Combating Climate Fake News Using NLP

Rayyan Mohamed

student

rayyanym17@gmail.com

ABSTRACT

As fake news becomes more prevalent across the US, important issues become harder to solve. One such issue is climate change, where climate misinformation has worsened viewer's abilities to distinguish between fake information and real information. This project's objective is to tackle climate misinformation using an artificial intelligence model. The model utilizes a BERT model tested on the "climate_fever" dataset to classify whether climate-related claims are true based on pieces of evidence.

1. INTRODUCTION

Over the past decade, misinformation has been involved in some of America's most pressing issues. Social media in particular has aggravated the situation. According to PBS, MIT researchers concluded that fake news on sites such as Twitter is "70 percent more likely to be retweeted than information that faithfully reports actual events" (Greenemeier 2018). This misinformation has extremely negative implications as it causes viewers to have false beliefs of the world around them. It also increases polarization, as viewers only seek information that affirms their beliefs, in effect creating an unwillingness to listen to opposing viewpoints. Thus, misinformation has magnified polarization and has exacerbated the nation's ability to resolve issues at hand.

This project combats climate misinformation through a machine learning model trained to classify the validity of real-world claims. In practice, users would be able to input a statement and the model would verify whether the statement is true or not. Classifying climate statements induces awareness, helping viewers properly understand information from an unbiased perspective.

2. LITERATURE REVIEW

There are various models already in existence that fact-check climate statements. These models train using different sources, such as online articles, news stories, and climate claims. But the main difference is how they are put into practice. Most models scan and fact-check whole articles while others are limited to fact-checking certain sites. But there isn't a more user oriented site that allows one to fact-check individual claims themself.

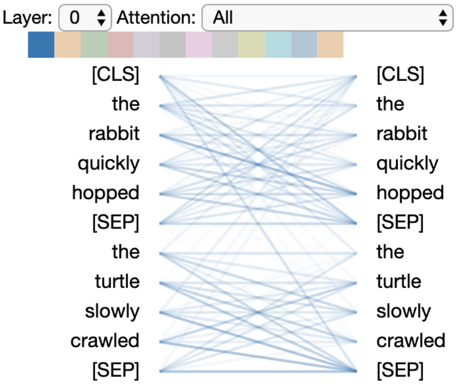
Another limitation is the problem of a limited dataset. One reason why fake information is a worsening problem is because fake information is becoming harder to distinguish from real information. Since AI models are trained using similarities found in previous articles, they have no understanding of what the "truth" actually is, thus making it harder to adapt to changes never seen before.

3. METHODS

3.1 BERT

In order to build the classifier, I used the model DistilBERT Base Uncased, a smaller, optimized version of BERT. BERT, which stands for Bidirectional Encoder Representations from Transformers, was developed by Google to understand human language, otherwise known as Natural Language Processing. BERT consists of two training methods, Next Sentence Prediction and Masked Language Modeling.

Mask Language Model uses the context of words in a sentence to predict a hidden word. This method allows the computer to learn the meaning of words and what they mean in the context of the sentence. Through understanding the meaning of words, the computer is also able to recognize how similar or different certain words are depending on their ability to replace each other in a sentence (for example, the word "fun"is replaceable by "enjoyable" in a sentence, whereas the word "boring" isn't). Given an input of one or more sentences, Next Sentence Prediction can then predict the next logical sentence. These two methods combined train the computer to understand relationships between words, sentences, and ideas.



Layer: 0 Attention: All

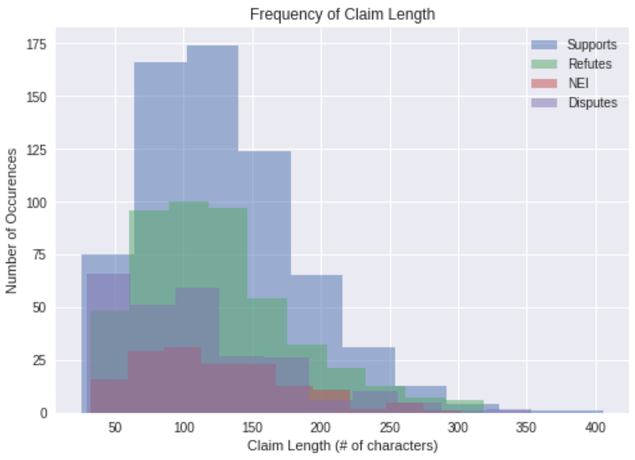
The image above displays how BERT processes language through finding relationships between words. In this example, BERT associates the word "rabbit" with "hopped". In contrast, "quickly" and "slowly" display no connection.

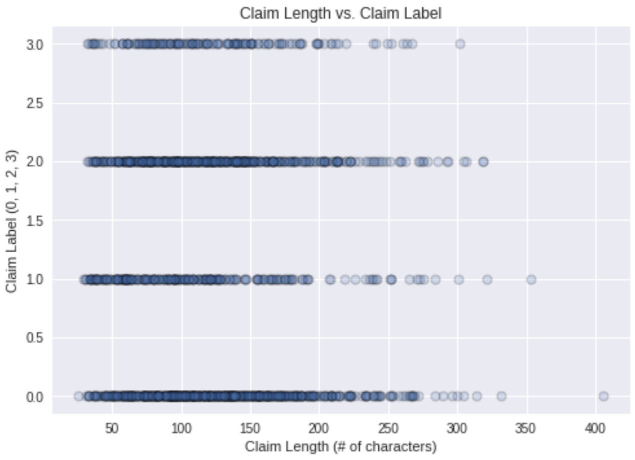
Rather than building an NLP model from scratch, this project utilized the language model DistilBERT Base Uncased and tweaked it to better classify climate related claims.

3.2 Dataset

The dataset I used for the project was the "climate_fever" dataset. It consists of 1,535 real-world climate related claims. Each claim is also accompanied by five evidence pieces for a total of 7,675 claim-evidence pairs. In order to classify whether evidence supports the assigned claim, there are four levels of classification, called the “claim_label”: 0 (supports), 1 (refutes), 2 (not enough information), and 3 (disputed).

Graphing the article lengths to their respective claim label was interesting. The claim label is the overall label assigned to a claim based on whether it is supported by the majority of its 5 evidence pieces. When plotting the graphs, as shown below, we notice that some claim labels occur more often than others. The "supports" label occurs most frequently, followed by "not enough information", "refutes", and "disputed". Specifically, for all 1535 claims, there are 654 "supports", 253 "refutes", 474 "not enough information", and 154 "disputed". Besides this, the claim length of all claim labels are relatively similar.





3.3 Model Development

Since this project was to classify climate-related claims, the model received claims as input and classified them to their respective claim label. For the best results, the model trained for 5 epochs with a batch size of 16. This means the model processed 16 samples each cycle for a total of 80 samples processed. Choosing the optimal number of epochs is crucial to avoid having an overfit or underfit model. Underfitting occurs when the model underperforms on the train and test dataset, whereas overfitting occurs when the model exhibits good performance on the training data, but not the testing data. Running for too many epochs leads to an overfit model which causes the model to make irrelevant connections within the data.

4. RESULTS

The model’s accuracy is 51.14% on the test set. This is relatively okay considering claims were classified among four categories. If the model were only guessing, the accuracy would be around 25%. The precision, a metric that compares the amount of true positives versus false positives, was 0.4573. This means under half of all positive samples predicted were predicted correctly. The recall score reports whether the algorithm correctly classified all claims, or labeled some wrong. In this case, the recall score was 0.5114 as well. To see whether precision or recall is a more important metric, one should look at whether false negatives are more costly than false positives or vice versa. In this case, since it would be more harmful to classify a fake news article as not being fake, recall is the more important metric to focus on. Finally, the F1 score, a combination of both precision and recall, was 0.4808.

This version of the model predicts the claim label given only the claim itself. The most probable reason why the model received low metrics is because not enough input was given in order to make a prediction. One suggestion to improve the model would be to utilize both the given claim and its 5 evidence pieces as input to predict the claim label.

5. MODEL IMPROVEMENT

5.1 Model Redevelopment

To improve the model, the model needed to take more information as input. Rather than inputting just the claim to classify the claim label, both the claim and its assigned evidence pieces were combined to form one cohesive input to classify the claim label. In order to utilize only the claims, evidence, and claim labels, the unnecessary elements in the dataset were deleted before the rest were combined. Although each claim in the dataset is accompanied by 5 evidence pieces, one input only consisted of one evidence and one claim combined. This meant that to use all the claims and evidences, claims had to be reused 5 times each, which in turn maximized the size of both the training and testing dataset by 5 times.

The model trained using very similar methods to the last version. The main difference, aside from the varied input, was the increased epoch rate. Instead of training for 5 epochs, the model trained for 10. This epoch size was found to yield better results, most likely because both the training and testing dataset increased by 5 times, meaning that the epoch rate would increase as well.

5.2 New Results

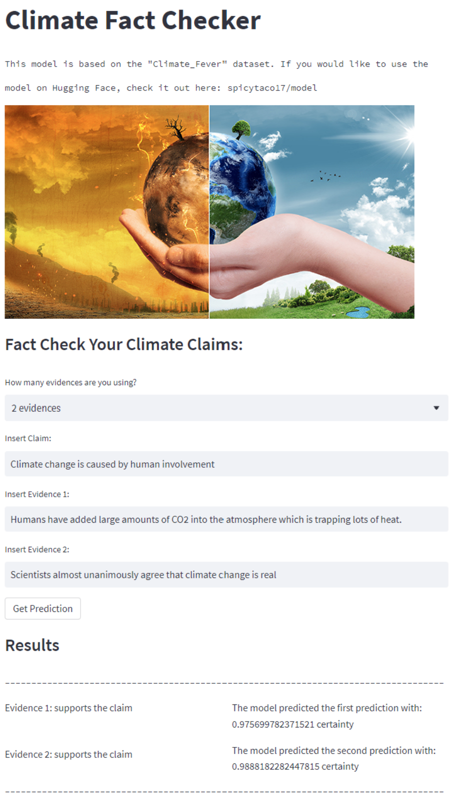
The model performance significantly increased after utilizing the evidence. The accuracy improved by 20% and became 71.27%. The precision was 0.7157 and the recall was 0.7127. Finally, the F1 score increased to 0.7141. This version of the model is available on Hugging Face with the tag: spicytaco17/model.

Furthermore, the model was developed as a web app using Streamlit, a Python app framework. Users are able to enter a claim with the choice of up to 5 supporting evidences. The model also displays the claim label and an explanation of the given label. This can be accessed through Streamlit using this link:

https://spicytaco17-climate-fake-news-app-57cfkv.streamlitapp.com/

It is also accessible as a Colab notebook through this link:

https://colab.research.google.com/drive/1a1gEA5eKsvce3WdvmSDBi5n5kQHqLMeP?usp=sharing



Climate Fact Checker

This model is based on the "Climate_Fever" dataset. If you would like to use the model on Hugging Face, check it out here: spicytaco17/model

Fact Check Your Climate Claims:

How many evidences are you using?

2 evidences

Insert Claim:

Climate change is caused by human involvement

Insert Evidence 1:

Humans have added large amounts of CO2 into the atmosphere which is trapping lots of heat.

Insert Evidence 2:

Scientists almost unanimously agree that climate change is real

Get Prediction

Results

Evidence 1: supports the claim — The model predicted the first prediction with: 0.975699782371521 certainty

Evidence 2: supports the claim — The model predicted the second prediction with: 0.9888182282447815 certainty

An image of the web app is shown above. Users are able to select how many pieces of evidence they want to use (from 1 to 5). The results are shown with the assigned label and model certainty.

6. LIMITATIONS OF THE STUDY

The main limitation to the model is the results of its metrics. It is difficult to improve these metrics because input is classified across 4 claim label categories. Another limitation is the data the model is trained from. As stated previously, false information is becoming harder to distinguish from true information. Rather than training just on the "Climate_Fever" dataset, it would be possible to collect statements from other sources such as news articles and social media. Separately, the model could possibly be further improved by introducing the evidence label category or votes. These display whether the individual evidence supports the claim.

7. CONCLUSION

Overall, this model utilizes a large dataset of 7,675 claim-evidence pairs to provide valuable insight on distinguishing climate-related statements. In testing, the model improved over time and received adequate results. If desired, the model could be further improved by using a wider variety of data and by expanding the use of such input labels. Another way would be to reach a recall of 1. This is preferable because it means all fake news is getting flagged, even though some statements may be genuinely true. Reporting false statements as true statements is harmful because it validates false information, which is why recall needs to be higher.

This model would be best suited for flagging content rather than evaluating it for a definitive answer. If improved to the highest possible recall, all false information could be flagged and later evaluated on whether it is true or not by a more accurate system. Using this operation answers the original question at hand: whether an NLP model could evaluate misleading climate-related information effectively. In the case that all false information is flagged, viewers aren't susceptible to untrue climate information, meaning that an NLP model can evaluate climate statements to some degree with slight assistance.

8. REFERENCES

[1] Diggelmann, Thomas. "Datasets: climate_fever." *Hugging Face*, 07 Aug. 2022. https://huggingface.co/datasets/climate_fever

[2] Patel, Kasha. "New artificial intelligence tool detects most common climate falsehoods." *Washington Post*, 07 Aug. 2022. https://www.washingtonpost.com/weather/2021/12/09/climate-change-study-misinformation/

[3] Diggelmann Thomas, Boyd-Graber Jordan, Bulian Jannis, Ciaramita Massimiliano, and Leippold Markus. "climate-fever: A Dataset for Verification of Real-World Climate Claims." *arXiv*, 07 Aug. 2022. https://arxiv.org/pdf/2012.00614v2.pdf

[4] Thanh, Dang. "Object Detection With Transformer: DETR." *Viblo Asia*, 07 Aug. 2022.https://viblo.asia/p/paper-explain-object-detection-with-transfromer-detr-eW65GpmjKDO