SRLFactQA at Factify5WQA: Composite Claim-Evidence Consistency Aware Semantic Role Labelling based Question-Answering Entailment

Hariram Veeramani[1,*], Surendrabikram Thapa[2], Rajaraman Kanagasabai[3] and Usman Naseem[4]

[1]University of California Los Angeles, USA
[2]Virginia Tech, Blacksburg, Virginia, 24060, USA
[3]Institute for Infocomm Research, Agency for Science Technology and Research, Singapore
[4]School of Computing, Macquarie University, Sydney, NSW, 2113, Australia

Abstract

In today's digital age, the rampant spread of fake news and misinformation poses significant challenges. Fact-checking, a crucial tool in combatting misinformation, has benefited from advances in machine learning and NLP techniques. There is also a growing emphasis on making fact-checking models transparent and interpretable. **FACTIFY-5WQA** 2024, a shared task within DeFactify, is built upon a 5W framework (who, what, when, where, and why) for question-answer-based fact explainability. In the task, participants are tasked with developing fact-checking systems that can automatically process claims, use the associated evidence and questions to determine the veracity of the claims, and assign the correct labels. The task uses the **FACTIFY-5WQA** dataset, a semi-automatically generated fact verification dataset designed to determine the veracity of a claim based on the evidence / documents provided and the questions and answers generated from both the claim and the evidence. In this paper, we report our approach using a twin semantic frame-based fine-grained inconsistency detection model, achieving an accuracy of 45.51%, a significant improvement over the baseline securing the second position in the shared task.

Keywords

Automatic Fact Verification, 5W Framework, Interpretability, Semantic Role Labeling

1. Introduction

In an era dominated by digital communication, fake news, and misinformation have become omnipresent challenges with far-reaching consequences [1, 2]. The rapid dissemination of false or misleading information can profoundly impact individuals, communities, and societies at large [3, 4, 5]. Fact-checking, a valuable weapon in combating misinformation, involves the

De-Factify 3: 3rd Workshop on Multimodal Fact Checking and Hate Speech Detection, co-located with AAAI 2024. Vancouver, Canada

*Corresponding author.
hariram@ucla.edu (H. Veeramani); sbt@vt.edu (S. Thapa); kanagasa@i2r.a-star.edu.sg (R. Kanagasabai); usman.naseem@jcu.edu.au (U. Naseem)
https://linktr.ee/therealthapa (S. Thapa); https://sites.google.com/view/krajaraman/ (R. Kanagasabai); https://usmaann.github.io/ (U. Naseem)
0000-0003-4119-8239 (S. Thapa); 0000-0003-0191-7171 (U. Naseem)

systematic evaluation of claims, ensuring accuracy, and dispelling falsehoods. Advances in machine learning and natural language processing (NLP) techniques have helped the internet to become a safer, and civil place. These methods have improved the ability of fact-checking algorithms to understand and analyze complex language structures, improving accuracy in detecting misinformation [6, 7]. More recently, researchers have also explored multimodal approaches, combining textual information with analysis of images and videos, to verify stance [8, 9, 10], content moderation [11] and also claims that involve various media types, providing a more comprehensive fact-checking solution [12].

There is an increasing emphasis on making fact-checking models more transparent and interpretable. Understanding how AI systems arrive at their conclusions is crucial for building trust and avoiding human oversight. DeFactify is an initiative towards building an aspect-based (delineating which part(s) are true and which are false) explainable system that can assist human fact-checkers in asking relevant questions related to a fact, which can then be validated separately to reach a final verdict. FACTIFY-5WQA 2024 is a part of this shared task, which is based on a 5W framework (who, what, when, where, and why) for question-answer-based fact explainability [13]. It presents a semi-automatically generated dataset called FACTIFY-5WQA. The dataset is a fact verification dataset that aims to detect whether a claim is false or not based on a given evidence/document, and specifically from questions and answers generated respectively from the claim and evidence.

In this paper, we report our system for addressing this task. Our approach is inspired by the fine-grained inconsistency detection model proposed by Chan et al. [14]. Given a claim - evidence pair, we employ a twin semantic frame model to first identify fine-grained facts and associated Semantic Role Labels. Then, a custom fact attention module is designed to align facts within the claim that are relevant to the facts presented in the evidence. This is in turn used to predict the probability of different factual error types based on the representations of evidence and claim, and deduce the output labels: support, neutral, or refute. Through benchmarking experiments, we observe that the proposed approach outperforms standard baselines. In particular, our methodology achieves an accuracy of 45.51% which is over 32.88% relative increase from the baseline accuracy. Our system ranked second in this shared task.

2. Task Description

Participants in this task are tasked with developing fact-checking systems that can automatically process claims, use the associated evidence and questions to determine the veracity of the claims and assign the correct labels. Each claim-evidence pair is labeled as 'Support', 'Neutral', or 'Refute' to indicate whether the evidence supports, is neutral towards, or refutes the claim's veracity.

2.1. Evaluation

The prediction is considered correct if both the BLEU score for the answers generated from the claim and the BLEU score for the answers from the evidence are greater than a predefined threshold. Additionally, the label assigned to the prediction must match the true label (i.e., Support, Neutral, or Refute) for the claim. Only when these two conditions are met—satisfactory

BLEU scores and correct label assignment—is a prediction deemed accurate. The final accuracy is then calculated as the percentage of such accurate predictions out of the total predictions made. This evaluation approach ensures that predictions are both linguistically coherent and consistent with the true veracity of the claims.

3. Dataset

FACTIFY-5WQA is a dataset designed to address the limitations of contemporary fact-checking systems that often provide numerical truthfulness scores and hence may not be easily interpretable by humans [15]. The primary goal of **FACTIFY-5WQA** is to bridge the gap between automated fact checking and the human-centered logical fact checking process. **FACTIFY-5WQA**, is semi-automatically generated and consists of 391,041 facts, making it a substantial and comprehensive resource. Each fact is associated with a set of 5W QAs, focusing on the "who, what, when, where, and why" aspects of the fact. These QAs provide a detailed and multidimensional perspective on each claim. To generate these QAs, a semantic role labeling system is utilized to locate the relevant 5Ws in the facts. This process involves using a masked language model, which is a powerful tool for natural language understanding (NLU). The dataset thus includes not only the claims themselves, but also the associated questions that can guide the fact-checking process and highlight different aspects of each claim.

As shown in Figure 1, the key components of the dataset are claims, evidence/documents, questions, answers, and labels. To explain in detail, the key components of dataset are:

- **Claims:** The dataset contains various claims that need to be fact-checked.
- **Evidence/Documents:** Each claim is associated with supporting evidence or documents that can be used to verify the claim's veracity.
- **Questions:** For each claim, a set of questions is provided, covering different aspects of the claim. These questions are intended to guide the fact-checking process and help fact-checkers focus on specific details.
- **Answers:** The dataset includes expected answers to the questions for both the claim and the evidence. These answers serve as reference points for evaluating the accuracy of fact-checking systems.
- **Labels:** Each claim-evidence pair is labeled as 'Support', 'Neutral', or 'Refute' to indicate whether the evidence supports, is neutral towards, or refutes the claim's veracity.

4. Method

Our model, as shown in Figure 2, is inspired by the fine-grained inconsistency detection model proposed by Chan et al. [14]. We tailored the original twin architecture consisting of document and summary branches, to our task making improvements via a novel question claim/evidence attention mechanism. The proposed approach leverages five major components along with a novel training objective.

- "claim": "Andre Agassi won seven titles.",
  "evidence": "Andre Kirk Agassi born April 29 , 1970 -RRB- is an American retired professional tennis player and former World No. 1 who was one of the sport 's most dominant players from the early 1990s to the mid-2000s . Generally considered by critics and fellow players to be one of the greatest tennis players of all time ...,
  "question": ["How many titles did andre agassi win?", "Who won seven titles?"],
  "claim_answer": ["seven titles", "Andre Agassi"],
  "evidence_answer": ["eight-time Grand Slam champion", "Agassi"],
  "label": "Refute"
- "claim": "London police officer seriously injured in machete attack during vehicle stop. https://t.co/tnCa0MK6R9",
  "evidence": "By Julia Hollingsworth, CNNUpdated 0758 GMT (1558 HKT) August 8, 2019 (CNN)A London police officer is in a critical condition after a driver he pulled over attacked him with a machete. ",
  "question": ["How was a london police officer seriously injured?", "Who was seriously injured in a machete attack?", "When was the london police officer attacked?"],
  "claim_answer": [": in machete attack", "London police officer", "during vehicle stop"],
  "evidence_answer": ["a driver he pulled over attacked him with a machete", "A London police officer", "August 8, 2019"],
  "label": "Support"

Figure 1: Samples of dataset from FACTIFY-5WQA

4.1. Fact Extraction

In the Fact Extraction module, we extract facts from the input claim and evidence by utilizing a Longformer-based semantic role labeling (SRL) tool [16], which assigns semantic frames to tokens in the text. These semantic frames consist of predicates and their corresponding arguments.

4.2. Fact Encoder

The Fact Encoder module focuses on representing the tokens in both the input claim and evidence. We achieve this by utilizing Adapter-BERT [17] and applying max pooling to fuse hidden states across all layers. The goal is to create a comprehensive representation of facts by employing attentive pooling across tokens within semantic frames, as different tokens within a fact may contribute differently to its representation.

4.3. Question Claim/Evidence Attention

In our implementation, we propose a question claim/evidence attention module [18] that incorporates attention mechanisms to process questions and claims. By dynamically attending to relevant parts of the evidence context tokens [19], we aim to improve the fact verification process. This attention-based approach enhances the understanding of the context surrounding a claim, leading to more accurate verification outcomes.

4.4. Document Fact Attention Module

The Document Fact Attention module is designed to identify facts within the claim that are relevant to the facts presented in the evidence. It leverages attention mechanisms to align summary facts with document facts, ultimately producing a claim context vector that encapsulates information from the claim related to a given evidence. This module enhances the model's ability to connect and interpret facts across the claim and evidence.

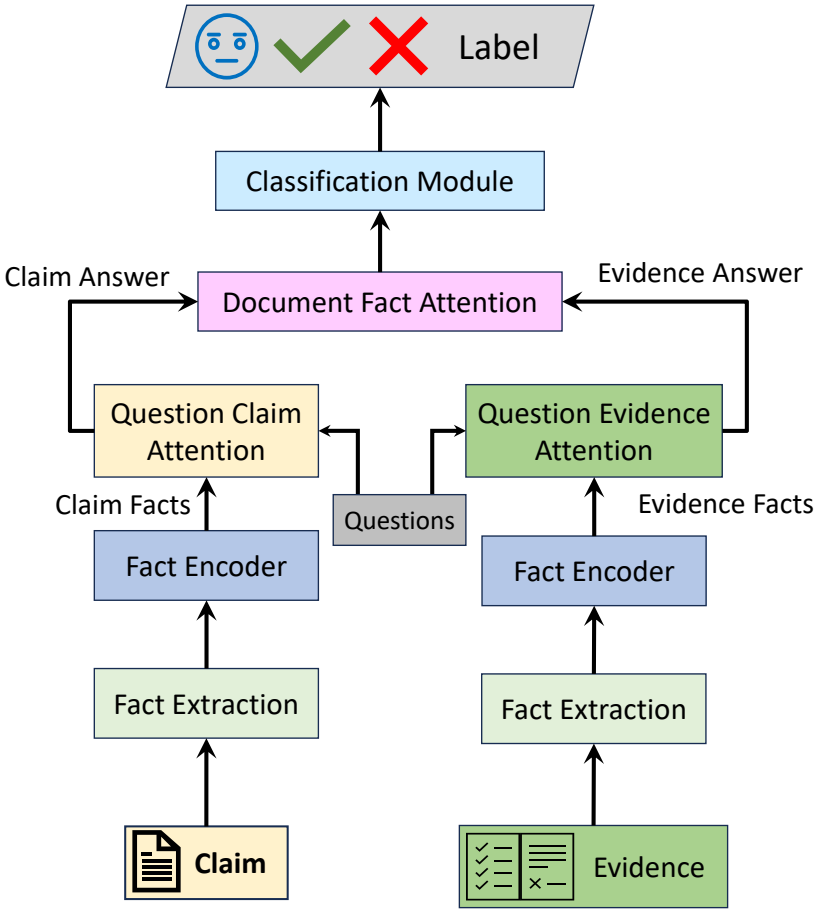

Figure 2: Proposed approach based on fine-grained inconsistency detection model [14].

4.5. Classification Module

In the Classification Module, we employ a linear classifier to predict the probability of different factual error types and question-answer triplets based on the representations of evidence answer and claim answer vectors obtained in response to the same question. By pooling the information from all evidence and claims relevant to the summary, we create fixed-size vectors that encapsulate the essential information. These vectors are then used as input for the linear classification layer, which predicts the probability of each factual error type/entailment [20]. This module facilitates the classification of errors and contributes to the overall accuracy of the fact verification process. In this module, we output labels corresponding to support, neutral, or refute.

4.6. Training Objective

We train our model with a joint objective of cross entropy (for classifier task) and SRLScore (for answer quality). SRLScore is a novel reference-free evaluation metric designed for text summarization tasks, utilizing Semantic Role Labels to assess factuality and interpretability of generated text [21]. It demonstrates competitive performance with state-of-the-art methods and exhibits stable generalization across datasets, offering an accessible and effective tool for text generation evaluation. The accuracy of our classifier in this context thus depends upon the quality of the answers provided. High-quality answers enhance our ability to predict entailment more effectively between the claim answer and the evidence answer.

5. Results

For this task, we implemented various baseline models along with the best-performing approach as discussed in section 4. As our first baseline model, we used AllenNLP-SRL[1] to use AllenNLP's pre-trained model [22] on Semantic Role Labeling (SRL) to make predictions. Similarly, as the second baseline, we use the twin architecture shown in Figure 2 without question claim/evidence attention blocks. With the architecture as shown in Figure 2, we were able to get an accuracy of 45.51%. The accuracy for the first baseline is 8.98% whereas the accuracy for second baseline is 37.91%. These results highlight the significant improvement achieved through the integration of the question claim/evidence attention layer. Apart from these baselines, the official baseline provided by the organizers is 34.22%. Our method thus attains a relative 32.99% increase in accuracy than the official baseline provided in FACTIFY-5WQA.

6. Conclusion

In this paper, we presented our participation in the FACTIFY-5WQA 2024 shared task, which focuses on developing interpretable and aspect-based fact-checking systems to combat misinformation. We introduced an innovative approach inspired by the interpretable fine-grained inconsistency detection model, specifically tailored for this task. Our model incorporates a question claim/evidence attention mechanism to dynamically attend to relevant evidence, enhancing the fact verification process. Through a combination of fact extraction, fact encoding, attention mechanisms, and a classification module, our approach achieved an impressive accuracy of 45.51%, and ranked second in the leaderboard. This accuracy surpassed both the official baseline (34.22%) and other baseline models, highlighting the efficacy of our proposed methodology. Further research could focus on refining the question claim/evidence attention mechanism and exploring techniques to improve the quality of answers, ultimately leading to even more accurate fact-checking systems.

[1] https://github.com/masrb/allenNLP-SRL

References

[1] R. Oshikawa, J. Qian, W. Y. Wang, A survey on natural language processing for fake news detection, in: Proceedings of the Twelfth Language Resources and Evaluation Conference, 2020, pp. 6086–6093.

[2] K. Pelrine, J. Danovitch, R. Rabbany, The surprising performance of simple baselines for misinformation detection, in: Proceedings of the Web Conference 2021, 2021, pp. 3432–3441.

[3] P. Patwa, M. Bhardwaj, V. Guptha, G. Kumari, S. Sharma, S. Pykl, A. Das, A. Ekbal, M. S. Akhtar, T. Chakraborty, Overview of constraint 2021 shared tasks: Detecting english covid-19 fake news and hindi hostile posts, in: Combating Online Hostile Posts in Regional Languages during Emergency Situation: First International Workshop, CONSTRAINT 2021, Collocated with AAAI 2021, Virtual Event, February 8, 2021, Revised Selected Papers 1, Springer, 2021, pp. 42–53.

[4] V. Qazvinian, E. Rosengren, D. Radev, Q. Mei, Rumor has it: Identifying misinformation in microblogs, in: Proceedings of the 2011 conference on empirical methods in natural language processing, 2011, pp. 1589–1599.

[5] K. Rauniyar, S. Poudel, S. Shiwakoti, S. Thapa, J. Rashid, J. Kim, M. Imran, U. Naseem, Multi-aspect annotation and analysis of nepali tweets on anti-establishment election discourse, IEEE Access (2023).

[6] Z. Guo, M. Schlichtkrull, A. Vlachos, A survey on automated fact-checking, Transactions of the Association for Computational Linguistics 10 (2022) 178–206.

[7] H. Veeramani, S. Thapa, U. Naseem, Knowtellconvince at araieval shared task: Disinformation and persuasion detection in arabic using similar and contrastive representation alignment, in: Proceedings of ArabicNLP 2023, 2023, pp. 519–524.

[8] K. Rajaraman, H. Veeramani, S. Rajamanickam, A. M. Westerski, J. J. Kim, Semantists at imagearg-2023: Exploring cross-modal contrastive and ensemble models for multimodal stance and persuasiveness classification, in: Proceedings of the 10th Workshop on Argument Mining, 2023, pp. 181–186.

[9] S. Thapa, K. Rauniyar, F. A. Jafri, S. Shiwakoti, H. Veeramani, R. Jain, G. S. Kohli, A. Hürriyetoğlu, U. Naseem, Stance and hate event detection in tweets related to climate activism - shared task at case 2024, in: Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE), 2024.

[10] S. Thapa, K. Rauniyar, F. A. Jafri, H. Veeramani, R. Jain, S. Jain, F. Vargas, A. Hürriyetoğlu, U. Naseem, Extended multimodal hate speech event detection during russia-ukraine crisis - shared task at case 2024, in: Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE), 2024.

[11] H. Veeramani, S. Thapa, R. Kanagasabai, U. Naseem, Unitetomoderate at dehate: The winning approach for segmentation-based content moderation with vision-text-mask modality fused large multimodal models, in: Proceedings of DeFactify 3.0: Third Workshop on Multimodal Fact-Checking and Hate Speech Detection, CEUR, 2024.

[12] M. Chakraborty, K. Pahwa, A. Rani, S. Chatterjee, D. Dalal, H. Dave, R. G, P. Gurumurthy, A. Mahor, S. Mukherjee, A. Pakala, I. Paul, J. Reddy, A. Sarkar, K. Sensharma, A. Chadha, A. Sheth, A. Das, FACTIFY3M: A benchmark for multimodal fact verification

with explainability through 5W question-answering, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 15282–15322. URL: https://aclanthology.org/2023.emnlp-main.945. doi:10.18653/v1/2023.emnlp-main.945.

[13] S. Suresh, A. Rani, P. Patwa, A. Reganti, V. Jain, A. Chadha, A. Das, A. Sheth, A. Ekbal, Overview of factify5wqa: Fact verification through 5w question-answering, in: Proceedings of DeFactify 3.0: Third Workshop on Multimodal Fact-Checking and Hate Speech Detection, CEUR, 2024.

[14] H. P. Chan, Q. Zeng, H. Ji, Interpretable automatic fine-grained inconsistency detection in text summarization, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 6433–6444. URL: https://aclanthology.org/2023.findings-acl.402. doi:10.18653/v1/2023.findings-acl.402.

[15] A. Rani, S. T. I. Tonmoy, D. Dalal, S. Gautam, M. Chakraborty, A. Chadha, A. Sheth, A. Das, FACTIFY-5WQA: 5W aspect-based fact verification through question answering, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 10421–10440. URL: https://aclanthology.org/2023.acl-long.581. doi:10.18653/v1/2023.acl-long.581.

[16] P. Shi, J. Lin, Simple bert models for relation extraction and semantic role labeling, arXiv preprint arXiv:1904.05255 (2019).

[17] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, S. Gelly, Parameter-efficient transfer learning for NLP, in: K. Chaudhuri, R. Salakhutdinov (Eds.), Proceedings of the 36th International Conference on Machine Learning, volume 97 of *Proceedings of Machine Learning Research*, PMLR, 2019, pp. 2790–2799. URL: https://proceedings.mlr.press/v97/houlsby19a.html.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[19] H. Veeramani, S. Thapa, U. Naseem, Temporally dynamic session-keyword aware sequential recommendation system, in: 2023 International Conference on Data Mining Workshops (ICDMW), 2023.

[20] H. Veeramani, S. Thapa, U. Naseem, Lowresourcenlu at blp: Enhancing sentiment classification and violence incitement detection through aggregated language models, in: Proceedings of the 1st International Workshop on Bangla Language Processing (BLP-2023), Singapore. Association for Computational Linguistics, 2023.

[21] J. Fan, D. Aumiller, M. Gertz, Evaluating factual consistency of texts with semantic role labeling, in: A. Palmer, J. Camacho-collados (Eds.), Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 89–100. URL: https://aclanthology.org/2023.starsem-1.9. doi:10.18653/v1/2023.starsem-1.9.

[22] M. Gardner, J. Grus, M. Neumann, O. Tafjord, P. Dasigi, N. F. Liu, M. Peters, M. Schmitz, L. Zettlemoyer, AllenNLP: A deep semantic natural language processing platform, in: E. L. Park, M. Hagiwara, D. Milajevs, L. Tan (Eds.), Proceedings of Workshop for NLP

Open Source Software (NLP-OSS), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 1–6. URL: https://aclanthology.org/W18-2501. doi:10.18653/v1/W18-2501.

A. Limitations

Despite the effectiveness of our approach in the FACTIFY-5WQA 2024 shared task, it faces several challenges, including a heavy reliance on the accuracy of semantic role labeling (SRL) for fact extraction, which can impact claim verification accuracy if errors occur. The model's computational intensity may limit its real-time application and use on lower-resource devices. Its generalizability across different domains, languages, and misinformation types remains untested, which could affect its applicability in specialized fields. Additionally, while aiming for interpretability, the model's complex neural network components still present a "black box" challenge, complicating the provision of clear explanations to non-expert users. The semi-automatically generated FACTIFY-5WQA dataset may introduce biases that influence model training and performance. The model may also struggle with complex claims and evidence that require nuanced understanding or implicit knowledge. Lastly, the current framework does not address scalability or updating mechanisms for incorporating new information over time, which is crucial for adapting to evolving misinformation landscapes. Addressing these limitations requires further research into more sophisticated NLP techniques, computational efficiency improvements, domain adaptation, enhanced interpretability, bias mitigation, and the development of scalable, update-friendly architectures.