Fraunhofer SIT at CheckThat! 2024: Adapter Fusion for Check-Worthiness Detection

Notebook for the CheckThat! Lab at CLEF 2024

Inna Vogel[1,2,*], Pauline Möhle[1]

[1]Fraunhofer Institute for Secure Information Technology SIT | ATHENE - National Research Center for Applied Cybersecurity, Rheinstrasse 75, Darmstadt, 64295, Germany, url=https://www.sit.fraunhofer.de/

[2]Advisori FTC GmbH, Kaiserstraße 44, 60329 Frankfurt am Main, Germany, url=https://www.advisori.de/

Abstract

This paper describes the Fraunhofer SIT team's third-place approach for CLEF-2024 CheckThat! lab Challenge Task 1 for English. The "Check-Worthiness Estimation" task is to determine whether a text snippet from a political debate should be prioritised for fact-checking. Identifying check-worthy statements aims to facilitate manual fact-checking by prioritising claims that fact-checkers should consider first. It can also be considered as the primary step of a fact-checking system. Our proposed system is an adapter fusion model that integrates a task adapter with a Named Entity Recognition (NER) adapter. Adapters offer a resource-efficient alternative to fully fine-tuning transformer models. Our submitted model achieves a F_1 score of 0.78 on the English test set and was ranked as the third best model in the competition.

Keywords

check-worthiness detection, fact-checking, adapter fusion, task adapter, NER

1. Introduction

The fact-checking process typically involves three main steps. The first step is to identify statements or claims within a text that need to be fact-checked, as not all claims are equally important or contain pertinent information that needs to be verified. This can include false claims, statistics or other objectively verifiable inaccuracies. Fact-checkers prioritise claims for verification based on their potential impact, factual consistency or public interest. Once a claim has been selected, the second step is to gather credible evidence to support or refute it by consulting reliable sources such as academic journals, official reports, reputable news organisations, subject matter experts and primary sources such as original documents or statistics. To ensure consistency and accuracy, fact-checkers and journalists cross-reference information from multiple sources. The main challenge is that the majority of fact-checkers' work remains manual. As a result, there is an urgent need to develop technologies that can facilitate, accelerate and improve journalists' fact-checking and fake news and misinformation detection tasks.

The first step in the fact-checking pipeline, automatically identifying statements worthy of verification, has the potential to assist fact-checkers and journalists by locating and highlighting statements within a text that warrant further verification. This process could streamline the fact-checking workflow and reduce the potential for human bias in selecting claims for verification. Check-worthy sentences or statements are usually those that contain factual information such as dates, definitions, statistics or descriptions of events or laws.

The CheckThat! Lab has been tackling this scientific problem for the past several years. The aim of this year's CheckThat! Lab Task 1 "Check-Worthiness Estimation" is to determine whether a claim in a tweet and/or a political debate/speech is worth fact-checking. The task is considered a binary classification task with data available in Arabic, English and Spanish [1]. Frauhofer SIT participated

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

*Corresponding author.

inna.vogel@advisori.de (I. Vogel)

www.linkedin.com/in/inna-vogel-nlp (I. Vogel)

© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

CEUR-WS.org/Vol-3740/paper-62.pdf

in Task 1 of the CLEF 2024 CheckThat! Lab Challenge for the English language identifying relevant claims in political debates.

In this paper, we propose an adapter fusion approach that integrates a task adapter with a Named Entity Recognition (NER) adapter. Adapters are a resource-efficient alternative to fully fine-tuned transformer models [2]. Initially, we trained a task adapter to effectively detect check-worthy statements. As check-worthy claims often contain facts in the form of named entities — such as personal names, dates, financial and percentage values - we combined the task adapter with a NER adapter. With a F_1 score of 0.78, our proposed adapter fusion model placed third in the competition.

2. Related Work

While early approaches focused on a fixed set of features (such as sentiment, word count, part of speech (PoS) tags and named entities (NE)) and utilized traditional machine learning models (Naive Bayes, SVM and Random Forest) [3], recent work focuses on pre-trained language models such as BERT [4, 5].

The CLEF CheckThat! challenge, which was introduced in 2018 and is still ongoing, has contributed a considerable amount of research in recent years. Despite the diversity of models and representations employed in the initial years of the challenge, including k-nearest neighbors [6] and recurrent neural networks [7] for models, and character n-grams [6] and word embeddings [8] for representation, neural approaches utilizing word embeddings demonstrated superior performance compared to classical methods [9]. This trend continued in the 2019 challenge, where the top-performing team used an LSTM model trained with dual token embeddings (domain-specific word embeddings and syntactic dependencies) after pre-training on previous debates [10].

After the emergence of transformers in 2019 [11], there was a shift in contributions towards utilizing transformers for check-worthiness detection in subsequent years [12, 13]. Following the introduction of GPT-3, the best-performing approach for the English subtask 2023 was to fine-tune GPT-3 with 7.7k examples from pre-existing datasets. However, subsequent experiments by the same group using DeBERTaV3 yielded almost identical results to GPT-3 [14].

Schlicht et al. [15] conducted an investigation into the cross-training of adapter fusion models across various world languages, including Arabic, English, and Spanish, for the purpose of multilingual check-worthiness detection. They used mBERT and XLM-R and adapter fusion models on multilingual datasets from the CLEF CheckThat! Lab 2022 and 2021 challenges. They showed that the models outperformed monolingual task adapters and fully tuned models. A F_1 score of 0.51 was achieved for the detection of English check-worthy claims. Vogel et al. [16] combined a task adapter and a NER adapter and achieved state-of-the-art results on two challenging check-worthiness benchmarks. The best-performing model achieved a F_1 score of 0.92 on the CheckThat! Lab 2023 dataset. In addition, the authors interpreted the fusion attentions, demonstrating the effectiveness of their approach.

3. Data Set Description

The data for this year's CheckThat! 2024 Challenge Task 1 "Check-Worthiness Estimation" is available in Arabic, English, and Dutch[1] [1]. However, our approach focuses only on the English data set. While the methodology employed could theoretically be applied to other languages, specific modifications to the model would be required to account for linguistic differences.

For the English task, the data set consists of political debates collected from the US presidential general election debates. Examples from the data set are shown in Table 1.

The aim goal of Task 1 is to identify entries that contain check-worthy claims. The data set was annotated by human labelers. The label distributions and data set splits were provided by the organisers and are shown in Table 2. As can be seen, the data set consists of 23,851 entries, divided into two classes: check-worthy and non-check-worthy, labeled "YES" and "NO" respectively. The data set is

[1]Spanish was only offered for training

Table 1
Instances of check-worthy (Yes) and non-check-worthy (No) sentences for Task 1

	Instance	Class
1.	It called for an increase in the production of energy in the United States.	Yes
2.	There are 9 countries that spend more than we do on public education.	Yes
3.	I'd like to mention one thing.	No
4.	"And for that to happen, we have to strengthen our economy here at home."	No

significantly unbalanced, with approximately 25% of the entries labeled as check-worthy and 75% labeled as non-check-worthy.
To train and test the system, the data set is divided into three subsets: training (Train: 22,501 entries), development (Dev: 1,032 entries), and development test (Dev Test: 318 entries). The development test set contains a slightly higher proportion of check-worthy entries (33%) compared to the other data sets. The unlabeled test set (Test) was provided for evaluation purposes and consists of 341 sentences.

Table 2
Class distribution of the CheckThat! Lab 2024 Task 1 English data set ("Test" set is unlabelled for evaluation purposes.)

	Total	Yes	No
Train	22,501	5,413	17,088
Dev	1,032	238	794
Dev Test	318	108	210
Sum	23,851	5,759	18,092
Test	341	-	-

4. Methodology and Results

In this section, we present our submitted adapter fusion approach, which combines a task adapter with a NER adapter. Adapters are a lightweight alternative to full model fine-tuning, consisting of a small set of re-initialised weights at each layer of the pre-trained model [2]. These newly introduced weights are updated during fine-tuning, while the pre-existing parameters of the model remain fixed. This feature makes adapters parameter-efficient, speeds up training iterations and, due to their compact and modular nature, enables their modular sharing and composition without compromising model performance.

Adapter fusion is a method that combines the knowledge derived from different pre-trained adapters that were trained for distinct tasks. This process incorporates an attention module, which adeptly merges knowledge from various task adapters dynamically. Consequently, it fuses the knowledge acquired from diverse adapters into a unified representation. Various fusion techniques, including weighted summation, gating mechanisms, or attention mechanisms, can be employed for this purpose [2]. The goal of the adapter fusion method is to harness the synergies between different tasks and adapters.

Initially, we trained a task adapter on the CheckThat! Lab 2024 [1] dataset to effectively identify check-worthy sentences. No data pre-processing or cleaning was applied to the dataset. To train the task adapter, we applied adapter transformers from the "Adapter Hub" repository for pre-trained adapter models [17]. We used the pre-trained RoBERTa model [18] to tokenize the input data using the maximum sequence length of 512 (truncation=True, padding="max_length"). The task adapter model was trained for 6 epochs with a learning rate of 1e-4 and a batch size of 32.

The task adapter was trained on the "Train" dataset containing 22,501 instances, while the performance of the models during training was evaluated on the "Dev" set containing 1,032 instances. Finally, the "Dev Test" set of 318 samples was used to evaluate the trained model (Table 2). Our model achieves a F_1 of 0.866 over the positive (check-worthy) class. The results of the evaluation are shown in Table 3.

Table 3
Evaluation scores Precision (P), recall (R), F_1 score and accuracy for the task adapter model.

	P	R	F_1	Accuracy
Dev	0.943	0.975	0.959	0.981
Dev Test	0.977	0.778	**0.866**	0.918

We chose to use the adapter fusion approach, combining a task adapter with a NER adapter, to effectively detect named entities in the dataset. Previous studies have shown that check-worthy sentences tend to contain more named entities than non-check-worthy sentences [16]. This is due to the fact that factual information often arises in the form of names and numerical data, encompassing personal names, company names, geographical locations, dates, years, and percentages. Table 4 gives examples of sentences from the dataset containing named entities.

Table 4
Examples of check-worthy sentences with name entities

	Instance	Class
1.	"Today, 47 million people are on food stamps."	Yes
2.	"Of the nine million people put to work in new jobs since I've been in office, 1.3 million of those has been among black Americans, and another million among those who speak Spanish."	Yes
3.	If you take the tax cut that the president of the United States has given – President Bush gave to Americans in the top 1 percent of America – just that tax cut that went to the top 1 percent of America would have saved Social Security until the year 2075.	Yes

The adapter fusion model takes as input the representations generated by multiple adapters, each trained for distinct tasks, and learns a parameterized mixer of the encoded information. The previously trained task adapter was fused with the fine-tuned version of the DistilRoBERTa [19] based NER model. The NER model was trained and evaluated on the CoNLL 2003 dataset and achieves an F_1 score of 0.92 [20].

We trained our adapter fusion model for 6 epochs with a learning rate of 5e-5 and a batch size of 32 with a maximum sequence length of 512. The model was evaluated on the "Dev Test" and achieves a F_1 of 0.916 over the check-worthy class. The results of the approach are shown in Table 5.

Table 5
Evaluation scores Precision (P), recall (R), F_1 score and accuracy for the adapter fusion model. Comparison with the top best models and the baseline.

	P	R	F_1	Accuracy
Fraunhofer SIT (Dev)	0.983	0.967	0.975	0.989
Fraunhofer SIT (Dev Test)	0.979	0.861	0.916	0.947
Fraunhofer SIT (Test)	-	-	**0.780**	-
Baseline (Test)	-	-	0.307	-
FactFinders (Test)	-	-	0.802	-
teamopenfact (Test)	-	-	0.796	-

Since the adapter fusion model outperformed the adapter model in terms of F_1 score, we used the former to classify the private test set of this year's CheckThat! 2024 competition. Our model achieves an F_1 score of 0.78 over the positive check-worthy class.

5. Conclusion and Future Work

Identifying check-worthy statements can be seen as a first step in detecting the spread of false information online. Used as a pre-filter, this approach can significantly reduce the amount of data requiring

manual evaluation by human experts. In this paper, we presented an adapter fusion method that combines a task-specific adapter and a NER adapter.

Initially, we trained a task adapter to detect check-worthy statements effectively. Given that check-worthy statements often contain named entities (such as references to persons, locations, or dates), we integrated this task adapter with a pre-trained NER adapter. This integration aimed to exploit the synergies between different tasks. Our approach achieves a F_1 score of 0.78 on the CheckThat! Lab 2023 test dataset and was ranked third in the competition.

Future research may explore the integration of additional task-specific or pre-trained adapters. In our current approach, we utilized a pre-trained NER adapter developed to detect four NER classes. Subsequent work could investigate the use of a NER classifier trained to identify a broader range of NER classes.

Acknowledgements

This work was supported by the German Federal Ministry of Education and Research (BMBF) and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of "ATHENE – CRISIS".

References

[1] A. Barrón-Cedeño, F. Alam, T. Chakraborty, T. Elsayed, P. Nakov, P. Przybyła, J. M. Struß, F. Haouari, M. Hasanain, F. Ruggeri, X. Song, R. Suwaileh, The clef-2024 checkthat! lab: Check-worthiness, subjectivity, persuasion, roles, authorities, and adversarial robustness, in: N. Goharian, N. Tonellotto, Y. He, A. Lipani, G. McDonald, C. Macdonald, I. Ounis (Eds.), Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2024, pp. 449–458.

[2] 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.), ICML, volume 97 of *Proceedings of Machine Learning Research*, PMLR, 2019, pp. 2790–2799. URL: http://dblp.uni-trier.de/db/conf/icml/icml2019.html#HoulsbyGJMLGAG19.

[3] N. Hassan, C. Li, M. Tremayne, Detecting check-worthy factual claims in presidential debates, in: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, CIKM '15, Association for Computing Machinery, New York, NY, USA, 2015, p. 1835–1838. URL: https://doi.org/10.1145/2806416.2806652. doi:10.1145/2806416.2806652.

[4] F. Alam, A. Barrón-Cedeño, G. S. Cheema, S. Hakimov, M. Hasanain, C. Li, R. Míguez, H. Mubarak, G. K. Shahi, W. Zaghouani, P. Nakov, Overview of the CLEF-2023 CheckThat! lab task 1 on check-worthiness in multimodal and multigenre content, in: Working Notes of CLEF 2023—Conference and Labs of the Evaluation Forum, CLEF '2023, Thessaloniki, Greece, 2023.

[5] K. Meng, D. Jimenez, F. Arslan, J. D. Devasier, D. Obembe, C. Li, Gradient-based adversarial training on transformer networks for detecting check-worthy factual claims, ArXiv abs/2002.07725 (2020). URL: https://api.semanticscholar.org/CorpusID:211146392.

[6] B. Ghanem, M. Montes, F. Rangel Pardo, P. Rosso, Upv-inaoe-autoritas - check that: Preliminary approach for checking worthiness of claims, 2018.

[7] C. Hansen, C. Hansen, J. Simonsen, C. Lioma, The copenhagen team participation in the check-worthiness task of the competition of automatic identification and verification of claims in political debates of the clef-2018 checkthat! lab, in: L. Cappellato , N. Ferro , J. Nie, L. Soulier (Eds.), CLEF 2018 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, 2018. 19th Working Notes of CLEF Conference and Labs of the Evaluation Forum, CLEF 2018 ; Conference date: 10-09-2018 Through 14-09-2018.

[8] R. Banerjee, C. Zuo, A. Karakaş, A hybrid recognition system for check-worthy claims using heuristics and supervised learning, 2018.

[9] P. Atanasova, A. Barron-Cedeno, T. Elsayed, R. Suwaileh, W. Zaghouani, S. Kyuchukov, G. D. S. Martino, P. Nakov, Overview of the clef-2018 checkthat! lab on automatic identification and verification of political claims. task 1: Check-worthiness, 2018. arXiv:1808.05542.

[10] C. Hansen, C. Hansen, J. Simonsen, C. Lioma, Neural weakly supervised fact check-worthiness detection with contrastive sampling-based ranking loss, volume 2380, ceur workshop proceedings, 2019. 20th Working Notes of CLEF Conference and Labs of the Evaluation Forum, CLEF 2019 ; Conference date: 09-09-2019 Through 12-09-2019.

[11] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[12] A. Barron-Cedeno, T. Elsayed, P. Nakov, G. D. S. Martino, M. Hasanain, R. Suwaileh, F. Haouari, N. Babulkov, B. Hamdan, A. Nikolov, S. Shaar, Z. S. Ali, Overview of checkthat! 2020: Automatic identification and verification of claims in social media, 2020. arXiv:2007.07997.

[13] P. Nakov, G. D. S. Martino, T. Elsayed, A. Barrón-Cedeño, R. Míguez, S. Shaar, F. Alam, F. Haouari, M. Hasanain, W. Mansour, B. Hamdan, Z. S. Ali, N. Babulkov, A. Nikolov, G. K. Shahi, J. M. Struß, T. Mandl, M. Kutlu, Y. S. Kartal, Overview of the clef–2021 checkthat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news, 2021. arXiv:2109.12987.

[14] M. Sawinski, K. Węcel, E. Księżniak, M. Stróżyna, W. Lewoniewski, P. Stolarski, W. Abramowicz, Openfact at checkthat! 2023: Head-to-head gpt vs. bert - a comparative study of transformers language models for the detection of check-worthy claims, 2023.

[15] I. B. Schlicht, L. Flek, P. Rosso, Multilingual detection of check-worthy claims using world languages and adapter fusion, in: J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2023, pp. 118–133.

[16] I. Vogel, P. Möhle, M. Meghana, M. Steinebach, Adapter fusion for check-worthiness detection – combining a task adapter with a ner adapter., ROMCIR 2024: The 4th Workshop on Reducing Online Misinformation through Credible Information Retrieval, held as part of ECIR 2024: the 46th European Conference on Information Retrieval, March 24, 2024, Glasgow, UK (2024). URL: https://romcir.disco.unimib.it/wp-content/uploads/sites/151/2024/03/Paper6_Vogel.pdf.

[17] C. Poth, H. Sterz, I. Paul, S. Purkayastha, L. Engländer, T. Imhof, I. Vulić, S. Ruder, I. Gurevych, J. Pfeiffer, Adapters: A unified library for parameter-efficient and modular transfer learning, 2023. arXiv:2311.11077.

[18] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, ArXiv abs/1907.11692 (2019). URL: https://api.semanticscholar.org/CorpusID:198953378.

[19] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, ArXiv abs/1910.01108 (2019).

[20] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, R. Vollgraf, FLAIR: An easy-to-use framework for state-of-the-art NLP, in: NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations), 2019, pp. 54–59.