NEWSCLAIMS: A New Benchmark for Claim Detection from News with Attribute Knowledge

Revanth Gangi Reddy[1], **Sai Chetan Chinthakindi**[1], **Zhenhailong Wang**[1], **Yi R. Fung**[1], **Kathryn Conger**[2], **Ahmed Elsayed**[2], **Martha Palmer**[2], **Preslav Nakov**[3], **Eduard Hovy**[4,5], **Kevin Small**[6], **Heng Ji**[1]
[1]UIUC [2]CU Boulder [3]MBZUAI [4]DARPA [5]CMU [6]Amazon
{revanth3,hengji}@illinois.edu

Abstract

Claim detection and verification are crucial for news understanding and have emerged as promising technologies for mitigating misinformation and disinformation in the news. However, most existing work has focused on *claim sentence* analysis while overlooking additional crucial attributes (e.g., the claimer and the main object associated with the claim). In this work, we present NEWSCLAIMS, a new benchmark for attribute-aware claim detection in the news domain. We extend the claim detection problem to include extraction of additional attributes related to each claim and release 889 claims annotated over 143 news articles. NEWSCLAIMS[1] aims to benchmark claim detection systems in emerging scenarios, comprising unseen topics with little or no training data. To this end, we see that zero-shot and prompt-based baselines show promising performance on this benchmark, while still considerably behind human performance.

1 Introduction

The internet era has ushered in an explosion of online content creation, resulting in increased concerns regarding misinformation in news, online debates, and social media. A key element of identifying misinformation is detecting the claims and the arguments that have been presented. In this regard, news articles are particularly interesting as they contain claims in various formats: from arguments by journalists to reported statements by prominent public figures.

Check-worthiness estimation aims to decide if a piece of text is worth fact-checking, i.e., whether it contains an important verifiable factual claim (Hassan et al., 2017a). Most current approaches (Jaradat et al., 2018; Shaar et al., 2021) largely ignore relevant attributes of the claim (e.g., the claimer and the primary object associated with the claim).

News Text: With the coronavirus pandemic continuing to spread around the globe, people are panicked, and they're looking for answers and explanations. *One wild theory that has made its way around the web is that the virus came from space*. Recently, Chandra Wickramasinghe, known for his work in astronomy and astrobiology, spread the idea that the virus was living on a comet and a piece of that space rock may have fallen to Earth.

Topic: Origin of the virus
Stance: Affirm
Claim Object: space
Claimer: Chandra Wickramasinghe

Figure 1: A news article containing a claim regarding the origin of COVID-19 with the *claim sentence* in italics, the *claim span* in red, and the *claimer* in blue. Also shown are the *claimer stance* and the *claim object*.

Moreover, current claim detection tasks mainly identify claims in debates (Gencheva et al., 2017), speeches (Atanasova et al., 2019a), and social media (Nakov et al., 2022), where the claim source (i.e., the claimer) is known.

News articles, on the other hand, have more complex arguments, requiring a deeper understanding of what each claim is about and identifying where it comes from. Thus, here we introduce the notion of *claim object*, which we define as an entity that identifies what is being claimed with respect to the topic of the claim. Figure 1 shows a claim about the origin of COVID-19, suggesting that the virus came from *space*, which is the claim object. We further identify the *claimer*, which could be useful for fact-checking organizations to examine how current claims compare to previous ones by the same person/organization. In this regard, we extend the claim detection task to ask for the extraction of more attributes related to the claim. Specifically, given a news article, we aim to extract all *claims* pertaining to a set of *topics* along with the corresponding *claim span*, the *claimer*, the *claimer's stance*, and the *claim object* for each claim. The claim attributes enable comparing claims at a more fine-grained level: claims with the same topic, object and stance can be considered equivalent

[1]The code and data have been made publicly available here: https://github.com/blender-nlp/NewsClaims

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 6002–6018
December 7-11, 2022

whereas those with similar claim objects but opposing stance could be contradicting. We note that while identifying the claim span and stance have been explored independently in prior work (Levy et al., 2014; Hardalov et al., 2021a), we bring them into the purview of a unified claim detection task.

To promote research in this direction, we release NEWSCLAIMS, a new evaluation benchmark for claim detection. We consider this in an evaluation setting since claims about new topics can emerge rapidly[2], requiring systems that are effective under zero/few-shot settings. NEWSCLAIMS aims to study how existing NLP techniques can be leveraged to tackle claim detection in emerging scenarios and regarding previously unseen topics. We explore multiple zero/few-shot strategies for our subtasks including topic classification, stance detection, and claim object detection. This is in line with recent progress in using pre-trained language models in zero/few-shot settings (Brown et al., 2020; Liu et al., 2021). Such approaches can be adapted to new use cases and problems as they arise without the need for large additional training data.

In our benchmark, all news articles are related to the COVID-19 pandemic, motivated by multiple considerations. First, COVID-19 has gained extensive media coverage, with the World Health Organization coining the term *infodemic*[3] to refer to disinformation related to COVID-19 (Naeem and Bhatti, 2020) and suggesting that "fake news spreads faster and more easily than this virus". Second, this is an emerging scenario with limited previous data related to the virus, making it a suitable candidate for evaluating claim detection in a low-resource setting. NEWSCLAIMS covers claims about four COVID-19 topics, namely the origin of the virus, possible cure for the virus, the transmission of the virus, and protecting against the virus.

Our contributions include (*i*) extending the claim detection task to include more attributes (claimer and object of the claim), (*ii*) releasing a manually annotated evaluation benchmark for this new task, NEWSCLAIMS, which covers multiple topics related to COVID-19 and is the first dataset with such extensive annotations for claim detection in the news, with 889 claims from 143 news articles, and (*iii*) demonstrating promising performance of various zero-shot and prompt-based few-shot approaches for the claim detection task.

[2]harmful-content-blog-post
[3]COVID-19 Infodemic

2 Related Work

Automatic fact-checking has a number of sub-tasks such as detecting check-worthy claims (Jaradat et al., 2018; Vasileva et al., 2019), comparing them against previously-fact checked claims (Shaar et al., 2020; Nakov et al., 2021), retrieving evidence relevant to a claim (Karadzhov et al., 2017; Augenstein et al., 2019) and finally inferring the veracity of the claim (Karadzhov et al., 2017; Thorne et al., 2018; Atanasova et al., 2019b). Our work here is positioned in the space of identifying check-worthy claims, also known as *check-worthiness estimation*. In this work, we show that identifying the topic of the claim is beneficial, by leveraging it towards stance detection (Section 5.3) and claim object detection (Section 5.2).

Argumentation mining (Palau and Moens, 2009; Stab and Gurevych, 2014; Stab et al., 2018) includes context-dependent claim detection (Levy et al., 2014, 2017), which entails detecting claims specifically relevant to a predefined topic. However, claims in the context of argumentation are neither necessarily factual nor verifiable. Moreover, prior work on both check-worthiness estimation and argumentation mining did not deal with identifying additional claim attributes, such as the claimer, or the source of the claim, and the claim object.

The claimer detection subtask is related to attribution in the news. Current attribution methods are mainly sentence-level (Pareti, 2016a) or only involve direct quotations (Elson and McKeown, 2010). In contrast, we require cross-sentence reasoning for identifying the claimer as it may not be present in the claim sentence (see Figure 1).

There has been recent work addressing claims related to COVID-19. Saakyan et al. (2021) proposed a new FEVER-like (Thorne et al., 2018) dataset, where given a claim, the task is to identify relevant evidence and to verify whether it refutes or supports the claim; however, this does not tackle identifying the claims or the claimer. There has also been work on identifying the check-worthiness of tweets related to COVID-19 (Alam et al., 2020; Jiang et al., 2021); however, unlike news articles, tweets do not require attribution for claimer identification.

3 Proposed Claim Detection Task

Our task is to identify claims related to a set of topics in a news article along with corresponding attributes such as the claimer, the claim object, and the claim span and stance, as shown in Figure 2.

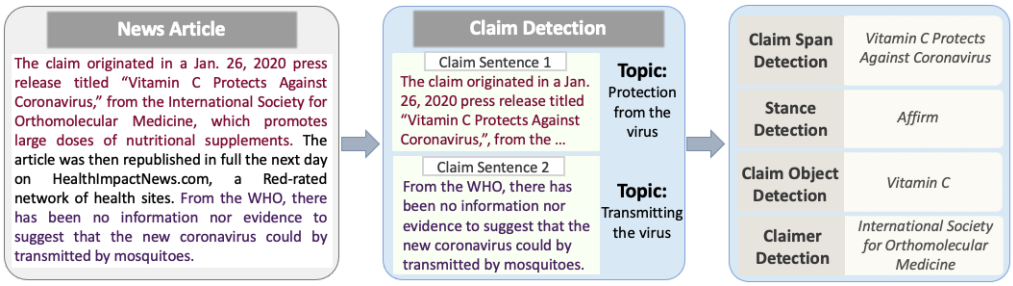


Figure 2: An example demonstrating our proposed claim detection task, and its subtasks. The following attributes are to be extracted for each claim: the *claimer*, *claimer's stance*, *claim object*, and *claim span*.

Claim Sentence Detection: Given a news article, the first subtask is to extract claim sentences relevant to a set of pre-defined topics. This involves first identifying sentences that contain *factually verifiable claims*, similar to prior work on check-worthiness estimation, and then selecting those that are related to the target topics. To address misinformation in an emerging real-world setting, we consider the following topics related to COVID-19: ***Origin of the virus***: claims related to the origin of the virus (i.e., location of first detection, zoonosis, 'lab leak' theories); ***Transmission of the virus***: claims related to who/what can transmit the virus or conditions favorable for viral transmission; ***Cure for the virus***: claims related to curing the virus, (e.g., via medical intervention after infection); and ***Protection from the virus***: claims related to precautions against viral infection.

Claimer Detection: Claims within a news article can come from various types of sources such as an entity (e.g., person, organization) or published artifact (e.g., study, report, investigation). In such cases, the claimer identity can usually be extracted from the news article itself. However, if the claim is asserted by the article author or if no attribution is specified or inferrable, then the article author, i.e. the journalist, is considered to be the claimer. The claimer detection subtask involves identifying whether the claim is made by a *journalist* or whether it is *reported* in the news article, in which case the source is also extracted. Moreover, sources of such reported claims need not be within the claim sentence. In our datatset NEWSCLAIMS, the claimer span was extracted from outside of the claim sentence for about 47% of the claims. Thus, the claimer detection subtask in our benchmark requires considerable document-level reasoning, thus making it harder than existing attribution tasks (Pareti, 2016b; Newell et al., 2018), which require only sentence-level reasoning.

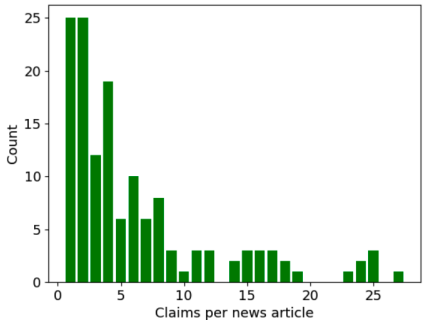
Claim Object Detection: The claim object relates to what is being claimed in the claim sentence with respect to the topic. For example, in a claim regarding the virus origin, the claim object could be the species of origin in zoonosis claims, or who created the virus in bioengineering claims. Table 1 shows examples of claim objects from each topic. We see that the claim object is usually an extractive span within the claim sentence. Identifying the claim object helps to better understand the claims and potentially identify claim–claim relations, since two claims with the same object are likely to be similar.

Topic	Claim Sentence
Origin	The genetic data is pointing to this virus coming from a **bat reservoir**, he said.
Transmission	The virus lingers in the **air indoors**, infecting those nearby
Cure	**Vitamin C** is an effective treatment for COVID-19.
Protection	Taking a **hot bath** prevents you from getting COVID-19.

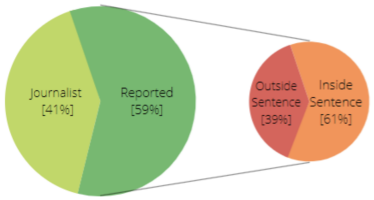
Table 1: Examples showing the claim object in **bold** for claims corresponding to NEWSCLAIMS topics.

Stance Detection: This subtask involves outputting whether the claimer is asserting (*affirm*) or refuting (*refute*) a claim within the given claim sentence. We note that stance detection in NEWSCLAIMS differs from the task formulation used in other stance detection datasets (Stab et al., 2018; Hanselowski et al., 2019; Allaway and McKeown, 2020) as it involves identifying the claimer's stance within a claim sentence – whereas prior stance detection tasks, as described in a recent survey by Hardalov et al. (2021b), involve identifying the stance for *target–context* pairs. For example, given pairs such as claim–evidence or headline–article, it involves identifying whether the evidence/article at hand supports or refutes a given claim/headline.

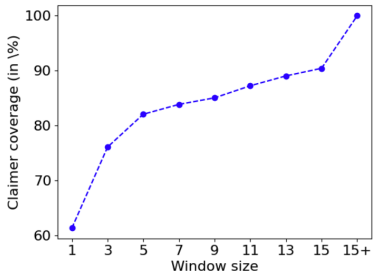
Claim Span Detection: Given a claim sentence, this subtask aims to identify the exact claim bound-



(a) claim counts per news article



(b) claims by journalists vs. reported ones, along with claimer coverage for reported claims



(c) coverage of claimer within a window size based on number of sentences around the claim sentence

Figure 3: Statistics about our claim detection benchmark: (a) number of claims per news article, (b) claims by journalists vs. reported claims, and (c) claimer coverage by window size within the news article for reported claims.

aries within the sentence, including the actual claim content, usually without any cue words (e.g., *asserted*, *suggested*) and frequently a contiguous subspan of the claim sentence. Identifying the precise claim conveyed within the sentence can be useful for downstream tasks such as clustering claims and identifying similar or opposing claims.

4 The NEWSCLAIMS Dataset

In this work, we build NEWSCLAIMS, a new benchmark dataset for evaluating the performance of models on different components of our claim detection task. Specifically, we release an evaluation set based on news articles about COVID-19, which can be used to benchmark systems on detecting claim sentences and associated attributes including claim objects, claim span, claimer, and claimer stance. NEWSCLAIMS uses news articles from the LDC corpus *LDC2021E11*, from which we selected those related to COVID-19. We describe below the annotation process (Section 4.1) and provide statistics about NEWSCLAIMS (Section 4.2).

4.1 Annotation

Given a news article, we split the annotation process into two phases: (*i*) identifying claim sentences with their corresponding topics, and (*ii*) annotating the attributes for these claims.[4] In the first phase, the interface displays the entire news article with a target sentence highlighted in red. The annotators are asked whether the highlighted sentence contains a claim associated with the four pre-defined COVID-19 topics and to indicate the specific topic if that is the case. In the second phase, the interface displays the entire news article with a claim sentence highlighted in red. The annotators are asked to identify the claim span, the claim object, and the claimer from the news article. The annotators are also asked to indicate the claimer's stance regarding the claim. We provide a checkbox to use if there is no specified claimer, in which case the journalist is considered to be the claimer.

For the first stage of annotation, which involves identifying claim sentences (and their topics) from the entire news corpus, we used 3 annotators per example hired via Mechanical Turk (Buhrmester et al., 2011). Only sentences with unanimous support were retained as valid claims. For the second stage, which involves identifying the remaining attributes (claim object, span, claimer, and stance), we used expert annotators to ensure quality, with 1 annotator per claim sentence. Annotators took ∼30 seconds per sentence in the first phase and ∼90 seconds to annotate the attributes of a claim in phase two. For claim sentence detection, the inter-annotator agreement had a Krippendorff's kappa of 0.405, which is moderate agreement; this is on par with previous datasets that tackled identifying topic-dependent claims (Kotonya and Toni, 2020; Bar-Haim et al., 2020), which is more challenging than topic-independent claim annotation (Thorne et al., 2018; Aly et al., 2021).

4.2 Statistics

NEWSCLAIMS consists of development and test sets with 18 articles containing 103 claims and 125 articles containing 786 claims, respectively. The development set can be used for few-shot learning or for fine-tuning model hyper-parameters. Figure 3a shows a histogram of the number of claims in a news article where most news articles contain up to 5 claims, but some have more than 10

[4] Detailed annotation guidelines and screenshots of the interface are provided in Section A.1 in the appendix.

claims. Claims related to the origin of the virus are most prevalent, with the respective topic distribution being 35% for origin, 22% for cure, 23% for protection, and 20% for transmission. Figure 3b shows the distribution of claims by journalists vs. reported claims: we can see that 41% of the claims are made by journalists, with the remaining 59% coming from sources mentioned in the news article. Moreover, for reported claims, the claimer is present outside of the claim sentence 39% of the time, demonstrating the document-level nature of this task. Figure 3c shows the claimer coverage (in %) based on a window around the claim by the number of sentences and indicates that document-level reasoning is required to identify the claimer, with some cases even requiring inference beyond a window size of 15. Note that the 61% inside-sentence coverage in Figure 3b corresponds to a window size of 1 in Figure 3c.

5 Baselines

In this section, we describe various zero-shot and prompt-based few-shot learning baselines for the claim detection subtasks outlined in Section 3. We describe a diverse set of baselines with each chosen to be relevant in an evaluation-only setting.

5.1 Claim Sentence Detection

Given a news article, we aim to detect all sentences that contain claims related to a pre-defined set of topics regarding COVID-19. We use a two-step procedure that first identifies sentences that contain claims and then selects those related to COVID-19.

Step 1. ClaimBuster: To identify sentences containing claims, we use ClaimBuster (Hassan et al., 2017b),[5] a claim-spotting system trained on a dataset of check-worthy claims (Arslan et al., 2020). As ClaimBuster has no knowledge about topics, we use zero-shot topic classification, as described below.

Step 2. ClaimBuster+Zero-shot NLI: Following Yin et al. (2019), we use pre-trained NLI models as zero-shot text classifiers: we pose the claim sentence to be classified as the NLI premise and construct a hypothesis from each candidate topic. Figure 4a shows the hypothesis corresponding to each of the topics. We then use the entailment score for each topic as its topic score and choose the highest topic score for threshold-based filtering.

[5] https://idir.uta.edu/claimbuster/api/

5.2 Claim Object Detection

Given the claim sentence and a topic, claim object detection seeks to identify what is being claimed about the topic, as shown in Table 1. We explore this subtask in both zero-shot and few-shot settings by converting it into a prompting task for pre-trained language models as described below:

In-context learning (few-shot): This setting is similar to (Brown et al., 2020), where the few-shot labeled examples are inserted into the context of a pre-trained language model. The example for which a prediction is to be made is included as a prompt at the end of the context. We refer the reader to Section A.3 in the appendix for an example. We use GPT-3 (Brown et al., 2020) as the language model in this setting.

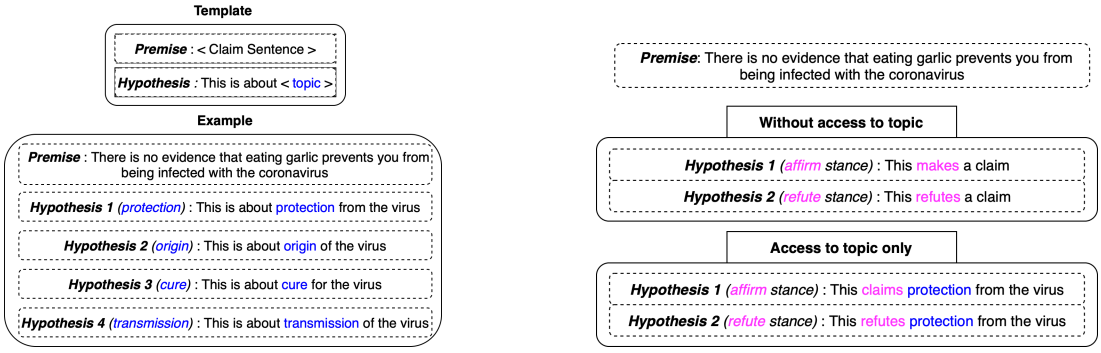
Prompt-based fine-tuning (few-shot): Following Gao et al. (2021), we fine-tune a pre-trained language model, base-T5 (Raffel et al., 2020), to learn from a few labeled examples. We convert the examples into a prompt with a format similar to the language model pre-training, which for this model involves generating the target text that has been replaced with a <MASK> token in the input. Thus, we convert the few-shot data into such prompts and generate the claim object from the <MASK> token. For example, given the claim sentence: *Research conducted on the origin of the virus shows that it came from bats*, and its topic (origin of the virus), the prompt would be: *Research conducted on the origin of the virus shows that it came from bats. The origin of the virus is <MASK>.*

Prompting (zero-shot): We consider the language models that were used in few-shot settings above with the same prompts but in zero-shot settings here. In this case, GPT-3 is not provided with any labeled examples in the context and T5 is used out-of-the-box without any fine-tuning.

5.3 Stance Detection

Given the claim sentence, stance detection identifies if the claimer is asserting or refuting the claim.

Zero-shot NLI: We leverage NLI models for zero-shot classification. Here, we construct a hypothesis for the *affirm* and the *refute* labels and we take the stance corresponding to a higher entailment score. We consider two settings while constructing the hypothesis based on claim topic availability. Examples are shown in Figure 4b.



(a) zero-shot NLI for topic classification

(b) zero-shot NLI for stance detection

Figure 4: Diagram (a) shows the template and an example for leveraging a pre-trained NLI model for zero-shot topic classification; the topic corresponding to the hypothesis with the highest entailment score is taken as the claim sentence topic. Diagram (b) shows examples for leveraging a pre-trained NLI model for zero-shot stance detection. Each example shows how the hypothesis is constructed based on the class label (in pink) and the topic (in blue).

5.4 Claim Span Detection

Given a claim sentence, claim span detection identifies the exact claim boundaries within the sentence.

Debater Boundary Detection: Our first baseline uses the claim boundary detection service from the Project Debater[6] APIs (Bar-Haim et al., 2021). This system is based on BERT-Large, which is further fine-tuned on 52K crowd-annotated examples mined from the Lexis-Nexis corpus.[7]

PolNeAR-Content: Our second baseline leverages PolNeAR (Newell et al., 2018), a popular news attribution corpus of annotated triples comprising the *source*, a *cue*, and the *content* for statements made in the news. We build a claim span detection model from it by fine-tuning BERT-large (Devlin et al., 2019) to identify the *content* span, with a start classifier and an end classifier on top of the encoder outputs, given the sentence as an input.

5.5 Claimer Detection

This subtask identifies if the claim is made by the journalist or a reported source, in addition to identifying the mention of the source in the news article.

PolNeAR-Source: We leverage the PolNeAR corpus to build a claimer extraction baseline. Given a statement, we use the *source* annotation as the claimer and mark the *content* span within the statement using special tokens. We then fine-tune a BERT-large model to extract the source span from the statement using a start classifier and an end classifier over the encoder outputs. At evaluation time, we use the news article as an input, marking the claim span with special tokens and using the sum of the start and the end classifier scores as a claimer span confidence score. This is thresholded to determine if the claim is by the journalist, with the claimer span used as an output for reported claims.

SRL: We build a Semantic Role Labeling (SRL) baseline for claimer extraction. SRL outputs the verb predicate-argument structure of a sentence such as who did what to whom. Given the claim sentence as an input, we filter out verb predicates that match a pre-defined set of cues[8] (e.g., *say*, *believe*, *deny*). Then, we use the span corresponding to the ARG-0 (agent) of the predicate as the claimer. As SRL works at the sentence level, this approach cannot extract claimers outside of the claim sentence. Thus, the system outputs *journalist* as the claimer when none of the verb predicates in the sentence matches the pre-defined set of cues.

6 Experiments

In this section, we evaluate various zero-shot and few-shot approaches for the subtasks of our claim detection task. To estimate the upper bounds, we also report the human performance for each subtask computed over ten random news articles.

6.1 Claim Sentence Detection

Setup: For zero-shot MNLI, we use BART-large[9] (Lewis et al., 2020) trained on the MultiNLI corpus

[6] Project Debater
[7] http://www.lexisnexis.com/en-us/home.page
[8] Appendix A.2 contains the complete set of cues.
[9] http://huggingface.co/facebook/bart-large-mnli

(Williams et al., 2018). ClaimBuster and the topic-filtering thresholds are tuned on the development set. For evaluation, we use precision, recall, and F1 scores for the filtered set of claims relative to the ground-truth annotations.

Results and Analysis: Table 2 shows the performance of various systems for identifying claim sentences about COVID-19. We use ClaimBuster, which does not involve topic detection, as a low-precision high-recall baseline. We can see that the performance improves by leveraging a pre-trained NLI model as a zero-shot filter for claims that are not related to the topics at hand. We also report results for both single-human performance and for 3-way majority voting. Note that even humans have relatively lower precision, demonstrating the difficulty of identifying sentences with claims. Nevertheless, the model performance is still considerably worse compared to human performance, showing the need for better models.

Model	P	R	F1
ClaimBuster	13.0	**86.5**	22.6
ClaimBuster + Zero-shot NLI	**21.8**	53.3	**30.9**
Human (single)	52.7	70.0	60.1
Human (3-way majority voting)	60.2	83.5	70.0

Table 2: Performance (in %) for various systems for detecting claims related to COVID-19.

6.2 Claim Object Detection

Setup: We use the development set to get the few-shot examples, sampling[10] five examples per topic. To account for sampling variance, we report numbers averaged over three runs. For language model sizes to be comparable, we use the Ada[11] version of GPT-3 and the base version of T5. We fine-tune T5-base for five epochs using a learning rate of 3e-5. We score using string-match F1, as done for question answering (Rajpurkar et al., 2016).

Results and Analysis: Table 3 shows the F1 score for extracting the claim object related to the topic. In zero-shot settings, we see that GPT-3 performs considerably better than T5, potentially benefiting from the larger corpus it was trained on. However, in a few-shot setting, T5 is competitive with GPT-3, showing the promise of prompt-based fine-tuning, even with limited few-shot examples.

[10]We will release the few-shot examples for reproducibility.

[11]https://blog.eleuther.ai/gpt3-model-sizes/

Approach	Model	Type	F1
Prompting	GPT-3	Zero-shot	15.2
Prompting	T5	Zero-shot	11.4
In-context learning	GPT-3	Few-Shot	**51.9**
Prompt-based fine-tuning	T5	Few-Shot	51.6
Human	-	-	67.7

Table 3: F1 score (in %) for various zero-shot and few-shot systems for the claim object detection sub-task.

6.3 Stance Detection

Setup: We use the same BART-large model trained for NLI as in Section 6.1. In the setting with access to the topic, we take the topic from the gold-standard annotation.

Results and Analysis: We also consider a majority class baseline that always predicts *affirm* as the stance. Table 4 shows the performance of stance detection approaches. We can see that the the NLI model with access to the topic performs the best, with considerable improvement in performance for the *refute* class. Thus, access to additional attribute information helps here as the topic of the claim can be used to come up with a more relevant hypothesis, as is evident from Figure 4b.

Model	Affirm F1	Refute F1	Acc.
Majority class	82.5	0.0	70.3
NLI (no topic)	89.1	68.0	83.8
NLI (with topic)	**91.1**	**78.8**	**87.5**
Human	97.0	84.2	94.9

Table 4: F1 score (in %) for the *affirm* and the *refute* classes along with overall accuracy for stance detection. The zero-shot NLI system is shown separately as it could access the topic while constructing the hypothesis.

6.4 Claim Span Detection

Results and Analysis: The evaluation measure in this setting is character-span F1. From Table 5, we see that the Debater claim boundary detection system considerably outperforms the attribution-based system. This could be because the former is trained on arguments, which are more similar to claims compared to statement-like attributions.

Model	Prec.	Recall	F1
PolNeAR-Content	67.0	42.8	52.3
Debater Boundary Detection	**75.7**	**77.7**	**76.7**
Human	82.7	90.9	86.6

Table 5: Performance (in %) of different systems for identifying the boundaries of the claim.

6.5 Claimer Detection

Setup: For the PolNeAR-Source system, the threshold for confidence score is tuned on the dev set. The claim span output from the Debater boundary system is used for marking the claim content in the context. For the SRL system, we leverage the parser[12] provided by AllenNLP (Gardner et al., 2018), which was trained on OntoNotes (Pradhan et al., 2013). The evaluation involves scores for the journalist (classification F1) and for reported (string-match F1), along with overall F1.

Model	F1	Reported	Journalist
SRL	41.7	23.5	**67.2**
PolNeAR-Source	**42.3**	**25.5**	65.9
Human	85.8	81.3	88.9

Table 6: Claimer detection. Reported are F1 scores for journalist claims and for reported claims, along with the overall F1.

Results and Analysis: Table 6 shows that automatic models perform considerably worse than humans for claimer detection. While the performance is relatively better for identifying whether a journalist is making the claim, models perform poorly for reported claims, which involves extracting the claimer mentions. For reported claims, Table 7 shows that the performance depends on whether the claimer is mentioned inside or outside of the claim sentence. Specifically, we see that these attribution models are able to handle claimer detection for reported claims only when the claimer mention is within the claim sentence. The need for cross-sentence reasoning for the claimer detection sub-task is evident from the low out-of-sentence F1 score for these sentence-level approaches.

Model	In-sentence	Out-of-sentence
SRL	35.8	2.4
PolNeAR-Source	**38.9**	**2.7**

Table 7: F1 score (in %) in terms of reported claims for extracting the claimer when it is present within or outside the claim sentence.

6.6 Error Analysis and Remaining Challenges

News articles have a narrative structure when presenting claims, by backing them up with some evidence. We observed that humans, when considering sentences without looking at the context, tend to identify such statements providing evidential information as claims too. Figure 5 shows some examples of errors corresponding to false positives from the human study. The human study identified the sentences in red as claims, in addition to the ones in green. In Figure 5a, the sentences in green contain concrete claims regarding the origin of the virus, with the first sentence claiming that it came from natural selection and the second sentence refuting that the virus was a laboratory manipulation. The sentence in red, on the other hand, simply provides evidence for natural evolution. In Figure 5b, the sentence in green contains a claim that refutes that these medicines can cure the virus. On the other hand, the sentence in red does not contain a claim because it simply asserts that these medicines are being used for treating patients, without any clear claim on whether they can actually cure the virus.

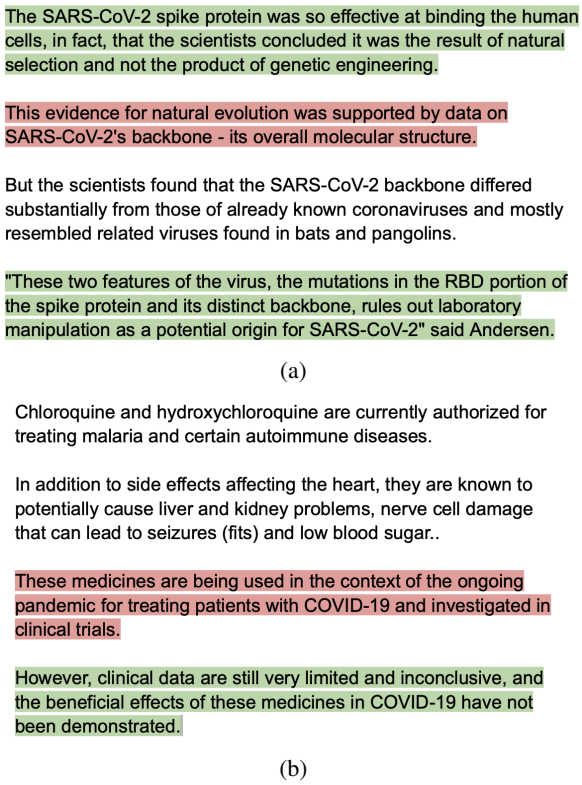


Figure 5: Some examples from the human study with the gold-standard claims highlighted in green and false positives from humans highlighted in red.

We investigated the NLI model performance for topic classification. Given the gold-standard claim sentence, the accuracy is 46.6% over these four topics. Topic-wise F1 was relatively poor for *Cure* (3.3%) compared to the other topics: *Origin* is 56.9%, *Protection* is 54.5%, and *Transmission* is 45.1%. Figure 6 shows the confusion matrix for

[12] AllenNLP SRL Parser

Claim Sentence	Gold topic	Predicted topic
This novel coronavirus was believed to have started in a large seafood or wet market, suggesting animal-to-person spread.	Origin	Transmission
A Wuhan laboratory official has denied any role in spreading the new coronavirus, after months of speculation about how the previously unknown animal disease made the leap to humans.	Origin	Transmission
One medication, an antiviral drug called Remdesivir, has been shown in certain studies to improve symptoms and shorten hospital stays.	Cure	Protection
Studies show hydroxychloroquine does not have clinical benefits in treating COVID-19.	Cure	Protection

Table 8: Some topic classification error examples from the zero-shot NLI model.

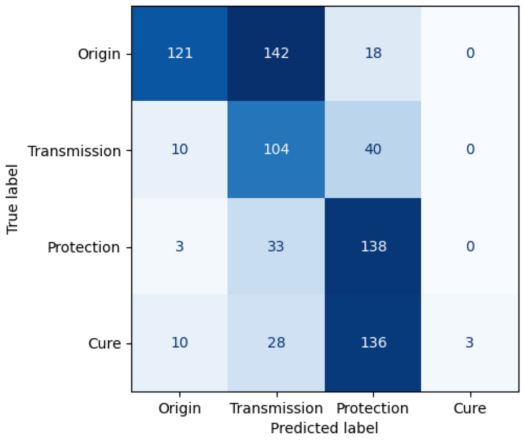


Figure 6: Confusion matrix for the topic classification predictions from the zero-shot NLI model.

the topic classification predictions. We see two dominant types of errors. First, most claims corresponding to the topic *Cure* are under *Protection*. This is potentially due to these two topics being related and the NLI model unable to differentiate that *Protection* corresponds to prevention measures before contracting COVID-19, while *Cure* refers to treatments after contracting COVID-19. Second, we see that a considerable number of claims related to *Origin* were classified as *Transmission*. This could be due to a statement about the virus originating in animals and then jumping to humans, which suggests that a claim about the origin of the virus was being misconstrued as one regarding the transmission of the virus. Some representative examples for both of these types of errors are shown in Table 8. Given the low topic classification performance of the NLI model, we need better zero-shot approaches for selecting claims related to COVID-19. This is important as the claim topic is crucial to claim object detection and it can help stance detection.

Stance detection performance could likely be improved by also leveraging claim objects while formulating the NLI hypothesis. For example, the stance for *"An Oxford University professor claimed that the coronavirus may not have originated in China."* was predicted as *affirm* even though it refutes that the virus originated in China. By leveraging the extracted claim object, the NLI hypothesis for the *refute* class could be better formulated as *"China is not the origin of the virus"*. The existing formulation, shown in Figure 4b, only uses the claim topic to put it as *"This refutes the origin of the virus"*. We leave this for future work.

The claimer detection subtask requires incorporating stronger cross-sentence reasoning when the mention is outside the claim sentence. This requires building attribution systems that are document-level. Moreover, the same news article can have similar claims but from different claimers. To prevent misattribution in such cases, it would be beneficial to identify the context within the news article that is relevant to the given claim, so as to remove noise from other related claims.

7 Conclusion and Future Work

We proposed a new benchmark, NEWSCLAIMS, which extends the current claim detection task to extract more attributes related to each claim. Our benchmark comprehensively evaluates multiple aspects of claim detection such as identifying the topics, the stance, the claim span, the claim object, and the claimer in news articles from emerging scenarios such as the COVID-19 pandemic. We showed that zero-shot and prompt-based few-shot approaches can achieve promising performance in such low-resource scenarios, but still lag behind human performance, which presents opportunities for further research. In future work, we plan to explore extending this to build claim networks by identifying relations between the claims, including temporal connections. Another direction is build a unified framework that can extract claims and corresponding attributes together, without the need for separate components for each attribute.

Acknowledgement

This research is based upon work supported by U.S. DARPA AIDA Program No. FA8750-18-2-0014. 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 DARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

Limitations

NEWSCLAIMS exclusively consists of claims regarding COVID-19, which were intentionally chosen in order to sufficiently study a quickly emerging subject. However, the performance on this dataset might likely not be representative of the performance on a broader set of topics. NEWSCLAIMS is not intended as a training dataset and a system using NEWSCLAIMS in this way should be carefully evaluated before being used to annotate a larger dataset aimed at deriving journalism-centric conclusions. In the future, these risks can be mitigated by a larger dataset that can be more reliable to study these phenomena and to draw conclusions about the underlying media content.

Ethics and Broader Impact

Annotator payment and approval Our annotation process involved using both Turkers and expert annotators. For the first stage of annotation, Turkers were paid 15 cents per example (each example takes 30-35 seconds on average, meaning $15 per hour). For the second stage, expert annotators were paid at an hourly rate, which was dependent on prior experience, but was always more than the usual rate of $14 USD per hour. As per regulations set up by our organization's IRB, this work was not considered to be human subjects research because no data or information about the annotators was collected, and thus it was IRB approval exempt.

Misuse Potential The intended use of NEWSCLAIMS is to evaluate methodological work regarding our augmented definition of claim detection, motivated by mitigating the spread of misinformation and disinformation in news media. However, given NEWSCLAIMS is a smaller dataset over a set of hand-chosen topics, there is also potential for misuse. Specifically, NEWSCLAIMS is not intended to directly make conclusions regarding the journalism quality nor quantify disagreement regarding the coverage of COVID-19 related topics. As there has been continued controversy regarding media coverage of COVID-19, a bad faith or misinformed actor could produce artifacts that result in sensational, but potentially inaccurate, conclusions regarding COVID-19 claims in news media.

Environmental Impact We would also like to warn that the use of large-scale Transformers requires a lot of computations and the use of GPUs for training, which contributes to global warming (Strubell et al., 2019). This is a bit less of an issue in our case, as we do not train such models from scratch; rather, we mainly use them in zero-shot and few-shot settings, and the ones we fine-tune are on relatively small datasets. All our experiments were run on a single 16GB V100.

References

Firoj Alam, Shaden Shaar, Fahim Dalvi, Hassan Sajjad, Alex Nikolov, Hamdy Mubarak, Giovanni Da San Martino, Ahmed Abdelali, Nadir Durrani, Kareem Darwish, et al. 2020. Fighting the covid-19 infodemic: modeling the perspective of journalists, fact-checkers, social media platforms, policy makers, and the society. *arXiv preprint arXiv:2005.00033*.

Emily Allaway and Kathleen McKeown. 2020. Zero-Shot Stance Detection: A Dataset and Model using Generalized Topic Representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8913–8931, Online. Association for Computational Linguistics.

Rami Aly, Zhijiang Guo, Michael Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. Feverous: Fact extraction and verification over unstructured and structured information. *arXiv preprint arXiv:2106.05707*.

Fatma Arslan, Naeemul Hassan, Chengkai Li, and Mark Tremayne. 2020. A benchmark dataset of check-worthy factual claims. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 821–829.

Pepa Atanasova, Preslav Nakov, Georgi Karadzhov, Mitra Mohtarami, and Giovanni Da San Martino. 2019a. Overview of the clef-2019 checkthat! lab: Automatic identification and verification of claims. task 1: Check-worthiness. *CLEF (Working Notes)*, 2380.

Pepa Atanasova, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, Georgi Karadzhov, Tsvetomila Mihaylova, Mitra Mohtarami, and James Glass. 2019b.

Automatic fact-checking using context and discourse information. *Journal of Data and Information Quality (JDIQ)*, 11(3):1–27.

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. Multifc: A real-world multi-domain dataset for evidence-based fact checking of claims. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4685–4697.

Roy Bar-Haim, Lilach Eden, Roni Friedman, Yoav Kantor, Dan Lahav, and Noam Slonim. 2020. From arguments to key points: Towards automatic argument summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4029–4039.

Roy Bar-Haim, Yoav Kantor, Elad Venezian, Yoav Katz, and Noam Slonim. 2021. Project Debater APIs: Decomposing the AI grand challenge. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 267–274, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Michael Buhrmester, Tracy Kwang, and Samuel D Gosling. 2011. Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, pages 3–5.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

David K Elson and Kathleen R McKeown. 2010. Automatic attribution of quoted speech in literary narrative. In *Twenty-Fourth AAAI Conference on Artificial Intelligence*.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Association for Computational Linguistics (ACL)*.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. Allennlp: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6.

Pepa Gencheva, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, and Ivan Koychev. 2017. A context-aware approach for detecting worth-checking claims in political debates. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, Varna, Bulgaria. INCOMA Ltd.

Andreas Hanselowski, Christian Stab, Claudia Schulz, Zile Li, and Iryna Gurevych. 2019. A richly annotated corpus for different tasks in automated fact-checking. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 493–503.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2021a. Cross-domain label-adaptive stance detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9011–9028, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2021b. A survey on stance detection for mis-and disinformation identification. *arXiv preprint arXiv:2103.00242*.

Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. 2017a. 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*, pages 1803–1812.

Naeemul Hassan, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, Aaditya Kulkarni, Anil Kumar Nayak, et al. 2017b. Claimbuster: The first-ever end-to-end fact-checking system. *Proceedings of the VLDB Endowment*, 10(12):1945–1948.

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 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, New Orleans, Louisiana. Association for Computational Linguistics.

Ye Jiang, Xingyi Song, Carolina Scarton, Ahmet Aker, and Kalina Bontcheva. 2021. Categorising fine-to-coarse grained misinformation: An empirical study of covid-19 infodemic. *arXiv preprint arXiv:2106.11702*.

Georgi Karadzhov, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, and Ivan Koychev. 2017. Fully automated fact checking using external sources. *arXiv preprint arXiv:1710.00341*.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical*

Methods in Natural Language Processing (EMNLP), pages 7740–7754.

Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. 2014. Context dependent claim detection. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1489–1500.

Ran Levy, Shai Gretz, Benjamin Sznajder, Shay Hummel, Ranit Aharonov, and Noam Slonim. 2017. Unsupervised corpus–wide claim detection. In *Proceedings of the 4th Workshop on Argument Mining*, pages 79–84.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

Salman Bin Naeem and Rubina Bhatti. 2020. The covid-19 'infodemic': a new front for information professionals. *Health Information & Libraries Journal*, 37(3):233–239.

Preslav Nakov, Alberto Barrón-Cedeño, Giovanni Da San Martino, Firoj Alam, Julia Maria Struß, Thomas Mandl, Rubén Míguez, Tommaso Caselli, Mücahid Kutlu, Wajdi Zaghouani, Chengkai Li, Shaden Shaar, Gautam Kishore Shahi, Hamdy Mubarak, Alex Nikolov, Nikolay Babulkov, Yavuz Selim Kartal, and Javier Beltrán. 2022. The CLEF-2022 checkthat! lab on fighting the COVID-19 infodemic and fake news detection. In *Advances in Information Retrieval - 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10-14, 2022, Proceedings, Part II*, volume 13186 of *Lecture Notes in Computer Science*, pages 416–428. Springer.

Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021. Automated fact-checking for assisting human fact-checkers. *arXiv preprint arXiv:2103.07769*.

Edward Newell, Drew Margolin, and Derek Ruths. 2018. An attribution relations corpus for political news. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Raquel Mochales Palau and Marie-Francine Moens. 2009. Argumentation mining: the detection, classification and structure of arguments in text. In *Proceedings of the 12th international conference on artificial intelligence and law*, pages 98–107.

Silvia Pareti. 2016a. Parc 3.0: A corpus of attribution relations. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3914–3920.

Silvia Pareti. 2016b. PARC 3.0: A corpus of attribution relations. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3914–3920, Portorož, Slovenia. European Language Resources Association (ELRA).

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards robust linguistic analysis using OntoNotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152, Sofia, Bulgaria. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. 2021. COVID-fact: Fact extraction and verification of real-world claims on COVID-19 pandemic. 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)*, pages 2116–2129, Online. Association for Computational Linguistics.

Shaden Shaar, Nikolay Babulkov, Giovanni Da San Martino, and Preslav Nakov. 2020. That is a known lie: Detecting previously fact-checked claims. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3607–3618.

Shaden Shaar, Maram Hasanain, Bayan Hamdan, Zien Sheikh Ali, Fatima Haouari, Alex Nikolov, Mücahid Kutlu, Yavuz Selim Kartal, Firoj Alam, Giovanni Da San Martino, et al. 2021. Overview of the clef-2021 checkthat! lab task 1 on check-worthiness estimation in tweets and political debates. In *CLEF (Working Notes)*.

Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 46–56.

Christian Stab, Tristan Miller, Benjamin Schiller, Pranav Rai, and Iryna Gurevych. 2018. Cross-topic argument mining from heterogeneous sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3664–3674.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, ACL '19, pages 3645–3650, Florence, Italy.

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: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819.

Slavena Vasileva, Pepa Atanasova, Lluís Màrquez, Alberto Barrón-Cedeño, and Preslav Nakov. 2019. It takes nine to smell a rat: Neural multi-task learning for check-worthiness prediction. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1229–1239.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3914–3923.

A Appendix

A.1 Annotation Interface

In this section, we list the annotation guidelines and provide screenshots of the interface for both phases of annotation. Phase 1 of annotation involves identifying sentences which contain claims relating to a set of pre-defined topics about COVID-19. Phase 2 consists of annotating the attributes such as claimer, claimer's stance, claim object and the claimer span for each of the claims identified in phase 1. Figure 9 and 10 show screenshots of the annotation interface for phase 1 and 2 respectively. Below are some guidelines which we provide for detecting the claim sentences:

- The highlighted sentence should be considered individually when deciding whether it contains a claim. The sentences around it are shown to provide context.
- Claims are usually statements made without presenting evidence or proof, and usually require further evidence to verify them. Sentences that just assert evidence or present facts should not be considered as claims.
- The claim sentences usually should also mention the object relating to the topic, i.e., which animal type the virus came from, what conditions can transmit the virus, what can cure the virus or what can protect from the virus.
- Only those claims should be considered for which these topics can be directly inferred without any need for additional knowledge.
- Sentences that contain both claims as well as refute statements should be considered. For example, a sentence that contains a statement that something cannot cure the coronavirus should be considered as containing a claim relating to the topic: Cure for the virus.

A.2 SRL cue words

Here, we list various cue words that we use to match against the verb predicates from the semantic role labeling system. These are categorized as affirming and refuting cue words, which are shown in tables 9 and 10 respectively.

A.3 GPT-3 prompt

In this section, we share more details of our approach for prompting GPT-3 for the claim object detection. In the in-context learning setting, we choose four examples from each topic as the few-shot examples. These labeled examples are then added to the context that is fed as input to GPT-3. The test example is added at the end of the context, in the form of a prompt, with the claim object to be generated by the system. Figure 7 shows an example input along with the prompt.

accuse, affirm, allege, announce, argue
assert, aver, avouch, avow, blame
broadcast, claim, comment, confirm, contend
credit, declare, defend, describe, disclose
discuss, express, find, hint, imply
insinuate, insist, intimate, maintain, proclaim
profess, publish, purport, reaffirm, reassert
remark, repeat, report, restate, reveal
say, state, suggest, tell, write

Table 9: Cue words corresponding to affirming a claim.

challenge, controvert, contradict,disagree
discredit, dispute, deny, disavow, discount
protest, purport, reaffirm, question, repudidate
reject, repudiate, rebut, suppress, disaffirm

Table 10: Cue words corresponding to refuting a claim.

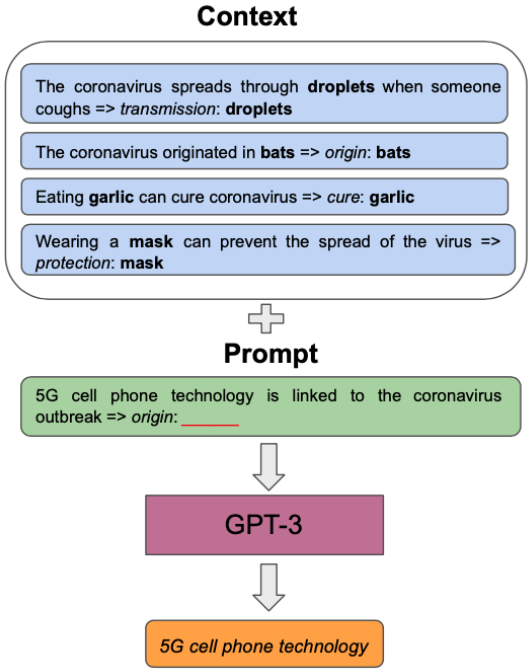


Figure 7: Figure showing the claim object detection sub-task input for GPT-3, with the few-shot labeled examples in context and the test example in the form of a prompt.

News Text: Another recent paper in Nature Medicine underscores that point. *"By comparing the available genome sequence data for known coronavirus strains, we can firmly determine that SARS-CoV-2 originated through natural processes," Kristian Andersen, PhD, an associate professor of immunology and microbiology at Scripps Research and corresponding author on the paper, said in a statement.* Andersen and colleagues' research implicates bats and possibly pangolins.

Topic: Origin of the virus
Stance: Affirm
Claim Object: natural processes
Claimer: Kristian Andersen

News Text: That the virus has natural origins is also apparent from its molecular structure. *Scientists writing in Nature Medicine journal on March 17 made clear that "all notable SARS-CoV-2 features" were also observed "in related coronaviruses in nature" and that therefore "we do not believe that any type of laboratory-based scenario is plausible."* As Josie Golding, epidemics lead at the UK-based Wellcome Trust, pointed out: The findings "are crucially important to bring an evidence-based view to the rumors that have been circulating about the origins of the virus."

Topic: Origin of the virus
Stance: Refute
Claim Object: laboratory
Claimer: Scientists

News Text: MYTH: Antibiotics are effective at treating coronavirus. *"Antibiotics do not treat viruses of any kind, including coronaviruses," Dr. Kenney states unequivocally*. Antibiotics target bacteria, not viruses.

Topic: Cure of the virus
Stance: Refute
Claim Object: Antibiotics
Claimer: Dr. Kenney

News Text: Gilead is working with the U.S. government on the logistics of remdesivir distribution and will provide more information when the company begins shipping the drug under the EUA. *"This EUA opens the way for us to provide emergency use of remdesivir to more patients with severe symptoms of COVID-19," said Daniel O'Day, Chairman and Chief Executive Officer of Gilead Sciences*. "We will continue to work with partners across the globe to increase our supply of remdesivir while advancing our ongoing clinical trials to supplement our understanding of the drug's profile.

Topic: Cure of the virus
Stance: Affirm
Claim Object: remdesivir
Claimer: Daniel O'Day

News Text: Does chlorine in pool water inactivate the virus? *"The good news is that the average amount of chlorine that's in a pool is going to kill the virus," Lavin says.* Assuming that your pool is properly maintained, the disinfecting chemicals in the water should be enough to render the virus inactive.

Topic: Protection from the virus
Stance: Affirm
Claim Object: chlorine
Claimer: Lavin

News Text: We reached out to him for a comment, but we haven't heard back. *On its website, the World Health Organization says that, while it recommends eating plenty of fruits and vegetables to stay healthy, there is no scientific evidence that lemon treats or prevents COVID-19 infection.* To prevent coronavirus infection, officials advise people to regularly wash their hands, avoid touching their face, disinfect surfaces in their homes daily and avoid people who are sick.

Topic: Protection from the virus
Stance: Refute
Claim Object: lemon
Claimer: World Health Organization

News Text: Airborne Transmission in Tight Spaces. *Medical professionals from the preeminent organizations on public health Centers for Disease Control and Prevention (CDC) and the World Health Organization have started changing their stance that COVID-19 is airborne.* This is important news.

Topic: Transmission of the virus
Stance: Affirm
Claim Object: air
Claimer: Medical professionals

News Text: If you are not staying in the same spot, like moving through a grocery store or walking, then your rate of infection decreases. *Doctors are concerned with the dosage of droplets that leads to infection.* As of the publication of this blog, doctors have not specified a dosage rate required for infection.

Topic: Transmission of the virus
Stance: Affirm
Claim Object: droplets
Claimer: Droplets

Figure 8: Some examples from the NEWSCLAIMS benchmark.

Please read the sentences for a news article snippet below, noting you will be answering questions regarding the red sentence.

There's currently no strong evidence that supplementing with vitamin C will prevent or cure COVID-19.

Most adults will also meet their vitamin C requirements from a diet that includes a variety of fruits and vegetables.

Myth 4: alkaline foods

Misinformation spread on social media suggests the virus can be cured by eating foods with a pH (level of acidity) that is higher than the virus's pH.

A pH below 7.0 is considered acidic, a 7.0 pH is neutral, and above pH 7.0 is alkaline.

Some of the "alkaline foods" said to "cure" coronavirus were lemons, limes, oranges, turmeric tea and avocados.

However, many of these online sources give incorrect pH values to these foods.

Consider the topics related to COVID-19 in the table below to determine if any *claims* are made regarding these topics:

Topics related to the virus	Example claims copied from instructions (same for all HITs)
Origin of the virus	• Illinois Senate Majority Leader Kimberly Lightford said the novel coronavirus was "man-made." • Research shows the genetic features of the virus rule out the possibility it was created or manipulated in a lab. • "Our analyses clearly show that SARS-CoV-2 is not a laboratory construct or a purposefully manipulated virus" • But to leading experts, the research is clear: the genetic structure of the virus shows it originated in bats"
Transmission of the virus	• Myth: Pets can spread the new coronavirus. July 14, 2020. • Coronavirus can be transmitted through mosquito bites. • COVID-19 cannot be transmitted in hotter, more humid climates. • The virus lingers in the air indoors, infecting those nearby.
Cure for the virus	• Vitamin C is an effective treatment for COVID-19. • Garlic does not cure COVID-19. • Colloidal silver has not been shown effective against new virus from China. • Convalescent plasma isn't quite the coronavirus miracle treatment it was supposed to be.
Protection from the virus	• Taking a hot bath prevents you from getting COVID-19. • Drinking alcohol reduces the risk of infection • Spraying chlorine or alcohol on the skin kills viruses on the body. • There's no evidence that taking vitamin C regularly can help prevent coronavirus or COVID-19.

Does the sentence in red contain a claim about any of the above four topics relating to the coronavirus? If yes, choose the appropriate topic. If no, choose None.

Note: Choose "None" if the sentence does not contain a claim. Please refer **Instructions** for definitions and detailed examples of such claims.

- ◯ Origin of the virus
- ◯ Transmission of the virus
- ◯ Cure for the virus
- ◯ Protection from the virus
- ◯ None

Figure 9: Screenshot of the phase 1 annotation interface.

Figure 10: Screenshot of the phase 2 annotation interface.