Explaining Veracity Predictions with Evidence Summarization: A Multi-Task Model Approach

Anonymous ACL submission

Abstract

The rapid dissemination of misinformation through social media increased the importance of automated fact-checking. Furthermore, studies on what deep neural models pay attention to when making predictions have increased in recent years. While significant progress has been made in this field, it has not yet reached a level of reasoning comparable to human reasoning. To address these gaps, we propose a multi-task explainable neural model for misinformation detection. Specifically, this work formulates an explanation generation process of the model's veracity prediction as a text summarization problem. Additionally, the performance of the proposed model is discussed on publicly available datasets and the findings are evaluated with related studies. [1]

1 Introduction

Fake news is considered as media content that contains misinformation and can mislead people (Shu et al., 2017; Zhou and Zafarani, 2020). Advancements in social networking and social media not only facilitate information accessibility but also cause the rapid spread of fake news on social media (Vosoughi et al., 2018). Consequently, fake news becomes a powerful tool for manipulating public opinion, as observed during influential events like the 2016 US Presidential Election and the Brexit referendum (Pogue, 2017; Allcott and Gentzkow, 2017). To address this issue, automated fake news detection methods have emerged, aiming to determine the veracity of claims while minimizing human effort (Oshikawa et al., 2020).

Multi-task learning (MTL) is a technique in machine learning to train similar tasks at the same time by leveraging their differences and commonalities (Crawshaw, 2020; Chen et al., 2021; Zhang and Yang, 2021). Additionally, MTL allows data utilization as the model can transfer knowledge between tasks. Notably, the insights gained while learning one task can benefit other related tasks, leading to better generalization across tasks. Moreover, from the business point of view, deploying a single multi-task model may reduce the complexity of maintenance and resource requirements.

This paper primarily focuses on designing a multi-task explainable misinformation detection model. To be more specific, a fact-checking model is trained on veracity prediction and text summarization tasks simultaneously. The generated summaries are derived from evidence documents and serve as justifications for the model's veracity prediction. Therefore, it should not be considered as a post-hoc explainability model. The contribution of the work lies in the use of multi-task learning for fact-checking and text summarization together, particularly through a new architecture including different neural models. The tasks, fact-checking and summarization, complement each other such that one does misinformation detection while the other explains the reason for the model's decision.

2 Related Work

Automated fake news detection studies have been studied from data mining (Shu et al., 2017) and natural language processing (Oshikawa et al., 2020; Guo et al., 2022; Vladika and Matthes, 2023) perspectives. Zhou and Zafarani (Zhou and Zafarani, 2020) classify the previous studies into four groups: knowledge-based (Pan et al., 2018; Cui et al., 2020), style-based (Zhou et al., 2020; Pérez-Rosas et al., 2018; Jin et al., 2016; Jwa et al., 2019), propagation-based (Hartmann et al., 2019; Zhou and Zafarani, 2019), and source-based (Sitaula et al., 2020).

Kotonya and Toni (Kotonya and Toni, 2020a) present a survey on explainable fact-checking that categorized the studies based on their methods for

[1] A GitHub link to the source code will be available at the camera-ready stage.

generating explanations. These methods include exploiting neural network artifacts (Popat et al., 2017, 2018; Shu et al., 2019; Lu and Li, 2020; Silva et al., 2021), rule-based approaches (Szczepański et al., 2021; Gad-Elrab et al., 2019; Ahmadi et al., 2020), summary generation (Atanasova et al., 2020a; Kotonya and Toni, 2020b; Stammbach and Ash, 2020; Brand et al., 2022), adversarial text generation (Thorne et al., 2019; Atanasova et al., 2020b; Dai et al., 2022), causal inference for counterfactual explanations (Cheng et al., 2021; Zhang et al., 2022; Li et al., 2023; Xu et al., 2023), neurosymbolic reasoning (Pan et al., 2023) and question-answering (Ousidhoum et al., 2022; Yang et al., 2022).

The most related study in the literature was the E-BART model (Brand et al., 2022) that was trained for both classification and summarization by introducing a joint prediction head on top of the BART (Lewis et al., 2020) language model. In other words, the encoder and decoder of the BART model are shared for both tasks. In contrast to this approach, this work incorporates the T5 Encoder as a shared module. For summarization, a T5 Decoder is trained while feed-forward layers are employed for classification. We also measured the effect of using two loss weighting strategies and evaluated the impact of instruction fine-tuning by switching the T5 model with the Flan-T5 (Chung et al., 2022) version.

3 Method

In this study, a multi-task model that is based on the T5 (Raffel et al., 2020) transformer is proposed. The model is trained on text summarization and veracity prediction tasks jointly. T5 transformer is an encoder-decoder model that converts each task to a text-to-text problem. Google AI released the Flan-T5 (Chung et al., 2022) model that employs instructional fine-tuning to further improve the T5 model that is also utilized in the evaluation.

The model architecture is given in Figure 1. Both summarization and classification tasks share a T5 Encoder during training. At first, the T5 Encoder encodes the claim and evidence sentences in a latent space. Afterwards, the T5 Decoder produces a summary using the T5 Encoder's representation. Simultaneously, for the veracity prediction, the encoder's output is processed by two feed-forward layers respectively. We employ the ReLU activation function and apply dropout between two linear

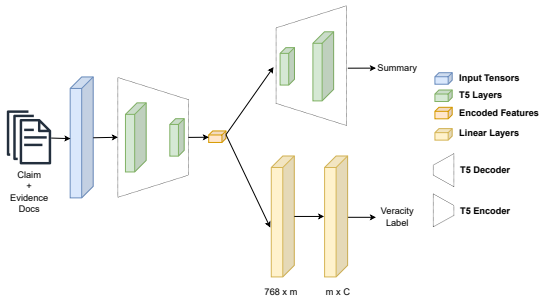

Figure 1: The multi-task model architecture

layers and the sigmoid activation function after the second linear layer. Besides, the cross entropy loss is used for measuring summary and classification losses.

Two loss weighting strategies are employed: i) static loss coefficients and ii) uncertainty weighting. For the static loss coefficients, constant weights are set for the classification and summarization losses prior to training. To determine the optimal weights, grid search-based validation experiments are performed. In addition to the static loss coefficients, this paper also utilizes the uncertainty weighting strategy (Kendall et al., 2018) that enables dynamic adjustment of the weights based on prediction confidence. Subsequently, the overall loss is calculated by taking the weighted sums of the summary loss and the classification loss.

Figure 2 presents an example claim alongside our model's predictions. Based on supplementary information provided under the "Evidence" section, the claim has been verified by a reviewer. The gold standard summary was also authored by human annotators, while the abstractive summary was generated by a T5-based multi-task model. The generated summary not only aligns with the veracity label but can be considered as an explanation of the model's reasoning behind its decision.

4 Experimental Results

In this section, the proposed model was evaluated on three benchmark datasets. Note that we employed the T5-large model in the Huggingface's transformer library [2] and only the best results obtained during the validation experiments for each model are presented. Note that the experiments were conducted using Nvidia RTX A6000 GPUs.

PUBHEALTH Results: The PUBHEALTH (Kotonya and Toni, 2020b) dataset consists of

[2] https://huggingface.co/docs/transformers/model_doc/t5

Claim: Study says too many Americans still drink too much.
Evidence: ... The researchers found that 64 percent of men and 79 percent of women said they drank no alcohol at all that day, and another 18 percent of men and 10 percent of women drank within the recommended amounts. Nine percent of men said they had three to four drinks the day before and 8 percent of women said they drank two to three alcoholic beverages, the researchers said. The heaviest drinkers of all were the 8 percent of men who had five or more drinks, and 3 percent of women who had four or more. "**Overall the study confirms that rates of unhealthy alcohol use in the U.S. are significant**," said Jennifer Mertens, a research medical scientist at Kaiser Permanente Division of Research in Oakland, ...
Gold Summary: On any given day in the United States, 18 percent of men and 11 percent of women drink more alcohol than federal guidelines recommend, according to a study that also found that 8 percent of men and 3 percent of women are full-fledged "heavy drinkers."
Our Model's Summary: Americans are still drinking too much alcohol, even if they don't drink at all on any given day, according to a new study.
Label: True

Figure 2: A sample claim from PUBHEALTH (Kotonya and Toni, 2020b) with our model's outputs

Table 1: Summarization results on PUBHEALTH

Model	Rouge-1	Rouge-2	Rouge-L
Oracle (Kotonya and Toni, 2020b)	39.24	14.89	32.78
Lead-3 (Kotonya and Toni, 2020b)	29.01	10.24	24.18
EXPLAINERFC-EXPERT (Kotonya and Toni, 2020b)	32.30	13.46	26.99
T5 single-task	30.90	13.40	27.16
T5 multi-task	**32.55**	**14.54**	**28.60**
Flan-T5 multi-task	32.38	14.03	28.41

Table 2: Veracity results on PUBHEALTH

Model	Precision	Recall	F1-macro	Accuracy
BERT (rand. sentences) (Kotonya and Toni, 2020b)	38.97	39.38	39.16	20.99
BERT (all sentences) (Kotonya and Toni, 2020b)	56.50	56.50	56.50	56.40
BERT (top-k) (Kotonya and Toni, 2020b)	77.39	54.77	63.93	66.02
SCIBERT (Kotonya and Toni, 2020b)	75.69	66.20	**70.52**	69.73
T5 single-task	**78.24**	71.05	61.08	71.05
T5 multi-task	77.62	70.32	60.93	70.32
Flan-T5 multi-task	76.46	**76.64**	65.18	**76.64**

health-related claims with justifications which were written by journalists were considered as gold explanations to evaluate the correctness of claims. Each claim was annotated as *True*, *False*, *Mixture* or *Unproven*. The training set consists of 9466 claims and 1183 claims exist in validation and test sets.

Table 1 displays the summarization outcomes of our proposed models in comparison to the baseline and Oracle models. Lead-3 (Kotonya and Toni, 2020b) served as the baseline that utilized the first three sentences as a summary. Oracle (Kotonya and Toni, 2020b) was an extractive summary model that served as an upper bound. Additionally, EXPLAINERFC-EXPERT (Kotonya and Toni, 2020b) was a state-of-the-art single-task abstractive summary generator model which performed slightly better than our single-task model. Note that the T5 single-task and the T5 multi-task models were almost identical to the model architecture given in Figure 1 but the classification head of the T5 single-task model was set to 0.

Furthermore, the Flan-T5 multi-task model represents an instruction fine-tuned variant of T5 that performed slightly less effectively than T5 for summarization, but both models outperformed the state-of-the-art model.

The results for veracity prediction using the precision, recall, F1-macro and accuracy metrics were presented in Table 2. The first two rows indicated the baselines. BERT (top-k) and SCIBERT models applied a sentence selection based on the sentences' semantic similarity with the claim sentences. For evidence selection, the authors employed the S-BERT (Reimers and Gurevych, 2019) model. Therefore, we followed a similar approach and selected the top-5 evidence sentences and the claim statement as input for these models.

The results indicate that the Flan-T5 variant outperformed the T5-based models for veracity detection but on the F1-macro metric the state-of-the-art SCIBERT model performed significantly better than our models. The main reason for this difference can be attributed to the considerable imbalance in label distribution. For instance, the ratio of claims labeled as *Unproven* is approximately 3.2%, while the *Mixture* cases constitute around 15.2% of the dataset. Our post-evaluation analysis, in Appendix A, revealed that despite the usage of additional coefficients for the *Unproven* and *Mixture* instances, our models suffered from the data imbalance problem. Overall, the joint training of veracity detection and evidence summarization has a positive influence on the performance of our model in both tasks.

FEVER Results: FEVER (Thorne et al., 2018) is a benchmark dataset that includes 185K claims with related evidence documents from Wikipedia. The dataset was published for the FEVER shared task in 2018 and the task consists of claim identification, evidence retrieval and fact-checking subtasks. For the fact-checking task, the claim statements were annotated as *Supports*, *Refutes* and *Not enough info*.

Table 3: Veracity and summarization results on e-FEVER

Model	Dataset	Acc. (w/o N.E.I)	Acc.	Rouge-1	Rouge-2	Rouge-L
E-BARTSmall (Brand et al., 2022)	eFever_Small	87.2	78.2	73.58	**64.37**	71.43
E-BARTFull (Brand et al., 2022)	eFever_Full	85.2	77.2	65.51	57.60	64.07
T5-Full (Only Summarization)	eFever_Full	-	-	65.94	57.53	65.09
T5-Full (Only Classification)	eFever_Full	91.12	73.61	-	-	-
T5-Small	eFever_Small	91.11	74.75	74.00	63.64	72.78
T5-Small (uncertainty weighting)	eFever_Small	90.66	74.57	**74.46**	64.32	**73.19**
T5-Full	eFever_Full	90.91	75,26	68,16	59,96	67,26
T5-Full (uncertainty weighting)	eFever_Full	90.90	74,28	67,30	59,36	66,49
Flan-T5	eFever_Full	**94.36**	**79.91**	66.75	58.42	65.88
Flan-T5 (uncertainty weighting)	eFever_Full	93.94	79.02	68.84	60.89	67.97

Since the FEVER test set did not contain the true labels, the multi-task model's veracity prediction performance was evaluated using the development set. We employed the DOMLIN system (Stammbach and Neumann, 2019) to retrieve evidence documents. DOMLIN retrieved evidence documents for 17K out of the 20K claims in the development set, while labeling the remaining instances as "not enough info." With this supporting information, our multi-task model achieved an accuracy score of 76.18%. However, its Flan-T5-based counterpart outperformed it with a score of 80.44%. It's worth noting that the DOMLIN model (Stammbach and Neumann, 2019) achieved an accuracy of 71.44%, DOMLIN++ (Stammbach and Ash, 2020) achieved 77.48%, and the E-BART (Brand et al., 2022) model reached an accuracy of 75.10% by utilizing the similar evidence retrieval method.

e-FEVER Results: The e-FEVER dataset (Stammbach and Ash, 2020) is a subset of the original FEVER dataset and consists of 67687 claims with evidence documents retrieved using the DOMLIN system (Stammbach and Neumann, 2019). In addition to claims and evidence documents, the authors published the summaries generated by the GPT-3-based model (Brown et al., 2020) for each claim. Hence, these summaries were leveraged as ground-truth explanations to compare our model's decision-making process with the GPT-3-based model.

The authors pointed out that the GPT-3-based model generated null summaries for certain claims. To address this issue, similar to Brand et al. (Brand et al., 2022), two variations of the dataset were utilized: *e-FEVER_Full* and *e-FEVER_Small*. The former contains all claims, while the latter excluded instances with null summaries. The *e-FEVER_Small* consists of 40702 instances. Moreover, Brand et al. (Brand et al., 2022) provided some examples labeled as *Not enough info* that could be either refuted or supported based on the provided evidence documents. Therefore, the binary veracity prediction performance of the multi-task model was measured by ignoring the *Not enough info* instances. Likewise, similar to Brand et al. (Brand et al., 2022) two variations of the multi-task model were trained: *T5-Small* and *T5-Full* where the former was trained on *e-FEVER_Small* and the latter was trained on *e-FEVER_Full*.

Table 3 demonstrated the summarization and veracity prediction results on the e-FEVER dataset. To the best of our knowledge, only Brand et al. (Brand et al., 2022) reported results on this dataset. The baseline models were outlined in the third and fourth rows that were trained specifically for either summarization or classification. Therefore, we did not report the classification results for the summarization model, and vice versa. The fifth and sixth rows indicated the models that utilized *e-FEVER_Small* dataset. Both of the models outperformed the E-BARTSmall model for summarization and binary classification. However, E-BARTSmall achieved higher accuracy than the proposed models in three-class classification.

Similarly, on the initial data, *eFever_Full*, the multi-tasked T5 models also achieved higher binary classification accuracy and summarization scores but performed worse than the E-BARTFull model in multi-class classification. On the other hand, replacing T5 with the Flan-T5 version led to the highest accuracy scores in both binary and multi-class classification. Moreover, the Rouge scores of the T5 and Flan-T5 models were higher than the E-BART model on the initial e-FEVER data. These results on the dataset indicate a mutual influence between veracity detection and evidence summarization due to joint training. Furthermore, it was observed that the overall performance was not significantly affected by the selection of the loss weighting strategy.

5 Conclusion

In this paper, a T5-based explainable multi-task fact-checking model is introduced. The results revealed that leveraging multi-task learning yields significant improvements in text summarization and veracity detection performance.

6 Limitations

First, the T5 and Flan T5 models were pre-trained massively on English corpora. Consequently, the performance of these models on languages with limited resources may not be satisfactory. Secondly, the validation experiments revealed significant fluctuations in the model's performance when utilizing certain hyperparameter sets. Therefore, the hyperparameter optimization was a critical part of the evaluation process. Furthermore, the interpretability of the generated explanations may vary depending on the complexity of the text. Transformer-based language models demand significant computational and hardware resources. However, some recent parameter-efficient fine-tuning techniques, such as LoRA (Hu et al., 2022), have demonstrated their effectiveness. Therefore, future research should address these limitations to enhance the robustness and applicability of our approach.

References

Naser Ahmadi, Thi-Thuy-Duyen Truong, Le-Hong-Mai Dao, Stefano Ortona, and Paolo Papotti. 2020. Rulehub: A public corpus of rules for knowledge graphs. *Journal of Data and Information Quality (JDIQ)*, 12(4):1–22.

Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–36.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020a. Generating fact checking explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7352–7364.

Pepa Atanasova, Dustin Wright, and Isabelle Augenstein. 2020b. Generating label cohesive and well-formed adversarial claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3168–3177.

Erik Brand, Kevin Roitero, Michael Soprano, Afshin Rahimi, and Gianluca Demartini. 2022. A neural model to jointly predict and explain truthfulness of statements. *ACM Journal of Data and Information Quality*, 15(1):1–19.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Shijie Chen, Yu Zhang, and Qiang Yang. 2021. Multi-task learning in natural language processing: An overview. *arXiv preprint arXiv:2109.09138*.

Lu Cheng, Ruocheng Guo, Kai Shu, and Huan Liu. 2021. Causal understanding of fake news dissemination on social media. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 148–157.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Michael Crawshaw. 2020. Multi-task learning with deep neural networks: A survey. *arXiv preprint arXiv:2009.09796*.

Limeng Cui, Haeseung Seo, Maryam Tabar, Fenglong Ma, Suhang Wang, and Dongwon Lee. 2020. Deterrent: Knowledge guided graph attention network for detecting healthcare misinformation. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 492–502.

Shih-Chieh Dai, Yi-Li Hsu, Aiping Xiong, and Lun-Wei Ku. 2022. Ask to know more: Generating counterfactual explanations for fake claims. *arXiv preprint arXiv:2206.04869*.

Mohamed H Gad-Elrab, Daria Stepanova, Jacopo Urbani, and Gerhard Weikum. 2019. Exfakt: A framework for explaining facts over knowledge graphs and text. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 87–95.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Mareike Hartmann, Yevgeniy Golovchenko, and Isabelle Augenstein. 2019. Mapping (dis-) information flow about the mh17 plane crash. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 45–55.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Zhiwei Jin, Juan Cao, Yongdong Zhang, Jianshe Zhou, and Qi Tian. 2016. Novel visual and statistical image features for microblogs news verification. *IEEE transactions on multimedia*, 19(3):598–608.

Heejung Jwa, Dongsuk Oh, Kinam Park, Jang Mook Kang, and Heuiseok Lim. 2019. exbake: Automatic fake news detection model based on bidirectional encoder representations from transformers (bert). *Applied Sciences*, 9(19):4062.

Alex Kendall, Yarin Gal, and Roberto Cipolla. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491.

Neema Kotonya and Francesca Toni. 2020a. Explainable automated fact-checking: A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5430–5443.

Neema Kotonya and Francesca Toni. 2020b. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Yichuan Li, Kyumin Lee, Nima Kordzadeh, and Ruocheng Guo. 2023. What boosts fake news dissemination on social media? a causal inference view. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 234–246. Springer.

Yi-Ju Lu and Cheng-Te Li. 2020. Gcan: Graph-aware co-attention networks for explainable fake news detection on social media. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 505–514.

Ray Oshikawa, Jing Qian, and William Yang Wang. 2020. A survey on natural language processing for fake news detection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6086–6093.

Nedjma Ousidhoum, Zhangdie Yuan, and Andreas Vlachos. 2022. Varifocal question generation for fact-checking. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2532–2544, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jeff Z Pan, Siyana Pavlova, Chenxi Li, Ningxi Li, Yangmei Li, and Jinshuo Liu. 2018. Content based fake news detection using knowledge graphs. In *International semantic web conference*, pages 669–683. Springer.

Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. 2023. Fact-checking complex claims with program-guided reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6981–7004, Toronto, Canada. Association for Computational Linguistics.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic detection of fake news. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3391–3401.

David Pogue. 2017. How to stamp out fake news. *Scientific American*, 316(2):24–24.

Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2017. Where the truth lies: Explaining the credibility of emerging claims on the web and social media. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 1003–1012.

Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 2018. Declare: Debunking fake news and false claims using evidence-aware deep learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 22–32.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 395–405.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36.

Amila Silva, Yi Han, Ling Luo, Shanika Karunasekera, and Christopher Leckie. 2021. Propagation2vec: Embedding partial propagation networks for explainable fake news early detection. *Information Processing & Management*, 58(5):102618.

Niraj Sitaula, Chilukuri K Mohan, Jennifer Grygiel, Xinyi Zhou, and Reza Zafarani. 2020. Credibility-based fake news detection. In *Disinformation, Misinformation, and Fake News in Social Media*, pages 163–182. Springer.

Dominik Stammbach and Elliott Ash. 2020. e-fever: Explanations and summaries for automated fact checking. *Proceedings of the 2020 Truth and Trust Online (TTO 2020)*, pages 32–43.

Dominik Stammbach and Guenter Neumann. 2019. Team domlin: Exploiting evidence enhancement for the fever shared task. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, pages 105–109.

Mateusz Szczepański, Marek Pawlicki, Rafał Kozik, and Michał Choraś. 2021. New explainability method for bert-based model in fake news detection. *Scientific Reports*, 11(1):1–13.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2019. Evaluating adversarial attacks against multiple fact verification systems. Association for Computational Linguistics.

Juraj Vladika and Florian Matthes. 2023. Scientific fact-checking: A survey of resources and approaches. *arXiv preprint arXiv:2305.16859*.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *science*, 359(6380):1146–1151.

Weizhi Xu, Qiang Liu, Shu Wu, and Liang Wang. 2023. Counterfactual debiasing for fact verification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6777–6789, Toronto, Canada. Association for Computational Linguistics.

Jing Yang, Didier Vega-Oliveros, Taís Seibt, and Anderson Rocha. 2022. Explainable fact-checking through question answering. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8952–8956. IEEE.

Weifeng Zhang, Ting Zhong, Ce Li, Kunpeng Zhang, and Fan Zhou. 2022. Causalrd: A causal view of rumor detection via eliminating popularity and conformity biases. In *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*, pages 1369–1378. IEEE.

Yu Zhang and Qiang Yang. 2021. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 34(12):5586–5609.

Xinyi Zhou, Atishay Jain, Vir V Phoha, and Reza Zafarani. 2020. Fake news early detection: A theory-driven model. *Digital Threats: Research and Practice*, 1(2):1–25.

Xinyi Zhou and Reza Zafarani. 2019. Network-based fake news detection: A pattern-driven approach. *ACM SIGKDD explorations newsletter*, 21(2):48–60.

Xinyi Zhou and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, 53(5):1–40.

A Confusion Matrices

Table 4: Confusion Matrix

Model		Unproven	False	Mixture	True	Accuracy
T5	Unproven	27	8	5	5	60.00
single	False	31	244	94	19	62.89
task	Mixture	17	41	131	12	65.17
	True	21	8	96	474	79.13
T5	Unproven	26	10	4	5	57.78
multi	False	31	236	106	15	62.43
task	Mixture	13	37	137	14	68.16
	True	15	17	99	468	78.13
Flan-T5	Unproven	25	14	1	5	55.56
multi	False	14	307	48	19	79.12
task	Mixture	9	61	87	44	43.28
	True	9	25	39	526	87.81

The confusion matrices of the models given in Table 2 are demonstrated in Table 4. Confusion matrices revealed that the margins between the state-of-the-art model's and our models' F1-macro scores are attributed to the class distributions. More specifically, the dataset is highly imbalanced and despite boosting the *Unproven* and *Mixture* instances, the models suffered from the class imbalance problem. Moreover, another takeaway is that boosting the *Mixture* instances decreased the accuracy of *False* claims, particularly for T5 models.

B Grid Search of Static Loss Coefficients

We performed an ablation study to explore candidate values to find an optimal set of hyper-parameters for our multi-task model. We performed a grid search using PUBHEALTH (Kotonya and Toni, 2020b) dataset to determine the optimal set of loss coefficients. The experimental results are presented in Table 5. Note that, we kept the linear layers' size (for veracity prediction), dropout probability, batch size and number of epoch constant.

C Grid Search of Hidden Layer Dimensions for Veracity Prediction

We also performed another ablation study to discover the optimal hidden layer size of the classification head of our multi-task model using the PUBHEALTH (Kotonya and Toni, 2020b) dataset. The experimental results are presented in Table 6. Note that, we kept the dropout probability, batch size and number of epochs constant.

Table 5: Grid search of loss coefficients

Veracity (a), Summary (b) loss coefficients	Veracity label coefficients	Rouge-1	Rouge-2	Rouge-L	F1-macro	F1-weighted
a=0.7, b=0.3	mixture_coeff=1.75, unproven_coeff=5	31,99	14,14	28,18	51,14	66,66
a=0.7, b=0.3	mixture_coeff=1.75, unproven_coeff=7	31,93	14,26	28,46	**60,76**	**73,16**
a=0.7, b=0.3	mixture_coeff=1.75, unproven_coeff=9	31,87	14,16	28,13	48,62	63,93
a=0.6, b=0.4	mixture_coeff=1.75, unproven_coeff=5	31,25	13,81	27,59	54,71	69,92
a=0.6, b=0.4	mixture_coeff=1.75, unproven_coeff=7	32,36	**14,59**	28,67	57,22	71,47
a=0.6, b=0.4	mixture_coeff=1.75, unproven_coeff=9	31,85	14,21	28,16	54,14	68,48
a=0.5, b=0.5	mixture_coeff=1.75, unproven_coeff=5	32,52	14,50	**28,74**	56,71	69,86
a=0.5, b=0.5	mixture_coeff=1.75, unproven_coeff=7	31,87	13,94	27,09	52,00	67,25
a=0.5, b=0.5	mixture_coeff=1.75, unproven_coeff=9	31,71	13,88	28,19	51,12	65,78
a=0.7, b=0.3	mixture_coeff=2.5, unproven_coeff=5	31,02	13,50	27,53	50,94	65,48
a=0.7, b=0.3	mixture_coeff=2.5, unproven_coeff=7	31,82	14,00	28,12	55,87	68,57
a=0.7, b=0.3	mixture_coeff=2.5, unproven_coeff=9	31,96	14,42	28,40	56,52	69,93
a=0.6, b=0.4	mixture_coeff=2.5, unproven_coeff=5	31,43	14,03	27,75	50,59	65,76
a=0.6, b=0.4	mixture_coeff=2.5, unproven_coeff=7	31,96	14,38	28,28	55,62	68,57
a=0.6, b=0.4	mixture_coeff=2.5, unproven_coeff=9	**32,54**	14,48	28,69	60,07	72,50
a=0.5, b=0.5	mixture_coeff=2.5, unproven_coeff=5	31,78	13,86	28,04	58,73	72,20
a=0.8, b=0.2	mixture_coeff=1.75, unproven_coeff=5	32,27	14,32	28,64	58,02	72,19
a=0.8, b=0.2	mixture_coeff=1.75, unproven_coeff=7	31,05	13,44	27,49	50,96	65,48
a=0.8, b=0.2	mixture_coeff=1.75, unproven_coeff=9	32,03	13,74	28,06	57,59	70,41
a=0.5, b=0.5	mixture_coeff=1.75, unproven_coeff=5	32,00	14,29	28,38	56,31	70,19
a=0.5, b=0.5	mixture_coeff=1.75, unproven_coeff=7	31,82	14,16	28,14	55,05	69,52
a=0.5, b=0.5	mixture_coeff=1.75, unproven_coeff=9	31,87	14,15	28,22	58,33	72,34
a=0.8, b=0.2	mixture_coeff=2.5, unproven_coeff=5	32,42	14,11	28,50	54,14	67,34
a=0.8, b=0.2	mixture_coeff=2.5, unproven_coeff=7	32,03	14,20	28,31	58,87	71,89
a=0.8, b=0.2	mixture_coeff=2.5, unproven_coeff=9	31,84	13,93	28,07	58,10	71,95
a=0.5, b=0.5	mixture_coeff=2.5, unproven_coeff=5	31,85	14,25	28,13	52,58	66,45
a=0.5, b=0.5	mixture_coeff=2.5, unproven_coeff=7	32,33	14,18	28,48	60,33	73,11
a=0.5, b=0.5	mixture_coeff=2.5, unproven_coeff=9	31,90	14,14	28,27	55,56	70,32

Table 6: Grid search of hidden layer size

Veracity (a), Summary (b) loss coefficients	Veracity label coefficients	Hidden Dim	Rouge-1	Rouge-2	Rouge-L	F1-macro	F1-weighted
a=0.7, b=0.3	mixture_coeff=2.5, unproven_coeff=7	16	31,82	14,00	28,12	55,87	68,57
a=0.7, b=0.3	mixture_coeff=2.5, unproven_coeff=9	16	31,96	14,42	28,40	56,52	69,93
a=0.6, b=0.4	mixture_coeff=2.5, unproven_coeff=7	16	31,96	14,38	28,28	55,62	68,57
a=0.6, b=0.4	mixture_coeff=2.5, unproven_coeff=9	16	32,54	14,48	28,69	60,07	72,50
a=0.5, b=0.5	mixture_coeff=2.5, unproven_coeff=7	16	32,33	14,18	28,48	60,33	**73,11**
a=0.5, b=0.5	mixture_coeff=2.5, unproven_coeff=9	16	31,90	14,14	28,27	55,56	70,32
a=0.7, b=0.3	mixture_coeff=2.5, unproven_coeff=7	32	31,97	14,21	28,23	51,14	65,77
a=0.7, b=0.3	mixture_coeff=2.5, unproven_coeff=9	32	31,83	14,00	28,05	57,25	68,34
a=0.6, b=0.4	mixture_coeff=2.5, unproven_coeff=7	32	31,82	14,21	28,14	58,96	60,78
a=0.6, b=0.4	mixture_coeff=2.5, unproven_coeff=9	32	32,08	14,09	28,34	52,47	65,67
a=0.5, b=0.5	mixture_coeff=2.5, unproven_coeff=7	32	32,07	14,33	28,32	59,18	71,91
a=0.5, b=0.5	mixture_coeff=2.5, unproven_coeff=9	32	31,79	14,13	28,29	49,99	61,82
a=0.7, b=0.3	mixture_coeff=2.5, unproven_coeff=7	64	32,55	14,54	28,60	**60,93**	72,51
a=0.7, b=0.3	mixture_coeff=2.5, unproven_coeff=9	64	**32,69**	**14,71**	**28,84**	49,08	62,63
a=0.6, b=0.4	mixture_coeff=2.5, unproven_coeff=7	64	31,97	14,28	28,30	44,73	57,52
a=0.6, b=0.4	mixture_coeff=2.5, unproven_coeff=9	64	31,98	14,19	28,33	57,78	72,52
a=0.5, b=0.5	mixture_coeff=2.5, unproven_coeff=7	64	31,78	13,95	28,01	59,22	72,20
a=0.5, b=0.5	mixture_coeff=2.5, unproven_coeff=9	64	31,63	13,99	27,89	53,21	66,03
a=0.7, b=0.3	mixture_coeff=2.5, unproven_coeff=7	128	31,97	14,21	28,23	51,14	65,77
a=0.7, b=0.3	mixture_coeff=2.5, unproven_coeff=9	128	31,83	14,00	28,05	57,25	68,34
a=0.6, b=0.4	mixture_coeff=2.5, unproven_coeff=7	128	31,82	14,21	28,14	48,42	60,78
a=0.6, b=0.4	mixture_coeff=2.5, unproven_coeff=9	128	32,08	14,09	28,34	52,47	65,67
a=0.5, b=0.5	mixture_coeff=2.5, unproven_coeff=7	128	31,79	14,13	28,29	49,99	61,82
a=0.5, b=0.5	mixture_coeff=2.5, unproven_coeff=9	128	32,55	14,54	28,60	**60,93**	72,51
a=0.7, b=0.3	mixture_coeff=2.5, unproven_coeff=7	256	32,07	14,33	28,32	59,18	71,91
a=0.7, b=0.3	mixture_coeff=2.5, unproven_coeff=9	256	**32,69**	**14,71**	**28,84**	49,08	62,63
a=0.6, b=0.4	mixture_coeff=2.5, unproven_coeff=7	256	31,97	14,28	28,30	44,73	57,52