Overview of the CLEF-2022 CheckThat! Lab Task 2 on Detecting Previously Fact-Checked Claims

Preslav Nakov[1], Giovanni Da San Martino[2], Firoj Alam[1], Shaden Shaar[4], Hamdy Mubarak[3] and Nikolay Babulkov[5]

[1]Mohamed bin Zayed University of Artificial Intelligence, UAE
[2]University of Padova, Italy
[3]Qatar Computing Research Institute, HBKU, Qatar
[4]Cornell University, USA
[5]Sofia University, Bulgaria

Abstract

We describe the fourth edition of the CheckThat! Lab, part of the 2022 Conference and Labs of the Evaluation Forum (CLEF). The lab evaluates technology supporting three tasks related to factuality, and it covers seven languages such as Arabic, Bulgarian, Dutch, English, German, Spanish, and Turkish. Here, we present the *task 2*, which asks to detect previously fact-checked claims (in two languages). A total of six teams participated in this task, submitted a total of 37 runs, and most submissions managed to achieve sizable improvements over the baselines using transformer based models such as BERT, RoBERTa. In this paper, we describe the process of data collection and the task setup, including the evaluation measures, and we give a brief overview of the participating systems. Last but not least, we release to the research community all datasets from the lab as well as the evaluation scripts, which should enable further research in detecting previously fact-checked claims.

Keywords

Check-Worthiness Estimation, Fact-Checking, Veracity, Verified Claims Retrieval, Detecting Previously Fact-Checked Claims, Social Media Verification, Computational Journalism, COVID-19

1. Introduction

There has been a surge in research to develop systems for automatic fact-checking. However, such systems suffer from credibility issues. Hence, it is important to reduce the manual effort by detecting when a claim has already been fact-checked. Work in this direction includes [1] and [2]: the former developed a dataset for the task and proposed a ranking model, while the latter proposed a neural ranking model using textual and visual modalities.

To address this, the CheckThat! lab initiative features a number of tasks aiming to help automate the fact-checking process and to reduce the spread of disinformation and misinformation. The CheckThat! 2022 lab was held in the framework of CLEF 2022 [3].[1] Figure 1 shows the full CheckThat! identification and verification pipeline, highlighting the three tasks targeted

CLEF 2022: Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy

preslav.nakov@mbzuai.ac.ae (P. Nakov); dasan@math.unipd.it (G. D. S. Martino); fialam@hbku.edu.qa (F. Alam); ss2753@cornell.edu (S. Shaar); hmubarak@hbku.edu.qa (H. Mubarak); nbabulkov@gmail.com (N. Babulkov)

CEUR Workshop Proceedings (CEUR-WS.org)

[1] http://sites.google.com/view/clef2022-checkthat/

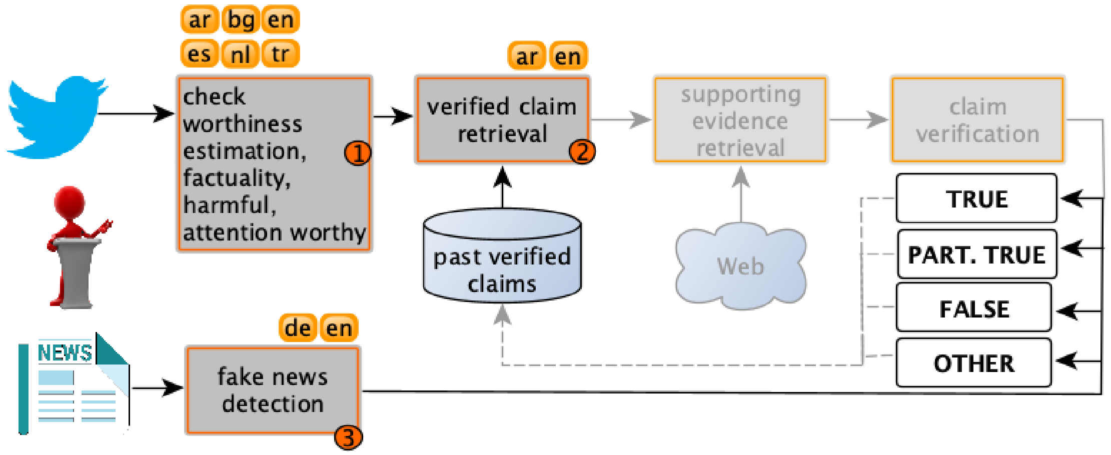

Figure 1: The full verification pipeline. The 2022 lab covers three tasks from that pipeline: (*i*) check-worthiness estimation, (*ii*) verified claim retrieval, and (*iii*) fake news detection. The gray tasks were addressed in previous editions of the lab [6, 7].

in this fifth edition of the lab: Task 1 on detecting relevant claims in tweets (this paper), Task 2 on retrieving relevant previously fact-checked tweets [4], and Task 3 on predicting the veracity of news [5].

In this paper, we describe in detail the second task, *detecting previously fact-checked claims*, of the CheckThat! lab tasks.[2] The second task is defined as follows: *"given a check-worthy input claim and a set of verified claims, rank the previously verified claims in order of usefulness to fact-check the input claim."* It consists of the following two subtasks:

Subtask 2A: Detecting previously fact-checked claims in tweets. Given a tweet, detect whether the claim it makes was previously fact-checked with respect to a collection of fact-checked claims. This is a ranking task, offered in Arabic and English, where the systems need to return a list of top-n candidates.

Subtask 2B: Detecting previously fact-checked claims in political debates or speeches. Given a claim in a political debate or a speech, detect whether the claim has been previously fact-checked with respect to a collection of previously fact-checked claims. This is a ranking task, and it was offered in English.

For Subtask 2A, we focused on tweets, and it was offered in Arabic, and English. The participants were free to work on any language(s) of their interest, and they could also use multilingual approaches that make use of all datasets for training. Subtask 2A attracted six teams, and the most successful approaches used transformers or a combination of embeddings, manually engineered features. More details are discussed in Section 3.

For Subtask 2B, we focused on political debates and speeches, and we used PolitiFact as the main data source. The task attracted one team, and a combination of transformers, prepossessing, and augmentation approaches performed the best.

[2] Refer to [3] for an overview of the full CheckThat! 2022 lab.

The remainder of the paper is organized as follows: Section 2 discusses related work. Sections 3 and 4 describe the dataset, the evaluation results, and the participating systems for subtasks 2A and 2B, respectively, and Section 5 concludes with final remarks.

2. Related Work

There has been a significant research focused on developing automatic systems for fact-checking [8, 9, 10, 11, 12]. Studies includes the development of datasets [13, 14], and evaluation campaigns [15, 6, 16, 17, 18]. However, such fully automatic systems suffer from credibility issues, e.g., in the eyes of journalists, and manual checking is still the norm. Thus, it is important to reduce that manual effort by detecting whether a claim has already been fact-checked [19]. Hence, a reasonable solution is to build tools to facilitate human fact-checkers, e.g., by detecting previously fact-checked claims.

Relevant work in this direction include [1, 20, 21]. In this work, we use their annotation setup and one of their datasets: PolitiFact. Previous work has mentioned the task as an integral step of an end-to-end automated fact-checking pipeline, but there was very little detail provided about this component and it was not evaluated [22].

There has been a number of tools developed such as *Fact Check Explorer*,[3], which allows users to search a number of fact-checking websites. However, the tool cannot handle a complex claim, as it uses the standard Google search functionality, which is not optimized for semantic matching of long claims.

A very recent survey reports what AI technology can offer to assist the work of professional fact-checkers [23], and has pointed out several research problems such as identifying claims worth fact-checking, detecting relevant previously fact-checked claims, retrieving relevant evidence to fact-check a claim, and verifying the claim.

Other recent work include memory-enhanced transformers for matching (MTM) to rank fact-checked articles [24], topic-aware evidence reasoning and stance-aware aggregation [25], claim matching [26], sequence-to-sequence transformer models [27] and deep Q-learning network [28].

3. Subtask 2A: Detecting Previously Fact-Checked Claims in Tweets

Given a tweet, the task asks to detect whether the claim the tweet makes was previously fact-checked with respect to a collection of fact-checked claims. The task is offered in Arabic and English. This is a ranking task, where the systems are asked to return a list of top-n candidates.

3.1. Dataset

Arabic For Arabic, we have 908 tweets, matching 1,089 verified claims (some tweets match more than one verified claim) in a collection of 30,379 previously fact-checked claims. The latter

[3] http://toolbox.google.com/factcheck/explorer

Table 1
Task 2: Statistics about the CT–VCR–22 corpus, including the number of *Input–VerClaim* pairs and the number of *VerClaim* claims to match the input claim against.

Partition	2A-Arabic	2A-English	2B-English
Input Claims	**908**	**1,610**	**752**
Training	512	999	472
Development	85	200	119
Dev-Test	261	202	78
Test	50	209	83
Input-VerClaims pairs	**1,089**	**1,610**	**869**
Training	602	999	562
Development	102	200	139
Dev-Test	335	202	103
Test	50	209	65
Verified claims (to match against)	**30,379**	**13,835**	**20,771**

include 5,921 Arabic claims from AraFacts [29] and 24,408 English claims from ClaimsKG [30], translated to Arabic using the Google Translate API.[4] The complete data collection process is discussed in [31].

English To develop the verified claims dataset, we used Snopes, a fact-checking website that targets rumors spreading in social media, and we collected 13,835 verified claims. Their fact-checking journalists often cite the tweet or the social media post that spreads the rumor when writing an article about a claim. We have 1,610 annotated tweets, each matching a single claim in a set of 13,835 verified claims from Snopes.

Data Statistics Table 1 shows statistics about the CT–VCR–22[5] corpus for Task 2, including both subtasks and languages. *Input–VerClaim* pairs represent input claims with their corresponding verified claims by a fact-checking source. The input for subtask 2A (2B) is a tweet (sentence from a political debate or a speech). More details about the corpus construction can be found in [31].

Data Split For Arabic, we provide 512 training, 85 dev, 261 dev-test and 50 test examples. In total, the Arabic dataset consists of 908 queries, 1,089 qrels, and a collection of 30,329 verified claims. For English, we provide 999 training, 200 dev, 202 dev-test and 209 test examples. In total, the English dataset consists of 1,610 queries, 1,610 qrels, and a collection of 13,835 verified claims.

[4] http://cloud.google.com/translate
[5] CT–VCR–22 stands for `CheckThat!` verified claim retrieval 2022.

3.2. Evaluation

For the ranking tasks, as in the two previous editions of the CheckThat! lab, we calculated *Mean Average Precision* (MAP), reciprocal rank, Precision@k ($P@k$) and MAP@k for $k \in \{1, 3, 5, 10, 20, 30\}$. We used MAP@5 as the official evaluation measure.

3.3. Overview of the Systems

A total of six teams participated in this task. One team participated in the Arabic task and six teams participated in the English task. Below, we discuss briefly the approaches used for system development by each team.

Team AI Rational [32] (2A-en: 2) experimented with a architecture that combines semantic, lexical and re-ranking modules and discovered that for the MAP@k and P@k measures the reranking task is equivalent to a classification one. Therefore, previously used re-rankers for this architecture like RankSVM and LambdaMART can be reduced to a classifier like a basic SVM. More specifically a pretrained SBERT, ElasticSearch and a SVM were used respectively as implementations of the stated above modules by the team.

Team BigIR [33] (2A-en: 3) used the same system proposed in [33] without any further fine-tuning. In other words, the pre-trained model was only fine-tuned on the CheckThat! 2021 dataset only [34], indicating the proposed system is performing well although it was not fine-tuned on the 2022 dataset. BigIR's system involves three steps. First, preprocessing in which the tweet is preprocessed and expanded with helpful information out of URLs, images, and videos. The second step is retrieving an initial list using a simple lexical retrieval model like BM25. Finally, reranking the initial list using a BERT-based model after fine-tuning it for this task. For English subtask, bigIR used MPNet model, and for Arabic, they used AraBERT. BigIR's system for Arabic did not perform better than random baseline, which is 0.0, therefore, we do not report the results for Arabic.

Team SimBa [35] (2A-en: 4 2B-en: 1) preprocessed the input claims by removing URLs, @-symbols and user information. They experimented with both unsupervised and supervised methods with blocking and balancing but found their primary submission, an unsupervised approach, to be most successful. For this, they generated sentence embeddings for all input claims and all verified claims using the sentence embedding models "Sentence-BERT" and "SimCSE", calculated the cosine similarity for all possible pairs of input and verified claims and averaged the two different similarity scores into one. Additionally, they computed the count of similar tokens without stop words and added it to the score. Finally, the five most similar verified claims for each input claim were computed based on the similarity score.

Team RIET Lab [36] (2A-en: 1) team created a pipeline for claim matching by using a sentence transformer (sentence-t5) for candidate selection and a generative model (gpt-neo)[37] for re-ranking. For finetuning the candidate selection model, they used an MNR loss with hard negatives via BM25. For the generative reranking step, they finetune an autoregressive language model using a new objective that heavily regularizes on mutual information from both a

Team	Languages		Transformers					Misc	
	Arabic	English	ARABERT	ColBERT	GPT-Neo	SBERT	ST5	Data augmentation	Preprocessing
AI Rational [32]		2				☑			
BigIR [33]		3	☑			☑			☑
Fraunhofer SIT [38]		6				☑		☑	☑
motlogelwan		5			☑				
RIET Lab [36]		1			☑		☑		☑
SimBa [35]	1	4				☑			☑

Table 2
Overview of the approaches to subtasks 2A. ☑=part of the official submission; ✔=considered in internal experiments.

likelihood and posterior perspective. The model yields high precision and due to its generative nature can also give analysts a better idea of confidence, which is important for fact-checking.

Team Fraunhofer SIT [38] (2A-en:6) proposed an ensemble classification approach. It uses state-of-the-art sentence transformers for estimating the semantic similarity between a given tweet and collection of previously fact-checked tweets with claims. Furthermore, it incorporates several preprocessing steps as well as back-translation as a data augmentation technique.

3.4. Results

Table 3 shows the official results for Task 2A English for all participated teams. We do not report results for Arabic as the scores are zero for both random baseline and the submitted system.

Arabic Team **bigIR** submitted a run for this subtask, however, they have not submitted working note. They used AraBERT to rerank a list of candidates retrieved by a BM25 model. Their approach consists of three main steps such as preprocessing, retrieving an initial list using BM25 and finally reranking the initial list using an AraBERT-based model. As with the random baseline, since the system did not match any input with the verified claims, the performance end up being 0.0.

English Six teams participated, submitting a total of thirty-two runs. All teams improved over the random baseline. Team **RIET Lab** [36] submitted the top run, based on a sentence transformer (sentence-t5) for candidate selection and a generative model (gpt-neo [37]) for re-ranking. Team **AI Rational** [32] ranked second, using a pretrained SBERT, ElasticSearch, and an SVM.

Table 3
Task 2A and 2B: Official evaluation results, in terms of MRR, MAP@k, and Precision@k. The teams are ranked by the official evaluation measure: MAP@5. Here, *Baseline* refers to the random baseline.

Team	MRR	MAP				Precision		
		@1	@3	@5	@10	@3	@5	@10
Task 2A: English								
1. RIET Lab [36]	0.957	0.943	0.955	0.956	0.956	0.322	0.194	0.098
2. AI Rational [32]	0.922	0.904	0.919	0.922	0.922	0.313	0.190	0.095
3. BigIR[33]	0.923	0.900	0.921	0.921	0.921	0.316	0.189	0.095
4. SimBa [35]	0.907	0.876	0.905	0.907	0.907	0.314	0.190	0.095
5. motlogelwan*	0.878	0.833	0.870	0.873	0.876	0.306	0.187	0.095
6. Fraunhofer SIT [38]	0.624	0.557	0.601	0.610	0.617	0.221	0.141	0.075
Task 2B: English								
SimBa [35]	0.475	0.408	0.446	0.459	0.459	0.190	0.126	0.063

4. Subtask 2B: Detecting Previously Fact-Checked Claims in Political Debates or Speeches

Given a claim in a political debate or a speech, the task asks to detect whether the claim has been previously fact-checked with respect to a collection of previously fact-checked claims. This is also a ranking task, and it was offered in English.

4.1. Dataset

We have 752 claims from political debates [1], matched against 869 verified claims (some input claims match more than one verified claim) in a collection of 20,771 verified claims in PolitiFact. We report some statistics about the dataset in the last column of Table 1.

4.2. Evaluation

Similarly to subtask-2A, we treat this as a ranking task, and we report the same evaluation measures. Once again, MAP@5 is the official evaluation measure.

4.3. Overview of the Systems

Team SimBa [35] (2B-en: 1) submitted a total of four runs. The computed different kinds of similarities between input and verified claims, including the cosine-similarity of sentence embeddings and different lexical similarity measures. They made use of a blocking approach to filter dissimilar pairs that can easily be excluded based on sentence embedding based similarity scores, training and applying their classifier only to distinguish between harder cases. For this, they considered the union of the 50 most similar pairs of input and verified claims regarding the similarity scores of four different sentence embedding methods (“Sentence-BERT”, “SimCSE”, “Universal Sentence Encoder” and “InferSent”). Their feature set consisted of “SimCSE”-similarity, Jaccard Distance, count and ratio of similar tokens and “WordNet”-synonyms. A linear support vector classifier was trained on the training data and predicted if a verified claim was relevant.

4.4. Results

Table 3 shows the official results for Task 2B, which was offered in English only. The table does not report the random baseline results as scores are zero for all metrics.

5. Conclusion and Future Work

We have provided a detailed overview of the CLEF 2022 CheckThat! lab task 2, which focused on detecting previously fact-checked claims in tweets (Subtask 2A), and in political debates or speeches (Subtask 2B). Inline with the general mission of CLEF, we promoted multi-linguality by offering the task in two different languages: Arabic and English. The participating systems fine-tuned transformer models, such as sentence BERT, ST5 and GPT-Neo, and some used data augmentation. For subtask 2A, six systems (one for Arabic and six for English) participated, and all outperformed a random baseline. For Subtask 2B, one participating team could beat the random baseline. In the future, we are considering targeting other tasks, which could play a relevant role in the analysis of journalistic and social media posts, besides the explicit factuality decision. We are considering both coverage bias in the news and subjectivity, among others.

Acknowledgments

This work is also part of the Tanbih mega-project,[6] developed at the Qatar Computing Research Institute, HBKU, which aims to limit the impact of “fake news”, propaganda, and media bias by making users aware of what they are reading, thus promoting media literacy and critical thinking.

References

[1] S. Shaar, N. Babulkov, G. Da San Martino, P. Nakov, That is a known lie: Detecting previously fact-checked claims, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL ’20, 2020, pp. 3607–3618.

[2] N. Vo, K. Lee, Where are the facts? searching for fact-checked information to alleviate the spread of fake news, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, 2020, pp. 7717–7731.

[3] 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, M. Wiegand, M. Siegel, J. Köhler, Overview of the CLEF-2022 CheckThat! lab on fighting the COVID-19 infodemic and fake news detection, in: Proceedings of the 13th International Conference of the CLEF Association: Information

[6] http://tanbih.qcri.org

Access Evaluation meets Multilinguality, Multimodality, and Visualization, CLEF '2022, Bologna, Italy, 2022.

[4] P. Nakov, G. Da San Martino, F. Alam, S. Shaar, H. Mubarak, N. Babulkov, Overview of the CLEF-2022 CheckThat! lab task 2 on detecting previously fact-checked claims, in: Working Notes of CLEF 2022—Conference and Labs of the Evaluation Forum, CLEF '2022, Bologna, Italy, 2022.

[5] J. Köhler, G. K. Shahi, J. M. Struß, M. Wiegand, M. Siegel, T. Mandl, M. Schütz, Overview of the CLEF-2022 CheckThat! lab task 3 on fake news detection, in: Working Notes of CLEF 2022—Conference and Labs of the Evaluation Forum, CLEF '2022, Bologna, Italy, 2022.

[6] A. Barrón-Cedeño, T. Elsayed, P. Nakov, G. Da San Martino, M. Hasanain, R. Suwaileh, F. Haouari, N. Babulkov, B. Hamdan, A. Nikolov, S. Shaar, Z. Sheikh Ali, Overview of CheckThat! 2020: Automatic identification and verification of claims in social media, LNCS (12260), Springer, 2020, pp. 215–236.

[7] T. Elsayed, P. Nakov, A. Barrón-Cedeño, M. Hasanain, R. Suwaileh, G. Da San Martino, P. Atanasova, Overview of the CLEF-2019 CheckThat!: Automatic identification and verification of claims, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, LNCS, 2019, pp. 301–321.

[8] Y. Li, J. Gao, C. Meng, Q. Li, L. Su, B. Zhao, W. Fan, J. Han, A survey on truth discovery, SIGKDD Explor. Newsl. 17 (2016) 1–16.

[9] K. Shu, A. Sliva, S. Wang, J. Tang, H. Liu, Fake news detection on social media: A data mining perspective, SIGKDD 19 (2017) 22–36.

[10] D. M. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild, M. Schudson, S. A. Sloman, C. R. Sunstein, E. A. Thorson, D. J. Watts, J. L. Zittrain, The science of fake news, Science 359 (2018) 1094–1096.

[11] S. Vosoughi, D. Roy, S. Aral, The spread of true and false news online, Science 359 (2018) 1146–1151.

[12] N. Vo, K. Lee, The rise of guardians: Fact-checking URL recommendation to combat fake news, in: Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2018, 2018, pp. 275–284.

[13] 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, 2015, pp. 1835–1838.

[14] I. Augenstein, C. Lioma, D. Wang, L. Chaves Lima, C. Hansen, C. Hansen, J. G. Simonsen, MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, 2019, pp. 4685–4697.

[15] J. Thorne, A. Vlachos, Automated fact checking: Task formulations, methods and future directions, in: COLING, Association for Computational Linguistics, 2018, pp. 3346–3359. URL: http://www.aclweb.org/anthology/C18-1283.

[16] S. Shaar, A. Nikolov, N. Babulkov, F. Alam, A. Barrón-Cedeño, T. Elsayed, M. Hasanain, R. Suwaileh, F. Haouari, G. Da San Martino, P. Nakov, Overview of CheckThat! 2020 English: Automatic identification and verification of claims in social media, in: [39], 2020.

[17] M. Hasanain, F. Haouari, R. Suwaileh, Z. Ali, B. Hamdan, T. Elsayed, A. Barrón-Cedeño,

G. Da San Martino, P. Nakov, Overview of CheckThat! 2020 Arabic: Automatic identification and verification of claims in social media, in: [39], 2020.

[18] P. Nakov, G. Da San Martino, T. Elsayed, A. Barrón-Cedeño, R. Míguez, S. Shaar, F. Alam, F. Haouari, M. Hasanain, N. Babulkov, A. Nikolov, G. Kishore Shahi, J. Maria Struß, T. Mandl, The CLEF-2021 CheckThat! Lab on detecting check-worthy claims, previously fact-checked claims, and fake news, in: ECIR, 2021, pp. 639–649.

[19] P. Arnold, The challenges of online fact checking, Technical Report, Full Fact, 2020.

[20] S. Shaar, F. Alam, G. D. S. Martino, P. Nakov, The role of context in detecting previously fact-checked claims, arXiv:2104.07423 (2021).

[21] S. Shaar, F. Alam, G. D. S. Martino, P. Nakov, Assisting the human fact-checkers: Detecting all previously fact-checked claims in a document, arXiv preprint arXiv:2109.07410 (2021). arXiv:2109.07410.

[22] N. Hassan, G. Zhang, F. Arslan, J. Caraballo, D. Jimenez, S. Gawsane, S. Hasan, M. Joseph, A. Kulkarni, A. K. Nayak, V. Sable, C. Li, M. Tremayne, ClaimBuster: The first-ever end-to-end fact-checking system, Proceedings of VLDB Endow. 10 (2017) 1945–1948.

[23] P. Nakov, D. Corney, M. Hasanain, F. Alam, T. Elsayed, A. Barrón-Cedeño, P. Papotti, S. Shaar, G. Da San Martino, Automated fact-checking for assisting human fact-checkers, in: Proceedings of the 30th International Joint Conference on Artificial Intelligence, IJCAI '21, 2021, pp. 4551–4558.

[24] Q. Sheng, J. Cao, X. Zhang, X. Li, L. Zhong, Article reranking by memory-enhanced key sentence matching for detecting previously fact-checked claims, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 5468–5481. URL: https://aclanthology.org/2021.acl-long.425. doi:10.18653/v1/2021.acl-long.425.

[25] J. Si, D. Zhou, T. Li, X. Shi, Y. He, Topic-aware evidence reasoning and stance-aware aggregation for fact verification, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 1612–1622. URL: https://aclanthology.org/2021.acl-long.128. doi:10.18653/v1/2021.acl-long.128.

[26] A. Kazemi, K. Garimella, D. Gaffney, S. Hale, Claim matching beyond English to scale global fact-checking, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 4504–4517. URL: https://aclanthology.org/2021.acl-long.347. doi:10.18653/v1/2021.acl-long.347.

[27] K. Jiang, R. Pradeep, J. Lin, Exploring listwise evidence reasoning with t5 for fact verification, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), Association for Computational Linguistics, Online, 2021, pp. 402–410. URL: https://aclanthology.org/2021.acl-short.51. doi:10.18653/v1/2021.acl-short.51.

[28] H. Wan, H. Chen, J. Du, W. Luo, R. Ye, A DQN-based approach to finding precise evidences

for fact verification, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Association for Computational Linguistics, Online, 2021, pp. 1030–1039. URL: https://aclanthology.org/2021.acl-long.83. doi:10.18653/v1/2021.acl-long.83.

[29] Z. S. Ali, W. Mansour, T. Elsayed, A. Al-Ali, AraFacts: The first large arabic dataset of naturally occurring claims, in: Proceedings of the Sixth Arabic Natural Language Processing Workshop, 2021, pp. 231–236.

[30] A. Tchechmedjiev, P. Fafalios, K. Boland, M. Gasquet, M. Zloch, B. Zapilko, S. Dietze, K. Todorov, ClaimsKG: A knowledge graph of fact-checked claims, in: International Semantic Web Conference, Springer, 2019, pp. 309–324.

[31] S. Shaar, F. Haouari, W. Mansour, M. Hasanain, N. Babulkov, F. Alam, G. Da San 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, 2021.

[32] V. Kostov, AI Rational at CheckThat! 2022: reranking previously fact-checked claims on semantic and lexical similarity, in: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, CLEF '2022, Bologna, Italy, 2022.

[33] W. Mansour, T. Elsayed, A. Al-Ali, Did i see it before? detecting previously-checked claims over twitter, in: European Conference on Information Retrieval, Springer, 2022, pp. 367–381.

[34] P. Nakov, D. S. M. Giovanni, 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, LNCS (12880), Springer, 2021.

[35] A. Hövelmeyer, K. Boland, S. Dietze, SimBa at CheckThat! 2022: lexical and semantic similarity based detection of verified claims in an unsupervised and supervised way, in: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, CLEF '2022, Bologna, Italy, 2022.

[36] S. D.-H. Michael Shliselberg, RIET Lab at CheckThat! 2022: improving decoder based re-ranking for claim matching, in: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, CLEF '2022, Bologna, Italy, 2022.

[37] S. Black, L. Gao, P. Wang, C. Leahy, S. Biderman, GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow, 2021. URL: https://doi.org/10.5281/zenodo.5297715. doi:10.5281/zenodo.5297715.

[38] R. A. Frick, I. Vogel, Fraunhofer SIT at CheckThat! 2022: ensemble similarity estimation for finding previously fact-checked claims, in: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, CLEF '2022, Bologna, Italy, 2022.

[39] L. Cappellato, C. Eickhoff, N. Ferro, A. Névéol (Eds.), CLEF 2020 Working Notes, CEUR Workshop Proceedings, 2020.