Fact and scientific claim verification with Natural Language Understanding and reasoning

Alan Song

amsong@andrew.cmu.edu

Meghana Balarama

mbalaram@andrew.cmu.edu

ABSTRACT

The spread of misinformation and disinformation over social media is a problem of growing concern, necessitating the use of automatic fact checking and fact verification about claims on vaccines to climate change. Fact-checked false claims can still spread on social media and thus it is also important to identify previously fact-checked claims. BERT classifier models fail to respond to changes in compositional and lexical meaning and hence the ability of such fact checking methods need to be tested on their ability to reason about real-world facts. In this project, we propose to study and improve the reasoning ability of fact verification models over structured and unstructured information by finding precise evidence to support or dismiss a claim.

INTRODUCTION

The goal of fact verification is to validate a claim in the context of evidence - given a claim C and evidence P as inputs, it predicts if the claim is supported, refuted, or can not be verified by the information in P.

The popular publicly available datasets for this task are FEVER (Fact Extraction and VERification) ([7]), FEVEROUS (Fact Extraction and VERification Over Unstructured and Structured information) ([1]), VitaminC (VitaminC Robust Fact Verification) ([6]), and CLIMATE-FEVER ([2]) each contain ing verified claims. Each claim is annotated with evidence in the form of sentences and/or cells from tables in Wikipedia, as well as a label indicating whether this evidence supports, refutes, or does not provide enough information to reach a verdict. However, large scale training data is not available for every new domain and human annotation of claims is expensive, thus necessitating the automatic generation of training data.

The existing models for fact verification propose pipeline models of abstract retrieval, rationale selection and stance prediction. Such works have the problems of error propagation among the modules in the pipeline and lack of sharing valuable information among modules [11]. Certain other methods use an ensemble of pre-trained models for multi-model fact verification to achieve better performance [10]. To successfully distinguish false articles from genuine ones, a model must be able to incorporate world knowledge into its computation, along with being proficient in natural language understanding. Models that detect false information can be fooled by carefully tweaked input and hence such a detector should base its classification on the semantic content of its input and its relation to real-world facts, and not on superficial features of the text [3]. The models developed for fact verification should be tested for their ability to reason about real-world facts - they should test "understanding": compositional semantics, lexical relations, and sensitivity to modifiers.

Another crucial aspect of fact verification is computing precise evidence - minimal sets of sentences that support or refute a given claim, rather than larger evidences - since larger evidences may contain conflicting pieces some of which support the claim while the other refute, thereby misleading the model. [9] proposes a DQN-based approach to retrieval of precise evidence and demonstrates improvements in achieving accurate claim verification.

LITERATURE REVIEW

Creating a dataset by manually writing claims and linking them to their supportive evidence for every new domain is expensive. To alleviate the need to create human-annotated training data, [5] proposes a method to automatically generate large-scale (evidence, claim) pairs to train the fact verification model. The proposed model Question Answering for Claim Generation (QACG) generates three types of claims from any given evidence - claims that are supported by the evidence, claims that are refuted by the evidence, claims that cannot be verified because there isn't enough evidence available. The claim generation is based on Question Generation which automatically asks questions from textual inputs. A Question Generator generates a question–answer pair (Q;A) for the evidence which is then converted into a claim C (QA-to-Claim) based on the following logical assumptions: a) if P can answer Q and A is the correct answer, then C is a supported claim; b) if P can answer Q but A is an incorrect answer, then C is a refuted claim; c) if P cannot answer Q, then C is a NEI claim. Crowd sourced human annotated datasets also suffer from undesired biases in the data that cause models to learn spurious patterns. [4] proposes a two-stage contrastive data-augmentation pipeline to generate new claims and evidence from existing samples that debiases fact verification models. Crosswise pairing of the generated samples with the original pair to form contrastive samples helps the model to rely less on spurious patterns and learn more robust representations.

A multi-modal fact verification network is proposed in [10], which extracts features from text and images using two pre-trained models and fuses different modalities and same modality but different sources using multiple co-attention networks. This model achieves a competitive performance without using auxiliary tasks or extra information. [11] aims to solve the error propagation problem and the lack of information sharing among modules with their ARSJoint model which jointly learns the three models for abstract retrieval, rationale selection and stance prediction. It is based on Machine Reading Comprehension (MRC) framework and learns additional information using the claim content as a query. It uses BioBERT and RoBERTa as its backbone networks. [8] performs fact-checking using a hierarchical multi-head attentive network that jointly combines multi-head word-level attention and multihead document-level attention to aid explanation in both word-level and evidence level.

Cleverly written fake articles can fool fact-verification systems and evade detection. A model should be able to incorporate world knowledge into its computation in addition to natural language understanding to successfully distinguish fake articles from genuine ones. The model should base its classification on the semantic content of its input and its relation to real-world facts, and not on superficial features of the text. [3] proposes an adversarial benchmark that targets three aspects of a model's "understanding" - whether it has the ability to employ semantic composition, whether it incorporates world knowledge of political parties, and whether adverb intensity is employed as a signal of fake news.

GAPS IDENTIFIED

Large scale training data is not available for every new domain. Much of the current state of the art large datasets are manually generated and annotated for their models to function properly. While trying to develop a different approach to tackle natural language processing, it is much more feasible to try to improve the generation of data to train pre-existing model structures to improve results. Much of the data involved in training models involved with fact verification requires complex human annotation and careful data labelling to successfully generate accurate models. Data in new domains not only requires numerous different claims for the models to be trained on, but these claims need to be fact-checked and classified by experts in the field to ensure accurate training data for the models. This human annotation of claims is expensive and time consuming and would be impossible to expand to new fields of interest in the future without wasting extensive manpower. Thus, the automatic generation of training data is necessary for the future of fact verification. Current model architecture seems to do a satisfactory job of fact verification for fields with ample claim data. Therefore, one large challenge will be to automatically generate new data for old and new fields that can be automatically verified and approved for use in training these verification models.

Additionally, models that detect false information can be fooled by carefully tweaked input. Natural language processing inherently needs to account for every single word in a claim to determine whether the entire claim is true or false, or if there is not enough information to determine this. Thus, even small semantic changes to the content of the claim inputs of the models can lead to varied output, even if the altered statement itself is essentially saying the same thing. Such a detector should base its classification on the semantic content of its input and its relation to real-world facts, and not on superficial features of the text. Currently, various adversarial attacks such as sentence negation and intensity reduction of statements for example, have a noticeable effect on the results of model accuracy. These attacks do not completely fool models but can lead to up to 20% of claim label flipping [3]. These existing models need to be tested for their ability to understand and reason about real-world facts and their robustness to adversarial attacks as well. Hopefully, with additional data to train the model, we can reduce underfitting in some datasets and further reduce the effect of adversarial attacks.

PROPOSED APPROACH

We propose:

(a) Automatically generating training data and data augmentations for domains for which large-scale training data is not available – CLIMATE-FEVER dataset

(b) Multi-task training of a joint model or fine-tuning an ensemble of pre-trained models on this mix of real-world synthetic data for fact verification

(c) Adversarial testing of the trained model to identify the vulnerabilities of the model

Our goal is to provide a framework that could be used for claim generation, evaluation of the effectiveness of the generated claims by fine-tuning transformer-based models on this dataset of generated claims along with the original dataset, and finally evaluating the adversarial robustness of the fact-verification models.

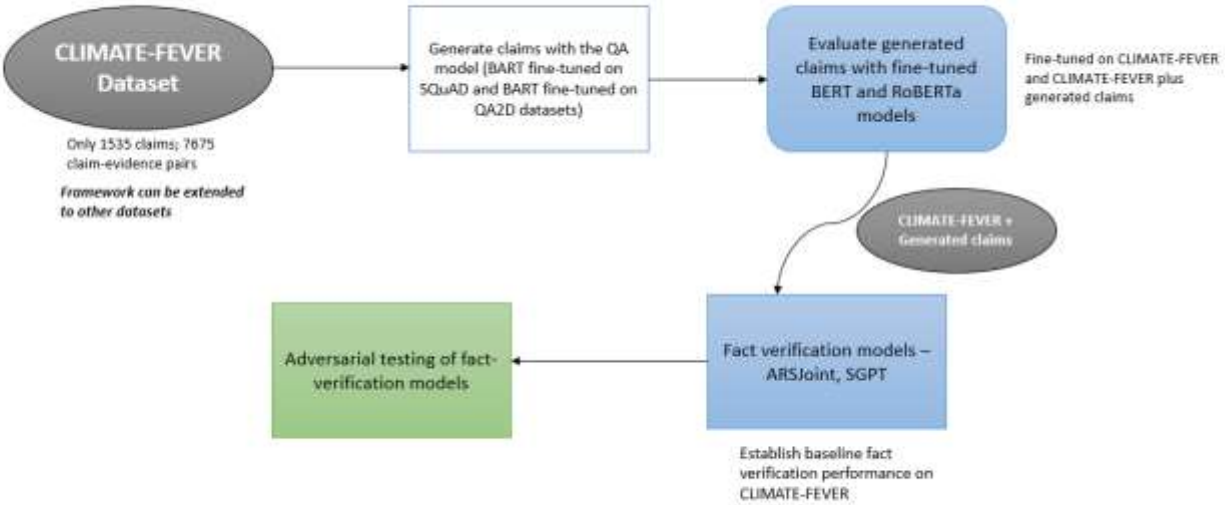

CLIMATE-FEVER Dataset
Only 1535 claims; 7675 claim-evidence pairs
Framework can be extended to other datasets
Generate claims with the QA model (BART fine-tuned on SQuAD and BART fine-tuned on QA2D datasets)
Evaluate generated claims with fine-tuned BERT and RoBERTa models
Fine-tuned on CLIMATE-FEVER and CLIMATE-FEVER plus generated claims
CLIMATE-FEVER + Generated claims
Fact verification models – ARSJoint, SGPT
Establish baseline fact verification performance on CLIMATE-FEVER
Adversarial testing of fact-verification models

Figure 1: Proposed framework

EXPECTED OUTCOMES

We expect that our proposed approach will improve the accuracy of claim verification and help in analysing the vulnerabilities of the model to adversarial attacks that are meant to fool the model. If we can achieve good performance on claim verification by augmenting a small dataset with automatically generated data and data augmentations, this would also lead to a framework that could then possibly be extended to multiple other domains.

EXPERIMENTS AND PRELIMINARY RESULTS

(a) CLIMATE-FEVER is a small dataset with only 1535 real-world claims regarding climate-change and, as compared to the FEVER dataset which has 185,445 manually verified claims.

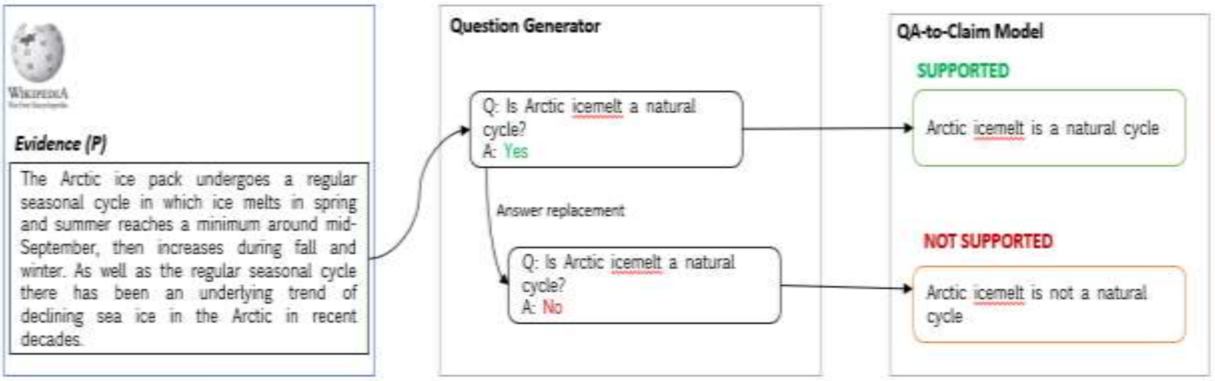
(b) Though there are various baselines for fact verification on FEVER dataset, there are no such baseline specifically for CLIMATE-FEVER

1.CLAIM GENERATION ON CLIMATE-FEVER DATASET

The **Question Generator** generates a question Q with A as the answer by taking as input an evidence P and a text span A. BART model, a large transformer-based sequence-to-sequence model pretrained on 160GB of text and fine-tuned on the SQuAD dataset is used for Question Generation, where the model encodes the concatenation of the SQuAD passage and the answer text and then learns to decode the question.

The **QA-to-Claim Model** takes as inputs Q and A, and outputs the declarative sentence C for the (Q,A) pair. BART model fine-tuned on the QA2D dataset is used here.

For the preliminary experiments, we have generated 200 claims: 100 each of SUPPORTED and REFUTED category.



The Arctic ice pack undergoes a regular seasonal cycle in which ice melts in spring and summer reaches a minimum around mid-September, then increases during fall and winter. As well as the regular seasonal cycle there has been an underlying trend of declining sea ice in the Arctic in recent decades.

Q: Is Arctic icemelt a natural cycle?
A: Yes

Q: Is Arctic icemelt a natural cycle?
A: No

Arctic icemelt is a natural cycle

Arctic icemelt is not a natural cycle

Figure 2: Claim Generation process using Question Generator and QA-to-claim model to generate SUPPORTED and REFUTED claims

Examples of generated claims:

1. Supported claims

```
{
        “id”:”28”
        “claim”:”Water vapor is the most powerful greenhouse gas”
        “label”:”SUPPORTED”
}
```

```
{
        “id”:”34”
        “claim”:”Arctic sea ice has been retreating over the past 30 years”
        “label”:”SUPPORTED”
}
```

2. Refuted claim

```
{
        “id”:63
        “claim”:”Humans are too insignificant to affect global climate”
        “label”:”REFUTED”
}
```

2. EVALUATION OF THE GENERATED CLAIMS

We fine-tune BERT and RoBERTa models first on the original CLIMATE-FEVER dataset and then on the combination of CLIMATE-FEVER and the generated claims for sequence classification. The models are trained to predict only SUPPORTED or REFUTED and do not consider the NEI category

	CLIMATE-FEVER (F1)	CLIMATE-FEVER with Generated Claims (F1)
BERT-base	48.3	49.6
RoBERTa-large	50.2	51.1

Table 1: Fact verification performance of BERT and RoBERTa models fine-tuned on CLIMATE-FEVER dataset and CLIMATE-FEVER dataset with the generated claims

DISCUSSION AND ROADMAP

1. Though the F1 scores are low in absolute terms, we do see an improvement in the F1 score when using the dataset augmented with generated claims, indicating the effectiveness of the claim generation model and the fact that a good and larger dataset could help improve the fact verification model on the CLIMATE-FEVER dataset.

2. We have currently evaluated the effectiveness of the generated claims using fine-tuned BERT and RoBERTa models on only generated claims of SUPPORTED and REFUTED category, excluding the NEI (Not enough information) category.

ROADMAP

1. NEI claim generation
   Claim generation on the FEVER dataset uses additional context that is available in the FEVER dataset and wiki-dumps. However, there is no such additional context available for CLIMATE-FEVER and hence we will be looking into methods to generate NEI claims for CLIMATE-FEVER.

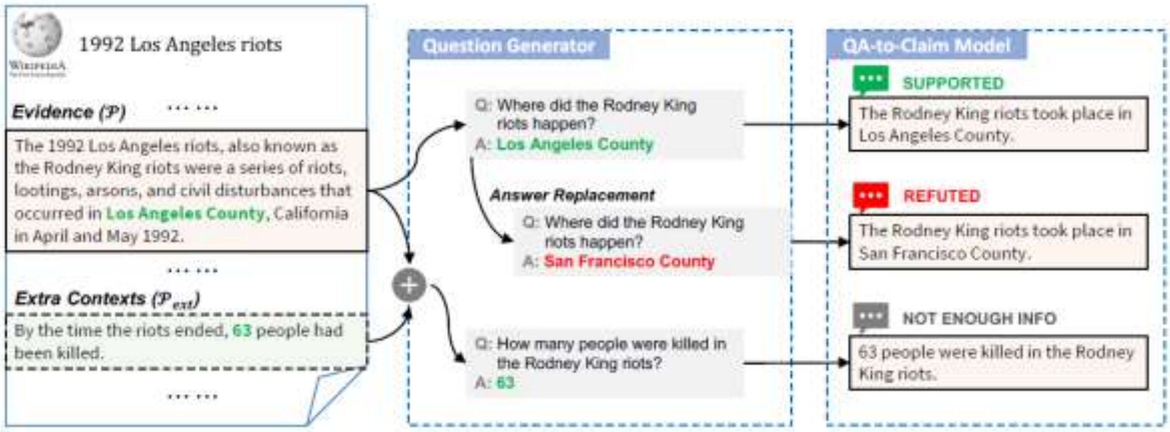

1992 Los Angeles riots
Wikipedia
Evidence ($\mathcal{P}$) ……
The 1992 Los Angeles riots, also known as the Rodney King riots were a series of riots, lootings, arsons, and civil disturbances that occurred in Los Angeles County, California in April and May 1992.
……
Extra Contexts ($\mathcal{P}_{ext}$)
By the time the riots ended, 63 people had been killed.
……
Question Generator
Q: Where did the Rodney King riots happen?
A: Los Angeles County
Answer Replacement
Q: Where did the Rodney King riots happen?
A: San Francisco County
Q: How many people were killed in the Rodney King riots?
A: 63
QA-to-Claim Model
SUPPORTED
The Rodney King riots took place in Los Angeles County.
REFUTED
The Rodney King riots took place in San Francisco County.
NOT ENOUGH INFO
63 people were killed in the Rodney King riots.

Figure 3: NEI claim generation on FEVER dataset using addition contexts [5]

2. Experiment with data augmentation along with generating more claims (In Progress)

   We will also use a data-augmentation technique based on negative claim generation. This generative process involves transforming of a positive claim, such as inserting a negation or replacing a word with antonyms. The generated claim has a different meaning so that it is refuted by the evidence *e* supporting a positive.

3. Fine-tuning and analysis of ARSJoint and SGPT models to establish a baseline for fact verification on CLIMATE-FEVER dataset

   Once we have a complete dataset of generated claims, we will fine-tune the BioBERT and RoBERTa based ARSJoint model and the SGPT model on the original CLIMATE-FEVER and the generated datasets to establish a baseline performance on CLIMATE-FEVER.

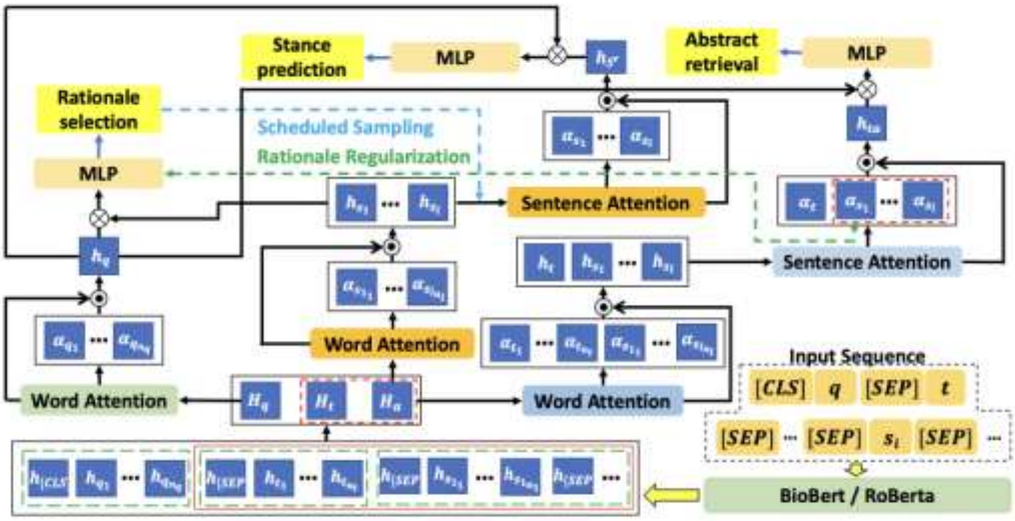


Figure 4: Fact Verification Model 1 - ARSJoint [11]

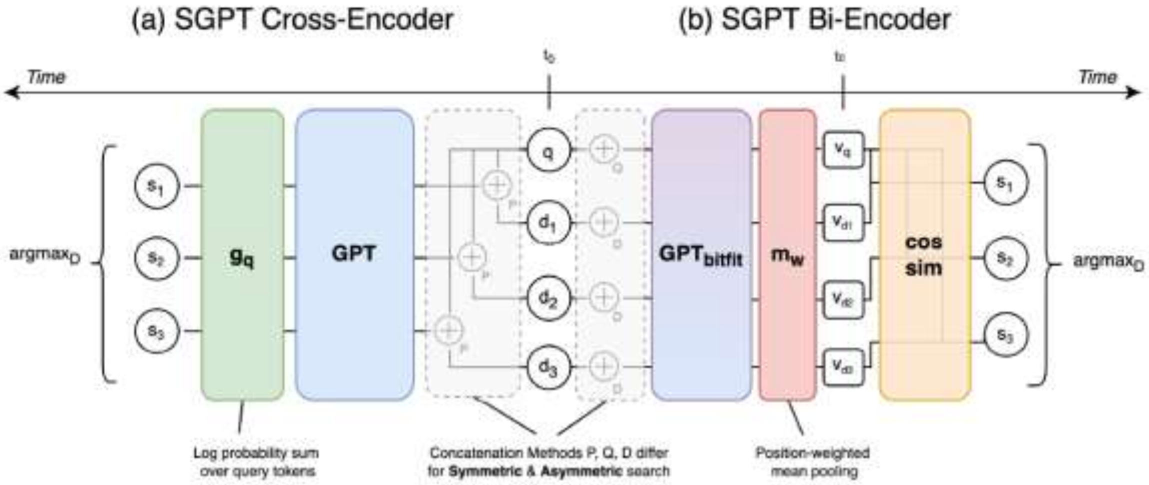

(a) SGPT Cross-Encoder
(b) SGPT Bi-Encoder
Time
t0
t0
Time
argmax_D
s1
s2
s3
g_q
GPT
P
P
P
q
d1
d2
d3
Q
D
D
D
GPT_bitfit
m_w
v_q
v_d1
v_d2
v_d3
cos sim
s1
s2
s3
argmax_D
Log probability sum over query tokens
Concatenation Methods P, Q, D differ for Symmetric & Asymmetric search
Position-weighted mean pooling

Figure 5: Fact Verification Model 2 – SGPT [12]

4. Design precise adversarial attacks, specific to climate change facts and claims to evaluate the vulnerabilities and adversarial robustness of the baseline models (In Progress)

Going forward, we intend to use known adversarial attacks as well as develop potentially new adversarial attacks to test the robustness of our models. We want our models to be resilient to these attacks and retain similar accuracy and performance on both legitimate claim data as well as data that has been altered by adversarial attacks. Currently, a well-known attack is sentence negation, where negative terms are inserted into claims to see if the model can properly reverse the classification of these claims [3]. Generally, a negative term inserted into a claim should reverse the truth value of the claim, and we plan on measuring how robust our system is to this kind of change. Facts in the world are generally a mix of both positive and negative statements, so it is important that our model is capable of properly classifying both types.

Another attack that has been tested before is intensity alteration [3]. By changing or removing words that have strong emphasis such as “totally” or “for sure,” claims may be differently classified by existing models. We hope to make our models more robust to these intensity alterations in the claim statements and instead have our models classify based on truth value of the underlying facts and information in the statement instead of terms that strengthen meaning. Our model should provide the same classification regardless of the intensity of the claims that we feed it.

For our primary domain of climate change, it is impossible to use attacks involving changing speaker names. Currently, there are also political party affiliation attacks that augment politically charged statements by changing the name of the person making the statement within the claim, in order to try to trick the model into giving an incorrect classification. This could be applied to climate change, since it is a politically charged topic, however we want to instead focus on the correctness of the factual information within the statements primarily. Instead, we also intend to investigate substituting synonyms or words that have equal meaning into our claims to see if our model can continue to give accurate classifications.

Another adversarial attack we plan on testing is a chaos attack, where we just inject random words or phrases into the claims to see if the model breaks or loses classification accuracy. We do not expect our model to be completely resistant to these attacks, but we hope that augmentations with random unrelated terms do not change the general classifications of our models too much. We also want to examine how ill-formatted sentences are classified by the model to see if we can still derive meaning and factual verification from statements littered with gibberish.

For any of our adversarial attacks, we plan on testing an unmodified existing model with pre-existing data, our augmented and generated data, pre-existing adversarially modified data, and our augmented and generated data with attacker modifications. We will measure our model accuracy on all of these datasets and compare to see if we managed to improve the robustness of our models.

REFERENCES

[1] Rami Aly et al. “FEVEROUS: Fact Extraction and VERification Over Unstructured and Structured information”. In: *CoRR* abs/2106.05707 (2021). arXiv: 2106.05707. URL: https://arxiv.org/abs/2106.05707.

[2] Thomas Diggelmann et al. “CLIMATE-FEVER: A Dataset for Verification of Real-World Climate Claims”. In: *CoRR* abs/2012.00614 (2020). arXiv: 2012.00614. URL: https://arxiv.org/abs/2012.00614.

[3] Lorenzo Jaime Yu Flores and Yiding Hao. *An Adversarial Benchmark for Fake News Detection Models*. 2022. arXiv: 2201.00912 [cs.CL].

[4] Minwoo Lee et al. “CrossAug: A Contrastive Data Augmentation Method for Debiasing Fact Verification Models”. In: *CoRR* abs/2109.15107 (2021). arXiv: 2109 . 15107. URL: https://arxiv.org/abs/2109.15107.

[5] Liangming Pan et al. “Zero-shot Fact Verification by Claim Generation”. In: *CoRR* abs/2105.14682 (2021). arXiv: 2105.14682. URL: https://arxiv.org/abs/2105.14682. [6] Tal Schuster, Adam Fisch, and Regina Barzilay. “Get Your Vitamin C! Robust Fact Verification with Contrastive Evidence”. In: *CoRR* abs/2103.08541 (2021). arXiv: 2103.08541. URL: https://arxiv.org/abs/2103.08541.

[7] James Thorne et al. “FEVER: a large-scale dataset for Fact Extraction and VERification”. In: *CoRR* abs/1803.05355 (2018). arXiv: 1803.05355. URL: http://arxiv.org/abs/1803. 05355.

[8] Nguyen Vo and Kyumin Lee. *Hierarchical Multi-head Attentive Network for Evidence-aware Fake News Detection*. 2021. arXiv: 2102.02680 [cs.AI].

[9] Hai Wan et al. “A DQN-based Approach to Finding Precise Evidences for Fact Verification”. In: Jan. 2021, pp. 1030–1039. DOI: 10.18653/v1/2021.acl-long.83.

[10] Wei-Yao Wang and Wen-Chih Peng. *Team Yao at Factify 2022: Utilizing Pre-trained Models and Co-attention Networks for Multi-Modal Fact Verification*. 2022. arXiv: 2201.11664 [cs.CV].

[11] Zhiwei Zhang et al. “Abstract, Rationale, Stance: A Joint Model for Scientific Claim Verifica tion”. In: *CoRR* abs/2110.15116 (2021). arXiv: 2110.15116. URL: https://arxiv.org/abs/2110.15116.

[12] Niklas Muennighoff. SGPT: GPT Sentence Embeddings for Semantic Search. 2022. DOI: 10.48550/ARXIV.2202.08904. URL: https://arxiv.org/abs/2202.08904.