Evidence based Automatic Fact-Checking for Climate Change Misinformation

Gengyu Wang Lawrence Chillrud
Kathleen McKeown
Columbia University
gengyu.wang@columbia.edu, lgc2139@columbia.edu, kathy@cs.columbia.edu

Abstract

Misinformation surrounding climate change (CC) proliferates across the internet at such rapid speeds and in such large quantities that human fact-checkers are unable to feasibly verify the veracity of most online CC-related information. While automatic fact-checking algorithms can supplement human fact-checking efforts, existing models suffer from a lack of domain-specific training data to robustly fact-check CC information. To address this problem, we tailor an existing automatic fact-checking system to the CC domain by introducing the popular semi-supervised training method, Unsupervised Data Augmentation (UDA), into our system's pipeline, allowing us to leverage large amounts of unlabeled CC-related claims. We evaluate our fact-checking model on the CC fact-checking dataset CLIMATE-FEVER, yielding a state-of-the-art (SotA) F1 score of 0.7182, improving upon the previously reported SotA F1 score of 0.3285.

Claim:	Global warming is driving polar bears toward extinction.
Claim Label:	`Supports`
Evidence #1	
Evidence Label:	`Supports`
Wikipedia Article:	Global Warming
Evidence Sentence:	
Environmental impacts include the extinction or relocation of many species as their ecosystems change, most immediately the environments of coral reefs, mountains, and the Arctic.	
Evidence #2	
Evidence Label:	`NotEnoughInfo`
Wikipedia Article:	Polar Bear
Evidence Sentence:	
Bear hunting caught in global warming debate.	

Table 1: Example of an annotated claim and two of its five evidence sentences from the CLIMATE-FEVER dataset. The overall claim label is assigned after the five evidence labels "vote" with their own labels.

Introduction

Along with the rise of social media platforms and online media content has come a wave of misinformation regarding climate change (CC). Such misinformation has been shown to manipulate the public's understanding and perception of CC, negatively affecting efforts made to combat the worsening global climate crisis (Anderegg et al. 2010; Ding et al. 2011; Benegal and Scruggs 2018; Van der Linden et al. 2017). In an attempt to help identify and flag incorrect or misleading online CC content, several fact-checking platforms[1] have been created to manually fact-check CC-related claims. However, manual fact-checking (also known as claim verification) is time-consuming and unable to cope with the increasingly vast amounts of misinformation on the internet. Automated fact-checking is a useful tool that can supplement human-driven fact-checking.

The ultimate goal of an automated fact-checking system is to achieve fully automatic real-time fact-checking. However, in the foreseeable future, fact-checking systems may not be able to replace experts to achieve such completely autonomous inspections. Nevertheless, the proposed fact-checking system can still help human experts search for evidence and provide suggestions, thereby significantly improving fact-checking efficiency.

Automated fact-checking systems require a large corpus of labeled training data which would ideally include a large quantity of CC-related claims, annotated with labels indicating if a given claim is true, false, or if the given claim cannot be verified.

To enable our fact-checking system to handle real-world, online CC content, we design a fact-checking pipeline comprised of five modules: 1) *Claim Detection*: detect verifiable claims from a given text; 2) *Evidence Retrieval*: query the claims through Google Search to retrieve relevant evidence, 3) *Evidence Selection*: select related evidence from search results; 4) *Label Prediction*: classify the rationale relation between each claim and evidence pair into one of three categories {`Supports, Refutes, NotEnoughInfo`}; and lastly, 5) classify articles into one of the same three categories when checking articles containing multiple claims.

For the evidence selection and label prediction modules (modules 3 and 4 in the pipeline outlined above), we utilize separate pre-trained RoBERTa (Liu et al. 2019) models and fine-tune each of them on two publicly available fact-checking datasets: FEVER(Thorne et al. 2018) and CLIMATE-FEVER (Diggelmann et al. 2021). FEVER is an influential fact-checking task that contains 145k synthetic claims and relevant evidence from Wikipedia. Unlike

[1]Such as climatefeedback.org, skepticalscience.com

FEVER, CLIMATE-FEVER is a recently published fact-checking dataset that focuses on climate change and contains only 1.5k claims collected from the real world.

Our experiments demonstrate that the in-domain training dataset is critical for the CC fact-checking task in evaluating the evidence retrieval and label prediction modules. To better train the model with the limited size of the in-domain dataset, we amassed our own unlabeled dataset of 4,127 CC-related claims. We then introduce Unsupervised Data Augmentation (UDA) (Xie et al. 2019), a popular semi-supervised training method, to the label prediction module, allowing us to leverage the unlabeled data during training to supplement the CLIMATE-FEVER dataset.

The RoBERTa model that makes use of the unlabeled data via UDA during training recorded an F1 score of 0.7182 on the label prediction task, improving upon the previous state-of-the-art F1 score of 0.3285 reported in Diggelmann et al. (2021).

Our contributions can be summarized as follows: 1) our fact-checking pipeline is unique in providing a system specifically to fact-check climate change texts and achieves state-of-art performance on the CLIMATE-FEVER dataset; 2) we propose a novel integration of UDA with fact-checking models and significantly improve performance when fine-tuning on small datasets; 3) we implement an approach that fact-checks climate change text against the latest documents on the open internet instead of against a pre-defined local corpus such as Wikipedia.

Related Work

There are many existing datasets designed to train fact-checking models. Most notably PolitiFact (Vlachos and Riedel 2014), Emergent (Ferreira and Vlachos 2016), (Wang 2017), SemEval 2017 Task 8 RumorEval (Derczynski et al. 2017), Snopes (Popat et al. 2017), CLEF-2018 Check- That! (Barrón-Cedeno et al. 2018), Verify (Baly et al. 2018), FEVER (Thorne et al. 2018), UKP Snopes (Hanselowski et al. 2019), and CLIMATE-FEVER(Diggelmann et al. 2021). As CLIMATE-FEVER is the only one to focus on climate change fact-checking, we decide to use it along with one of the largest general corpora, FEVER, to train our CC fact-checking models.

Many approaches developed for the FEVER task mainly focus on the claim verification step (Nie, Chen, and Bansal 2019; Luken, Jiang, and de Marneffe 2018; Yoneda et al. 2018; Hanselowski et al. 2018; Yin and Roth 2018; Lewis et al. 2020). GEAR (Zhou et al. 2019) and Zhong et al. (2019) formulates claim verification as a graph reasoning task. Many fact verification systems leverage Natural Language Inference (NLI) techniques (Chen et al. 2016; Peters et al. 2018; Li et al. 2019) to verify the claim. Diggelmann et al. (2021) leveraged a label predictor that was trained on FEVER to evaluate on the CLIMATE-FEVER dataset, yielding a 32.85 F1 score. Our CC-specific fact-checking system is one of the first climate-specific claim verification models to the best of our knowledge. It also is a unique approach to fact-checking in leveraging unlabeled data using the semi-supervised UDA training approach.

Task Formulation and Dataset

We address the climate change fact-checking problem in an extended context of CLIMATE-FEVER, a task that requires a system to verify climate change-related claims on the internet by retrieving evidence from Wikipedia. We begin by introducing in detail the CLIMATE-FEVER task and then elaborating upon our adaptations.

In the CLIMATE-FEVER task, given a CC-related claim, a fact-checking system needs to retrieve related articles from Wikipedia, select a set of sentences to be used as relevant evidence, and finally predict the veracity relationship between the given claim and the selected evidence from a set of three labels, {`Supports, Refutes, NotEnoughInfo`}. Claims in CLIMATE-FEVER were collected from scientifically-informed sources as well as sources skeptical or in denial of CC, to ensure a variety of real-world claims were collected. In total, the dataset contains 1,535 claims and five evidence sentences per claim. Each claim and evidence sentence pair is annotated with a veracity label. The five labels between each evidence sentence and the claim are then used to "vote" on the overall claim's veracity. An example of one claim with two of its five evidence sentences is presented in Table 1.

To enable fact-checking of general internet text, such as long articles or short tweets, we extended the task described above to fact-check isolated, verifiable claims located within the climate-related text (e.g., a news article) instead of fact-checking the whole text. In addition to recognizing that the real-world text may contain information that may not be recorded in Wikipedia, this extended task also requires the fact-checking system to retrieve evidence from sources including, but not limited to, Wikipedia.

Fact-checking System

Our fact-checking pipeline is adapted from the fact-checking systems presented in DeYoung et al. (2019) and Wadden et al. (2020) to fit our extended fact-checking task outlined above. Wadden's system consists of three components. Provided with an input claim to be fact-checked against a corpus as potential evidence, the system includes 1) Evidence Retrieval (using Term Frequency–Inverse Document Frequency (TF*IDF) (Salton and McGill 1986) to retrieve related evidence from a local corpus), 2) Rationale Selection (identifying rationales for each evidence and claim pair), and 3) Label Prediction (making the final prediction).

We make the following enhancements to accommodate the system for our extended task: 1) add a claim detection module to find verifiable claim sentences, 2) replace the TF*IDF-based evidence retrieving module with Google Programmable Search Engine[2]; 3) create an unsupervised dataset for climate change fact-checking and implement the semi-supervised learning method UDA for the label prediction module; 4) add a summary step that evaluates the overall veracity of a CC-related article containing multiple claims.

In the remainder of this section, we begin by introducing the claim detection module, followed by the adapted fact-checking system. In the subsequent section, we discuss the

[2]https://programmablesearchengine.google.com/about/

methodology used in curating the unlabeled dataset, and our implementation of the semi-supervised learning method UDA.

Claim Detection

The claim detection module aims to identify sentences within a body of text that are verifiable and check-worthy. To build the model, we fine-tune RoBERTa on a claim detection dataset ClaimBuster (Arslan et al. 2020), adding a mean pooling layer and a dropout layer before the linear classification head to mitigate any possible overfitting, following the architecture developed by Barrón-Cedeno et al. (2019). ClaimBuster contains 23,533 statements extracted from all U.S. general election presidential debates (1960-2016) which were then annotated by human coders.

Given a multi-sentence text, we feed the $[CLS]$ token of each single sentence's RoBERTa encoding to a linear classifier and consider as claims only those sentences where the confidence score is higher than 0.9. We choose this threshold because we observed that out of 50 samples above 0.9 confidence score, 47 are good well-defined claims. Our fine-tuned model achieved an F1 score on the ClaimBuster test set of 0.9447.

Evidence Retrieval

Instead of retrieving k text pieces from a corpus with the highest TF*IDF similarity, we integrate Google Search into the system to retrieve text evidence from the open internet. By doing so, we enable the system to retrieve evidence from continually updated sources instead of a quickly outdated corpus and verify new claims published on the internet in real-time.

Given a claim as input, the system collects text snippets from the top 10 search results as potential evidence. Snippets are automatically generated, designed to summarize page content that best relates to the given query[3].

To ensure we only collect reliable snippets as candidate evidence sentences, we restrict the search results to Wikipedia, along with 55 other websites that we, the authors, judge as credible. We include our complete list of credible sources in the appendix. Given inevitable disputes over which online organizations can be deemed credible, the modularity of our fact-checking system allows us to easily update and modify our list of credible sources without affecting the other components in the pipeline.

Evidence Selection

Given a claim, c and candidate evidence e (the search snippet), we train a model to predict whether the snippet is an evidence sentence for the claim. We use RoBERTa to encode the concatenated sequence [c SEP e], and then feed the $[CLS]$ token obtained from the encoding to a linear classification layer. To fine-tune the model pre-trained with FEVER, we pair each claim with its labeled evidence in CLIMATE-FEVER dataset as positive examples and use non-rationale sentences in the introduction section of the Wikipedia article as negative examples.

[3]Details are described here: https://developers.google.com /search/docs/advanced/appearance/good-titles-snippets

Label Prediction for Claim

Sentences identified by the evidence selector are passed to a separate RoBERTa model to make the final label prediction. To train the label prediction model, we fine-tune the FEVER-trained RoBERTa model on claim-evidence sentence pairs and their gold labels from the CLIMATE-FEVER dataset. Specifically, given a claim, c and evidence sentence e', We use RoBERTa to encode the concatenated sequence [c SEP e'], and then feeds the output $[CLS]$ token obtained from the encoding to a linear classification layer, and finally, a cross-entropy loss is employed. Each pair is then assigned a label from the set of {`Supports`, `Refutes`, `NotEnoughInfo`}. For claims with multiple evidence sentences, the overall label for the claim is determined by the majority vote, where each evidence sentence "votes" on the final label with its own label.

Label Prediction for Long Text

Given a piece of long text containing multiple claims, such as a long article or twitter thread, a veracity label for the entire piece can be assigned by fact-checking all of the individual claims identified in the text. However, making sense of the multiple resulting (potentially conflicting) evidence labels is not trivial. Instead of using a majority vote system such as the one used at the claim-level scale outlined above, we introduce a log-likelihood ratio (LLR) calculation as an alternative.

The confidence score obtained from RoBERTa for a single claim's label lies in the range of $[0, 1]$. If we simply sum all of the confidence scores of every claim's label within the long text document, we would have to choose a threshold. Matters are complicated by the fact that there could be different numbers of scores to sum depending on the document. Because the LLR metric lies in the range of $(-\infty, \infty)$, summing the LLRs can avoid the problems of setting a threshold which may be biased. LLR is also advantageous as it explicitly accounts for class priors in the training dataset, which may differ from future querying inputs while using the system. Accounting for the class priors makes predicting scores for the long text from different sources more consistent and comparable.

Calculating LLR from the confidence score output of the RoBERTa model is straightforward. We let a positive LLR indicate the evidence refutes its accompanying claim, and a negative LLR indicates the evidence supports its claim. Given a claim and evidence pair (c, e), we denote the the confidence score (i.e the probability) of `Refutes` label y_r as

$$P(y_r \mid (c, e)) \tag{1}$$

We denote N_r as the number of claim-evidence pairs labeled `Refutes`, N_s as the number of `Supports` pairs and N_{nei} as number of `NotEnoughInfo` pairs. The total number of samples in the training dataset is $N = N_r + N_s + N_{nei}$. Then

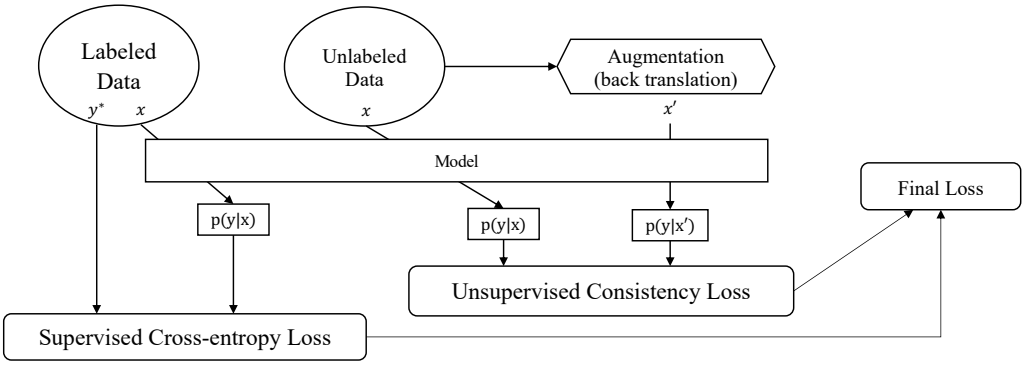

Figure 1: Unsupervised Data Augmentation (UDA) for Climate Change Fact-checking

the prior probabilities of each class can be computed as:

$$\begin{aligned} P(y_s) &= N_s/N \\ P(y_r) &= N_r/N \\ P(y_{nei}) &= N_{nei}/N \end{aligned} \tag{2}$$

Then the LLR for each claim and evidence pair is given by

$$LLR_{(c,e)} = \log_{10} \frac{P(y_r \mid (c,e))}{1 - P(y_r \mid (c,e))} + \log_{10} \frac{P(y_s)}{P(y_r)} \tag{3}$$

An LLR of 0 results in a `NotEnoughInfo` label. Then the final prediction for the text is based on the sign of the LLR sum over all claims:

$$\widetilde{y}_{\texttt{article}} = sign(\sum LLR_{(c,e)}) \tag{4}$$

where a positive sign indicates that the overall content in the long text is classified as refuted by the retrieved evidence, while a negative sign indicates it is supported.

Semi-supervised Training

As the only CC-specific dataset for fact-checking, CLIMATE-FEVER is invaluable for this task. However, given only $1,535$ examples of CLIMATE-FEVER for fine-tuning RoBERTa, we expect to achieve higher performance when training the models with additional data.

We tackle this issue by introducing the semi-supervised training method UDA into RoBERTa's training process, allowing RoBERTa to learn from vast amounts of unlabeled data as well as the labeled data in CLIMATE-FEVER. In this section, we introduce UDA in detail, then discuss the methodology with which we created the unlabeled dataset.

Unsupervised Data Augmentation (UDA)

Typically, unsupervised data augmentation (UDA) involves the use of data augmentation methods, such as back translation, which are applied to each unlabeled data point to obtain a noised paraphrase. For our CC-specific fact-checking pipeline, we also introduce noise into our unlabeled dataset via back-translation. Given an unlabeled CC-related claim c_{en} in English, we first translate the pair to German, yielding c_{de}, and then back into English, giving the noisily paraphrased claim c'_{en}. By doing so, we can generate diverse paraphrases while retaining the original sentences' semantics (Yu et al. 2018). Back-translation was done using the pre-trained fairseq English-German, and German-English translation models (Ott et al. 2019).

During training, labeled data, x, and unlabeled data, u, are loaded simultaneously in different batch-sizes. We refer to the ratio between the batch-size of labeled data and the supervised batch size of unlabeled data as the UDA ratio. In each training step, the cross-entropy loss is computed between the predicted output, y, and the golden label, y^*, for each x, comprising the supervised learning arm of training. Within the same training step, a much larger batch of unlabeled data, u, is fed to the same model along with its back-translated counterparts, u'. The model predicts a label for u, denoted $p(y|u)$, and separately predicts a label for u', denoted $p(y|u')$. Then the unsupervised consistency loss is computed between $p(y|u)$ and $p(y|u')$. The overall objective of UDA is to minimize the sum of the supervised cross-entropy loss and the unsupervised consistency loss. This method is illustrated in Figure 1.

Unsupervised Training Dataset Creation

The unlabeled dataset was collected by identifying check-worthy claims in a topically related set of news articles and then obtaining related evidence for the claims from the open internet. Collecting the unlabeled dataset of claims for UDA began by using a keyword-matching method to find CC-related news articles. The claim detection module described in the fact-checking pipeline above then runs through each news article, identifying verifiable and check-worthy claims. Only those claims that were assigned a confidence threshold $\alpha \geq 0.9$ were then retrieved for the unlabeled dataset. Each claim collected in this way was then put through the Evidence Retrieving module. The resulting snippets were given to our trained evidence selection module to find the subset of sentences within the snippets that would serve as helpful evidence for fact-checking.

This resulted in an unlabeled dataset of 4,127 claims, with 13,513 claim-evidence pairs. On average, each unlabeled claim gathered this way has 3.27 corresponding evidence sentences. As explained in the UDA section above, the unlabeled claim-evidence pairs are then back-translated to produce the

Evidence Selection	Precision	Recall	F1
RoBERTa Large (FEVER)	0.066	0.015	0.024
+ CLIMATE-FEVER	**0.981**	**0.981**	**0.981**

Table 2: Evaluation Result for Evidence Selection

augmented half of the unlabeled dataset.

Experiments

Our experiments evaluate the two primary components of our fact-checking system: the evidence selection module and the label prediction module. We also demonstrate the effectiveness of introducing the UDA-driven semi-unsupervised training regimen for the label prediction module. As the CLIMATE-FEVER dataset does not contain a predefined training-testing split, we created our own. Our held-out evaluation dataset consists of 95 randomly chooses examples from CLIMATE-FEVER, while the remaining data points were used as the training data. [4]

Models

The baseline models are two separate RoBERTa models pre-trained with the FEVER dataset for evidence selection and label prediction module. We then fine-tune the two pre-trained models with CLIMATE-FEVER and evaluate their performance. Finally, we apply the UDA training method in the fine-tuning process for the label prediction module and compare the results with previous methods.

Results

In Table 2, we present the evaluation results for the evidence selection module. The model fine-tuned with CLIMATE-FEVER boosts the F1 score from 0.024 to 0.981. Considering that the results for this module are close to the upper bound, we decided not to apply UDA to the evidence selection module.

We present the evaluation results for the claim-level label prediction module in Table 3. Compared to the RoBERTa Large fine-tuned with FEVER, one additional step of fine-tuning with CLIMATE-FEVER improved the macro F1 from 0.492 to 0.642, which is 0.15 absolute increase. With the unlabeled dataset, the models integrated with UDA in the fine-tuning step boosted macro F1 from 0.642 to 0.718, and the precision from 0.672 to 0.747. We conduct the Statistical Hypothesis Tests for *RoBERTa Large (FEVER)* and the model with *+ CLIMATE-FEVER + UDA ratio 18*, the significance level is 0.0039, which proves the statistical difference between the two models. The baseline macro F1 for label predictor trained on FEVER and evaluated on the CLIMATE-FEVER is reported as 0.3285 from Diggelman et. al., significantly lower than our best performing model. [5]

[4]We followed the CLIMATE-FEVER paper to remove examples with the DISPUTE label to fit the three-way classification task.

[5]We do not include the 0.3285 baselines from Diggelman et. al. in Table 3 as their evaluation was done on the whole CLIMATE-FEVER dataset.

Label Prediction (Claim)	Precision	Macro F1	Macro F1 w/o Not EI
RoBERTa Large (FEVER)	0.579	0.492	0.400
+ CLIMATE-FEVER	0.674	0.642	0.599
+ CLIMATE-FEVER + UDA ratio 8	**0.747**	0.717	0.675
+ CLIMATE-FEVER + UDA ratio 18	0.737	**0.718**	**0.691**

Table 3: Evaluation Result for Label Prediction (claim level). The results for two different UDA ratio settings are presented. "w/o Not EI" stands for without Not Enough Information class.

Conclusion

In this paper, we propose a unique fact-checking pipeline that can verify the veracity of climate change-specific text. We enable this pipeline to fact-check climate change claims against the latest documents on the open internet instead of the static local corpus. The improvement of the model fine-tuned on the only in-domain dataset, CLIMATE-FEVER, demonstrates the value of the in-domain dataset. However, we are able to significantly improve performance even more through the novel introduction of the semi-supervised training method, UDA, into the fine-tuning of fact-checking models with unlabeled in-domain data and obtain 0.3893 absolute increase from the previous state-of-the-art F1 score reported in Diggelmann et al. (2021).

Acknowledgments

This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001120C0123. The views, opinions and/or findings expressed are those of the author and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

References

Anderegg, W. R.; Prall, J. W.; Harold, J.; and Schneider, S. H. 2010. Expert credibility in climate change. *Proceedings of the National Academy of Sciences* 107(27): 12107–12109.

Arslan, F.; Hassan, N.; Li, C.; and Tremayne, M. 2020. A benchmark dataset of check-worthy factual claims. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, 821–829.

Baly, R.; Mohtarami, M.; Glass, J.; Màrquez, L.; Moschitti, A.; and Nakov, P. 2018. Integrating stance detection and fact checking in a unified corpus. *arXiv preprint arXiv:1804.08012* .

Barrón-Cedeno, A.; Da San Martino, G.; Jaradat, I.; and Nakov, P. 2019. Proppy: A system to unmask propaganda in online news. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 9847–9848.

Barrón-Cedeno, A.; Elsayed, T.; Suwaileh, R.; Màrquez, L.; Atanasova, P.; Zaghouani, W.; Kyuchukov, S.; Da San Martino, G.; and Nakov, P. 2018. Overview of the CLEF-2018

CheckThat! Lab on Automatic Identification and Verification of Political Claims. Task 2: Factuality. *CLEF (Working Notes)* 2125.

Benegal, S. D.; and Scruggs, L. A. 2018. Correcting misinformation about climate change: The impact of partisanship in an experimental setting. *Climatic change* 148(1): 61–80.

Chen, Q.; Zhu, X.; Ling, Z.; Wei, S.; Jiang, H.; and Inkpen, D. 2016. Enhanced lstm for natural language inference. *arXiv preprint arXiv:1609.06038* .

Derczynski, L.; Bontcheva, K.; Liakata, M.; Procter, R.; Hoi, G. W. S.; and Zubiaga, A. 2017. SemEval-2017 Task 8: RumourEval: Determining rumour veracity and support for rumours. *arXiv preprint arXiv:1704.05972* .

DeYoung, J.; Jain, S.; Rajani, N. F.; Lehman, E.; Xiong, C.; Socher, R.; and Wallace, B. C. 2019. Eraser: A benchmark to evaluate rationalized nlp models. *arXiv preprint arXiv:1911.03429* .

Diggelmann, T.; Boyd-Graber, J.; Bulian, J.; Ciaramita, M.; and Leippold, M. 2021. CLIMATE-FEVER: A Dataset for Verification of Real-World Climate Claims.

Ding, D.; Maibach, E. W.; Zhao, X.; Roser-Renouf, C.; and Leiserowitz, A. 2011. Support for climate policy and societal action are linked to perceptions about scientific agreement. *Nature Climate Change* 1(9): 462–466.

Ferreira, W.; and Vlachos, A. 2016. Emergent: a novel dataset for stance classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*, 1163–1168.

Hanselowski, A.; Stab, C.; Schulz, C.; Li, Z.; and Gurevych, I. 2019. A richly annotated corpus for different tasks in automated fact-checking. *arXiv preprint arXiv:1911.01214* .

Hanselowski, A.; Zhang, H.; Li, Z.; Sorokin, D.; Schiller, B.; Schulz, C.; and Gurevych, I. 2018. Ukp-athene: Multi-sentence textual entailment for claim verification. *arXiv preprint arXiv:1809.01479* .

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; Riedel, S.; and Kiela, D. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M. F.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 9459–9474. Curran Associates, Inc.

Li, T.; Zhu, X.; Liu, Q.; Chen, Q.; Chen, Z.; and Wei, S. 2019. Several experiments on investigating pretraining and knowledge-enhanced models for natural language inference. *arXiv preprint arXiv:1904.12104* .

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* .

Luken, J.; Jiang, N.; and de Marneffe, M.-C. 2018. QED: A fact verification system for the FEVER shared task. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, 156–160.

Nie, Y.; Chen, H.; and Bansal, M. 2019. Combining fact extraction and verification with neural semantic matching networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 6859–6866.

Ott, M.; Edunov, S.; Baevski, A.; Fan, A.; Gross, S.; Ng, N.; Grangier, D.; and Auli, M. 2019. fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Peters, M. E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365* .

Popat, K.; Mukherjee, S.; Strötgen, J.; and Weikum, G. 2017. Where the truth lies: Explaining the credibility of emerging claims on the web and social media. In *Proceedings of the 26th International Conference on World Wide Web Companion*, 1003–1012.

Salton, G.; and McGill, M. J. 1986. Introduction to modern information retrieval .

Thorne, J.; Vlachos, A.; Christodoulopoulos, C.; and Mittal, A. 2018. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355* .

Van der Linden, S.; Leiserowitz, A.; Rosenthal, S.; and Maibach, E. 2017. Inoculating the public against misinformation about climate change. *Global Challenges* 1(2): 1600008.

Vlachos, A.; and Riedel, S. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 workshop on language technologies and computational social science*, 18–22.

Wadden, D.; Lin, S.; Lo, K.; Wang, L. L.; van Zuylen, M.; Cohan, A.; and Hajishirzi, H. 2020. Fact or Fiction: Verifying Scientific Claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 7534–7550.

Wang, W. Y. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648* .

Xie, Q.; Dai, Z.; Hovy, E.; Luong, M.-T.; and Le, Q. V. 2019. Unsupervised data augmentation for consistency training. *arXiv preprint arXiv:1904.12848* .

Yin, W.; and Roth, D. 2018. Twowingos: A two-wing optimization strategy for evidential claim verification. *arXiv preprint arXiv:1808.03465* .

Yoneda, T.; Mitchell, J.; Welbl, J.; Stenetorp, P.; and Riedel, S. 2018. Ucl machine reading group: Four factor framework for fact finding (hexaf). In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, 97–102.

Yu, A. W.; Dohan, D.; Luong, M.-T.; Zhao, R.; Chen, K.; Norouzi, M.; and Le, Q. V. 2018. Qanet: Combining local convolution with global self-attention for reading comprehension. *arXiv preprint arXiv:1804.09541* .

Zhong, W.; Xu, J.; Tang, D.; Xu, Z.; Duan, N.; Zhou, M.; Wang, J.; and Yin, J. 2019. Reasoning over semantic-level graph for fact checking. *arXiv preprint arXiv:1909.03745* .

Zhou, J.; Han, X.; Yang, C.; Liu, Z.; Wang, L.; Li, C.; and Sun, M. 2019. GEAR: Graph-based evidence aggregating and reasoning for fact verification. *arXiv preprint arXiv:1908.01843* .

Appendix: List of Evidence Sources for Evidence Retrieval

wsj.com, washingtonpost.com, nytimes.com, bbc.com, economist.com, newyorker.com, ap.org, reuters.com, bloomberg.com, foreignaffairs.com, theatlantic.com, ceasefiremagazine.co.uk, canadiandimension.com, aljazeera.com, taipeitimes.com, france24.com, indiatimes.com, straitstimes.com, egypttoday.com, trtworld.com, thelancet.com, sciencemag.org, journals.plos.org/plosmedicine, journals.plos.org/plosbiology, academic.oup.com/nar, nature.com, embopress.org, jamanetwork.com, cell.com, ahajournals.org, ashpublications.org, bmj.com, biorxiv.org, medrxiv.org, ncbi.nlm.nih.gov, cdc.gov, who.int, theguardian.com, un.org, sciencedirect.com, unesco.org, nationalgeographic.org, nationalgeographic.com, abc.net.au, forbes.com, pewresearch.org, pewtrusts.org, emerald.com, unicef.org, bbc.co.uk, climatecentral.org, ipcc.ch, carbonbrief.org, climateinterpreter.org