COMP90042 Project 2023: Automated Fact Checking For Climate Science Claims

1084678

Abstract

This report outlines the construction of an evidence retrieval and claim verification system using Roberta and FAISS as well as justifications for decision made and the resulting findings.

1 Introduction

For the Project of COMP90042, I have been given the task of creating a system for validating claims related to climate change. More specifically, the system when presented with a claim related to climate change must retrieve and present up to five related evidence passages from an evidence corpus, which it will then use to categorise the claim as either "REFUTED", "SUPPORTED", "DISPUTED" or "NOT ENOUGH INFO".

The final system is implemented in a two stage process which first uses a Roberta encoding based semantic search to retrieve related evidence passages from the evidence corpus and then uses another Roberta model to determine the label of each piece of evidence in relation to the original claim. The final label assigned to the claim is then computed by aggregating across all claim and evidence pair labels.

The evidence corpus was supplied as part of the project materials as well as labelled training and development sets. No other training data was used. Code references are also included in their respective Jupyter Notebook files.

2 Method

Development and implementation of the system was split into two separate tasks of evidence retrieval and claim verification which each have their own dedicated subsystem. Relevant evidence passages are first retrieved from the evidence corpus using the evidence retrieval subsystem, which then passes the retrieved evidences onto the claim verification system in order to determine the final label. The final results are then outputted to a json file.

The development set was used to gauge model performances during development and to find best parameters, the distilled version of Roberta was also used during development to save time and computational resources. For the final model, the development set was incorporated into the training data for all models and the base version of Roberta was used.

2.1 Evidence Retrieval

The evidence retrieval system consists of a sentence transformer and an encoded evidence dataset with a semantic search index. When a claim is given, the claim is encoded using the sentence transformer and used to search the dataset's index for semantically similar evidences, retrieving the 5 highest scoring pieces of evidence.

2.1.1 Sentence Transformer

The sentence transformer is created using Roberta base and is fine tuned on the given training dataset. Roberta was chosen over similar models such as Bert, due to its better performance on NLP benchmarks such as GLUE.

To preserve the original semantic meaning and take full advantage of Roberta's self-attention mechanism during the fine-tuning process, the training data was only extracted, repackaged into appropriately shaped datasets or lists and fed to the model's appropriate tokenizer. No other data preprocessing steps were undertaken.

The model is fine tuned for two epochs using triplet loss on training examples formed from the provided training evidence which has been shaped into triples of claims, related evidence (positive) and unrelated evidence (negative). Two epochs were used as any more resulted in reduced performance, likely due to overfitting.

At first, the unrelated evidence passages were chosen randomly from the evidence corpus in order to build the initial evidence retriever system, however the final system leverages the initial system(s) to retrieve unrelated evidences with a high similarity score to each claim, thus creating 'harder' training examples. This led to a modest increase in evidence retrieval performance.

2.1.2 Semantic Search

The fine tuned sentence transformer is then used to encode the entire evidence passage dataset, which can then be used by FAISS to build a similarity index. FAISS was chosen over other similar libraries such as ElasticSeach due to it's ease of use and GPU support. Finally, similarity search can then be performed on the evidence dataset with the encoded claim in order to retrieve related evidence.

2.2 Claim validation

The claim validation consists of a multiclass classifier which uses a fine tuned Roberta model in order to classify claim and evidence sentence pairs according to whether they support, refute or neither support nor refute a claim. This classification is performed for all claim and evidence pairs and the final resultant label is decided by aggregating all labels according to a formula, it is only at this step that the 'DISPUTED' label is potentially created at this step.

2.2.1 Classifier

The classifier uses a Roberta base model which is fine tuned on training examples of claim and evidence pairs with labels corresponding to the three classes. Training examples were not formed using any training data with the 'DISPUTED' label, as this would introduce noise into the training process, due to the uncertainty of whether the evidence supports or refutes a claim.

The model was trained for one epoch with a weighted decay of 0.01 using cross entropy loss, as this seemed to maximise classifier performance. A custom weighting based on reciprocally normalized class size to the total number of samples was applied to the loss function to account for class label imbalances, this moderately improved classification accuracy.

2.2.2 Aggregation

The aggregation algorithm is based on additional clarification given by the subject. A claim is categorised as 'SUPPORTED' if there is at least one supported label and no refuted labels, conversely, it is 'REFUTED' if there is at least one refuted and no supported labels. If there are both supported and refuted, the claim is 'DISPUTED'. Finally, if there is none of either supported or refuted, the label will be classified as 'NOT ENOUGH INFORMATION'.

3 Results

In this section I will discuss the performance of the system throughout its development and how it was affected by certain modifications, as well as the final system.

	Base system	Weighted loss	Hard examples	Combined system
Evidence Retrieval F-score	0.07359	0.07359	0.07760	0.07760
Claim Classification Accuracy	0.51948	0.55195	0.48701	0.52597
Harmonic Mean of F and A	0.12891	0.12986	0.13387	0.13525

Table 1: System performance with certain features

As seen above, use of a weighted loss function and hard examples raises the performance of their respective subsystems and intended metrics, however the improved evidence retrieval system appears to negatively impact claim verification system performance. This may be the case due to the retrieval of more relevant evidence which are harder to classify accurately, as they cannot simply be labeled as 'NOT ENOUGH INFO' by the classifier.

Fortunately this loss in performance is partially mitigated by improvements from the use of a weighted loss function, which attempts to encourage the classification system to rely more on semantic relations of the data and less on probabilistic class size, hence resulting in improved classification even against harder evidence.

3.1 Final System

The final system has mediocre performance in comparison to most classifiers on Codalab, achieving a Harmonic Mean of F and A score of 0.16840 based on ongoing evaluation. This score is hampered mostly by the subpar evidence retrieval system, scoring only 0.1007 for evidence retrieval F-score, the label classification system, however, seems to perform very well, with an accuracy of 0.51320, being competitive with most other systems on Codalab.

The difference in performance compared to development systems is likely due to the use of the full training data and the base Roberta model, there may also be other unaccounted for variations in model training and variations in development and test data.