters of claim extraction models.

	NBOW	RoBERTa/CT-BERT
learning rate	1e-4,5e-5,1e-3,5e-3	8e-6,1e-5,3e-5
batch size	4,16	8,16
epoch	50	12

Table 10: Hyperparameters of stance classification models.

11.5 times per day on average while 56% of total treatments encountered were mentioned less than one time in the period studied. Therefore, we require a method that takes into account the historical frequency of treatments to calculate the strength of the association between the trendiness of treatment and the date.

To do this, we use a one-tailed Fisher's Exact Test (Fisher, 1922), which has been shown to be effective in rare event detection applications in the domain of statistical natural language processing (Moore, 2004; Johnson et al., 2007; Ritter et al., 2011).

To apply this test, we first calculate the hypergeometric probability, the probability of a particular distribution of treatment frequencies assuming independence between the treatment and the date. We define T and D as the events when a tweet's extracted treatment is t and when a tweet is published on date d respectively. Also, we let $C(X)$ be the observed frequency of event X and $C(X, Y)$ be

the joint frequency of event X and event Y. Given these definitions, we can calculate the hypergeometric probability, $p_{T,D}$, as follows:

$$p_{T,D} = \frac{C(T)!C(\neg T)!C(D)!C(\neg D)!}{N!C(T,D)!C(\neg T,D)!C(T,\neg D)!C(\neg T,\neg D)!}$$

where N is the sample size.

Given this formula of the hypergeometric probability for a distribution with a treatment t and date d, we calculate the p-value of the test by summing the hypergeometric probabilities of this distribution and all more extreme distributions. In our case, more extreme distributions are hypothetical distributions where the joint frequency of a tweet with a specific treatment published on a specific date is greater than $C(T, D)$.

A treatment is flagged by our system if its p-value is less than the threshold or α-value as set by the content moderator and it has not previously broken this threshold.

F Policy Violations Per Hour Calculation

Here we detail the calculation of the 124.2 policy violations per hour statistic reported in §4.2. First, we calculate the total amount of time required by each of the phases of the human annotation:

1. **Stage 1:** Detecting Misleading COVID-19 Treatments (§4.2.1)
   - # of claims = 300 claims
   - time of verifying a claim $= 89.7s/$claim
   - time spent on claim annotation $= 300 * 89.7s = 7.5$ hours
2. **Stage 2:** Identifying COVID-19 Policy Violations (§4.2.2)
   - # of tweets mentioning an unapproved claim in the full batch $= 10246$ claims
   - time to annotate an individual tweet $= 16.1s/$tweet
   - estimated time spent on tweet annotation $= 10246 * 16.1s = 45.8$ hours

Next, we detail the calculation steps using these calculated times:

1. Total annotation time $= 7.5 + 45.8 = 53.3$ hours
2. Estimated # of tweets (considering the full batch - see scores 4 and 5 bars in Figure 4) containing policy violation $= 3151 + 3467 = 6618$ tweets
3. # policy violated identified per hour of annotation time $= 6618/53.3$ hours $= 124.2$ tweets/hours
4. total # tweets judged per hour of annotation time $= 10246/53.3$ hours $= 192.2$ tweets/$hour$

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 6

☑ A2. Did you discuss any potential risks of your work?
Section 6, Section 7

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 4.1.1, Section 4.1.2

☑ B1. Did you cite the creators of artifacts you used?
Section 4.1.1

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Section 7

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section 4.1.1

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Section 7

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 4.1.2

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 4.1.2

C ☑ Did you run computational experiments?

Section 4.1.1, Section 4.1.2

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix D1, Appendix D2

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendix D1, Appendix D2

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 4

☐ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Not applicable. Left blank.

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Section 4.1.2 and Appendix B

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Appendix B

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Section 4.1.2

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Appendix B

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Appendix B

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Appendix B