Reproducibility of Investigating Whether Fact Checking Models Learn to Reason

Yunyi Xu[1], Xiangyu Liu[2], Jiayang Lyu[3]

***Abstract*— The results of a research should be reliable and reproducible to enable future studies. Therefore, reproducibility is a major theme in the scientific method. We investigated a study of fact checking based on reasoning [1] in this mini-project. The authors claimed that the models performance best when training only with the evidence. The results we reproduced mostly align with the results of the original research, which support the claim.**

I. Introduction and Motivation

Reproducibility is an essential part of scientific research for several reasons. Academic integrity wise, by reproducing the results, one can show whether the claimed results are true. Another reason is that reproducibility enables future studies build on this research to carry out.

For the project, we focused on the reproducibility of the article "Automatic Fake News Detection: Are Models Learning to Reason?" [1].

We chose this paper for the following reasons: Fake News Detection is a very interesting topic in natural language processing and it has been widely applied in our real world. Now most of these detection models for fake news are based on reasoning: given a claim with associated evidence, the models would give prediction of claim veracity based on the supporting or refuting content within the evidence. But in this paper we chose, it investigates the relationship and importance of both claim and evidence, and after implementing three different models: Random Forest(RF), Long Short Memory Model(LSTM) and Bidirectional Encoder Representations from Transformers(BERT) on the extracted MultiFC similar dataset [2], it is found that using only the evidence can most often obtain the best performance.

[1]Yunyi Xu (260820529): yunyi.xu@mail.mcgill.ca
[2]Xiangyu Liu (260848712): xiang.yu.liu@mail.mcgill.ca
[3]Jiayang Lyu (260842471): jiayang.lyu@mail.mcgill.ca

II. Summary of the investigated paper

The paper[1] focuses on one question: provided a claim and associated evidence, whether the model determines the claim's veracity by reasoning over the evidence. They train three models (Term-frequency-based Random Forest, GloVe-based LSTM, BERT) on the MultiFC dataset[2] to compare the results of three different input types: claim only, evidence only, and cliam+evidence. They find that the best performance can most often be obtained using only evidence, which shows that the face-checking models are not using reasoning but an inherent signal in the evidence itself.

III. Experiment Setup

	#Claims	Labels
PolitiFact	13,581	pants on fire! (10.6%), false (19.2%), mostly false (17.0%), half-true (19.8%), mostly true (18.8%), true (14.8%)
Snopes	5,069	false (64.3%), mostly false (7.5%), mixture (12.3%), mostly true (2.8%), true (13.0%)

Fig. 1: Dataset Statistics

Setup

There are 2 datasets, namely PolitiFact and Snopes. Both of them consists of political claim-evidence pairs. The training/validation/test sets are splited into a ratio of 7:1:2. The classification of the 2 datasets are shown in Fig.1.

The models leveraged in the experiments vary in complexity, namely Term-frequency based Random Forest (RF), GloVe-based LSTM model (LSTM), and BERT-based model (BERT). For the tuning details, we followed the guidelines in the original work. We also tried some alterations to the hyperparameters and we found the tunings suggested by the authors were reliable and worked the best.

Challenges

First, we found that BERT nodels requires large amount of time and space to train. In our experiments, 1 BERT model (6 in total) involves over 100M parameters, and

takes up to 6 hours to train. We encountered a few failures as it hit the RAM limit in Google Colab Pro after running for hours. To solve this problem, we adapted early stopping by reducing the patience from 10 to 2. We are aware that this will hurt the test results.

Second, a similar dataset for this paper is hard to obtain as the original dataset consists of snippets, which are returned by Google searching API. After searching for many similar projects and databases, We found one json file containing evidence with urls but we failed at data prepocessing and it could not fit into our models.

Model		Train: Snopes				Train: PolitiFact			
		Within dataset Eval: Snopes		Out-of dataset Eval: PolitiFact		Within dataset Eval: PolitiFact		Out-of dataset Eval: Snopes	
		F1micro	F1macro	F1micro	F1macro	F1micro	F1macro	F1micro	F1macro
RF	Claim	0.472	0.236	**0.271**	0.225	0.258	0.250	0.540	0.218
	Evidence	**0.559**	0.295	0.256	0.199	0.309	0.302	**0.586**	0.230
	Claim+Evidence	0.513	0.265	0.240	0.194	0.305	0.302	0.584	0.218
LSTM	Claim	0.401	0.243	0.247	0.213	0.250	0.249	0.542	0.228
	Evidence	0.464	0.269	0.244	0.195	0.250	0.252	0.542	0.256
	Claim+Evidence	0.469	0.277	0.257	0.204	0.259	0.263	0.539	**0.286**
BERT	Claim	0.519	0.299	0.252	0.197	0.284	0.283	0.580	0.244
	Evidence	0.529	0.297	0.257	**0.228**	**0.341**	**0.350**	0.561	0.268
	Claim+Evidence	0.555	**0.329**	0.246	0.187	0.312	0.317	0.574	0.250

Fig. 2: Evaluation using micro and macro F1. Per column, the best score per method is underlined and the best score across all methods is highlighted in bold, in the same fashion of the paper.

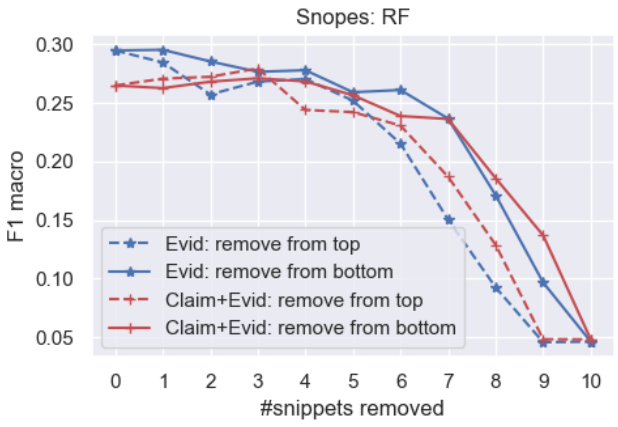

Fig. 3: Snopes-RF:evidence only VS evidence+claim

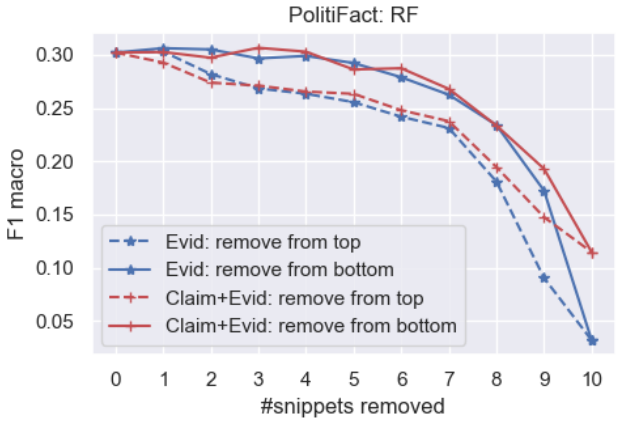

Fig. 4: PolitiFact-RF:evidence only VS evidence+claim

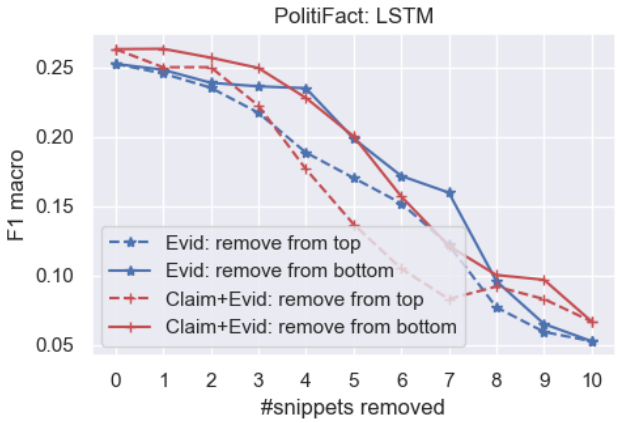

Fig. 5: PolitiFact-LSTM:evidence only VS evidence+claim

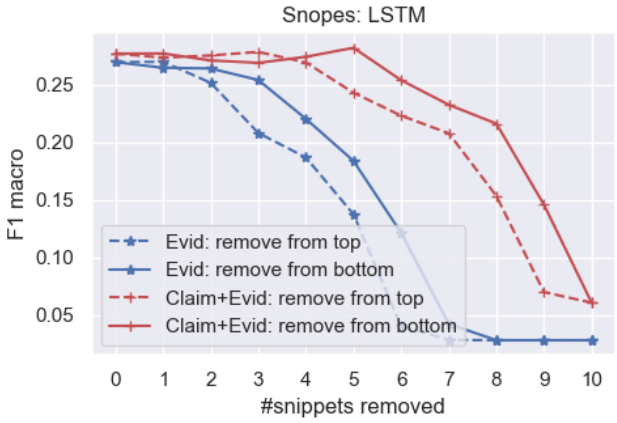

Fig. 6: Snopes-LSTM:evidence only VS evidence+claim

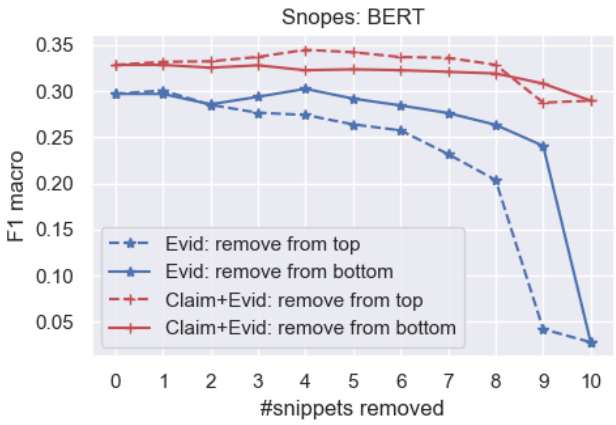

Fig. 7: Snopes-BERT:evidence only VS evidence+claim

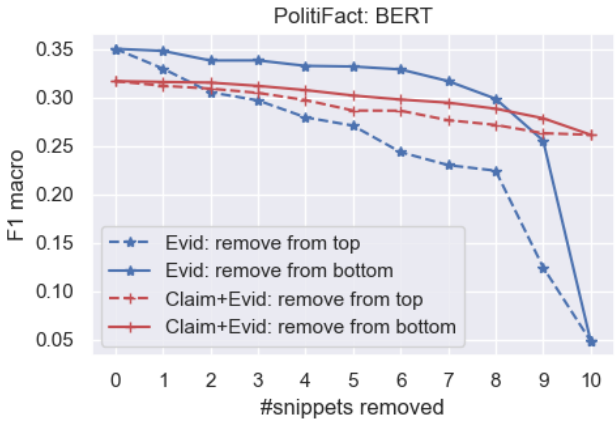

Fig. 8: PolitiFact-BERT:evidence only VS evidence+claim

IV. Results

Compared our results and the results given in the paper, the overall performance is very similar. There is little difference between the numerical results and plots generated from our trained models and those given in the paper. The best performances of the three models and three different input types are indeed most often obtained by using evidence only.

However,some details from our result show weaker support to the conclusion in the investigate paper.

Influence of the models

First, the choice of models can seriously influence the result of which input type has the best performance (i.e. Not all best performances are obtained by using evidence only). In Random Forest model, the best performance are obtained by using evidence only. However, in the LSTM model, the best performance are obtained by using evidence+claim and its performance fig**??** fur surpasses using evidence only, which provides completely opposite evidence for the argument in the paper. Another opposite evidence will be fig**??**, in which the performance of using claim and evidence is better.

Insignificant Adavantages

Second,the advantages of using evidence only is not significant. For example, in the BERT model and dataset Snopes, the F1macro is only 2.5% higher than using claim+evidence, which cannot be viewed as a convincing evidence. Meanwhile, according to plots of Random Forest model and PolitiFact dataset, the performance of two different input types is almost the same fig8.

V. Discussion and Conclusion

According to results, the conclusion in the paper are not strongly supported by the existing evidence. Although the use of evidence only outperforms the use of claim and evidence in the two of three tested models, the numerical advantage of using evidence only is not significant enough. Furthermore, other factors like randomness in the systems can also lead to the same effect rather than directly concluding that the models do not reason. Therefore, a more decisive numerical advantage of using evidence only is needed.

Another adverse evidence is the better performance of using claims and evidence in the LSTM. The results make us doubt the universality of the outperformance of using evidence only. The conclusion in the paper may be too general.

We are also skeptical of one argument of precondition: the increase of effectiveness when including the evidence in the model is the sign of reasoning. In our opinion, reasoning involves understanding and inference, which are the essence of intelligence, and it is not easy to achieve as discussed in the paper[3]. The term of reasoning should be explicitly defined.

To further explore the reproducibility of this paper, we tried to implement our models on similar datasets to see the performance. Due to time issue and the complexity of the dataset, we could not have tested all models. But as we found in the paper Hierarchical Multi-head Attentive Network for Evidence-aware Fake News Detection[4], the authors designed their experiment on the similar dataset which is also from PolitiFact and Snopes(claim and evidence provided). In this paper, different models are tested using claim and evidence+claim to compare performances, which could partially reflect the irrelevance of reasoning.

As pointed at the end of paper, the claim leads us to a potential problem of how exactly the evidences are obtained, we believe that a way to improve our current experiment is to adding extending representations into evidence. Taking example of political news we are using in the paper, an statement published by the White House is more likely to be true compared to a post regarding the same topic on Twitter. Hence we could contain more information or weights on the source of evidence, which could possibly improve the performance.

VI. Statement of Contributions

1) Yunyi Xu: Running and debugging the source code. Recording results. Drawing tables. Writing the report.
2) Xiangyu Liu: Researching. Running and debugging the source code. Plotting graphs. Writing the report.
3) Jiayang Lyu: Researching and writing report.

References

[1] Hansen, C., Hansen, C., amp; Chaves Lima, L. (2021). Automatic fake news detection: Are models learning to reason? Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). https://doi.org/10.18653/v1/2021.acl-short.12

[2] Augenstein, I., Lioma, C., Wang, D., Chaves Lima, L., Hansen, C., Hansen, C., amp; Simonsen, J. G. (2019). MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). https://doi.org/10.18653/v1/d19-1475

[3] Bender, E. M., amp; Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.463

[4] Vo, N., amp; Lee, K. (2021). Hierarchical multi-head Attentive Network for evidence-aware fake news detection. Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. https://doi.org/10.18653/v1/2021.eacl-main.83