Evaluating Performance of Various Fact Checking Methods

Arvind Sreenivas
Sagi Shaier
Katharina Kann

University of Colorado, Boulder

Abstract

In this paper, we evaluate the performance of various claim verification models on a newly created dataset of medical articles from PubMed and their corresponding references. We compared MultiVers, the Fake News Challenge model, and sciKGAT, which represent a range of approaches to factuality evaluation, from structured knowledge graphs to neural networks for evidence aggregation and reasoning. Our evaluation metrics included precision, recall, and F1-score. Our results indicate that the Fake News Challenge model outperformed the others, achieving the highest F1 score and accuracy. However, the MultiVers and sciKGAT models, trained on abstract-based data, were affected by truncation due to large token sizes. We discuss the limitations of our study and propose directions for future research to improve claim verification in the medical domain.

1 Introduction

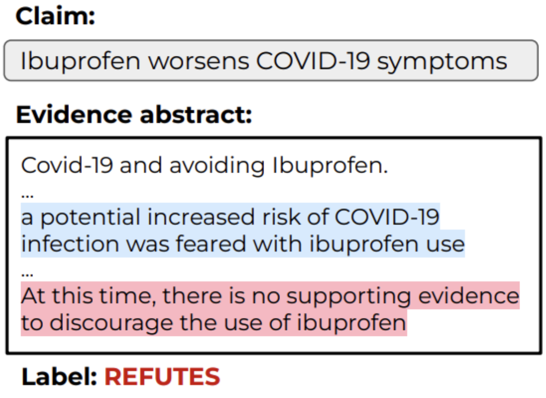
The accuracy and factuality of text generated by natural language processing (NLP) models have important implications for their utility in various domains, including medicine. Medical text generation has potential applications in areas such as clinical decision-making and patient education. Ensuring the accuracy of the information presented is critical for maintaining patient safety and achieving optimal healthcare outcomes.

Several methods for factuality evaluation in medical text have been proposed in recent years. These methods vary in their approach, from rule-based to machine learning-based techniques. Despite the availability of these methods, there is still a need to evaluate their effectiveness and identify areas for improvement.

In this paper, we evaluate a few existing methods for factuality evaluation in medical text, namely: MultiVers, and Fine-grained Fact Verification with Kernel Graph Attention Network (KGAT). These methods represent a range of approaches to factuality evaluation, from using structured knowledge graphs to leveraging neural networks for evidence aggregation and reasoning.

To evaluate the effectiveness of these methods, we created a new dataset of medical articles and their references. We annotated the dataset with factuality labels and used it to evaluate the performance of each method. Our evaluation metrics include precision, recall, and F1-score.

The remainder of the paper is organized as follows. Section 2 reviews the literature on factuality evaluation in medical texts and describes the four methods we evaluate. In Section 3, we describe our dataset and the evaluation metrics used. In Section 4, we present our experiments' results and analyze each method's performance. In Sections 5 and 6, we discuss the implications of our findings and directions for future research, and we conclude the paper.



Claim:

Ibuprofen worsens COVID-19 symptoms

Evidence abstract:

Covid-19 and avoiding Ibuprofen.
...
a potential increased risk of COVID-19 infection was feared with ibuprofen use
...
At this time, there is no supporting evidence to discourage the use of ibuprofen

Label: REFUTES

Figure 1: Sample input used in MultiVers from (Wadden et al., 2022)

2 Related Work

(Wadden et al., 2022) propose a model, MultiVers for full-context claim verification that predicts whether a claim is true or false based on the entire context of the claim and abstract. The model uses the Longformer as its encoder and assigns global attention to the <s> token, as well as all tokens in the claim and all </s> tokens. It predicts the fact-checking label directly based on an encoding of the entire claim and abstract, and enforces consistency of rationales with the predicted label during decoding.

(Riedel et al., 2017) use a system for stance detection that is an end-to-end approach that utilizes a MLP with one hidden layer to solve the Fake News Challenge. For text input, the system uses two basic BOW representations - TF and TF-IDF. The features extracted from the headline and body pairs include the TF vector of the headline, the TF vector of the body, and the cosine similarity between the 2-normalised TF-IDF vectors of the headline and body. The classifier used is also a MLP with one hidden layer consisting of 100 units, and a softmax on the output of the final linear layer. The system predicts the highest scoring label ('agree', 'disagree', 'discuss', or 'unrelated').

(Liu et al., 2019) used the FEVER dataset, which contains annotated claims with Wikipedia documents, to evaluate their method. They used two evaluation metrics: Label Accuracy (LA) and FEVER score, which considers the quality of the retrieved evidence. For document retrieval, they used a constituency parser in AllenNLP to extract phrases indicating entities and then used these phrases as queries to find relevant Wikipedia pages. For claim verification The training and development sets are built with golden evidence and higher ranked evidence with sentence retrieval. All claims are assigned with five pieces of evidence using RoBERTa and are evaluated. The authors also took part in the SCIFACT shared task using their KGAT model, this implementation was based on RoBERTa and additionally had an abstract retrieval, rerank and rationale selection stages. Previous works in fact verification have primarily utilized either the FEVER dataset consisting of Wikipedia documents, the SCIFACT dataset with biomedical data, or datasets from News articles. In contrast, our study introduces a new dataset based on Pubmed articles and their cited references supported by their full texts, allowing for a more comprehensive evaluation of fact verification models. We will employ similar methods and models used in previous works to assess their performance on our dataset.

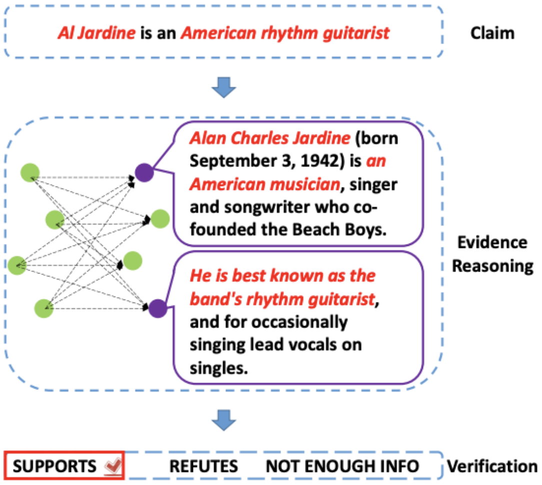

Al Jardine is an American rhythm guitarist
Claim
Alan Charles Jardine (born September 3, 1942) is an American musician, singer and songwriter who co-founded the Beach Boys.
He is best known as the band's rhythm guitarist, and for occasionally singing lead vocals on singles.
Evidence Reasoning
SUPPORTS
REFUTES
NOT ENOUGH INFO
Verification

Figure 2: Sample input format used in KGAT from (Liu et al., 2019)

3 Dataset Creation

[1] The dataset was generated through a three-step process using articles from PubMed and their OpenAccess API.

3.1 Citation Extraction

To create our dataset, we utilized the vast collection of articles available through PubMed's OpenAccess license. We extracted statements from these articles that contained a citation at the end of the sentence, as opposed to statements with citations in the middle or with multiple citations. This approach increased the likelihood that each statement pertained specifically to the cited material, which resulted in a higher quality of claims in our dataset. For example, we selected a statement such as "As demonstrated in artificial membrane systems, cholesterol facilitates the formation of sphingolipid-containing microdomains [31]." while excluding a statement such as "Some of these mutations occurred in the receptor binding sites of HA genes [7], [8] and in other segments of the virus [9]."

3.2 Full Text Gathering

To gather the full texts of the articles we used the PubMed API along with the PubMed ID which provided access to open access articles. This was then stored as claim:evidence pairs.

Table 1: Prediction Evaluation of Three Models

	Support			**Contradict / Not Related**			**Accuracy**
Model	**Precision**	**Recall**	**F1 Score**	**Precision**	**Recall**	**F1 Score**	**Overall**
MultiVers	0.51	0.17	0.26	0.49	0.83	0.62	0.50
FNC	0.66	0.44	0.53	0.84	0.93	0.88	0.81
sciKGAT	0.51	0.92	0.66	0.54	0.10	0.17	0.52

3.3 Data Formatting

The dataset was divided into two categories, positive and negative citations. The positive citations consisted of claim:evidence pairs where the evidence was the full text referred to by the citation. In contrast, the negative citation pairs were composed of claim:evidence pairs where the evidence was not entirely related to the citation. To generate negative citations, we split the overall data in half, randomly shuffled one half, and used the BM-25 ranking algorithm to select the other half based on a 70% similarity threshold with the citations. The input formatting for each model varied and was specified in their respective GitHub repositories. The data needed to be provided in claims and corpus format, with the full texts in the corpus and the citations, along with the article ID they cite, in the claims.

4 Evaluation Methods

4.1 MultiVers

The SCIFACT task utilizes four evaluation metrics for the SCIFACT data. However, (Wadden et al., 2022) uses only two of them for their experiments' significant findings. We use one of those methods for our results. The abstract-level label-only evaluation measures the model's F1 score in identifying abstracts that SUPPORT or REFUTE each claim. It is sufficient for the models to predict the correct label y(c, a), and rationales are not required. From (Wadden et al., 2022), given a claim c and a collection of candidate abstracts which may contain evidence relevant to c, the scientific claim verification task requires a system to predict a label y(c, a) ∈ {SUPPORTS, REFUTES, NEI1}, which indicates the relationship between c and a for each candidate a.

4.2 Fake News Challenge

The evaluation method for the Fake News Challenge models is a two-level scoring system that includes a weighted scoring system. The first level, which accounts for 25% of the score weighting, involves classifying the headline and body text as either related or unrelated. While this task is considered easier and less relevant to detecting fake news, it is still necessary to evaluate the models' ability to establish the relationship between the headline and body text.

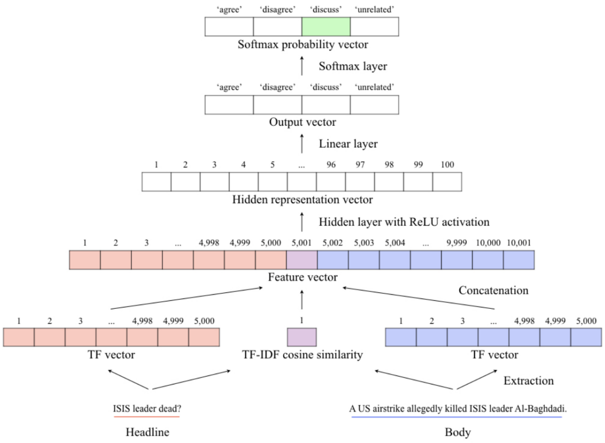


Figure 3: System architecture used in FNC (Riedel et al., 2017)

The second level, which accounts for 75% of the score weighting, involves classifying related pairs as agrees, disagrees, or discusses. This task is more difficult and more relevant to detecting fake news, and therefore is given much more weight in the evaluation metric. The models' ability to accurately classify related pairs as agrees, disagrees, or discusses is considered a critical component of their ability to detect fake news.

We use their output of 'agrees', 'disagrees' and 'discuss' and calculate the accuracy, F1 scores based on our gold labels.

4.3 sciKGAT

In their work on claim verification, (Liu et al., 2019) used Label Accuracy (LA) and FEVER Score as the primary evaluation metrics to assess the system's ability to classify claims and retrieve relevant evidence.

[1] https://github.com/sreenivasarvind/factualityeval

LA calculates the accuracy rate of claim classification without considering retrieved evidence. This metric provides a broad overview of the system's performance.

FEVER Score, on the other hand, considers whether the system has provided a complete set of golden evidence and better reflects the system's inference ability. The system's ability to retrieve relevant evidence and support its classification decision is evaluated using this metric.

To evaluate the evidence sentence retrieval accuracy, the authors used Precision, Recall, and F1 metrics. These metrics were used to assess the system's ability to retrieve relevant evidence sentences, based on the provided sentence-level labels. The sentence-level labels specified whether a sentence is evidence or not to verify the claim. In our work, we also used their model, sciKGAT, which they used for the SCIFACT shared task, and generated predictions using it for the LA metric.

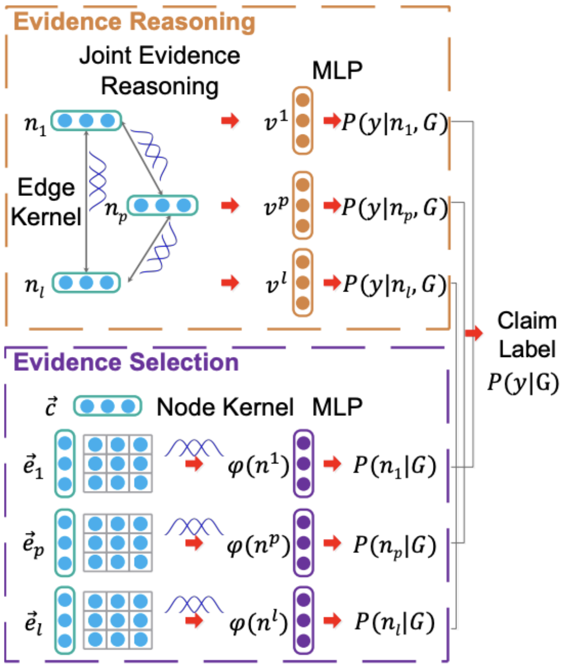


Figure 4: System architecture used in KGAT (Liu et al., 2019)

5 Results

To evaluate the performance of the claim verification models, we used a test dataset consisting of 1,347 positive and negative data-points. We generated predictions using the models proposed in previous works, and the results of each model are presented in Table 1. Overall, the model that performed the best on our test dataset was the Fake News Challenge model, which used a Bag of Words and TF-IDF based method to generate predictions. This model achieved the highest F1 score and accuracy among the three models evaluated. On the other hand, the MultiVers and sciKGAT models under-performed with low F1 scores and accuracy.

6 Discussion

It is important to note that the MultiVers and sciKGAT models were trained on abstract-based data, which is significantly smaller in size than full-text data. As a result, a large portion of the data was truncated due to large token sizes, leading to poor results for both models. The original MultiVers model tested on the SCIFACT dataset achieved an abstract-level F1 score of 67.6, while the sciKGAT model achieved an F1 score of 72.4. These findings suggest that addressing the truncation issue could potentially improve the prediction results.

Additionally, it is important to consider the limitations of the "Contradict/Not Related" scores reported in this study. The negative data generated for this study does not necessarily contradict the claims made, as it was not curated in this manner. Furthermore, the inclusion of the "discuss" and "not related" tags may add ambiguity to the classification process, as they were classified under the same category as contradict. These limitations should be taken into account when interpreting the results of this study.

7 Limitations & Future work

Moving forward, the next steps in this study will involve training each model using our train data to evaluate their performance on our specific dataset. To overcome the truncation issue, we plan to explore various methods, such as embedding the splices of the data and joining them during training or splitting the input and training it on the same claim separately.

Furthermore, additional models will be compared and evaluated, and the size of the dataset will be increased over time to improve training performance. We also intend to generate negative data more carefully, ensuring that it actually contradicts the statements or is not related. We will consider adding an additional class to signify no meaningful prediction at all to improve the evaluation of the class predictions.

Overall, the next steps of this study aim to refine the models and improve their performance in the context of claim verification. With continued investigation and experimentation, it may be possible to develop even more accurate models and techniques for verifying claims in this domain.

8 Conclusion

Our study evaluated the performance of three claim verification models on a new dataset of medical articles from PubMed and their corresponding references. The results revealed that the Fake News Challenge model performed the best, while the MultiVers and sciKGAT models underperformed due to truncation issues with abstract-based data. This study highlights the need for further research to improve the performance of claim verification models in the medical domain. Future work should address truncation issues, refine the models, and explore alternative techniques to enhance their performance. Additionally, the generation of negative data should be improved, and the size of the dataset should be increased to provide better training. With continued investigation and experimentation, we believe it is possible to develop more accurate models and techniques for claim verification in the medical domain, ultimately benefiting clinical decision-making and patient education.

References

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *CoRR*, abs/2005.11401.

Zhenghao Liu, Chenyan Xiong, and Maosong Sun. 2019. Kernel graph attention network for fact verification. *CoRR*, abs/1910.09796.

Benjamin Riedel, Isabelle Augenstein, Georgios P. Spithourakis, and Sebastian Riedel. 2017. A simple but tough-to-beat baseline for the fake news challenge stance detection task. *CoRR*, abs/1707.03264.

David Wadden, Kyle Lo, Lucy Lu Wang, Arman Cohan, Iz Beltagy, and Hannaneh Hajishirzi. 2022. MultiVerS: Improving scientific claim verification with weak supervision and full-document context. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 61–76, Seattle, United States. Association for Computational Linguistics.

Jie Zhou, Xu Han, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2019. GEAR: Graph-based evidence aggregating and reasoning for fact verification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 892–901, Florence, Italy. Association for Computational Linguistics.