Fact and scientific claim verification with Natural Language Understanding and reasoning

Meghana Balarama
mbalaram@andrew.cmu.edu

Alan Song
amsong@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.

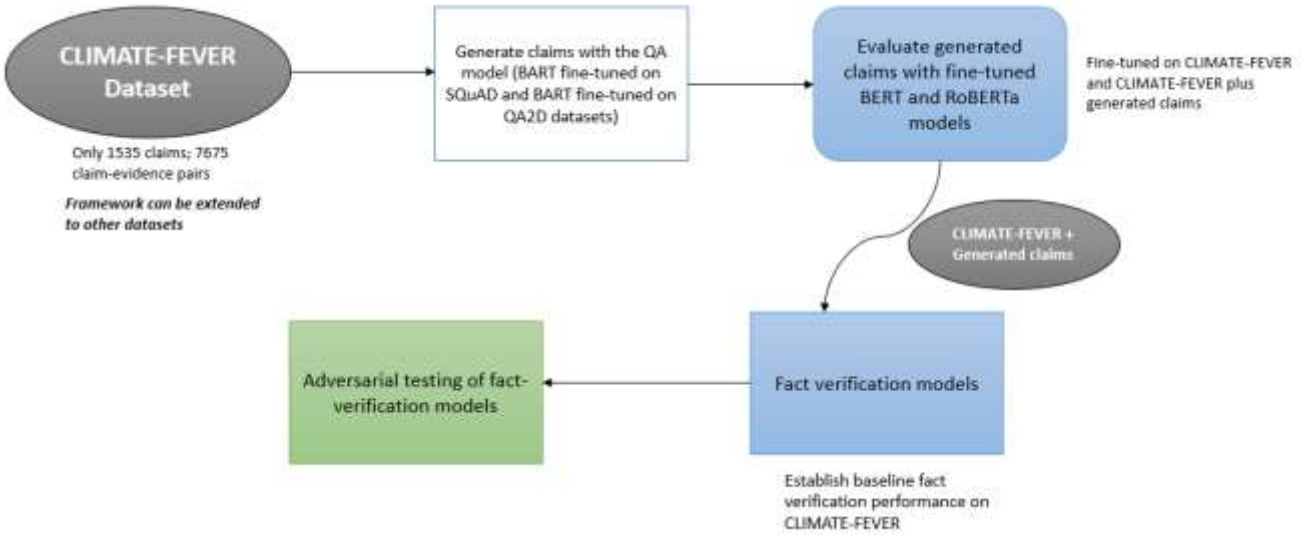

Figure 1: Proposed framework

CLAIM GENERATION ON THE CLIMATE-FEVER DATSET

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. The model used is trained to convert question answering datasets into natural language inference datasets.

To generate **SUPPORTED** claims, we use Named Entity Recognition to identify all entities within evidence P. For each entity a, we use the Question Generator to generate question q = G(P, a). The QA-to-claim model takes in the (q,a) pair to generate SUPPORTED claim M(q,a).

To generate **REFUTED** claim, answer a is replaced with another entity a' with the same type using answer replacement after generating (q,a) pair such that a' becomes an incorrect answer to question q. Entity a is used as query to retrieve top 5 most similar phrases in Sense2Vec as the replacing answer a'. The QA-to-claim model then takes in (q,a') to generate REFUTED claims.

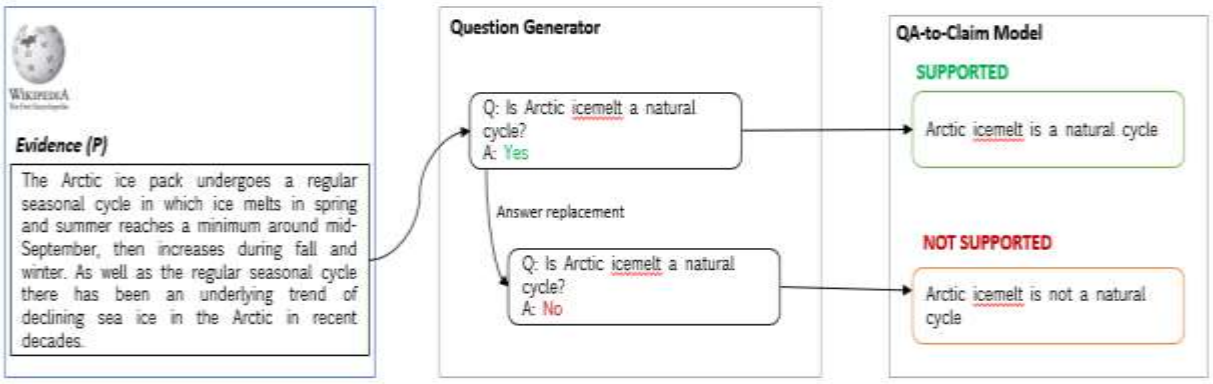

Figure 2: Claim Generation process using Question Generator and QA-to-claim model to generate SUPPORTED and REFUTED claims

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 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

We have generated 12200 claims: 6100 each of SUPPORTED and REFUTED category.

Examples of generated claims:

1. Supported claims

{

"id":10681

"claim":"Multiple lines of evidence indicate Greenland's ice loss is accelerating and will contribute sea level rise in the order of metres over the next few centuries"

"context":"The glaciers in Greenland are also contributing to a rise in the global sea level faster than was previously believed"

"label":"SUPPORTED"

}

{

"id":3648

"claim":"Hurricane Harvey gave Houston and the surrounding region a $125 billion lesson about the costs of misjudging the potential for floods"

"context":"Preliminary reporting from the National Oceanic and Atmos[heric administration set a more concrete total at $125 billion, making Harvey the 2nd costliest tropical cyclone on record, behind Hurrican Katrina with 2017 costs of $161 billion (after adjusting for inflation)"

"label":"SUPPORTED"

}

2. Refuted claim

{

"id":8330

"claim":"Scientists tried to 'hide the decline' in global temperature"

"context":"The email was widely misquoted as a trick to 'hide the decline' as though it referred to a decline in measured global temperatures, but this was obviously untrue as when the email was written temperatures were far from declining: 1988 had been the warmest year recorded"

"label":"REFUTED"

}

{

"id":7915

"claim":"Scientists project that the Arctic will be ice free in the summer of 2013"

"context":"Research shows that the Arctic may become ice-free in the summer for the first time in human history by 2040"

"label":"REFUTED"

}

2. DATA AUGMENTATION WITH NEGATIVE CLAIM GENERATION

Our original hypothesis was to use a data-augmentation technique based on negative claim generation. This generative process involves transforming 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.

However, while generating claims with question answering, we observed that negated claims were automatically being created by answer replacement and the claims generated by data augmentation were very similar to the QA claims. Hence, we decided to not use the data augmentation technique.

3. EVALUATION OF THE GENERATED CLAIMS

We fine-tuned 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 generated claims only belong to the SUPPORTED and REFUTED categories but the CLIMATE-FEVER dataset has SUPPORTED, REFUTED, and NEI claims. We concatenate CLIMATE-FEVER, generated supported claims and generated refuted claims and shuffle it to obtain our final dataset that we use for training. Before training and benchmarking fact verification models on the generated claims, it is important to evaluate the effectiveness of these claims in helping the model learn. The generated claims may overlap with the already existing claims and if the overlap with certain kind of claims is large, this could lead to an unbalanced and biased dataset. It is also possible that the generated claims aren't helping the model learn anything new – as detailed in the NEI claim generation section.

	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 200 generated claims (only SUPPORTED+ REFUTED without NEI claims)

	CLIMATE-FEVER (F1)	CLIMATE-FEVER with Generated Claims (F1)
BERT-base	54.1	58.6
RoBERTa-large	57.4	63.3

Table 2: Fact verification performance of BERT and RoBERTa models fine-tuned on CLIMATE-FEVER dataset and CLIMATE-FEVER dataset with 12200 generated claims (SUPPORTED, REFUTED and NEI claims)

We 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.

4. BASELINE PERFORMANCE ON CLIMATE-FEVER

We fine-tuned 5 different models on our dataset consisting of CLIMATE-FEVER and generated claims to establish a baseline fact-verification performance for CLIMATE-FEVER. Each model was fine-tuned for 10 epochs on the cumulative dataset consisting of CLIMATE-FEVER and 12200 generated SUPPORTED and REFUTED claims.

MODEL	**F1 Score**
BERT-base	58.6
RoBERTa-large	63.3
ALBERT	59.2
DistilBERT	58.4
ClimateBERT-base	**75.2**
ClimateBERT-fact-checking	**77.1**

Table 3: Fact verification performance of models fine-tuned CLIMATE-FEVER dataset with 12200 generated claims (SUPPORTED, REFUTED and NEI claims)

The ClimateBERT Language Model is a DistilRoBERTa based model additionally pretrained on a text corpus comprising climate-related research paper abstracts, corporate and general news and reports from companies. It has 82M parameters and is trained on 1.6M paragraphs. The training approach for CLIMATEBERT comprises all three phases—using a language model pretrained on a general domain, the domain-adaptive pretraining on the climate domain, and the training phase on climate related downstream tasks.

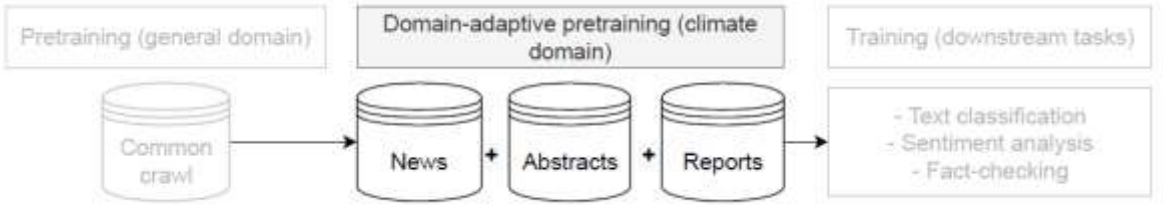

Figure 3: Sequence of training phases of ClimateBERT. The model is pre-trained on news articles, reports, and abstracts and evaluated on downstream tasks of text classification, sentiment analysis, and fact-checking [13]

ClimateBERT-base (no fine-tuning on CLIMATE-FEVER) has an F1 score of 75.2 and on fine-tuning it with our dataset, we achieved an F1 score of 77.1. Our dataset helps improve the fact-verification performance on CLIMATE-FEVER.

5. ADVERSARIAL ROBUSTNESS

We used known adversarial attacks as well as 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 the 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 these datasets and compare to see if we managed to improve the robustness of our models.

ATTACKS DESIGNED

Adversaial attacks were designed and tested only on correctly classified SUPPORTED or REFUTED claims. Testing on NEI claims resulted in essentially only NEI classifications regardless of the modifications of the claims. This makes intuitive sense since if a claim is unable to be classified, slightly modifying the semantics of the sentence would not be expected to suddenly have it be classified differently. The claims were carefully modified to preserve grammatical structure, and then passed into our model along with unchanged evidence to observe if the classification changed or stayed the same.

Below, we highlight examples of the adversarial attacks we generated and used in testing the robustness of our model.

EVIDENCE: Because water vapor is a greenhouse gas, this results in further warming and so is a \"positive feedback\" that amplifies the original warming.

CLAIM: The CO2 **amplifies** the warming and mixes through the atmosphere, spreading warming throughout the planet.

NEGATION: The CO2 **does not amplify** the warming and mix through the atmosphere, spreading warming throughout the planet.

INTENSITY: The CO2 **may amplify** the warming and mix through the atmosphere, spreading warming throughout the planet.

SYNONYM: The CO2 **increases** the warming and mixes through the atmosphere, spreading warming throughout the planet.

ANTONYM: The CO2 **decreases** the warming and mixes through the atmosphere, spreading warming throughout the planet.

REWORDING: The CO2 spreads warming through the planet by amplifying the warming and mixing through the atmosphere.

CHAOS: The CO2 **runs** the warming and mixes through the atmosphere, spreading warming throughout the planet.

Our attacks mainly target either the main verb in a sentence, as shown above, or primary adjectives that give the sentence their meaning. The negation attack is simply the insertion of the word “not” or “no” into a sentence to flip the meaning. The intensity attack is one that changes the certainty or intensity of a sentence by adding/removing/changing words such as may, will, definitely, etc. This attack was originally done on fake news articles, so there were a prevalence of words that added to the “clickbait” effect of those articles. However, in our context, since most of the claims are scientific in nature, most of the intensity attacks focus on changing certainty, such as adding the word “may” in the example given above. The synonym, antonym, and chaos attacks focus on substituting the main verb or adjective of a sentence with something else, while preserving the rest of the sentence. For the synonyms and antonyms, common verbs were manually selected to replace the main verb of the sentence while either preserving or flipping the meaning of the sentence. The chaos attack replaces the main verb or adjective with a common word that makes no sense in context. The rewording attack rearranges the existing words in a sentence while trying to preserve the meaning.

ATTACK RESULTS

	Correct	**Incorrect**	**NEI**
Negation	77% (95)	7% (9)	16% (20)
Intensity	97% (120)	2% (3)	1% (1)
Synonym	72% (90)	11% (14)	17% (21)
Antonym	60% (74)	22% (28)	18% (22)
Rewording	98% (122)	1% (1)	1% (1)
Chaos	17% (21)	11% (14)	72% (89)

Table 4: Results of adversarial attacks on 124 correctly classified SUPPORTED and REFUTED claims from the verification dataset

In the columns of the table, correct means that the modified claims were classified as they were expected to be. For example, for negation attacks and antonym attacks, the claims that were classified as correct were those that the model flipped the label for. For intensity, synonym, rewording, and chaos attacks, correct meant that the model maintained the same label for the modified claim. NEI means that the model classified the new modified claim as NEI. The rows are the different types of adversarial attacks we performed on our claims.

The model performs poorly on negation and intensity attack and is resilient to synonym attacks but somewhat struggles against antonym attacks. This is also expected since changing the main verb to something entirely different, especially ones that may not be present in the evidence used to classify the claim will definitely cause incorrect or NEI classifications. Our rewording results however showed very strong model robustness. The chaos attack is interesting as even while completely changing claim meaning by substituting bogus for the main verb, the model was still able to make some correct and incorrect classifications, rather than always defaulting to NEI.

What did not work

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.

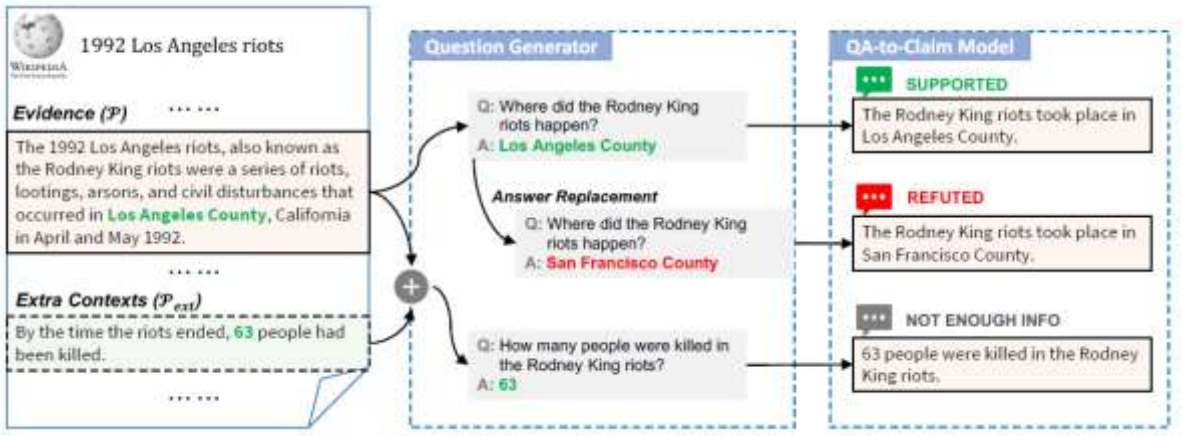

Figure 4: NEI claim generation on FEVER dataset using addition contexts [5]

Method: For NEI claim generation on CLIMATE-FEVER we used all the additional evidence as context when generating claim from one particular evidence. We then tried to use only a subset of evidence from similar claims as context.

Observation: We found that this led to a lot of claims being generated which were arbitrarily irrelevant to the evidence it was generated on, simply because named entities were picked and replaced from the context evidences. Though such (claim, evidence) pairs are technically NEI claims, they do not help the model learn how to figure out if an evidence is complete enough to make a decision about the claim – in this case, since the claims and evidences were totally unrelated, this distinction is very obvious.

Hence, we decided not to use the generated NEI claims during the training of the model.

FACT CHECKER DEPLOYMENT

In addition to providing a framework that can be used to pre-process and generate claims for any dataset for fact verification, test the effectiveness of these claims by fine-tuning BERT based sequence classification models, and perform adversarial testing on fact verification models, we have also built a MVP demo application, hosted on Streamlit, that will allow users to fact check climate-change related claims. Currently, the app needs both claim and evidence as input from users, but at deployment, the goal is to have users only input the claims and the application will look through all the climate-change related articles that were used to train ClimateBERT and display the authenticity of the claim along with the source that supports or refutes the claim.

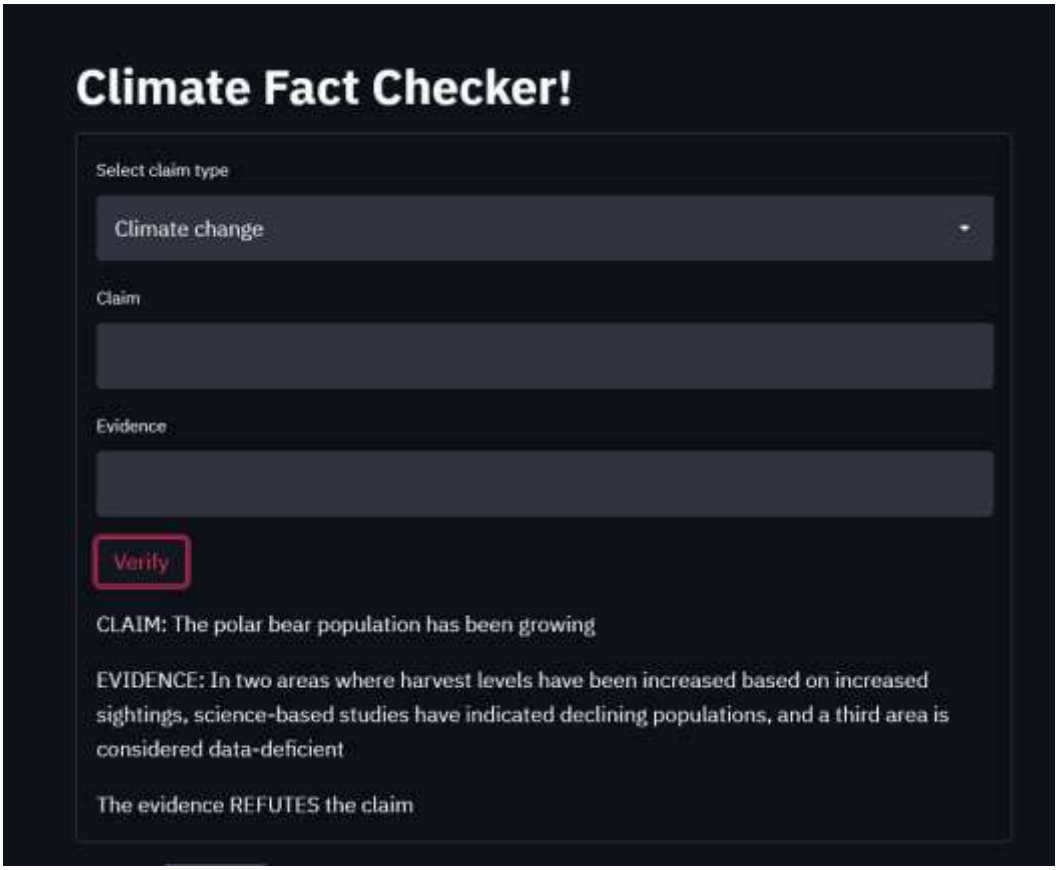

Figure 5: MVP demo application, hosted on Streamlit

CONCLUSION

In this project we have built a framework that can be used to pre-process and generate claims for any dataset for fact verification, test the effectiveness of these claims by fine-tuning BERT based sequence classification models, and perform adversarial testing on fact verification models. We created a dataset with SUPPORTED and REFUTED claims and established a baseline performance of fact verification on CLIMATE-FEVER augmented with our generated claims by fine-tuning five different sequence classification transformer models and found that fine-tuned ClimateBERT gives the best F1 score. We tested the fine-tuned ClimateBERT model for adversarial robustness using different adversarial attacks and found that, while our model performed slightly worse than existing benchmarks for certain adversarial attacks, the model was still overall robust to these attacks. Additionally, we found that the new attacks that we devised were effective in evaluating model robustness and could be extended to other fact verification work in the field. We also demonstrated a MVP application that can be deployed to help users verify climate-change (and potentially other claims) in real time.

FUTURE WORK

From our experiments it was clear that claim generation is bottle-necked by the lack of available evidences. There are only 7500 evidences and hence, there are only so many claims that can be generated by extracting named entities from the evidences. One possible way to overcome this is to use authentic and verified climate-change related articles, news, research papers, and reports from companies and generating precise evidences from them. Our experiment for NEI claim generation indicates that we need a method for NEI claim generation that does not rely on additional context. Additionally, we believe that adversarial testing should be incorporated as a framework for model robustness testing in any fact verification context. The development of automatic test case generation may be an important and valuable problem to look into for the future of fact verification.

GitHub Project Repo

https://github.com/meghana17/18662-project

Our generated claims and other data files can be found at

https://drive.google.com/drive/folders/1rc5hlOqOg5yXD5-NwChLDjcHtdeNuYUb

The MVP demo application is hosted on Streamlit here
https://share.streamlit.io/meghana17/18662-project/main/app.py

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 Verification”. 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.

[13] Nicolas Webersinke et al. “ClimateBert: A Pretrained Language Model for Climate-Related Text”. In: CoRR abs/2110.12010 (2021). arXiv: 2110.12010. URL: https://arxiv.org/abs/2110.12010.