Adapter fusion for check-worthiness detection - combining a task adapter with a NER adapter

Inna Vogel[1,*,†], Pauline Möhle[2], Meghana Meghana[3] and Prof. Dr. Martin Steinebach[3]

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

Abstract

Detecting check-worthy statements aims to facilitate manual fact-checking efforts by prioritizing claims that fact-checkers should prioritize first. It can also be considered as the first step of a fact-checking system. In this paper, we present an adapter fusion model that combines a task adapter with a NER adapter achieving state-of-the-art results on two challenging check-worthiness benchmarks. Adapters are a resource-efficient alternative to fully fine-tuning transformer models. Our best performing model obtains an $F1$ score of 0.92 on the CheckThat! Lab 2023 dataset. Additionally, we interpret the fusion attentions, demonstrating the effectiveness of our approach. The quantitative analysis of the fusion attentions shows that named entities contribute significantly to the prediction of the adapter fusion model.

Keywords

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

1. Introduction

Fact-checking online content is essential to ensure the reliability of information shared through various online communication channels, such as news websites and social media platforms. Fact-checkers and journalists are constantly working to identify and correct misinformation and communicate their work as quickly as possible. But with the amount of information published every day online and the limited resources available to journalists and fact-checkers, it is almost impossible to keep up with this critical work.

The fact-checking process consists usually of three main steps. The first step involves identifying statements or claims in a text that are worth fact-checking, as not all claims are equally important or contain relevant information that needs to be fact-checked. These can be false allegations, statistics or other objectively verifiable false information. Fact-checkers prioritize claims for verification based on their potential impact, the claim's factual coherence or the public interest in the claim. Once a claim has been selected, the second step is to gather trustworthy evidence to confirm or disprove it by researching reliable sources. These sources can include academic journals, official reports, reputable news organizations, subject matter

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

*Corresponding author.

inna.vogel@sit.fraunhofer.de (I. Vogel); pauline.moehle@sit.fraunhofer.de (P. Möhle); meghana.meghana@sit.fraunhofer.de (M. Meghana); martin.steinebach@sit.fraunhofer.de (Prof. Dr. M. Steinebach)

experts and primary sources such as original documents or statistics. To ensure consistency and accuracy, fact-checkers and journalists compare information from multiple sources. The main challenge is that the vast majority of the work of the fact-checker is still done manually. Therefore, there is a need to develop technologies that would facilitate, speed up and improve the task of fact-checking and journalists detecting fake news and misinformation.

The first step in the fact-checking pipeline, the automatic identification of check-worthy statements, could facilitate the work of fact-checkers or journalists by identifying and highlighting statements within a text that require further verification. This could streamline the fact-checking process and reduce the potential for human bias in claim selection.

We consider the check-worthiness detection task as a binary classification task. Check-worthy sentences or statements are usually those that contain factual information such as dates, definitions, statistics or descriptions of events or laws. They are usually of interest to or likely to affect the general public [1]. Nakov et al. [2] extend the definition and add that such statements can also be potentially damaging to society, public figures or a company. Non-check-worthy statements, on the other hand, contain subjective opinions and beliefs rather than factual claims [1, 2].

In this paper, we propose an adapter fusion approach that combines a task adapter with a NER (Named Entity Recognition) adapter. Adapters are a resource-efficient alternative to fully fine-tuned transformer models [3]. They are lightweight and modular neural networks, learning tasks with fewer parameters and transferring knowledge across tasks and languages [4]. We first trained a task adapter to effectively detect check-worthy statements. As we noticed that check-worthy claims increasingly contain facts in the form of named entities, such as personal names, dates, financial and percentage values or names of events or historical occurrences (e.g., "World War II" or "the Great Depression"), we fused the task adapter with a NER adapter. Our approach achieves state-of-the-art results on two challenging check-worthiness detection benchmarks.

Our contributions are summarized as follows:

- We are the first to propose an adapter fusion model that combines a task adapter with a NER adapter.
- Our model achieves state-of-the-art performance by a substantial margin on challenging check-worthiness benchmarks.
- We use an explainability tool to interpret the classification results, demonstrating the effectiveness of our approach.

2. Related work

The first methods of check-worthiness detection were based on the extraction of meaningful features from the text. Given a US presidential election transcript, ClaimBuster [5] predicts check-worthiness by using a support vector machine and extracting a set of 6,615 features in total (such as word count, sentiment, tf-idf weighted bag-of-words, part-of-speech tags or entity type). Gencheva et al. [6] extended the work of Hassan et al. [5] by including contextual features such as the position of the sentence, the size of a segment belonging to a speaker, topics or word embeddings. A MAP of 0.427 was achieved using a neural network with all features combined.

Meng et al. [7] used both the ClaimBuster dataset and the CLEF CheckThat! Lab 2019 dataset [8] on the detection of check-worthy factual claims using adversarial training on transformer models. Their model achieved an improvement in the $F1$ score of 4.7 points compared to other models on these datasets.

CheckThat! Lab organises multilingual check-worthiness detection tasks since 2020. They support more languages every year for multimodal and multigenre content [9]. The aim of the CheckThat! Lab challenge in 2023 [10] was to determine whether the information contained in the political debate is reliable and worthy of further fact-checking. Sawiński et al. [11] were the best performing team in English. They experimented with fine-tuning a variety of BERT models and found that fine-tuning DeBERTaV3 [12] yielded near-identical performance to GPT-3, achieving an $F1$ score of 0.89. Frick et al. [13] came second in the competition with an $F1$ score of 0.87 by fine-tuning BERT three times, starting with a different seed for model initialisation, resulting in three models. They combined these models into an ensemble using a model souping technique that adaptively adjusts the influence of each model based on its performance.

Schlicht et al. [9] investigated cross-training of adapter fusion models on world languages (such as Arabic, English and Spanish) to detect check-worthiness in multiple languages. Therefore, they used mBERT and XLM-R and adapter fusion models (combining task and language adapters) within a transformer as well as fully fine-tuned transformers. They showed that the models could perform better than monolingual task adapters and fully tuned models. For the detection of English check-worthy claims, an $F1$ score of 0.51 was achieved using the multilingual dataset from the CLEF CheckThat! Lab challenge 2022 [14] and 2021 [15] applying XLM-R in combination with a fully fine-tuned transformer.

3. Dataset description

We used the CheckThat! Lab 2023 dataset [10] and the ClaimBuster dataset [1] for our experiments. The CheckThat! Lab 2023 English dataset consists of political debates collected from the US presidential general election debates [10]. The aim of the CheckThat! Lab 2023 task was to predict whether a text snippet from a political debate needs to be assessed manually by an expert by estimating its check-worthiness. Examples from the dataset are shown in Table 1.

Table 1
Examples from the CheckThat! Lab 2023 dataset for check-worthy (Yes) and non-check-worthy (No) statements

	Instance	Class
1.	And that means 98 percent of American families, 97 percent of small businesses, they will not see a tax increase.	Yes
2.	I said we'd get tougher with child support and child support enforcement's up 50 percent.	Yes
3.	But I'm not going to do that.	No
4.	But the important thing is what are we going to do now?	No

The dataset is divided into four subsets. The "Train" subset consists of 16,876 entries. Each entry is labelled either "Yes" or "No" as to whether it is worth checking (YES) or not (No). The

Table 2
Class distribution of the CheckThat! Lab 2023 check-worthiness dataset for English.

	Total	Yes	No
Train	16,876	4,058	12,818
Dev	5,625	1,355	4,270
Dev Test	1,032	238	794
Test	318	108	210
Sum	23,851	5,759	18,092

development set "Dev" contains 5,625 statements, the development test set "Dev Test" contains 1,032 entries and the test set for the final evaluation "Test" contains 318 statements. The label distributions and dataset splits are shown in Table 2.

While the first three partitions primarily use the ClaimBuster dataset described in Arslan et al. [1], there have been some updates made by the CheckThat! Lab 2023 organizers to improve the quality of the annotations. The test set includes sentences that were not featured in the ClaimBuster dataset.

The ClaimBuster dataset was labelled by 101 annotators over a period of 26 months. The following three classes have been annotated. 1. Check-worthy factual sentences: These sentences contain statements of fact that the general public will have an interest in finding out whether they are true or not. Journalists and fact-checkers look for these kinds of statements to check their veracity. 2. Unimportant factual sentences: These are factual statements, but they are not verifiable or the general public is not interested in knowing whether these sentences are true or false. Fact-checkers do not consider these sentences to be worth checking. 3. Non-factual sentences: These sentences contain no factual claims. Subjective sentences such as opinions, and beliefs fall into this category [1].

To compare our results with the work of Meng et al. [7], the ClaimBuster dataset was used as a baseline. The dataset consists of 9,674 sentences, of which 6,910 are non-check-worthy and 2,764 are check-worthy [7]. The authors excluded the class of non-check-worthy factual sentences, having observed that this class was not really useful and could negatively affect the performance of models. To compare our results, we used the same split as the authors - 67.5% of the dataset was used to train, 7.5% for validation and 25% to test our model.

Both datasets are highly unbalanced, with about a quarter of the sentences being check-worthy. This is also due to the fact that attention-worthy sentences occur less frequently in the text than non-check-worthy sentences.

4. Methodology

4.1. Adapter fusion - combining a task adapter with a NER adapter

Transformer models, pre-trained on massive amounts of text data and then fine-tuned on target tasks, have led to significant advances in NLP, achieving state-of-the-art results in a variety of tasks. However, models such as RoBERTa [16] and BERT [17] consist of millions of parameters, making it prohibitively expensive to share and distribute fully tuned models for each individual

Table 3
Distribution of named entities in the CheckThat! Lab 2023 dataset using 5,759 sentences per class

	PER	LOC	ORG	MISC
Check-worthy sentences	1,226	1,739	973	1,240
Non-check-worthy sentences	708	1,015	356	630

downstream task. Adapters are a lightweight alternative to full model fine-tuning, consisting of only a tiny set of newly initialised weights at each layer of the pre-trained model [3]. These new weights are then updated during fine-tuning and the pre-trained parameters of the pre-trained model are frozen. This means that adapters are parameter efficient, speed up training iterations, and can be shared and composited due to their modularity and compact size without compromising the performance of the model. Adapters have been shown to work on par with full fine-tuning by adapting the representations at each level [4].

Adapter fusion is a method of combining the knowledge of multiple pre-trained adapters trained for different tasks. Adapter fusion consists of an attention module that learns how to dynamically combine knowledge from different task adapters. This means that it fuses the information learned by different adapters into a coherent representation. Different fusion strategies can be used, such as weighted summation, gating mechanisms, or attention mechanisms. The goal is to capture the synergies between different tasks and adapters. First, we trained a task adapter to efficiently detect and classify the sentences worth checking by journalists and fact-checkers. This model serves as our baseline. In a quantitative analysis counting the frequencies of named entities in the two classes, we found that check-worthy sentences tend to contain more named entities than non-check-worthy sentences. Table 3 shows the distribution of named entities in the CheckThat! Lab 2023 dataset.

Since the dataset contains less check-worthy sentences (5,759) than non-check-worthy sentences (18,092), we reduced the number of sentences in the negative class so that the dataset used for the analysis is balanced, i.e. 5,759 sentences per class. To count the frequencies of the named entities, we used the four-class NER model for English "Flair", which achieves an $F1$ score of 0.93 on the CoNLL 2003 dataset [18].

The four named entity classes are: Person Name (PER), Location Name (LOC), Organisation Name (ORG) and Miscellaneous (MISC). The distribution of the classes shows that all the classes of named entities are more present in the sentences that are worth fact-checking. We assume that named entities, as in journalism, are essential holders of information. The "Five Ws" in journalism "Who?, What?, Where?, When?, and Why?" are essential questions that aim to provide a complete understanding of a situation. The questions "Who?, Where?, When?" and partly also "What?" (e.g. *"30% of those vaccinated"*) can be detected by a NER recogniser.

This motivated us to experiment with the fusion of a trained task adapter with a NER adapter. The architecture of our final model is illustrated in Figure 1. The input to the architecture is a statement while the output is a probability score that determines the check-worthiness of the given input statement. The adapter fusion component takes as input the representations of multiple adapters trained on different tasks and learns a parameterized mixer of the encoded information.

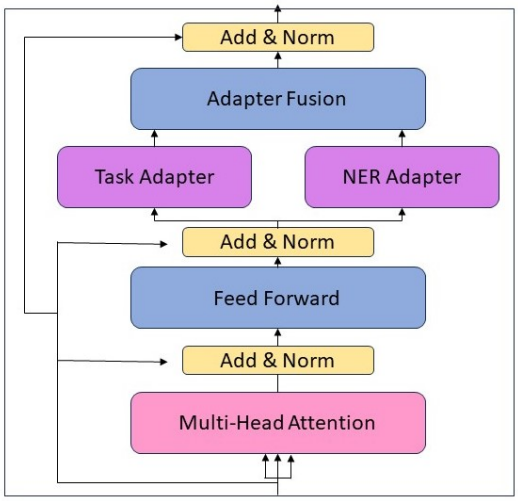

Figure 1: The architecture of the adapter fusion model combining a task adapter and a NER adapter.

4.2. Implementation details

The task adapter model "AF-TA" and the adapter fusion model "AF-TA-NER" were trained on the CheckThat! Lab 2023 [10] dataset. For this, adapter transformers from the "Adapter Hub" repository for pre-trained adapter modules were employed [4]. We used the pre-trained RoBERTa model as it showed significant improvements in various NLP benchmarks compared to the original BERT model [16].

Our models were trained on the "Train" dataset with 16,876 instances, while the performance of the models during training was evaluated on the "Dev"set with 5,625 instances. Finally, the "Dev Test" set of 1,032 samples was used to compare the performance of the different trained models. The overall best performing model was evaluated on the "Test" dataset (Table 2) using the $F1$ score metric over the positive (check-worthy) class. We tried a number of different parameters and report on the ones that performed best.The task adapter model "AF-TA" was trained for 6 epochs with a learning rate of 1e-4 and a batch size of 32, using a maximum sequence length of 512.

To train our "AF-TA-NER" model, we fused the task adapter model "AF-TA" and 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 $F1$ score of 0.92 [20]. The adapter fusion model "AF-TA-NER", which combines the task adapter and the NER adapter model performed best when the number of epochs was set to 5, the learning rate to 5e-5, the batch size to 8 and the maximum sequence length of 512.

5. Baselines

To compare the performance of our adapter fusion model, "AF-TA-NER", three different baselines were used. As the first baseline, we chose the best performing system of the CheckThat! Lab

2023 challenge "OpenFact" [11]. The highest $F1$ score of 0.89 was obtained by the GPT-3 Curie model fine-tuned with approximately 7,690 examples selected on label quality criteria.

Meng et al. [7] proposed in their work a method by applying adversarial perturbations using the BERT architecture. In order to detect verifiable factual claims, they used the ClaimBuster dataset [1]. To validate their model, they performed 4-fold cross-validation, selecting the best model from each fold using the weighted $F1$ score calculated on the validation set. We split the data into 25% test, 7.5% validation and 67.5% training and applied stratified 4-fold cross-validation according to the authors in order to compare the performance of our model. The reported $F1$ score is based on classifications across all folds. Their best performing model "CB-BBA" achieves an average $F1$ score of 0.83 on the positive (check-worthy) class. We used the same dataset split for our proposed models, applying stratified 4-fold cross-validation.

To determine whether the proposed NER adapter fusion model "AF-TA-NER" can improve classification results, we used the task adapter model "AF-TA" as a third baseline. The implementation details were given in section 4.2.

6. Evaluation results

Table 4 shows the performance results of the "AF-TA-NER" model and two baselines, the "AF-TA" model as well as the best performing system in the CheckThat! Lab 2023 challenge "OpenFact" Sawiński et al. [11].

The "AF-TA" task adapter model achieves an $F1$ score of 0.87 while the GPT-3 Curie model used in "OpenFact" achieves an $F1$ score of 0.89. The proposed "AF-TA-NER" outperforms both models by achieving an $F1$ score of 0.92 on the positive class (check-worthy). The negative class (not check-worthy) achieves an $F1$ score of 0.96.

Table 4
Precision (P), recall (R), $F1$ score and accuracy for the CheckThat! Lab 2023 dataset. The best model is marked in bold. The second best model is underlined.

Model	P	R	$F1$	Accuracy
OpenFact	0.95	0.85	<u>0.89</u>	0.93
AF-TA	0.96	0.79	0.87	0.92
AF-TA-NER	0.98	0.86	**0.92**	0.95

Table 5 shows the classification results of our two proposed models compared to the approach presented by Meng et al. [7]. Our models outperform "CB-BBA" ($F1$ score 0.84) by a substantial margin. The "AF-TA-NER" model ($F1$ score 0.89) performs slightly better than the "AF-TA" model ($F1$ score 0.88). Compared to the "CB-BBA" model, our best performing model achieves a 5 point improvement in $F1$ score. The "AF-TA-NER" model outperforms all three baselines and achieves state-of-the-art performance on different benchmarks. The confusion matrix of our "AF-TA-NER" model is shown in Figure 2.

Table 5
Precision (P), recall (R), $F1$ score and averaged across stratified 4-fold cross validation. The best model is marked in bold. The second best model is underlined.

Model	P	R	$F1$
CB-BBA	0.84	0.83	0.84
AF-TA	0.87	0.90	<u>0.88</u>
AF-TA-NER	0.88	0.90	**0.89**

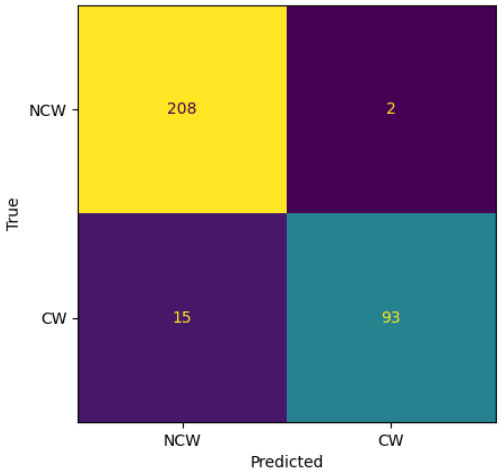

Figure 2: Cunfusion matrix of the "AF-TA-NER" model. CW refers to the check-worthy class, while NCW refers to the non-check-worthy class.

7. Interpretation of the fusion attentions

In this section, we present the interpretation of the fusion attentions using the "Transformers Interpret"[1] library. The core attribution methods on which Transformers Interpret is built are "Integrated Gradients" and a variant of them, "Layer Integrated Gradients". The feature attribution score is a summary or average of the attributions from each layer, explaining which features were most important in a model's prediction for a given input.

The aim of the interpretation of the classification results is to analyze whether our proposed adapter fusion model is capable of reliably classifying check-worthy statements in a text. This analysis can also be useful for journalists and fact-checkers to check the basis on which the model has made its decision. The following quantitative and qualitative analysis is based on the classification results of our trained adapter fusion model "AF-TA-NER".

To determine which named entity class contributes the most to the classification, we chose a NER model for the quantitative analysis that can recognize the highest number of classes. Therefore, we used spaCy's[2] NLP pipeline model "en_core_web_smpipeline". The model provides a NER system that can identify named entities and classify them into 18 predefined categories.

[1]Transformers Interpret: https://github.com/cdpierse/transformers-interpret.
[2]spaCy: https://spacy.io/.

The aim of the quantitative analyses was to investigate whether NER features were important for the predictions of the "AF-TA-NER" model.

Table 6 lists the classes that contributed most to the classification of the check-worthy class. To compare, we also show how the respective NER class relates to the negative class. Positive attribution numbers indicate whether a word contributes positively to the predicted class, while negative numbers indicate the opposite.

Table 6
Contribution of NER features to the prediction of the adapter fusion model "AF-TA-NER". Positive attribution numbers indicate that a word contributes positively to the predicted class, while negative numbers indicate that a word contributes negatively to the predicted class. CW refers to the check-worthy class, while NCW refers to the non-check-worthy class.

	CW	NCW
Money	0.87	-0.38
Percent	0.67	-0.97
Cardinal	0.48	0.11
Date	0.50	-0.07
Event	0.27	-0.18
Ordinal	0.23	-0.11

The quantitative analysis shows that named entities contribute significantly to the prediction of the adapter fusion model. The NER class "money" contributes most positively to the positive class (e.g. *"I paid $38 million one year..."*), while at the same time it has little relevance for the classification of the negative class. The same holds true for "Percent", "Cardinal" and "Date" classes - each of them contributing significantly to the positive class (e.g. *"When we were in office, there was 15% less violence in America..."*), while contributing negatively to the non-check-worthy class.

It is interesting to note that the classes "Person" (0.18), "Place" (0.04) and "Organisation" (0.14), although also relevant for the classification of the check-worthy class, are less significant than the other mentioned NER classes in Table 6. Even the class "Event" (0.27), which refers to mentions of events in the text such as hurricanes, battles, wars or sports events, contributes more to the model's attention. This suggests that the model would still give good classification results even if the names of people and places were removed from the dataset, e.g., for data protection reasons.

Using the Transformers Interpret heat map visualization, we analyzed the contribution of tokens to the model's prediction. This type of visualization is particularly useful for understanding and interpreting the decisions made by complex transformer models. Each token in the input text is colour-coded based on its contribution score. The colour intensity represents the magnitude of the contribution. Our qualitative analysis shows that action verbs (or dynamic verbs), which describe the action that a subject of a sentence performs (e.g. *"run"*, *"fight"*, *"sleep"*), contribute to the prediction of the check-worthy class. Examples are shown in Figure 7. The colour intensity shows that words like *"paid", "released", "reduced", and "beat"* contribute positively to the prediction of the adapter fusion model.

As there are only two examples of False Positive (FP) classifications (Figure 2), no reliable analysis can be made. However, we suspect that these two examples shown in Figure 4 were

Figure 3: Heat map using the Transformers Interpret library showing how different tokens contribute to the model's prediction of the TP (check-worthy) class.

incorrectly labelled as not check-worthy by the human annotators. In our opinion, these examples contain relevant facts that should be fact-checked. The same applies to False Negative cases. In the following example, we found no evidence for a verifiable statement: *"Yes, you said that"*. However, the annotation is difficult to evaluate, as it is also subjective.

Figure 4: Heat map using the Transformers Interpret library, showing the prediction of the model on the FP class.

8. Conclusion and future work

In this paper, we presented our work on detecting check-worthy factual claims employing an adapter fusion approach by combining a task adapter with a NER adapter. We first trained a task adapter to effectively detect check-worthy statements and used it as our first baseline. As we analyzed that check-worthy claims increasingly contain facts in the form of named entities, we fused the task adapter with a NER adapter. The goal was to capture the synergies between different tasks and adapters. Our approach achieves state-of-the-art results on two challenging benchmarks. Our best "AF-TA-NER" adapter fusion model achieves an $F1$ score of 0.92 on the CheckThat! Lab 2023 dataset and an $F1$ score of 0.89 on the ClaimBuster dataset.

Additionally, we used an explainability tool to interpret the fusion attentions, demonstrating the effectiveness of our approach. The quantitative analysis showed that named entities contribute significantly to the prediction of the adapter fusion model. To determine which NER

class contributes the most to the classification results, we used a NER system that can identify 18 predefined categories of named entities. By analysing the attention weights, we found that it is not the NER classes "Person", "Location" or "Organization" that contribute most to the positive class, but rather the classes "Money", "Percent", "Cardinal" and "Date". In the future, we therefore plan to employ a NER adapter that can classify more than four classes. Additionally, we want to investigate the fusion of different task adapters and interpret how fusion attention differs across adapters and fusion models.

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” and "Lernlabor Cybersicherheit" (LLCS).

References

[1] F. Arslan, N. Hassan, C. Li, M. Tremayne, A benchmark dataset of check-worthy factual claims, in: 14th International AAAI Conference on Web and Social Media, AAAI, 2020. URL: https://api.semanticscholar.org/CorpusID:216870066.

[2] P. Nakov, G. Da San 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, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction: 12th International Conference of the CLEF Association, CLEF 2021, Virtual Event, September 21–24, 2021, Proceedings, Springer-Verlag, Berlin, Heidelberg, 2021, p. 264–291. URL: https://doi.org/10.1007/978-3-030-85251-1_19. doi:10.1007/978-3-030-85251-1_19.

[3] 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.

[4] J. Pfeiffer, A. Rücklé, C. Poth, A. Kamath, I. Vulić, S. Ruder, K. Cho, I. Gurevych, Adapterhub: A framework for adapting transformers, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020): Systems Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 46–54. URL: https://www.aclweb.org/anthology/2020.emnlp-demos.7.

[5] N. Hassan, F. Arslan, C. Li, M. Tremayne, Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster, Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2017).

[6] P. Gencheva, P. Nakov, L. Màrquez, A. Barrón-Cedeño, I. Koychev, A context-aware approach for detecting worth-checking claims in political debates, in: Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP

2017, INCOMA Ltd., Varna, Bulgaria, 2017, pp. 267–276. URL: https://doi.org/10.26615/978-954-452-049-6_037. doi:10.26615/978-954-452-049-6_037.

[7] 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.

[8] P. Atanasova, P. Nakov, G. Karadzhov, M. Mohtarami, G. D. S. Martino, Overview of the CLEF-2019 checkthat! lab: Automatic identification and verification of claims. task 1: Check-worthiness, in: L. Cappellato, N. Ferro, D. E. Losada, H. Müller (Eds.), Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9-12, 2019, volume 2380 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2019. URL: https://ceur-ws.org/Vol-2380/paper_269.pdf.

[9] 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.

[10] 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.

[11] M. Sawiński, 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, in: Conference and Labs of the Evaluation Forum, 2023. URL: https://api.semanticscholar.org/CorpusID:264441775.

[12] P. He, J. Gao, W. Chen, Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing, in: The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023, OpenReview.net, 2023. URL: https://openreview.net/pdf?id=sE7-XhLxHA.

[13] R. A. Frick, I. Vogel, J. Choi, Fraunhofer SIT at checkthat!-2023: Enhancing the detection of multimodal and multigenre check-worthiness using optical character recognition and model souping, in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, September 18th to 21st, 2023, volume 3497 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 337–350. URL: https://ceur-ws.org/Vol-3497/paper-029.pdf.

[14] P. Nakov, A. Barrón-Cedeño, G. Da San Martino, F. Alam, J. M. Struß, T. Mandl, R. Míguez, T. Caselli, M. Kutlu, W. Zaghouani, C. Li, S. Shaar, G. K. Shahi, H. Mubarak, A. Nikolov, N. Babulkov, Y. S. Kartal, J. Beltrán, The clef-2022 checkthat! lab on fighting the covid-19 infodemic and fake news detection, in: M. Hagen, S. Verberne, C. Macdonald, C. Seifert, K. Balog, K. Nørvåg, V. Setty (Eds.), Advances in Information Retrieval, Springer International Publishing, Cham, 2022, pp. 416–428.

[15] S. Shaar, F. Haouari, W. Mansour, M. Hasanain, N. Babulkov, F. Alam, G. D. S. Martino, T. Elsayed, P. Nakov, Overview of the CLEF-2021 checkthat! lab task 2 on detecting previously fact-checked claims in tweets and political debates, in: G. Faggioli, N. Ferro,

A. Joly, M. Maistro, F. Piroi (Eds.), Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21st - to - 24th, 2021, volume 2936 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021, pp. 393–405. URL: https://ceur-ws.org/Vol-2936/paper-29.pdf.

[16] 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.

[17] 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.

[18] A. Akbik, D. Blythe, R. Vollgraf, Contextual string embeddings for sequence labeling, in: COLING 2018, 27th International Conference on Computational Linguistics, 2018, pp. 1638–1649.

[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.