# COVID-Fact: Fact Extraction and Verification of Real-World Claims on COVID-19 Pandemic

**Arkadiy Saakyan[1], Tuhin Chakrabarty [1], and Smaranda Muresan[1,2]**
[1]Department of Computer Science, Columbia University
[2]Data Science Institute, Columbia University
a.saakyan@columbia.edu, {tuhin.chakr,smara}@cs.columbia.edu

## Abstract

We introduce a FEVER-like dataset COVID-Fact of 4,086 claims concerning the COVID-19 pandemic. The dataset contains claims, evidence for the claims, and contradictory claims refuted by the evidence. Unlike previous approaches, we automatically detect true claims and their source articles and then generate counter-claims using automatic methods rather than employing human annotators. Along with our constructed resource, we formally present the task of identifying relevant evidence for the claims and verifying whether the evidence refutes or supports a given claim. In addition to scientific claims, our data contains simplified general claims from media sources, making it better suited for detecting general misinformation regarding COVID-19. Our experiments indicate that COVID-Fact will provide a challenging testbed for the development of new systems and our approach will reduce the costs of building domain-specific datasets for detecting misinformation.

## 1 Introduction

The proliferation of disinformation and misinformation on the web is increasing at a scale that calls for the automation of the slow and labor-intensive manual fact-checking process (Vosoughi et al., 2018). *New York Times* reports that "Physicians say they regularly treat people more inclined to believe what they read on Facebook than what a medical professional tells them." Disinformation is even more acute around the recent COVID-19 pandemic. As a result, there is a need for automated fact-checking tools to assist professional fact-checkers and the public in evaluating the veracity of claims that are propagated online in news articles or social media.

Ideally, a fact-checking pipeline will address several tasks: 1) Consider real-world claims, 2) Retrieve relevant documents not bounded to a known

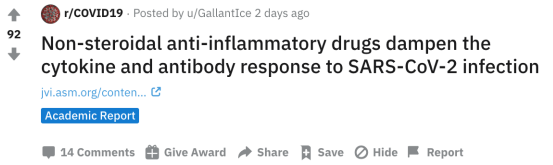


Figure 1: A claim from the *r/COVID19* subreddit with an academic report as an evidence source linked to it.

document collection (e.g., Wikipedia) and which contain information to validate the claim, 3) Select evidence sentences that can support or refute the claim and 4) Predict the claim veracity based on this evidence. Recent work on end-to-end fact-checking, including models and datasets, has advanced the field by addressing several tasks in the pipeline, but not all (Thorne et al., 2018, 2019; Hanselowski et al., 2019; Augenstein et al., 2019; Diggelmann et al., 2021; Wadden et al., 2020). One line of work that includes FEVER (Thorne et al., 2018, 2019) and SciFact (Wadden et al., 2020) addresses tasks 2, 3 and 4, but assumes a given document collection for task 2 (Wikipedia or CORD-19, respectively) and does not address task 1. Moreover, the refuted claims in these datasets are manually generated by asking humans to produce counter-claims for a given claim supported by a source document. Another line of work that includes Multi-FC (Augenstein et al., 2019) addresses tasks 1, 2 and 4, but not 3. It provides real-world claims collected from fact-checking websites and evidence documents and other meta-information, but it does not provide evidence sentences.

We propose a novel semi-automatic method to build a fact-checking dataset for COVID-19 (COVID-Fact) with the goal of facilitating all the above tasks. We make the dataset and code available for future research at https://github.com/asaakyan/covidfact. Our contributions are as



*Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 2116–2129
August 1–6, 2021. 

| | |
|---|---|
| Original Claim | Closed environments **facilitate** secondary transmission of coronavirus disease 2019 |
| Counter-Claim | Closed environments **prevent** secondary transmission of coronavirus disease 2019 |
| Gold Document | https://www.medrxiv.org/content/10.1101/2020.02.28.20029272v2 |
| Gold Evidence | It is plausible that closed environments contribute to secondary transmission of COVID-19 and promote superspreading events. |
| Original Claim | Oxford vaccine **triggers** immune response |
| Counter-Claim | Oxford vaccine **inhibits** immune response |
| Gold Document | https://www.bbc.com/news/uk-53469839 |
| Gold Evidence | They are injecting coronavirus RNA (its genetic code), which then starts making viral proteins in order to trigger an immune response. |

Table 1: Original and counter-claims from our dataset with gold documents and evidence sentences identified by our system supporting and refuting them, respectively.

follows:

- *Automatic real-world true claim and trustworthy evidence document selection* (Section 2.1). We start with the heavily moderated *r/COVID19* subreddit, that requires every claim/title post to be accompanied by a source evidence document from peer-reviewed research, pre-prints from established servers, or information reported by governments and other reputable agencies. Figure 1 shows one such claim with the associated source belonging to the *Academic Report* flair. We propose additional filtering methods to ensure source quality and that claims are well-formed. This step provides us with real-world true claims about COVID-19 and evidence documents not bounded to a known document collection. Moreover, the language of the claims could be both technical and lay (see Figure 1 and Table 1), unlike SciFact which is geared only towards scientific claims.

- *Automatic generation of counter-claims* (Section 2.2). An end-to-end fact-checking system requires both true and false claims for training. Following FEVER and SciFact, to obtain false claims, we aim to generate counter-claims of the original true claim. The advantage is that we obtain evidence documents/sentences for free. However, unlike FEVER and SciFact, we propose a novel approach to automatically generate counter-claims from a given claim using two steps: 1) select salient words from the true claim using attention scores obtained from a BERT (Devlin et al., 2019) model fine-tuned on the SciFact dataset, and 2) replace those words with their opposites using Masked Language Model infilling with entailment-based quality control. Table 1 shows examples of generated counter-claims.

- *Evidence sentence selection using text similarity and crowdsourcing* (Section 2.3). For evidence sentence selection, we calculate the semantic similarity between the original true claim and the sentences in source evidence documents using sentence-BERT (SBERT) (Reimers and Gurevych, 2019), retrieve top five sentences and use crowdsourcing for final validation. Table 1 shows examples of evidence sentences that support the true claims and refute the corresponding counter-claims.

- *COVID-Fact dataset of 4,086 real-world claims annotated with sentence-level evidence and a baseline on this task.* Our results show that models trained on current datasets (FEVER, SciFact) do not perform well on our data (Section 4). Moreover, we show the usefulness of our dataset through zero-shot performance on the scientific claim verification task on SciFact (Wadden et al., 2020) data (Section 4).

## 2 COVID-Fact Dataset Construction

The COVID-Fact dataset contains $4,086$ real-world claims with the corresponding evidence documents and evidence sentences to support or refute the claims. There are $1,296$ supported claims and $2,790$ automatically generated refuted claims. In this section, we present the three main steps to semi-automatically construct this dataset: **1)** real-world true claim and trustworthy evidence document selection (Section 2.1), **2)** automatic counter-claim generation (Section 2.2) and **3)** evidence sentence selection (Section 2.3).

### 2.1 Real-World Claim and Trustworthy Evidence Document Selection

The subreddit *r/COVID19* is a heavily moderated online discussion forum that seeks to facilitate scientific discussion around COVID-19. Each post on this subreddit has a title and needs to contain a link to a source, governed by several rules: posts linking to non-scientific sources will be removed; comments making a statement as fact or which include figures or predictions also need to be supported by evidence; allowed sources include peer-reviewed research, pre-prints from established servers, and information reported by governments and other reputable agencies. Moreover, the posts are annotated with "flairs", or short description of the posts' category such as Academic Report, Academic Comment, Preprint, Clinical, Antivirals, Government Agency, Epidemiology, PPE/Mask research, General. Having access to such flairs allows to select claims, for example, related to "Vaccine research" or "Epidemiology". This could further help in training models targeting even more specific types of disinformation, like disinformation about antivirals or PPE/masks. In our study, the titles of the post are considered *candidate claims* and the associated sources are considered *evidence documents*. Posts from the *r/COVID19* subreddit are extracted via the Pushshift Reddit API.[1] Two issues still need to be addressed: 1) ensure that titles are well-formed claims; 2) ensure the highest trustworthiness of the posts and their associated sources.

**Filtering for well-formed claims.** The definition of a claim can vary depending on domain, register or task (Daxenberger et al., 2017). For our work, we consider a claim to be a proposition whose truthfulness can only be determined by additional evidence. In addition, a well-formed claim has to be a full sentence. Thus, to filter out most of the titles that are not well-formed claims, we employ a simple syntax-based approach to remove questions and consider statements that have at least a main verb. This filtering steps allows us to remove titles such as "B cell memory: understanding COVID-19" and consider titles such as the ones in Figure 1 and Table 1. In addition, we ask three volunteer computer science students with background in argumentation and linguistics to manually verify that the entire resulting set does indeed contain only well-formed claims. While we could have employed more sophisticated claim detection methods, there are no large-scale datasets for COVID-19 to train a claim detection model. We therefore did not want to introduce additional noise in our dataset by using a machine learning approach.

**Filtering for trustworthiness.** To ensure high trustworthiness of posts (and thus our true claims) and the linked sources, we employ several filtering steps. First, the posts in this subreddit undergo moderation, and thus we discard titles/claims that belong to posts flagged as taken down by the moderators using the posts' "removed" flair. Moreover, users of the Reddit platform may upvote or downvote a post, and the ratio of upvotes can serve as a rough indication of the reliability of the source. Hence, posts (and thus claims/sources) with upvote ratio lower than 0.7 are rejected. We then reject claims where the linked source in the post has an Alexa Site Rank[2] lower than $50,000$, rejecting the outliers by the site rank (see the box plot in Appendix B.2). Finally, we reject a claim if the linked source in the post does not appear in the top 5 Google search results when querying the title of the post.

From an initial set of $22,646$ posts, automatic syntactic filtering for well-formed claims results in a set of $6,154$ claims, further reduced to $1,526$ after filtering for trustworthiness and finally reduced to $1,407$ through manual validation. Thus, the resulted dataset after all the filtering steps consists of $1,407$ true claims and the associated source evidence documents (an additional set of 111 claims are removed in the evidence sentence selection step in Section 2.3). Besides the linked source document in the post, we retrieve for each claim four additional sources from the top 5 Google search results. This is motivated by the fact that the same claim can be reported by various sources. For example, the second claim in Table 1 "*Oxford vaccine triggers immune response*" is reported, besides the *bbc.com* given in the original post, also by other trustworthy sources such as *usnews.com*, *medscape.com*, *cnbc.com*. Unlike FEVER and SciFact, which constrain their evidence document collection to Wikipedia or pre-selected scientific articles, we collect evidences from any of the websites linked to the Reddit post or appearing in the top 5 Google search results. Even though over time the Google search results may change, the collection of evidence documents for COVID-Fact is considered

[1] https://github.com/pushshift/api

[2] https://www.alexa.com/topsites

fixed and will be released for reproducibility.

Like SciFact (Wadden et al., 2020), our dataset contains several claims with scientific jargon such as "*Altered blood cell traits underlie a major genetic locus of severe COVID-19*". However, unlike SciFact, our dataset also contains scientific claims expressed in lay terms. For example, a claim like "*Loss of smell is a symptom of COVID-19*" is much simpler and can be understood by a wider audience compared to "*Emerging evidence supports recently acquired anosmia and hyposmia as symptoms of COVID-19*". This is important, as a lot of (dis)information is expressed in lay language intended for the general public not versed in scientific language. Another issue adding to the complexity of the task around COVID-19 (dis)information are non-scientific claims that focus on public health policies or statements from public health authorities. For example, a claim like "*CDC says new COVID strain in UK could already be circulating undetected in U.S*" would not occur in scientific literature, but occurs in media outlets linked as sources in the *r/COVID19* subreddit.

### 2.2 Automatic Counter-Claim Generation

An end-to-end fact-checking system requires both true and false claims. Following FEVER and SciFact, to obtain false claims we aim to generate counter-claims of the original true claims (from Section 2.1). However, in FEVER (Thorne et al., 2018) and SciFact (Wadden et al., 2020) the generation of counter-claims was done manually by human annotators, which is an expensive approach that might not scale well. We propose an approach to generate counter-claims automatically (see Table 1 for examples). Our counter-claim generation consists of two stages: 1) select salient words from the true claims, and 2) replace those words with their opposite using Mask Language Model infilling with entailment-based quality control. We discuss these steps below.

#### 2.2.1 Salient Words Selection

Salient words (keywords) are essential to the overall semantics of a sentence. For example, in the claim "*Oxford vaccine triggers immune response*", a salient word would be **"triggers"**. By changing the word "triggers" to "inhibits" we change the meaning of above claim to its opposite (counter-claim). Recently Zhang et al. (2020b) used YAKE (Campos et al., 2018, 2020), an unsupervised automatic keyword extraction method for selecting salient words to guide their text generation process. For selecting salient words from a claim, we experiment with YAKE as one of our methods. In addition, we explore an attention-based method described below.

**Attention-Based Salience.** Recently, Sudhakar et al. (2019) use self-attention scores from BERT (Devlin et al., 2019) to delete keywords from an input sequence for the task of Style Transfer. They use a novel method to extract a specific attention head and layer combination that encodes style information and that can be directly used as importance scores. Inspired by them, we use the same approach for our task. We fine-tuned BERT for a sentence classification task (veracity prediction) on the SciFact (Wadden et al., 2020) dataset, and extract the attention scores from the resulting model. Given the SciFact dataset $D = (x_1, y_1), ..., (x_m, y_m)$ where $x_i$ is a claim and $y_i \epsilon$ {SUPPORTED, REFUTED} is a veracity label we observe that the self-attention based classifier defines a probability distribution over labels: $p(y|x) = g(v, \alpha)$ where $v$ is a tensor such that $v[i]$ is an encoding of $x[i]$, and $\alpha$ is a tensor of attention weights such that $\alpha[i]$ is the weight attributed to $v[i]$ by the classifier in deciding probabilities for each $y_j$ . The $\alpha$ scores can be treated as importance scores and be used to identify salient words.

**Quality of Salient Words Selection.** We evaluate how well our salient word selection methods correlate with human judgement. We randomly select 150 original claims for an Amazon Mechanical Turk task. The annotators were asked to select a word that could potentially invert the meaning of the sentence if it were to be replaced. For every claim, three separate annotators were recruited which means that we would have at most three different chosen salient keywords. For each claim, we compute the set intersections between the three keywords selected by our automatic methods (YAKE and Attention-based) vs. the keywords selected by the annotators on AMTurk. We found that keywords selected using self-attention scores have a significantly higher recall (Two-Proportion Z-test with p-value $< .00001$) than YAKE ($68\%$ vs. $54\%$). The average number of words per claim in COVIDFact is 14, so the task of selecting one salient keyword is challenging even for humans. Given this, our Recall@3 scores demonstrate the reliability of automatic attention-based salient word selection.

#### 2.2.2 Masked Language Model Infilling with Entailment-based Quality Control

After selecting salient words from the true claims for replacement, we need to provide only paraphrases that are opposite in meaning and consider the context in which these words occur. Language models have been used previously for infilling tasks (Donahue et al., 2020) and have also been used for automatic claim mutation in fact checking (Jiang et al., 2020). Inspired by these approaches, we use the Masked Language Model (MLM) RoBERTa (Liu et al., 2019) fine-tuned on CORD-19 (Wang et al., 2020) for infilling. The fine-tuned RoBERTa is available on Huggingface [3]. We generate a large number (10-30) of candidate counter-claims with replaced keywords per each original claim.

After generating multiple candidate counter-claims based on MLM infilling, we select the ones that have the highest contradiction score with the original claim. To compute the contradiction score we use the RoBERTa (Liu et al., 2019) model trained on Multi-NLI (Williams et al., 2018) due to its size and diversity. The scores are in the range from 0 to 1. We first set the minimum score threshold and then select top three claims above the threshold.

To select the right threshold for contradiction score-based filtering we perform the following experiment. We presented 150 randomly selected claims to Amazon Mechanical Turk workers. Annotators were presented with the original claim and five generated candidate counter-claims from MLM infilling. They were then asked if those claims are implied by the original claim (hence, for example, noun shifts would be judged as "not implied"). We labeled claims as "contradictory" if the majority of the annotators agreed on the label. We observed a point-biserial correlation of 0.47 between dichotomous human judgement and continuous contradiction scores, indicating moderate agreement. We convert the contradiction scores to binary outcomes, assigning 1 if the score is above the threshold and 0 otherwise. We compute precision, recall, F1 score and accuracy for different thresholds. As threshold value increases, we see a steady increase in precision, indicating that taking a higher threshold value we are almost guaranteed to select a contradictory sentence (for example, for a threshold of 0.995, precision is 93%). Obviously, this comes at a cost of decreased recall. We selected a threshold of 0.9 (precision 76%), since we want to prioritize precision, but do not want to reduce our dataset too much due to the low recall. At this threshold, our 1,407 claims generate additional 4,042 false claims. An alternative approach of replacing salient words with antonyms from standard lexicons like WordNet (Miller, 1995) was considered. However, a suitable antonym was absent in several cases, most notably nouns. The RoBERTa model is able to provide domain-aware substitutions. For example, replacing the word "humans" by the word "mice" reverses the meaning of the claim the domain of clinical trial reports, yet the words human and mouse can hardly be considered antonyms. Lexical replacement without consideration of context can also cause grammatical issues.

Our method of counter-claim generation only changes a single word or a multi-word expression, since pre-trained MLMs like BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) do not allow for multiple word masking. However, this method can be extended to masking multiple words using recent pre-trained language models like BART (Lewis et al., 2020).

#### 2.2.3 Analysis of Counter-claim Generation

Upon deeper inspection we observe that the attention scores described in Section 2.2.1 were distributed across different parts of speech like *verbs* or *adjective modifiers* or *nouns*. We show the distribution of the most frequent parts of speech of salient words and replacement words in our dataset in Figure 2. This means our counter-claims were generated with more creativity than just the addition of obvious triggers like "not". The majority of claim negations involved a reversal of effect direction; for instance "*Suspicions grow that nanoparticles in Pfizer's COVID-19 vaccine trigger rare allergic reactions.*" was negated as "*Suspicions grow that nanoparticles in Pfizer's COVID-19 vaccine trigger* ***systemic*** *allergic reactions.*" where a simple adjective modifier changes the truthfulness. Similarly for a claim "*Electrostatic spraying will prevent the spread of COVID-19*" a negated claim is "*Electrostatic spraying will* ***facilitate*** *the spread of COVID-19*" which flips the main verb in the claim. In Table 2, one can see several examples of how the generated counter-claims reverse the meaning of the original sentence.

[3] https://huggingface.co/amoux/roberta-cord19-1M7k

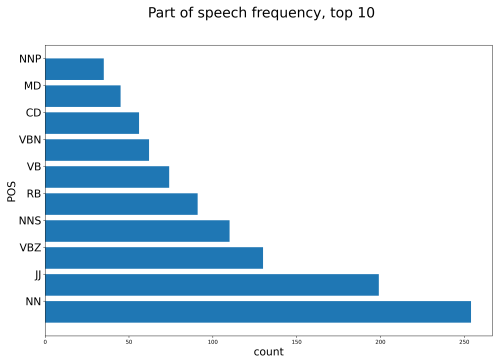


Figure 2: Most frequent POS tags of salient words.

| Original | Generated |
|---|---|
| *..**people** in UK receive ..* | *..**mice** in UK receive . . .* |
| *.. **human** ACE2.* | *.. **bat** ACE2.* |
| ***FDA** takes key action ..* | ***WHO** takes key action ..* |
| *..**improves** the effect ..* | *..**inhibits** the effect ..* |
| *..**blocks** SARS-CoV-2..* | *..**enchanced** SARS-CoV-2..* |
| *..are **not** fit for purpose ..* | *..are **good** fit for ..* |
| *.. the **final** stage ..* | *.. the **first** stage ..* |
| *.. shows **positive** results.* | *.. shows **no** results.* |

Table 2: A detailed look into what parts of speech are replaced, and in what direction the claims are reversed. We omitted full claims due to space constraint. The first 3 claims show nouns, the next 2 show verbs and the final 3 show adjective modifications.

### 2.3 Evidence Sentence Selection

To select evidence sentences we follow the approach proposed by Hidey et al. (2020). Given the true claims and the 5 evidence documents for each claim (Section 2.1) we use cosine similarity on SBERT sentence embeddings (Reimers and Gurevych, 2019) to extract the top 5 sentences most similar to the true claim. Note that we only need to do this step for true claims, as automatically the evidence sentences that support the true claim will be the evidence sentences that refute the corresponding counter-claims. Sentences containing the claim itself were discarded. The collected five sentences will serve as candidate evidence sentences for future human validation described below.

| Split | Supported | Refuted |
|---|---|---|
| **Train** | 1036 | 2227 |
| **Dev** | 130 | 289 |
| **Test** | 130 | 274 |

Table 3: Breakdown of claims by label for train, dev, test sets.

**Crowdsourcing for Final Evidence Sentence Selection.** Amazon Mechanical Turk workers were given a claim and the 5 automatically selected candidate evidence sentences. They were asked to select which of the evidence sentences support the claim (they could select several) or they could select that the evidence is absent. To discourage low quality responses, we used a trick sentence that would allow us to disqualify dishonest entries. For the trick we used a phrase "*It is not true that*" concatenated with the original sentence, and rejected entries that marked that option as evidence for the claim. In 111 cases, annotators could not agree on the evidence or agreed that the evidence was absent, where agreement is defined as the majority vote. We disregard these true claims from our COVID-Fact dataset as they would not have associated evidence sentences.

We asses the quality of the majority vote annotations by comparing the gold evidence label annotations with an independent re-annotation by three Amazon Mechanical Turk workers. We select a sample of 100 claims' evidences (7% of the 1,296 original claims). We observe a Cohen's kappa (Cohen, 1968) of 0.5 between majority votes of the two independent groups of Amazon Turk workers, indicating moderate agreement (Artstein and Poesio, 2008). We find this encouraging given the complexity of the task, especially considering that the workers did not have domain-specific knowledge.

## 3 Experimental setup

Table 3 shows the dataset statistics for train/dev/test split of COVID-Fact.

### 3.1 COVID-Fact Task Formulation

COVID-Fact Task follows the FEVER shared task definition. The set of all claims is denoted by $C$. The set of gold evidence sentences for a claim $c \in C$ is denoted by $E(c)$. The gold label for a given claim and evidence pair is defined as $v(c, E(c)) \in \{\text{SUPPORTED, REFUTED}\}$. The task consists of the following subtasks outlined below.

**Evidence Retrieval.** Given a claim $c$, a system must retrieve a set of up to five evidence sentences $\hat{E}(c)$. We evaluate the evidence retrieval system quality using precision, recall, and F1 scores. Evidence recall is computed as the number of evidence sets that contain a gold evidence over the total number of evidence sets.

**Veracity Prediction.** Given a claim $c$ and a set of evidence sentences $E(c)$, a system must determine a label $\hat{v}(c, E(c)) \in \{\text{SUPPORTED, REFUTED}\}$. We evaluate veracity prediction using F1 score and accuracy.

**Evidence Retrieval + Veracity Prediction (COVID-FEVER Score)** Given a claim $c$, a system must retrieve a set of evidence sentences $\hat{E}(c)$, and determine a label $\hat{v}(c, \hat{E}(c)) \in \{\text{SUPPORTED, REFUTED}\}$. A claim has a COVID-FEVER of 1 if it correctly predicts the veracity of the claim-evidence pair and if at least one of the predicted evidence from the predicted evidence matches the gold evidence selected by annotators (thus a stricter score than veracity prediction accuracy). This metric is similar to the FEVER score (Thorne et al., 2018).

### 3.2 Baseline Pipeline for COVID-Fact

Our end-to-end pipeline consists of the following steps: 1) Evidence retrieval using Google Search + SBERT 2) Veracity prediction using RoBERTa fine-tuned on fact-checking and entailment inference datasets.

**Baseline for Evidence Retrieval.** We use the same approach as was used for the construction of the dataset to provide a strong baseline for evidence retrieval on COVID-Fact. Google search was used to identify five potential source documents by querying the claim. This step is followed by selecting most similar sentences through computing cosine similarity between sentence embeddings of the claim and candidate sentences using SBERT (Reimers and Gurevych, 2019).

**Baseline for Veracity Prediction.** Our baseline for veracity prediction is a RoBERTa model. We concatenate all evidence sentences in the evidence set and use it as input for a binary classification task similar to the GLUE RTE task (Wang et al., 2018). We evaluate the models with gold evidence, as well as Top-5 and Top-1 evidences ranked by SBERT cosine similarity with the original claim.

### 3.3 Experiments

Besides evaluating our baseline pipeline on the COVID-Fact dataset, we perform several additional experiments outlined below. All hyperparameters can be found in Appendix A.

**Adequacy of Existing Datasets for COVID-Fact.** For the task of veracity prediction, we evaluate the performance of RoBERTa-large fine-tuned on FEVER, SciFact, MNLI and our COVID-Fact dataset. Moreover, we also experiment with fine-tuning RoBERTa-large on SciFact + COVID-Fact and on FEVER + COVID-Fact.

**Usefulness of COVID-Fact for Zero-Shot Scientific Fact-checking.** Even though not explicitly designed for COVID-19 related claims, Wadden et al. (2020) showed how models trained on the SciFact dataset could verify claims about COVID-19 against the research literature. COVID-Fact on the contrary was not explicitly designed for scientific fact-checking, although our resource contains a substantial number of scientific claims. This provides us the opportunity to test the generalizability and robustness of our dataset. To do so, we train models on COVID-Fact claims and gold evidence and evaluate the veracity performance on the SciFact dev set in a zero-shot setting. We remove the NOT ENOUGH INFO claims from the SciFact dataset.

## 4 Results and Analysis

Table 5 summarizes the results for the evidence retrieval evaluation. Our pipeline provides a strong baseline with F1 score of $\approx 32$. For comparison, the baseline system in FEVER (Thorne et al., 2019) achieves the F1 score of 18.26. Note Top 5 evidence retrieval performs worse than gold since we evaluate how the system performs with automatically negated claims as well, for which we re-run the Google+SBERT method.

Table 4 summarizes the results for the veracity prediction task using gold and retrieved evidence. We observe that, given the gold evidences, fine-tuning on COVID-Fact led to performance improvement of 25 F1-score and 35 F1-score compared to training solely on SciFact and FEVER respectively. This indicates that the COVID-Fact dataset is challenging and cannot be solved using popular fact-checking datasets like FEVER and SciFact. This could be explained by the fact that claims about COVID-19 are comprised of a mix of scientific and general-domain claims. The poor macro-F1 score for claim only baseline shows that the model does not learn spurious correlation between a claim and the veracity label. With Top 5 and Top 1 retrieved evidences, we observed that COVID-Fact is still difficult to outperform. The zero-shot performance is negligibly affected by the retrieved evidence. Our baseline pipeline achieves

| Veracity Prediction | | | | | | | | COVID-FEVER |
|---|---|---|---|---|---|---|---|---|
| | Gold | | Top 5 | | Top 1 | | | Top 5 |
| | Acc | F1 | Acc | F1 | Acc | F1 | | Score |
| MNLI (Williams et al., 2018) | 61.3 | 64.2 | 53.1 | 51.5 | 65.4 | 60.6 | | 35.1 |
| SciFact (Wadden et al., 2020) | 56.9 | 57.0 | 53.7 | 54.0 | 54.3 | 54.0 | | 36.9 |
| FEVER (Thorne et al., 2018) | 48.3 | 47.0 | 46.2 | 45.0 | 48.6 | 48.0 | | 35.4 |
| COVID-Fact | **83.5** | **82.0** | **84.7** | **83.0** | **83.2** | **81.0** | | **43.3** |
| SciFact + COVID-Fact | 82.2 | 81.0 | 83.0 | 82.0 | 80.2 | 79.0 | | 43.0 |
| FEVER + COVID-Fact | 74.8 | 70.0 | 78.2 | 73.0 | 73.3 | 68.0 | | 35.4 |
| COVID-Fact (Claim only) | 67.5 | 40.0 | - | - | - | - | | - |

Table 4: Performance of various training configurations of RoBERTa-large in the Veracity Prediction as well as Evidence Retrieval + Veracity Prediction (See Section 3.1). The top 3 rows under Veracity Prediction show a zero shot setting where models are trained on existing fact-checking datasets. We test the model performance on claims with gold evidence selected by humans VS claims with top 5 retrieved evidences and top 1 retrieved evidence on COVID-Fact test set. $p < .001$ using approximate randomization test.

| | **Evidence Retrieval** | | |
|---|---|---|---|
| | P | R | F1 |
| Top 5 | 22.27 | **52.37** | 31.25 |
| Top 3 | 24.77 | 45.14 | **31.99** |
| Top 1 | **29.68** | 29.93 | 29.80 |

Table 5: Performance of our system's Evidence Retrieval part (see Section 3.1). We compare the precision (P), Recall (R), and F1-score of top 5, top 3, top 1 retrieved sentences, ranked by SBERT cosine similarity score.

| Train Setting | Acc | F1 |
|---|---|---|
| COVID-Fact | 80.8 | 80.0 |
| Sci-Fact | 83.7 | 83.0 |

Table 6: Two-way Veracity prediction results on Sci-fact dev set by models trained on COVID-Fact data as well as Sci-Fact data.

| | |
|---|---|
| C1 | SARS-CoV-2 is **not detectable** in the vaginal fluid of **women** with severe COVID-19 infection |
| EV1 | **All 10 patients** were tested for SARS-CoV-2 in vaginal fluid,and all samples tested **negative** for the virus. |
| C2 | Baricitinib **restrains the immune dysregulation** in COVID-19 patients |
| EV2 | Here, we provide evidences on the efficacy of Baricitini, a JAK1/JAK2 inhibitor, in **correcting the immune abnormalities** observed in patients hospitalized with COVID-19. |

Table 7: Claims (C1 & C2) which are classified incorrectly as REFUTES in the light of SUPPORTing evidence by our best veracity models. Words crucial for correct verification are highlighted.

the COVID-FEVER score of 43.3 using Top 5 evidence sentences. Adding the FEVER and SciFact datasets deteriorates the results.

Table 6 shows a strong zero shot performance of COVID-Fact for scientific claim verification (training on COVID-Fact train set, testing on the SciFact dev set). SciFact only contains scientific claims, therefore the model trained only on SciFact does not generalize well to COVID-Fact, which also contains non-scientific claims. COVID-Fact, on the other hand, contains enough scientific claims so that the model generalizes well to SciFact. This result shows semi-automated COVID-Fact is not inferior to mostly manual SciFact.

**Error analysis.** We observe that errors in veracity prediction can be attributed to three factors: Cause and Effect, Commonsense or Scientific Background. For instance, in the first (C1, EV1) pair in Table 7, *not detectable* is the Cause while *testing negative* is the Effect. To verify this claim, the veracity model needs to have knowledge of counterfactuals. Furthermore, it should be understood that *All 10 patients* mention in EV1 should refer to *women* in C1, due to mention of "vaginal fluids" but this requires commonsense knowledge outside the text. Finally, it might be hard for veracity models to correctly classify claim evidence pairs which include knowledge of domain-specific or scientific lexical relationships. For instance in (C2, EV2) we see that both bolded phrases in red and blue refer to the same phenomena, but *immune dysregulation* is "a breakdown of immune system processes" and restraining it can be seen as the same concept as *correcting immune abormalities*, but the model is not able to capture such complex domain specific knowledge.

## 5 Related Work

**Fact-Checking.** Approaches for predicting the veracity of naturally-occurring claims have focused

on statements fact-checked by journalists or organizations such as PolitiFact.org (Vlachos and Riedel, 2014; Alhindi et al., 2018), news articles (Pomerleau and Rao, 2017), or answers in community forums (Mihaylova et al., 2018, 2019). Mixed-domain large scale datasets such as UKP Snopes (Hanselowski et al., 2019), MultiFC (Augenstein et al., 2019), and FEVER (Thorne et al., 2018, 2019) rely on Wikipedia and fact-checking websites to obtain evidences for their claims. Even though these datasets contain many claims, due do domain mismatch they may be difficult to apply for COVID-19 related misinformation detection. SciFact (Wadden et al., 2020) introduced the task of scientific fact-checking, generating a dataset of 1.4K scientific claims and corresponding evidences from paper abstracts annotated by experts. However, the dataset does not contain simplified scientific claims encountered in news and social media sources, making it difficult to optimize for a misinformation detection objective. Another approach to misinformation detection similar to ours is CLIMATE-FEVER (Diggelmann et al., 2021). They adapted FEVER methodology to create a dataset specific to climate change fact-checking. However, due to difficult and expensive methods employed for generation of FEVER, it can be difficult to extrapolate this method to assemble a COVID-19 specific dataset.

**COVID-19 related NLP tasks.** Numerous NLP approaches were employed to aid the battle with the COVID-19 pandemic. Notably Wang et al. (2020) released CORD-19, a dataset containing 140K papers about COVID-19 and related topics while Zhang et al. (2020a) created a neural search engine COVIDEX for information retrieval. To combat misinformation Lee et al. (2020) proposed a hypothesis that misinformation has high perplexity. Hossain et al. (2020) released COVIDLIES: a dataset of 6761 expert-annotated tweets matched with their stance on known COVID-19 misconceptions. The dataset provides a comprehensive evaluation of misconception retrieval but does not analyze evidence retrieval and prediction of veracity of claims based on presented evidence. Poliak et al. (2020) collected 24,000 Question with expert answers from 40 trusted websites to help NLP research with COVID related information. COVID-Fact, on the other hand, deals with real world claims and presents an end-to-end fact checking system to fight misinformation.

## 6 Conclusion

We release a dataset of 4,086 claims concerning the COVID-19 pandemic, together with supporting and refuting evidence. The dataset contains real-world true claims obtained from the *r/COVID19* subreddit as well as automatically generated counter-claims. Our experiments reveal that our dataset outperforms zero-shot baselines trained on popular fact-checking benchmarks like SciFact and FEVER. This goes on to prove how domain-specific vocabulary may negatively impact the performance of popular NLP benchmarks. Finally, we demonstrate a simple, scalable, and cost-efficient way to automatically generate counter-claims, thereby aiding in creation of domain-specific fact-checking datasets. We provide a detailed evaluation of the COVID-Fact task and hope that our dataset serves as a challenging testbed for end-to-end fact-checking around COVID-19.

## 7 Ethics

The data was collected from Reddit keeping user privacy in mind. Reddit is a platform where users post publicly and anonymously. For our dataset, only titles and links to external publicly available sources like news outlets or research journals were collected, as well as post metadata such as flairs, upvote ratio, and date of the post. User-identifying information, including, but not limited to, user's name, health, financial status, racial or ethnic origin, religious or philosophical affiliation or beliefs, sexual orientation, trade union membership, alleged or actual commission of crime, was not retrieved and is not part of our dataset. For all the crowdsourcing annotation work, we fairly compensate crowd workers in accordance with local minimum wage guidelines.

One significant concern might arise regarding the use of language models for counter-claim generation. Our model is a controlled generation system (word-level replacement) and is not suited for generation of entirely new and original claims. Neither it is the case that it can be used for generation of entire articles of false information, or generating false evidence for the counter-claims. The model for replacing keywords from original claims is trained on CORD-19 (Wang et al., 2020), a scientific corpus of high quality and trustworthy information about COVID-19. We generate counter-claims to create a resource that will help NLP models learn how to identify false information and provide evidence

for the predicted label leading to more explainable models. Consequently, our approach is suited for improving entailment and veracity prediction performance of fact-checking systems, rather than improving generative qualities of false-claim generation systems. The fact that we use our model to generate false claims also helps to address the concerns of biased language generation. In the unlikely event our model produces biased claims, they could serve as good examples of false claims containing bias, which would be an interesting topic for further research (bias in disinformation). We therefore believe the net positive impact of our work far outweighs the potential risks.

## References


Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. 2018. Where is your evidence: Improving fact-checking by justification modeling. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 85–90, Brussels, Belgium. Association for Computational Linguistics.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Comput. Linguist.*, 34(4):555–596.

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, Hong Kong, China. Association for Computational Linguistics.

Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Mário Jorge, Célia Nunes, and Adam Jatowt. 2018. Yake! collection-independent automatic keyword extractor. In *European Conference on Information Retrieval*, pages 806–810. Springer.

Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. 2020. Yake! keyword extraction from single documents using multiple local features. *Information Sciences*, 509:257 – 289.

Jacob Cohen. 1968. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological bulletin*, pages 213–220.

Johannes Daxenberger, Steffen Eger, Ivan Habernal, Christian Stab, and Iryna Gurevych. 2017. What is the essence of a claim? cross-domain claim identification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2055–2066, Copenhagen, Denmark. Association for Computational Linguistics.

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.

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2021. Climate-fever: A dataset for verification of real-world climate claims.

Chris Donahue, Mina Lee, and Percy Liang. 2020. Enabling language models to fill in the blanks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2492–2501, Online. Association for Computational Linguistics.

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, Hong Kong, China. Association for Computational Linguistics.

Christopher Hidey, Tuhin Chakrabarty, Tariq Alhindi, Siddharth Varia, Kriste Krstovski, Mona Diab, and Smaranda Muresan. 2020. DeSePtion: Dual sequence prediction and adversarial examples for improved fact-checking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8593–8606, Online. Association for Computational Linguistics.

Tamanna Hossain, Robert L. Logan IV, Arjuna Ugarte, Yoshitomo Matsubara, Sean Young, and Sameer Singh. 2020. COVIDLies: Detecting COVID-19 misinformation on social media. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.

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*, pages 3441–3460, Online. Association for Computational Linguistics.

Nayeon Lee, Yejin Bang, Andrea Madotto, and Pascale Fung. 2020. Misinformation has high perplexity.

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, Online. Association for Computational Linguistics.

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. *CoRR*, abs/1907.11692.

Tsvetomila Mihaylova, Georgi Karadzhov, Pepa Atanasova, Ramy Baly, Mitra Mohtarami, and Preslav Nakov. 2019. Semeval-2019 task 8: Fact checking in community question answering forums. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 860–869.

Tsvetomila Mihaylova, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, Mitra Mohtarami, Georgi Karadzhov, and James R. Glass. 2018. Fact checking in community forums. *CoRR*, abs/1803.03178.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019a. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019b. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Adam Poliak, Max Fleming, Cash Costello, Kenton W Murray, Mahsa Yarmohammadi, Shivani Pandya, Darius Irani, Milind Agarwal, Udit Sharma, Shuo Sun, Nicola Ivanov, Lingxi Shang, Kaushik Srinivasan, Seolhwa Lee, Xu Han, Smisha Agarwal, and João Sedoc. 2020. Collecting verified COVID-19 question answer pairs. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.

Dean Pomerleau and Delip Rao. 2017. Fake news challenge.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Akhilesh Sudhakar, Bhargav Upadhyay, and Arjun Maheswaran. 2019. "transforming" delete, retrieve, generate approach for controlled text style transfer. 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 3269–3279, Hong Kong, China. Association for Computational Linguistics.

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. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2019. The FEVER2.0 shared task. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, pages 1–6, Hong Kong, China. Association for Computational Linguistics.

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 18–22. Association for Computational Linguistics.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Michael Kinney, Yunyao Li, Ziyang Liu, William Merrill, Paul Mooney, Dewey A. Murdick,

Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Nancy Xin Ru Wang, Christopher Wilhelm, Boya Xie, Douglas M. Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020. CORD-19: The COVID-19 open research dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Edwin Zhang, Nikhil Gupta, Rodrigo Nogueira, Kyunghyun Cho, and Jimmy Lin. 2020a. Rapidly deploying a neural search engine for the COVID-19 Open Research Dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online. Association for Computational Linguistics.

Yizhe Zhang, Guoyin Wang, Chunyuan Li, Zhe Gan, Chris Brockett, and Bill Dolan. 2020b. POINTER: Constrained progressive text generation via insertion-based generative pre-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8649–8670, Online. Association for Computational Linguistics.

## A Model implementation details

We used fairseq library (Ott et al., 2019a) for RoBERTa model training.

### A.1 Salient word selection hyperparameters

We use the uncased BERT model since many titles contain words that are all capitalized. We train the model on the SciFact classification task using 15 epochs and batch size of 16. The training loss is $7.15e-03$. The rest of the parameters are set to default as in (Sudhakar et al., 2019).

### A.2 Veracity prediction hyperparameters

- **No of Parameters:** We use the RoBERTA-large checkpoint (355M parameters) and use the FAIRSEQ implementation (Ott et al., 2019b) [4].
- **No of Epochs:** We fine-tune pre-trained RoBERTa for 10 epochs for each model and save the best model based on validation accuracy on COVIDFact.
- **Training Time:** Our training time is 30 minutes for each model except for ones with FEVER which takes around 10 hours.
- **Hardware Configuration:** We use 2 RTX 2080 GPUs.
- **Training Hyper parameters:** We use the same parameters as the FAIRSEQ github repository where RoBERTa was fine-tuned for the RTE task in GLUE with the exception of the size of each mini-batch, in terms of the number of tokens, for which we used 1024. [5]

## B Dataset statistics

Figures below visualize most frequent flairs in the dataset, as well as word clouds with keywords and replaced words.

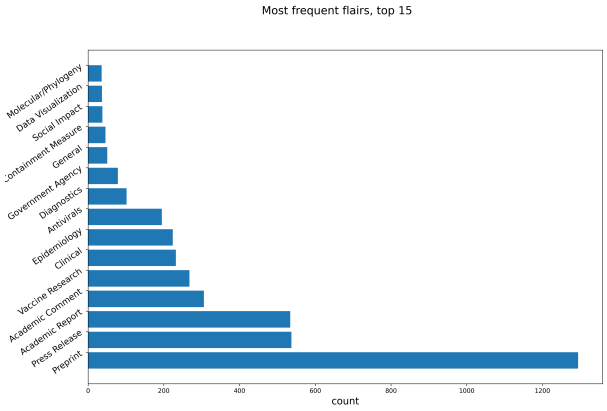


Figure 3: Most frequent flairs in the dataset.


Figure 4: Top image: word cloud of salient words. Bottom image: word cloud of replaced words.

[4] https://github.com/pytorch/fairseq/tree/master/examples/roberta

[5] https://github.com/pytorch/fairseq/blob/master/examples/roberta/README.glue.md

### B.1 Word replacement statistics

Figures below show most frequent salient words as well as most frequent words that were used to replace the salient words (replacement words). POS tags obtained using the flair python library tagger.

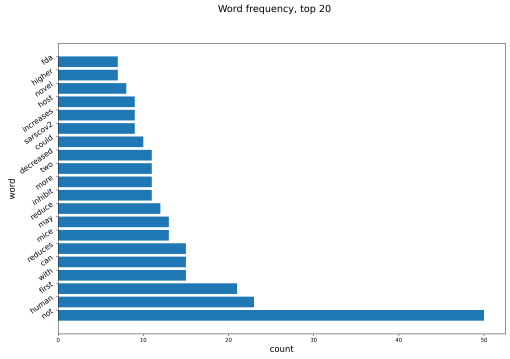


Figure 5: Most frequent salient words.

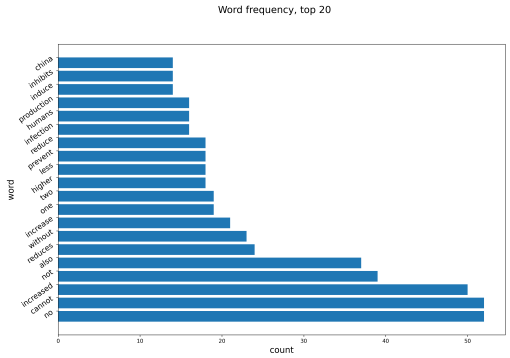


Figure 6: Most frequent replacement words.

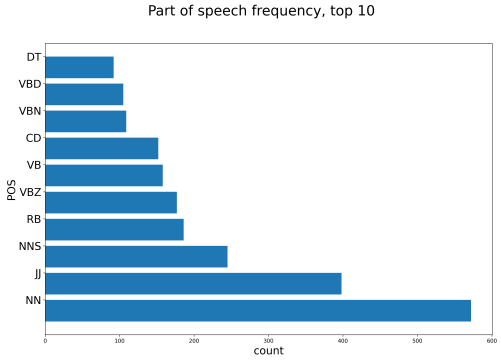


Figure 7: Most frequent POS tags of replacement words.

### B.2 Alexa threshold

A boxplot that helped us select the 50,000 Alexa siterank threshold. The plot shows site ranks for 2K initially scraped claims. Outliers (points outside of the whiskers of the plot) are all above the 50,000 threshold.

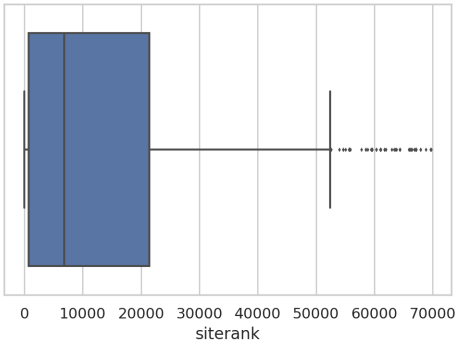


Figure 8: Alexa Site Rank boxplot.