Bridging Textual and Tabular Worlds for Fact Verification: A Lightweight, Attention-Based Model

Shirin Dabbaghi[1], Canasai Kruengkrai[2], Ramin Yahyapour[1], Junichi Yamagishi[2]
[1]Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG), Germany
[2]National Institute of Informatics (NII), Japan
sdabbag@gwdg.de, canasai@gmail.com, ramin.yahyapour@gwdg.de, jyamagis@nii.ac.jp

Abstract

FEVEROUS is a benchmark and research initiative focused on fact extraction and verification tasks involving unstructured text and structured tabular data. In FEVEROUS, existing works often rely on extensive preprocessing and utilize rule-based transformations of data, leading to potential context loss or misleading encodings. This paper introduces a simple yet powerful model that nullifies the need for modality conversion, thereby preserving the original evidence's context. By leveraging pre-trained models on diverse text and tabular datasets and by incorporating a lightweight attention-based mechanism, our approach efficiently exploits latent connections between different data types, thereby yielding comprehensive and reliable verdict predictions. The model's modular structure adeptly manages multi-modal information, ensuring the integrity and authenticity of the original evidence are uncompromised. Comparative analyses reveal that our approach exhibits competitive performance, aligning itself closely with top-tier models on the FEVEROUS benchmark.

Keywords: fact-checking, fact verification, natural language processing, tabular data, FEVEROUS

1. Introduction

The proliferation of the Internet and social media has led to an increase in potentially and deliberately misleading and inaccurate statements. With minimal constraints on sharing information, anyone can now effortlessly spread erroneous or biased statements to a wide and diverse audience (Saeed et al., 2021). Fact-checking has emerged as a crucial task to assess the veracity of claims and assertions, ensuring the dissemination of accurate and reliable information (Bouziane et al., 2021; Trokhymovych and Trumper, 2020; Sathe and Park, 2021).

Several benchmark datasets have been developed to evaluate the performance of fact-checking systems. While some focus solely on text, such as Fact Extraction and VERification (FEVER) (Thorne et al., 2018), others center on tabular data, like TabFact and INFOTABS (Chen et al., 2020; Gupta et al., 2020). Nevertheless, there is a pressing need for more comprehensive and inclusive benchmarks that reflect real-world fact-checking scenarios. Addressing this concern, Aly et al. (2021) introduced FEVER Over Unstructured and Structured information (FEVEROUS). FEVEROUS includes textual and tabular data to provide a more holistic representation of fact-checking tasks. In this benchmark, models are challenged to extract relevant evidence sentences or table cells from millions of unstructured passages and to incorporate this multi-modal information effectively to verify a given claim.

Previous works on FEVEROUS have often addressed the fact verification challenge by converting all pieces of evidence into a unified format, either plain text or several tables. However, such format conversions have downsides, leading to a loss of rich context information from the original evidence and often misleading information encoding (Hu et al., 2022).

Inspired by the work of Kruengkrai et al. (2021), we herein propose a straightforward model that aims to eliminate the need for extensive preprocessing or rule-based requirements to convert between textual and tabular data formats. Designed with a simple modular structure, our model leverages the strengths of existing models pre-trained on tabular and text datasets and even fine-tuned for different tasks to obtain the contextual embedding of each data type. Additionally, we utilize a lightweight attention-based mechanism to explore and capitalize on the latent connections between data types. This facilitates a more comprehensive and reliable prediction of claim veracity, utilizing the strengths of both data modalities without compromising the integrity of the original evidence. A comparative analysis against existing methods reveals that our model performs competitively, achieving results nearly equivalent to the top-tier models evaluated on the FEVEROUS benchmark.

2. Related Work

2.1. FEVER and FEVEROUS

The task of fact-checking has gained prominence as an essential process for validating the veracity of claims and for promoting the circulation of reliable information (Bouziane et al., 2021; Trokhy-

LREC-COLING 2024, pages 2515–2519

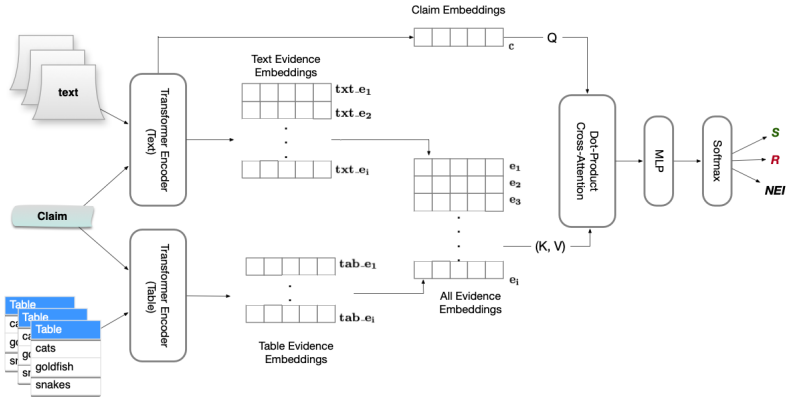

Figure 1: The proposed model.

movych and Trumper, 2020; Sathe and Park, 2021). A series of benchmark datasets and tasks have been developed to facilitate the advancement of automatic fact-checking algorithms. While each of these benchmarks has substantially contributed to the development of fact-checking models, they are often limited by the type of data they present: text-only, as in FEVER (Thorne et al., 2018), or table-only, as in TabFact, INFOTABS, and SEM-TABFACTS (Chen et al., 2020; Gupta et al., 2020; Wang et al., 2021). These singular foci fail to encapsulate the diverse and multi-modal nature of real-world fact-checking scenarios. This has led to the introduction of FEVEROUS, a benchmark that combines both textual and tabular data, presenting a more nuanced challenge for fact-checking algorithms (Aly et al., 2021).

Fact verification tasks often include two steps: evidence extraction and verdict prediction. In this research, we focus on the latter.

In the FEVEROUS benchmark, previous works have generally relied on transforming the multi-modal evidence—text and tables—into a unified text format (Aly et al., 2021; Malon, 2021) or into a unified table format (Bouziane et al., 2021). While this simplifies the problem to some extent, the process often results in the loss of context or nuance that is critical for accurate verdict prediction. Such an approach might limit the model's performance due to misleading information encoding. Hu et al. (2022) introduced a dual-channel unified format fact verification model (DCUF) in which data conversion happens in both directions simultaneously. Despite achieving a notable performance in the FEVEROUS task, their solution introduces new challenges, such as intensified preprocessing requirements, potential data redundancies, and increased computational demand.

2.2. Tabular Encoders

This subsection provides an overview of the three advanced neural tabular encoders — TAPAS, Tapex, and Pasta — we used in this research for proper understanding and handling of structured tabular data.

TAPAS (Herzig et al., 2020) is a weakly supervised model for parsing and extracting information from tabular data. Unlike typical question-answering systems, TAPAS does not require manually annotated logical forms for training. It extends bidirectional encoder representations from transformers (BERT) with additional structure-aware positional embeddings to represent tables, and it is suitable for tasks like table-based question answering and fact verification, where structured data are prevalent. Tapex (Liu et al., 2022) is a model proposed for table pre-training. It demonstrates that table pre-training can be achieved by learning a neural SQL executor over a synthetic corpus. Tapex focuses on training neural networks to understand and query tables effectively without needing real data. However, Pasta (Gu et al., 2022) focuses on numerical and statistical data within tables. Pre-trained on diverse tabular datasets, Pasta can reason using statistical operations, aggregations, and numerical relationships in tables. It is ideal for tasks involving numerical computations, value prediction, and verifying claims based on statistical data.

3. Method

Simplicity was one of our primary objectives in developing our model. Adopting a modular approach, we designed the model as a collection of independent components, each responsible for handling specific aspects of the process. This modular approach simplifies the model architecture and enables seamless adaptation to different fact-checking scenarios and datasets. Each module can be independently fine-tuned or replaced, making incorporating improvements or domain-specific enhancements easy without overhauling the entire system.

For the rest of this section, we first formalize the problem domain we aim to address, followed by providing a detailed explanation of the model components and workflow.

Given a claim c and sets of textual evidence

Models	Development		Test	
	FEVEROUS score	Label accuracy [%]	FEVEROUS score	Label accuracy [%]
Official Baseline (Aly et al., 2021)	19	53	17.73	48.48
EURECOM (Saeed et al., 2021)	19	53	20.01	47.79
Z team	-	-	22.51	49.01
CARE (Kotonya et al., 2021)	26	63	23	53
NCU (Gi et al., 2021)	29	60	25.14	52.29
Papelo	28	66	25.92	57.57
FaBULOUS (Bouziane et al., 2021)	30	65	27.01	56.07
DCUF (Hu et al., 2022)	35.77	72.91	33.97	63.21
Our Model	34.94	71.86	32.81	62.02

Table 1: Model performance on the development set and test set.

$TextEvids = \{txt\_e_i\}_{i=1}^{M}$ (with M being the number of pieces of text evidence) and tabular evidence $TabEvids = \{tab\_e_i\}_{i=1}^{N}$ (with N being the number of pieces of table evidence), the task is to determine the veracity of c. Each claim is classified into one of three categories: Supported (S), Refuted (R), or Not Enough Information (NEI).

Figure 1 depicts an abstract overview of our solution. The model is based on a dual transformer architecture, including text and table transformers, to obtain the evidence embeddings in their original format for the given claim. For notation, we refer to these modules as $\mathrm{BERT\_text}$ and $\mathrm{BERT\_tab}$. The next section shows our solution's performance considering different transformer-based models trained for textual and tabular datasets.

For a pair of a claim c and txt_e_i, we expect the outputted classification (CLS) token of BERT_text to summarize the key information from txt_e_i in the context of c:

$$\mathbf{txt\_e_i} = \mathrm{BERT\_text}_{\mathrm{CLS}}(c, txt\_e_i) \quad (1)$$

Likewise, for tabular evidence tab_e_i:

$$\mathbf{tab\_e_i} = \mathrm{BERT\_tab}_{\mathrm{CLS}}(c, tab\_e_i) \quad (2)$$

At the core of the model is a cross-attention module tasked with establishing the correlations between the textual and tabular modalities, thereby enabling the seamless fusion of their embeddings. This module aims to acquire an enriched representation of both types of evidence for the given claim, permitting a holistic and context-aware inference.

The cross-attention module is based on the scaled dot-product attention mechanism introduced by Vaswani et al. (2017). It is calculated as follows:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V, \quad (3)$$

where d_k represents the dimensionality of the keys and values used in the scaled dot-product attention mechanism. It is typically set to control the scale of the dot product. Here, Q represents the query, which is the claim representation $\mathbf{c}$, while K and V represent the keys and values constituted by the evidence set $E \in \mathbb{R}^{(M+N)\times D}$.

$$\mathbf{c} = \mathrm{BERT\_text}_{\mathrm{CLS}}(c) \quad (4)$$

$$E = \{\mathbf{e_i}\}_{i=1}^{M+N} = \{\mathbf{txt\_e_i}\}_{i=1}^{M} \oplus \{\mathbf{tab\_e_i}\}_{i=1}^{N} \quad (5)$$

The output of the cross-attention module is a weighted sum of the evidence embeddings, creating a context-rich representation of the claim.

The final stage of our model involves a classification task performed by a multi-layer perceptron (MLP) followed by a softmax layer. Let the output from the cross-attention module be denoted as $\mathbf{z} = Attention(\mathbf{c}, E, E)$. The probability distribution, P, over the three classes $[S, R, NEI]$ is then calculated as:

$$P(y|\mathbf{c}, E) = \mathrm{softmax}(\mathrm{MLP}(\mathbf{z})) \quad (6)$$

Given a ground truth label $\mathbf{y}^*$ and the predicted probability distribution $P(y|\mathbf{c}, E)$, we use the cross-entropy loss function to optimize the model:

$$\mathcal{L} = -\sum_i \mathbf{y_i^*} \log P(y_i|\mathbf{c}, E) \quad (7)$$

4. Experiments

4.1. Model Settings and Configurations

In this research, we focused on the fact verification task and relied on the retrieved evidence method presented in the work by (Hu et al., 2022)[1]. For each given claim, we considered a set of seven pieces of evidence, out of which five were in text format and the other two were tabular. Using the PyTorch framework, our model was trained on a machine with two NVIDIA Tesla A100 GPUs. We opted for the Adam optimizer with a learning rate of $1e-5$. The training was carried out for six epochs with an early stopping criterion based on the validation loss. Typically, training converged within two to three epochs, taking approximately two hours. Furthermore, we utilized an MLP comprising three layers, each followed by a ReLU activation function,

[1]The code and model checkpoints are available at: https://github.com/nii-yamagishilab/MLA-FEVEROUS-COLING24

and included a dropout layer with a rate of 0.2 for regularization.

Our base model for the $\mathrm{BERT\_text}$ was a large DeBERTa model[2] fine-tuned on multiple natural language inference (NLI) datasets. As for $\mathrm{BERT\_tab}$, we used the TAPAS large, fine-tuned specifically on the TabFact dataset[3]. Furthermore, we configured the cross-attention layer with 16 attention heads and a hidden size of 1024.

The experiments in this research were evaluated using a standard train-test split of the FEVEROUS dataset, in line with the evaluation metrics established for the claim verification task in this benchmark: FEVEROUS score and label accuracy in percentage.

4.2. Results and Analysis

Table 1 presents a comparative evaluation of our model against other fact-checking approaches on the FEVEROUS benchmark. Our solution achieved a FEVEROUS score of 34.94% and a label accuracy of 71.86%, comparable with the highest submitted model (Hu et al., 2022) in the FEVEROUS benchmark.

We further studied our solution using multiple transformer models fine-tuned on different datasets and tasks to show the effectiveness of the implementation design. Therefore, we evaluated the base $BERT\_tab$ (TAPAS), as well as the Tapex[4] (400M parameters) fine-tuned on the Tabfact dataset and the Pasta large[5] model with around 430M trainable parameters, which we fine-tuned on the FEVEROUS dataset. For the text encoders, we also tested RoBERTa[6] and BART[7], each with a size of 350M and 400M trainable parameters, respectively.

The results, presented in Table 2, showcase the consistency of the overall performance in different models. Although some differences emerged, our solution structure enables any combination of text and table encoders to be used for fact-checking. In other words, in this experiment, we aimed to demonstrate just the model's versatility and robustness. As such, we consciously refrained from utilizing additional optimizations like hyperparameter tuning, ensuring that the observed performance metrics reflect the model's inherent capabilities.

Table 3 shows how different data types contribute to our model's performance. We evaluated both text and tabular evidence separately and in combination. When the model was trained on text-only evidence, it yielded a FEVEROUS score of 33.53% and a label accuracy of 70.13%. In contrast, after using table-only evidence, the scores were slightly diminished. However, both modalities propelled the model to its peak performance. This collaborative effect underscores the significance of our model's ability to seamlessly integrate textual and tabular evidence, leveraging the full potential of multi-modal data.

Models	FEVEROUS score	Label accuracy [%]
TAPAS + DeBERTa	34.94	71.86
TAPAS + RoBERTa	34.13	70.31
TAPAS + BART	33.85	69.75
Tapex + DeBERTa	33.00	69.39
Tapex + RoBERTa	32.23	67.27
Tapex + BART	31.28	65.38
Pasta + DeBERTa	34.22	70.70
Pasta + RoBERTa	31.60	66.98
Pasta + BART	34.53	70.22

Table 2: Model performance with different pre-trained/fine-tuned models on tabular and textual datasets.

Data Type	Feverous score	Label accuracy [%]
Only Text	33.53	70.13
Only Table	32.00	67.04
Text + Table	34.94	71.86

Table 3: The importance of different data types to the model performance.

5. Conclusion

The FEVEROUS task comprises two distinct sub-tasks: evidence retrieval and verdict prediction. Researchers engaging with the FEVEROUS benchmark have the flexibility to choose between these sub-tasks, allowing them to use and evaluate separate independent models when contributing to the main assignment. In this paper, we introduced a modular, attention-based model for the verdict prediction sub-task.

The proposed model aims to effectively integrate both textual and tabular data and eliminate the need for cumbersome modality conversions, often resulting in the loss of crucial context and nuance. By leveraging pre-trained models for each data type and by utilizing a lightweight cross-attention mechanism, we could effectively exploit the latent relationships between text and table data. Compared with other approaches within the FEVEROUS benchmark, our model showcases a competitive performance, underscoring its potential for accurate and versatile fact verification.

Acknowledgement: This work was conducted during the first author's internship at NII, Japan. This work is supported by JST CREST Grants (JPMJCR18A6 and JPMJCR20D3) and MEXT KAKENHI Grants (21H04906), Japan.

[2] DeBERTa-v3-large-mnli-fever-anli-ling-wanli
[3] tapas-large-finetuned-tabfact
[4] tapex-large-finetuned-tabfact
[5] PASTA
[6] roberta-large-snli_mnli_fever_anli_R1_R2_R3-nli
[7] bart-large-snli_mnli_fever_anli_R1_R2_R3-nli

6. References

Rami Aly, Zhijiang Guo, Michael Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. The Fact Extraction and VERification Over Unstructured and Structured information (FEVEROUS) Shared Task. *FEVER 2021 - Fact Extr. Verif. Proc. 4th Work.*, (NeurIPS 2021):1–13.

Mostafa Bouziane, Hugo Perrin, Amine Sadeq, Thanh Nguyen, Aurélien Cluzeau, and Julien Mardas. 2021. FaBULOUS: fact-checking based on understanding of language over unstructured and utructured information. *FEVER 2021 - Fact Extr. Verif. Proc. 4th Work.*, pages 31–39.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2020. Tabfact: a Large-Scale Dataset for Table-Based Fact Verification. *8th Int. Conf. Learn. Represent. ICLR 2020*, pages 1–14.

In Zu Gi, Ting Yu Fang, and Richard Tzong Han Tsai. 2021. Verdict Inference with Claim and Retrieved Elements Using RoBERTa. *FEVER 2021 - Fact Extr. Verif. Proc. 4th Work.*, pages 60–65.

Zihui Gu, Ju Fan, Nan Tang, Preslav Nakov, Xiaoman Zhao, and Xiaoyong Du. 2022. PASTA: Table-Operations Aware Fact Verification via Sentence-Table Cloze Pre-training. *Proc. 2022 Conf. Empir. Methods Nat. Lang. Process. EMNLP 2022*, pages 4971–4983.

Vivek Gupta, Maitrey Mehta, Pegah Nokhiz, and Vivek Srikumar. 2020. INFOTABS: Inference on tables as semi-structured data. In *Proc. Annu. Meet. Assoc. Comput. Linguist.*, pages 2309–2324.

Jonathan Herzig, Paweł Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Martin Eisenschlos. 2020. TAPAS: Weakly supervised table parsing via pre-training. *Proc. Annu. Meet. Assoc. Comput. Linguist.*, pages 4320–4333.

Nan Hu, Zirui Wu, Yuxuan Lai, Xiao Liu, and Yansong Feng. 2022. Dual-Channel Evidence Fusion for Fact Verification over Texts and Tables. *NAACL 2022 - 2022 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. Proc. Conf.*, pages 5232–5242.

Neema Kotonya, Thomas Spooner, Daniele Magazzeni, and Francesca Toni. 2021. Graph Reasoning with Context-Aware Linearization for Interpretable Fact Extraction and Verification. *FEVER 2021 - Fact Extr. Verif. Proc. 4th Work.*, pages 21–30.

Canasai Kruengkrai, Junichi Yamagishi, and Xin Wang. 2021. A multi-level attention model for evidence-based fact checking. *Find. Assoc. Comput. Linguist. ACL-IJCNLP 2021*, pages 2447–2460.

Qian Liu, Bei Chen, Jiaqi Guo, Morteza Ziyadi, Zeqi Lin, Weizhu Chen, and Jian Guang Lou. 2022. Tapex: Table Pre-Training Via Learning a Neural Sql Executor. *ICLR 2022 - 10th Int. Conf. Learn. Represent.*, (Mlm):1–19.

Christopher Malon. 2021. Team Papelo at {FEVEROUS}: Multi-hop Evidence Pursuit. In *Proc. Fourth Work. Fact Extr. Verif.*

Mohammed Saeed, Giulio Alfarano, Khai Nguyen, Duc Pham, Raphaël Troncy, and Paolo Papotti. 2021. Neural re-rankers for evidence retrieval in the FEVEROUS task. *FEVER 2021 - Fact Extr. Verif. Proc. 4th Work.*, (Section 2):108–112.

Aalok Sathe and Joonsuk Park. 2021. Automatic fact-checking with document-level annotations using BERT and multiple instance learning. *FEVER 2021 - Fact Extr. Verif. Proc. 4th Work.*, pages 101–107.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: A large-scale dataset for fact extraction and verification. *NAACL HLT 2018 - 2018 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.*, 1:809–819.

Mykola Trokhymovych and Diego Saez Trumper. 2020. Natural language inference for fact-checking in wikipedia.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Adv. Neural Inf. Process. Syst.*, volume 2017-Decem, pages 5999–6009.

Nancy X.R. Wang, Diwakar Mahajan, Marina Danilevsky, and Sara Rosenthal. 2021. SemEval-2021 Task 9: Fact Verification and Evidence Finding for Tabular Data in Scientific Documents (SEM-TAB-FACTS). *SemEval 2021 - 15th Int. Work. Semant. Eval. Proc. Work.*, pages 317–326.