Verification of Claims in the COVID-19 Pandemic

Saptarashmi Bandyopadhyay
Yongle Zhang
Sarah Vahlkamp
saptab1@umd.edu
yongle@umd.edu
svahl@umd.edu
University of Maryland
College Park, Maryland, USA

ABSTRACT

Our project attempts to verify claims related to the COVID-19 pandemic by compiling evidence from existing scientific literature and justifying classifications. Extending the SciFact project, the project 1) establishes the scientific and political climate that necessitates such work, 2) then set up a baseline for the label prediction and rationale selection models, 3) finetunes the rationale selection model by distillation of the pre-trained RoBERTa model reducing the number of parameters, and 4) explains the builds on the work done previously with SciFact. It presents an argument for the use of models that incorporate both expert reviews and a corpus of knowledge with a wider breadth. We have shown in our project that the existing pre-trained models of SciFact face challenges while verifying COVID-19 claims involving racial bias, social bias, conspiracy theories, and other challenges. We have shown that reducing the number of parameters in the BERT model can be helpful to mitigate the bias to a slight extent.

Links to our code

Baseline and Finetuning code:
https://colab.research.google.com/drive/1mySXIeOQEtZntrBnZp2sVeTmroCfOCon?usp=sharing
Bias code:
https://colab.research.google.com/drive/1ZdJYqLbWVXDkxz0eUN8CTMBG95qKFFAg?usp=sharing
Logistic Regression code:
https://colab.research.google.com/drive/1AqkBJ3mm5xXhOPXtyEGnzSyk4jCsXdro?usp=sharing
Visualization code:
https://colab.research.google.com/drive/1Q2H7ceQ6ktQMqiW5gKRDZkV7CC5UdWMD?usp=sharing
Fine-tuned model zip file:
https://drive.google.com/file/d/1Z3Ak5ECM0icsiHlmNd3$_1Ui6BRHVUn70/view?usp=sharing_{eilts}$ = 5fd74c54
Generated report for social and racial bias claims in COVID-19 folder:
https://drive.google.com/drive/folders/1H4e-zZ58BV7LAQ2Qb6nQ4oLIU1tjV-EP?usp=sharing

CCS CONCEPTS

• **Computing methodologies** → **Classification and regression trees**; *Model verification and validation*; *Information extraction.*

KEYWORDS

datasets, natural language processing, SciFact, BERT, COVID-19, verification, bias, analysis

ACM Reference Format:
Saptarashmi Bandyopadhyay, Yongle Zhang, and Sarah Vahlkamp. 2018. Verification of Claims in the COVID-19 Pandemic. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/1122445.1122456

Woodstock '18, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Despite the relative recency of the COVID-19 epidemic, the ubiquitous nature of a pandemic has given rise to a substantial and growing body of literature from researchers interested in the topic. In particular, this is ushered by a few widely available open-access datasets, the existence of which allows for multi-institutional teams to attack researching various aspects of the problem. [1, 3, 4, 6, 8] The global nature of the challenge, coupled with varied but often robust data collection methodology, has provided a foundation for researchers to contribute their take on COVID-19 from the lens of their respective disciplines. The availability of data has led to near-real-time summary statistic dashboards, broad quantitative analysis, and especially relevant to this project, the application of machine learning techniques. Researchers are building on existing work to validate information veracity, and tailoring their work to address the COVID-19 epidemic more directly. Given the content the articles have previously addressed, (pandemic, bleach, conspiracy, science, etc.,) this work takes an already consequential area of study and addresses an area that should be of grave concern for the public.

Concurrent to the rise of the pandemic is a second challenge, the pervasive infodemic. A term first associated with SARS, it is now being applied to the widening gap between scientific knowledge and rapidly spreading information of various levels of legitimacy. [5] While groups like the Poynter Institute are doing their due diligence to fact-check, draw attention to the false claims, and educate about media literacy, the sheer number of deceptive or erroneous allegations necessitates a more automated look at the information. In attempting to address this, scholars are using several repositories and methods to draw, not conclusions per se, but suggestions as to the veracity of published news articles related to the Coronavirus. In building on recently created tools that verify established information about the pandemic, this work furthers the ability to combat erroneous, and often dangerous, claims through citation of evidence-based literature. This work extends existing work of

classifying a contentious claim by verifying the classification with evidence from the background knowledge base.

System Name	Research Group	References	Sentence-level F1	Abstract-level F1
VerT5erini (2-stage Neural Retrieval)	covidex.ai	paper	58.8	62.7
VerT5erini (BM25 Retrieval)	covidex.ai	paper	55.5	59.2
SciKGAT	THUNLP and MSR	code	50.5	58.3
VeriSci	Semantic Scholar	paper	39.5	46.5
Zero-Shot (trained on FEVER)	Semantic Scholar	paper	26.9	36.4

Table 1: Models and Research Groups as Presented on SciFact page.

2 BACKGROUND

Scientific claim verification forms the base of the SciFact model, which processes claims related to COVID-19 against a repository of research abstracts, leading to a determination that the literature *supports*, *refutes*, or is *unable to verify* a claim. We used this necessary step to evaluate supported claims of racial and economic disparities in both the direction of the pandemic and the related literature.

2.1 Credibility

Determining the credibility of articles presents a significant challenge in the quest to identify purveyors of the infodemic. As Fairbanks, et al conclude in their 2018 study on credibility assessment, "...there is no analogue for fake news words versus real news words." Going back to 2006, we see nascent credibility research pertaining to weblogs relied on the partially-automated combination of machine learning and expert evaluation. [7] This partial automation is very much still a requirement in the various models we found. However, scholars are sharing datasets like Sci-Fact that compile expert claims and annotated abstracts or wikipedia entries. [9] This same study presents a case for "formal(ized) scientific verification", and introduces Sci-Fact as a secondary training set to be used after something like FEVER in order to optimize domain-specific recognition. Similarly, ReCOVery was introduced at this year's CIKM Conference, as a repository of information to use as a baseline for continued study into fake news and Covid-19. This dataset relies heavily on source credibility, extrapolating determinations from this to future articles in a bid to scale the model. [10]

2.2 Controversy in AI Language Models

There has been renewed discussion about ethics in AI language models, given Timnit Gebru's most recent in-progress paper and her controversial exit from Google. Please note that this is being presented without value judgment, but is rather just a note of the controversy surrounding big language models in terms of environmental factors, illusions of meaning, research opportunity costs, and training models that may introduce "racist, sexist, and otherwise abusive language." [2] The paper has not been published, so the information is being pulled from an account by Karen Hao, whose interpretation of the article was written about in her MIT Technology Review article.

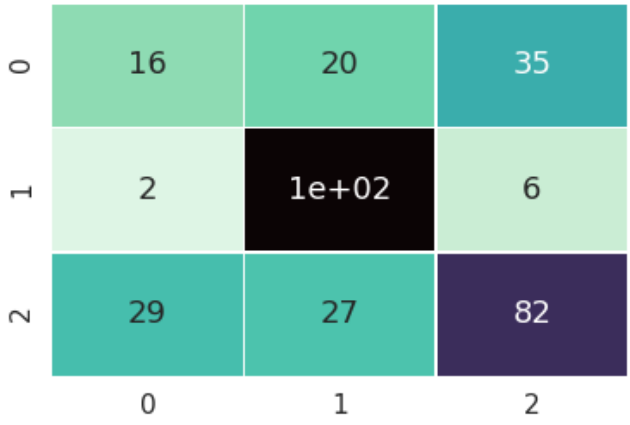

Figure 1: Confusion matrix

As discussed in that article, the paper presents a lot of information on language models, with its most controversial pieces likely tied to the risks. The authors point to four areas: Environmental and financial costs, massive data and inscrutable models, research opportunity costs, and illusions of meaning. This relates most directly at Google to the BERT language model, which has been incorporated into their search engine. This recent occurrence further underscores the necessity to evaluate tools like this against racial, economic, and other demographic variables, in an effort to identify disproportionate or misleading information.

For this paper, we used DistilBERT to run its Masked Language Model on phrases regarding demographics. We looked at the same phrase for women and men, black, asian, white, and hispanic, older and younger, and doctors and nurses with the goal of identifying any bias in the results.

3 MODEL

3.1 Baseline Implementation

Conducting baseline implementation began with cloning SciFact's github: https://github.com/allenai/SciFact, installing dependencies, and downloading the original dataset and best model as claimed by SciFact. They used 2 separate models for label prediction and rationale selection using the pre-trained BERT models. The RoBERTa-large model performed the best according to their paper. But their architecture is computationally intensive with 24 layers and 355 million parameters. We used Google Colab, so we mounted the drive to our repository, edited source code, and debugged. It was challenging given that Google Colab has strict usage limits which were insufficient to train huge pre-trained models. The BERT models were run using the Hugging face library. We created the results directory and then executed the verification script to verify the statement on SciFact's github: "Coronavirus droplets can remain airborne for hours." Then, we generated and stored the report that contained results and their relevant selections, along with a determination of "Support", "Refute", or "Not enough information". These results indicated five documents supported the claim, while one refuted it.

3.2 Logistic Regression Model

We implemented a stochastic gradient ascent for logistic regression as a comparison with the work by Wadden and colleagues[9]. We created a logistic regression model with a softmax activation function and applied a stochastic gradient optimizer. We used each claim's word frequency as the feature and only trained on claim corpus and split the data by the 70% vs. 30% training-testing ratio. The model reached an accuracy rate of 0.5309.

Claim

Coronavirus droplets can remain airborne for hours

Evidence

A Physics Modeling Study of SARS-CoV-2 Transport in Air

Decision: SUPPORT (score=0.99, evidence scores=0.49)

- Health threat from SARS-CoV-2 airborne infection has become a public emergency of international concern.
- During the ongoing coronavirus pandemic, people have been advised by the Centers for Disease Control and Prevention to maintain social distancing of at least 2 m to limit the risk of exposure to the coronavirus.
- We carry out a physics modeling study for SARS-CoV-2 transport in air.
- We show that if aerosols and droplets follow semi-ballistic emission trajectories, then their horizontal range is proportional to the particle's diameter.
- For standard ambient temperature and pressure conditions, the horizontal range of these aerosols remains safely below 2 m.
- We also show that aerosols and droplets can remain suspended for hours in the air, providing a health threat of airborne infection.
- The latter argues in favor of implementing additional precautions to the recommended 2~m social distancing, e.g. wearing a face mask when we are out in public.

Airborne/Droplet Infection Isolation

Decision: SUPPORT (score=0.99, evidence scores=0.21)

- Airborne/droplet infection is caused by infected agents in the air around a person.
- Microbial pathogenic agents that are mainly transmitted airborne are aerosols, re-aerosols, microbe-carrying particles, huge amounts of bacteria-carrying airborne skin cells, dust, droplets and droplet nuclei.

Figure 2: Example output from the initial validation report.

We summarized several reasons for the low-performance logistic regression model. One could be the sparse vectors because of the diverse topic of COVID-related claims. Another reason is that we did not consider enough number of features for training, such as the tf-idf. The results also indicate the direction of exploring evidence-based prediction on top of claims.

3.3 Adapting Logistic Regression

As seen in the table below, results across several text corpora showed higher than average precision, recall, and F1 scores. Given the output, those most concerned with reducing false positives would do to use the RoBERTa Large model with SciFact alone. If we are concerned with reducing false negatives, RoBERTa Large with Snopes had the best score at 0.877. Similar to our precision score, the F1 accuracy score, which takes both false positives and false negatives into account, is highest with the RoBERTa Large model with SciFact alone. Based on the results (as shown in the figure below,) it would seem the RoBERTa Large model with SciFact alone provides the greatest accuracy. It is not surprising that this training dataset would be more reliable than the dataset that includes only the rationale from the same corpus, but in fact, all testing that includes the SciFact corpus performed at higher than average levels. Interestingly, without SciFact, FEVER had good results for false positives, but performed poorly on false negatives and F1 score (0.221, and 0.331, respectively.) In the Snopes set, false positives accounted for a larger part of our error set at 0.267, less than the F1 of 0.409.

Model	Train	Test	P	R	F1
RoBERTa Large	SciFact	dev	0.737	0.705	0.721
RoBERTa Large	SciFact Only Rationale	dev	0.601	0.609	0.605
RoBERTa Large	FEVER SciFact	dev	0.724	0.672	0.697
RoBERTa Large	FEVER	dev	0.659	0.221	0.331
RoBERTa Large	Snopes	dev	0.267	0.877	0.409

Table 2: Model and Train Test Precision, Recall, and F1

3.4 Fine-tuned Model

The best performing models in the SciFact project were the RoBERTa-large model trained on SciFact dataset for the rationale selection model and the RoBERTa-large model trained on the FEVER dataset at first and then on SciFact dataset for the label prediction model. Four our task, we have considered to finetune the rationale selection model. The logistics constraint was that FEVER dataset was huge with 1.4GB dataset which was used in label prediction in the baseline implementation along with SciFact dataset which was of much lesser size of 8.3MB. That's why we fine-tuned the rationale selection model with the DistilRoBERTa-base model using only the SciFact's training and validation datasets. Also, the DistilRoBERTa-base model is more simplistic in computational resources using only 82 million parameters and 6 layers in comparison to the RoBERTa-large model with 355 million parameters and 24 layers. An intuition behind this idea was that if there was any bias in the dataset or the model architecture, we shall try to mitigate the bias by using a model with a simpler computational architecture.

3.5 Differences from Existing Work

A comparative analysis of the rationale selection models using RoBERTa large from the SciFact project and our implemented DistilRoBERTa-base model on the SciFact dataset has been presented in Table 3. It is observed that although the number of parameters is decreased significantly from 355 million parameters to 82 million parameters and the number of layers are reduced significantly from 24 BERT layers to 6 BERT layers, the Precision score is almost un-affected with 0.601 for RoBERTa large and 0.584 for DistilRoBERTa Base. So we understand making the pre-trained model unnecessarily bigger while making it out of bounds for implementation by ordinary students due to its sheer computational size may not have served its purpose. We also show in Section 4.4 how their baseline models can lead to problematic verification of claims in COVID-19 pandemic involving social or racial bias and conspiracy theories which can be mitigated to a slight extent while using the DistilRoBERTa-base model.

Model	Train	Test	P	R	F1
RoBERTa Large	SciFact Only Rationale	dev	0.601	0.609	0.605
DistilRoBERTa Base	SciFact Only Rationale	dev	0.584	0.503	0.5404

Table 3: Comparative analysis of the SciFact and our rationale selection models

4 RESULTS

4.1 Loss Function Plots, Accuracy Plots: Justifying Selection of Best Checkpoint

Epoch	Iteration	Loss	Precision Train	Precision Dev
0	671	0.4009	0	0
1	671	0.3876	0.5872	0.5263
2	671	0.386	0.7122	0.6120
3	671	0.3345	0.7430	0.6327
4	671	0.1355	0.7228	0.6222
5	671	0.6366	0.8685	0.7078

Table 4: Loss and Train and Dev Precision

Through all iterations, we ended up with six epochs, as shown in Table 4.1. Just one epoch would most likely lead to under-fitting, but we had

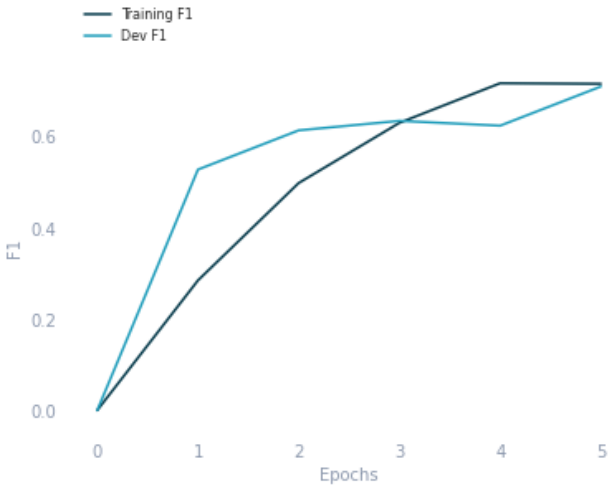

Figure 3: Training and Dev F1

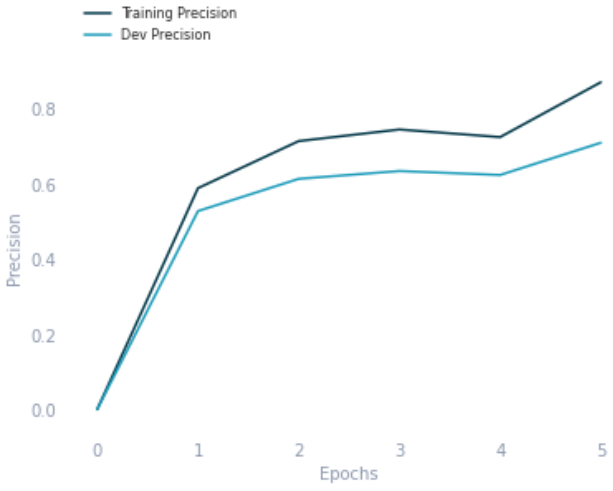

Figure 4: Training and Dev Precision.

several iterations to use. We suggest Epoch 4 as the model checkpoint for verification. Given the output statistics of loss, precision, recall and F-measure on the training and development datasets, we considered epochs 3, 4, and 5, but epoch 3 had more loss and less precision, and epoch 5 had very high loss, but very high precision, too, indicating that the model may have been over-fitting in epoch 5.

4.2 Analysis

Our model uses a format similar to the SciFact corpus. We use an integer as the document id, include the title of the paper, include the abstract, four lines from the paper, and then the label, as seen below.
Corpus: "doc id": An integer as the id,
"title": The title of the paper,
"abstract": Lines from the abstract of the paper,
"structured": false
An example of the claims training data follows:

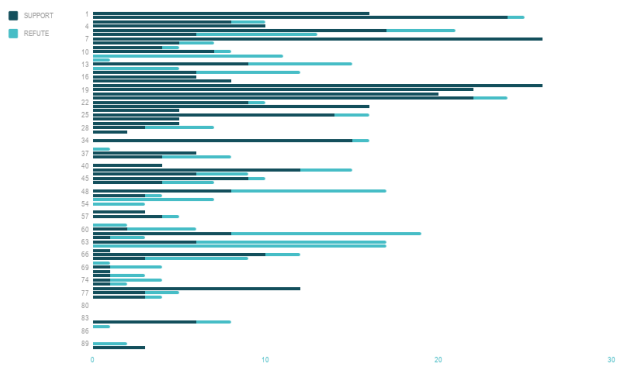

Figure 5: Claim Graph

"id": 2, "claim": "1 in 5 million in UK have abnormal PrP positivity.", "evidence": "13734012": [("sentences": [4], "label": "CONTRADICT")], "cited doc ids": [13734012]
Claims-dev uses a similar format to claims-train.
Claims-test "id": 7, "claim": "10-20% of people with severe mental disorder receive no treatment in low and middle income countries."

We ran 90 different claims through the model, but only 83 of these actually output reports. Of these 83, 66 had something to report, but the rest had no output in the report. There were 622 total decisions made off these claims, for an average of 8 decisions per claim, 7.1 *supports* per claim, and 2.7 *refutations* per claim. (Note that this includes articles that were found in multiple places, so they were repeated. The highest number of *supporting* documents for one claim was 26 (claim 7 and 18 each,) and the highest number of *refuting* documents for one claim was 17 (claim 64.)

Interestingly, if you look at report 38, you will see that the same article/abstract was given different evidence scores, depending on where the article was pulled. The claim: *Ibuprofin does not make COVID-19 worse.* The first iteration of the article, Ibuprofen use and clinical outcomes in COVID-19 patients, supported the claim with a score of 0.99 and evidence scores=1.00. The next article, the same article, supported the claim with a score of 0.97 and evidence scores = 0.28, and 1.00. Even stranger, when looking at the sentences split from each article, they all have the same information.

4.3 Comparative Study of Baseline and Fine-tuned Models

In order to compare the baseline and the fine-tuned model, we look at claim # 67, Covid-19 increases negative mental effects in first-responders.

In the initial report, this claim is refuted by five articles, and supported by three articles. In the new model, the same claim is refuted by seven articles, and supported by four articles. Already, we see that the new model has picked up extra articles, including more articles to support the model. Further, we see that three of the articles included in the baseline model, are no longer in the fine-tuned model. Finally, we see that those articles that are in both, have different score and evidence scores. The baseline model of SciFact shockingly refutes the claim with wrong evidences like "Do Quarantine Experiences and Attitudes Towards COVID-19 Affect the Distribution of Mental Health in China? A Quantile Regression Analysis" with a score of 0.99 which may be biased in itself given the availability of information and censorship. Our DistilRoberta-base model does use the same evidence to refute the claim but the score decreases to 0.97.

Figure 6: Article #1 in Claim #67

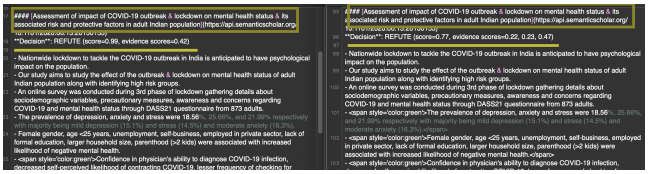

Figure 7: Another look at an article with different scores for the same article in Claim #67

4.4 Analysis of Racial and/or Social Bias Claims

We conducted an analysis investigating whether the proposed model can bring potential racial or social bias to particular groups. We collected 90 claims from social media[1], major fact-checking website[2], Johns Hopkins Coronavirus Resource Center[3], and the SciFact database[4], and verified these claims via the model. We found that the model showed poor performance on a specific type of claims, such as conspiracy theories. For example, given the claim "5G is not responsible for COVID-19 pandemic", the model failed to find any evidence to either support or disapprove the statement. The model failed to generate any result for 12 claims. Besides, we also found surprising results on particular racial groups. For instance, our current model could not verify the claim "US President Donald Trump's claim that COVID-19 is China-virus is correct" and returned null results.

It was also shocking to note that for claims like "The COVID-19 casualty list disproportionately features the Hispanic American community." or similar claims regarding the African American and the Native American communities the model generated evidence refuting these claims while the evidences themselves actually support those claims. For example for the Hispanic American community, the model used an article titled "The Disproportionate Impact of COVID-19 on Racial and Ethnic Minorities in the United States" to refute the claim with a score of 0.74 that Hispanic Americans are disproportionately dying in the COVID-19 pandemic. But the article actually supported the claim! We suggest that researchers may consider their future work about debiasing the model from these perspectives by investigating both the model bias and the dataset bias.

5 DISCUSSION

Our research in the COVID-19 fact verification shows that the use of huge pre-trained models over millions of parameters have to be done carefully given its challenges to verify claims involving racial and social bias. Reducing the number of parameters in the pre-trained model from RoBERTa to DistilRoBERTa can help to a slight extent in such situations if the training is done on biased datasets. There are also challenges in democratization of Deep Learning research with the resources needed to train these models concentrated at the hands of very few established people which does not allow implementation of diverse ideas from students like us. However, there is no going back as demonstrated in our implementation of logistic regression for classification of claims which led to a significantly lower test accuracy. As discussed in the previous section on analysis of racial and/or social bias claims, we found a number of the claims that we were not able to verify via our model had racial and/or social bias implications. Given the empirical importance of these areas in relation to Covid-19, this seems to be an issue in the corpus.

6 FUTURE RESEARCH

The provenance of the background knowledge base being used to verify claims can be under scrutiny. There has to be techniques to evaluate how unbiased the dataset and the knowledge base are to make sure that the model is performing fairly to verify claims regarding the COVID-19 pandemic. There is a possibility to explore the verification of claims in a granular level where a part of the claim is verified and a part of the claim is refuted. But it would involve generating probabilities for the sub-classes along with relevant annotation. Exploring the scalability of the huge pre-trained models is also a significant issue. These techniques can be used for domain adaptation in future.

REFERENCES

[1] Center for Disease Control and Prevention. [n.d.]. COVID-19 Data from the National Center for Health Statistics. https://www.cdc.gov/nchs/covid19/index.htm

[2] Karen Hao. [n.d.]. *We read the paper that forced Timnit Gebru out of Google. Here's what it says.* Retrieved December 15, 2020 from https://www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru/

[3] Chad W. Jennings and Shane Glass. [n.d.]. COVID-19 public dataset program: Making data freely accessible for better public outcomes. https://cloud.google.com/blog/products/data-analytics/free-public-datasets-for-covid19

[4] Kaggle. [n.d.]. COVID-19 Open Research Dataset Challenge (CORD-19). https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge

[5] Merriam-Webster. 2020. Words We're Watching: 'Infodemic'. https://www.merriam-webster.com/words-at-play/words-were-watching-infodemic-meaning. Accessed: 2020-12-16.

[6] National Institutes of Health. [n.d.]. Open-Access Data and Computational Resources to Address COVID-19. https://datascience.nih.gov/covid-19-open-access-resources

[7] Victoria Rubin and Elizabeth Liddy. 2006. Assessing Credibility of Weblogs. 187–190.

[8] Johns Hopkins University and Medicine. [n.d.]. Coronavirus Resource Center. https://coronavirus.jhu.edu/

[9] 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)*. Association for Computational Linguistics, Online, 7534–7550. https://doi.org/10.18653/v1/2020.emnlp-main.609

[10] Xinyi Zhou, Apurva Mulay, Emilio Ferrara, and Reza Zafarani. 2020. ReCOVery: A Multimodal Repository for COVID-19 News Credibility Research. In *Proceedings of the 29th ACM International Conference on Information amp; Knowledge Management* (Virtual Event, Ireland) *(CIKM '20)*. Association for Computing Machinery, New York, NY, USA, 3205–3212. https://doi.org/10.1145/3340531.3412880

[1]https://www.facebook.com
[2]https://www.factcheck.org/a-guide-to-our-coronavirus-coverage/
[3]https://coronavirus.jhu.edu/data
[4]https://SciFact.apps.allenai.org/