FeedbackFruits Peer Review

Review export of

Project Peer Review

Reviews written by

Chi Zhang

Downloaded at 22 May 2023, 14:02

Review for Anonymous user

Accessibility consideration

This file does not fully conform with all applicable guidelines for accessible digital documents

OPEN ALTERNATIVE

Summary

Comment criterion

1 comments

CZ Chi Zhang
21-05-2023 11:05:47 Compliment

The report introduces the Roberta and FAISS methods to construct an evidence retrieval model, followed by a Roberta for claims classification. Sentence-transformer is trained by inserting the 'triple' of claim, positive & negative evidence, and gradually introducing hardly distinguishable negative evidence for modest tuning. Once the sentence transformer is tuned, it was applied to encode the evidence and apply FAISS to compute the semantical similarities (retrieve top-5). Then, another fine-tuned Roberta with cross-entropy-loss based on claim-evidence pairs ("DISPUTED" pairs are excluded to reduce bias) is trained for classification. Once the model is trained, the aggregation algorithm is applied for classification by capturing the different types of evidence combinations. During development, the development set was used for evaluating the performance by adding additional operations to the evidence retrieval system and guide for the model's final construction.

Strengths

Comment criterion

1 comments

CZ Chi Zhang
21-05-2023 17:04:44 Compliment

The report has followed the standard structure of introduction, methodology, result, analysis and conclusion, especially by stating the applied methods / models at the beginning, which points out the main idea of the rest of the report content to the reader. For the Usage of the table, on Page 2, the result table has delivered a good visualisation for the reader to figure out the difference between the performance of the regular and the further improved models. Accompanied with the text elaboration below the table, the trade-off between the performance in retrieval and classification was revealed and indicates an overall improvement by adding 'Hard Example' to train the model.

The experiments were overall sound by testing the regular model first and improving it along with checking performance based on the development set. The workload has contained the standard fine-tuning BERT model and FAISS method for implementing the algorithm, both solid work. This project has two novelty points: the 'Hard Example' and 'Classification' rules, especially the 'Hard Example', which further tuned the BERT model and improved its capacity to distinguish the negative evidence for pairing.

Overall, the finding is convincing, with an around/above-average performance result and has been presented concisely.

Weaknesses

Comment criterion

1 comments

CZ Chi Zhang
21-05-2023 18:01:05 Suggestion

It would be even better to split the 'Results' part into 'Results' and 'Evaluation' parts (Page 2, Right Section), allowing a more detailed explanation of each implemented model and presenting the potential reason for the improvement or degradation. The table format can be adjusted to ensure it won't exceed the right margin boundary. Besides, if there are some further tuning of hyperparameters and you want to list them in the report, you could plot a line chart for visualisation.

The experiment is sound already; however, if some references further support models listed in the 'Method' section (Page 1, Right Section ~ Page 2, Left Section) would make your design more convincing. Further explanation of the Aggregation (Page 2, Bottom Left Section) could illustrate the reason for applying this rule for the reader's better understanding. The workload is sufficient; you might test more pre-trained BERT models (or listed in your report if you have already done that) for fine-tuning in the future study to allow you to illustrate the optimal algorithm that is closer to the global optimal.

FeedbackFruits Peer Review

Review export of

Project Peer Review

Reviews written by

Chi Zhang

Downloaded at 22 May 2023, 14:02

Review for Anonymous user

Accessibility consideration

This file does not fully conform with all applicable guidelines for accessible digital documents

OPEN ALTERNATIVE

Summary

Comment criterion

1 comments

CZ Chi Zhang
21-05-2023 20:57:33 Compliment

This research applied the fine-tuned RoBERTa to compare the similarity between claims and evidence for evidence retrieval and then similarly applied the fine-tuned RoBERTa to process claim classification.

Before applying the fine-tuned Roberta, both BM25 and word2vec were tested, while unsuitability was found based on data exploration. Then, Roberta was applied and fine-tuned with training data and self-generated negative claim-evidence pairs. Once Roberta was tuned, the top-5 pieces of evidence were retrieved based on the similarity between the test claim and the evidence from the corpus. After the evidence is retrieved, claim classification is conducted based on the evidence's labels obtained by Roberta (after comparing it to the other fine-tuned BERT models). Besides implementing the algorithm, this research highlighted the reason for lowering the RoBERTa's performance (overfitting), which links to the limited training data (the model recognises the same pattern too often during training, and little data cannot help further tune the complicated model).

Strengths

Comment criterion

1 comments

CZ Chi Zhang
21-05-2023 22:49:16 Compliment

This report follows a standard structure with a well-split introduction, EDA, preliminary experiment, methodology (combined with result and evaluation), critical analysis, and conclusion. All tables and plots in this report are informative by revealing the performance during the model's development and some essential data information.

The whole research is sound since it combines related research (references) and the author's implementation during algorithm designing. Besides, the process of judging the potential method for evidence retrieval is sound by recalling the data's characteristics to help discard unsuitable models (e.g. BM25). The workload of this research is enough as it considers multiple embedding methods (Evidence Retrieval Task based on similarity) and investigates various pre-trained BERT models for further fine-tuning (Classification Task). Furthermore, a deeper analysis was conducted to present the potential reason for the model's overfitting issue, showing a sufficient investigation of the problem. Last but not least, this research has its novelty, especially in classification. The author only uses the majority voting on the labels assigned to the retrieved evidence to distinguish the claim's fact type rather than taking the claims into account while still achieving high accuracy in classification.

Overall, the finding is convincing that the algorithm performed relatively well in the evidence retrieval task's F-score and had a high accuracy in the claim labelling task. The report has precisely presented how this algorithm was developed and performed.

Weaknesses

Comment criterion

1 comments

CZ Chi Zhang
21-05-2023 23:22:26 Suggestion

There are further improvements in providing tables and plots for better visualisation purposes though it is already good enough. For the exploratory analysis, including some tables describing the claims and evidence data can provide a clearer vision for the reader to understand the large gaps between the amount of accessible training data and the full-text corpus (quite helpful in revealing the potential reason for the model's overfitting). Besides, it would be nicer if the plots' resolution was clearer (e.g. Page 2&3 Plots).

The evaluation process based on the development set for helping to tune the model is well presented. However, the final evaluation result (exact value) based on the full-test-set released in Codalab was not specified. Presenting the final result directly in numbers would give the reader a more straightforward understanding of the difference between the model's local and actual performance. This final result section could be presented before the critical analysis, where the explanation could illustrate right afterwards.

FeedbackFruits Peer Review

Review export of

Project Peer Review

Reviews written by

Chi Zhang

Downloaded at 22 May 2023, 14:02

Review for Anonymous user

Accessibility consideration

This file does not fully conform with all applicable guidelines for accessible digital documents

OPEN ALTERNATIVE

Summary

Comment criterion

1 comments

CZ Chi Zhang
22-05-2023 10:01:53 Compliment

This report introduced models based on the TF-IDF, BM25, BERT or SimCSE algorithm(s) for processing the evidence retrieval and then classified claims based on the validated evidence(s). Besides, the solution to the sample imbalance issue was considered to achieve a more accurate fact-checking system.

Text embedding was processed before computing the similarity between claims and evidence. The author tested four main types of embedding methods, which were directly applied for retrieving evidence or building the combined retrieval model, and picked the optimal local model based on F-Score (TFIDF with BERT trained after SimCSE). Once the retrieval model was finalised, claim(s) and the matched evidence(s) were preprocessed (concatenation and embedding). Then, the classification model was built with BERT and additional logic (e.g. adding a Bi-LSTM). The author picked the BERT with BiLSTM & BiGRU, with the local highest accuracy in the test set. Besides the model, limitations of this research were listed (e.g. too-large evidence corpus issue and less generalisation issue), accompanied by potential solutions.

Strengths

Comment criterion

1 comments

CZ Chi Zhang
22-05-2023 13:17:14 Compliment

This report has followed a standard structure by introducing the overall research, data analysis, and theoretical explanation of methodologies. Then, the model detail related to the task is delivered, accompanied by the result illustration. The illustrations of plots and tables gave the reader a clear view, allowing a more straightforward understanding.

From the soundness perspective, this report has followed the correct logic, especially when investigating the potential method for the classification task, which built the model based on solid reference (e.g. Page 4 Section 4.4). Besides, the whole algorithm development process has included the evaluation based on the development/test set, which provided a guideline for making the decision. On the other hand, the author has also investigated enough by considering multiple embedding methods and different combinations of models to achieve a better result. The combined model also shows the novelty of this research by introducing the SimCSE to help improve the BERT's performance after using TFIDF to screen evidence initially.

Based on the sufficient and novel work presented above, the result of this research is convincing by obtaining an overall good result in both F-score and Accuracy in the development set and relatively good in the test set. Moreover, all results were illustrated clearly with tables.

Weaknesses

Comment criterion

1 comments

CZ Chi Zhang
22-05-2023 14:01:30 Suggestion

This report could be further improved though it has delivered the message concisely. It would be nicer to make the plot's resolution higher (Page 3 Section 4.1) to allow the reader to get the information directly from visualisation and use the text to assist understanding.

For the retrieve evidence model construction, it would be even better if tuning the TF-IDF's number of evidence retrieval / screening (Page 2 Section 3.5.1~3.5.2) can be revealed in plots. Besides, on the final evaluation of the retrieval model, if the F-score based on the development set was also specified (Page 3 Section 3.5.4), the reader could see if there is a difference between the local and the actual model's performance. Also, if the final evaluation based on the codalab 'full-test-set' was listed as required (or pointed out clearly in the report), the report could become more precise and convincing.