Overview of the CLEF–2022 CheckThat! Lab on Fighting the COVID-19 Infodemic and Fake News Detection

Nakov, Preslav; Barrón-Cedeño, Alberto; da San Martino, Giovanni; Alam, Firoj; Struß, Julia Maria; Mandl, Thomas; Míguez, Rubén; Caselli, Tommaso; Kutlu, Mucahid; Zaghouani, Wajdi

Published in:
Experimental IR Meets Multilinguality, Multimodality, and Interaction - 13th International Conference of the CLEF Association, CLEF 2022, Proceedings

DOI:
10.1007/978-3-031-13643-6_29

Overview of the CLEF–2022 CheckThat! Lab on Fighting the COVID-19 Infodemic and Fake News Detection

Preslav Nakov[1(✉)], Alberto Barrón-Cedeño[2], Giovanni da San Martino[3], Firoj Alam[4], Julia Maria Struß[5], Thomas Mandl[6], Rubén Míguez[7], Tommaso Caselli[8], Mucahid Kutlu[9], Wajdi Zaghouani[10], Chengkai Li[11], Shaden Shaar[12], Gautam Kishore Shahi[13], Hamdy Mubarak[4], Alex Nikolov[14], Nikolay Babulkov[15], Yavuz Selim Kartal[16], Michael Wiegand[17], Melanie Siegel[18], and Juliane Köhler[5]

[1] Mohamed Bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
preslav.nakov@mbzuai.ac.ae
[2] Università di Bologna, Forlì, Italy
a.barron@unibo.it
[3] University of Padova, Padova, Italy
dasan@math.unipd.it
[4] Qatar Computing Research Institute, HBKU, Ar-Rayyan, Qatar
{fialam,hmubarak}@hbku.edu.qa
[5] University of Applied Sciences Potsdam, Potsdam, Germany
{julia.struss,juliane.koehler}@fh-potsdam.de
[6] University of Hildesheim, Hildesheim, Germany
mandl@uni-hildesheim.de
[7] Newtral Media Audiovisual, Madrid, Spain
{ruben.miguez,javier.beltran}@newtral.es
[8] University of Groningen, Groningen, Netherlands
t.caselli@rug.nl
[9] TOBB University of Economics and Technology, Ankara, Turkey
m.kutlu@etu.edu.tr
[10] Hamad Bin Khalifa University, Ar-Rayyan, Qatar
wzaghouani@hbku.edu.qa
[11] University of Texas at Arlington, Arlington, USA
cli@uta.edu
[12] Cornell University, Ithaca, USA
ss2753@cornell.edu
[13] University of Duisburg-Essen, Duisburg, Germany
gautam.shahi@uni-due.de
[14] CheckStep, London, UK
alex.nikolov@checkstep.com
[15] Sofia University, Sofia, Bulgaria
[16] GESIS-Leibniz Institute for the Social Sciences, Mannheim, Germany
yavuzselim.kartal@gesis.org
[17] University of Klagenfurt, Klagenfurt, Austria
michael.wiegand@aau.at
[18] Darmstadt University of Applied Sciences, Darmstadt, Germany
melanie.siegel@h-da.de

A. Barrón-Cedeño et al. (Eds.): CLEF 2022, LNCS 13390, pp. 495–520, 2022.
https://doi.org/10.1007/978-3-031-13643-6_29

Abstract. We describe the fifth edition of the CheckThat! lab, part of the 2022 Conference and Labs of the Evaluation Forum (CLEF). The lab evaluates technology supporting tasks related to factuality in multiple languages: Arabic, Bulgarian, Dutch, English, German, Spanish, and Turkish. Task 1 asks to identify relevant claims in tweets in terms of check-worthiness, verifiability, harmfullness, and attention-worthiness. Task 2 asks to detect previously fact-checked claims that could be relevant to fact-check a new claim. It targets both tweets and political debates/speeches. Task 3 asks to predict the veracity of the main claim in a news article. CheckThat! was the most popular lab at CLEF-2022 in terms of team registrations: 137 teams. More than one-third (37%) of them actually participated: 18, 7, and 26 teams submitted 210, 37, and 126 official runs for tasks 1, 2, and 3, respectively.

Keywords: Fact-Checking · Disinformation · Misinformation · Check-Worthiness · Verified Claim Retrieval · Fake News · COVID-19

1 Introduction

The mission of the `CheckThat!` lab is to foster the development of technology to assist in the process of fact-checking claims made in political debates, social media posts and news articles. The five editions of the lab have been held in 2018–2022, targeting various natural language processing and information retrieval tasks related to factuality [11,12,24,25,61,62,65,66]. The aim is to develop systems that can be useful as supportive technology for investigative journalism, as they could provide help and guidance, thus saving time [30,35,37,42,63,76,97]. For example, a system could automatically identify check-worthy claims, make sure they have not been fact-checked already by a reputable fact-checking organization, and then present them to a journalist for further analysis in a ranked list [83]. In addition, we can develop systems to identify whether documents are potentially *useful* for human fact-checkers to verify a claim [63,100], and it could also estimate a *veracity score* supported by evidence to increase the journalist's understanding and trust in the system's decision [82].

`CheckThat!` at CLEF 2022 is the fifth edition of the lab [61], and aims to foster the technology on three timely problems in multiple languages: Arabic, Bulgarian, Dutch, English, German, Spanish, and Turkish. Task 1 asks to detect relevant tweets: check-worthy, verifiable, harmful, and attention-worthy. Task 2 aims at detecting previously fact-checked claims in tweets, political debates and speeches. Task 3 focuses on checking the veracity of news articles.

2 Previously on `CheckThat!`

The tasks in the current edition of `CheckThat!` are a follow up or reformulations of those from 2021 [27,65,66], where the focus was on (*i*) *tweets*, (*ii*) *political debates and speeches*, and (*iii*) *news articles*. It featured five languages: Arabic, Bulgarian, English, Spanish, and Turkish. Next, we include a brief overview of the most successful approaches explored in the tasks of that edition.

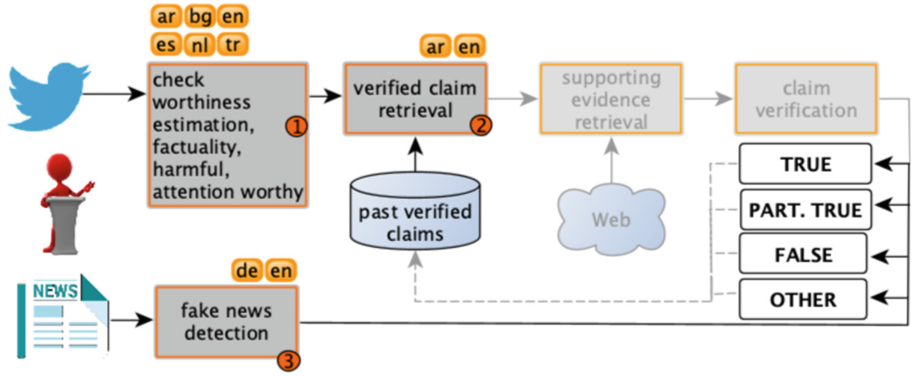

Fig. 1. The full verification pipeline. The 2022 lab covers three tasks from that pipeline: (*i*) check-worthiness estimation, (*ii*) verified claim retrieval, and (*iii*) fake news detection. The gray tasks were addressed in previous editions of the lab [12,25].

Task 1_{2021}. Determine whether a piece of text is worth fact-checking [85]. The most successful submissions used BERT, AraBERT, and RoBERTa [85,99], and some systems used WordNet [101] and LIWC [80].

Task 2_{2021}. Given a check-worthy claim in the form of a tweet, and a set of previously fact-checked claims, rank these previously fact-checked claims in order of their usefulness to fact-check that new claim [84]. The most successful approaches were based on AraBERT, RoBERTa, and Sentence-BERT [18,55,74].

Task 3_{2021}. Given the text and the title of a news article, determine whether the main claim it makes is true, partially true, false, or other. Also, identify the domain of the article: health, crime, climate, elections, or education [88]. The task was offered in English. The most successful pre-trained language model was RoBERTa [9,20,44]. Ensembles were also popular, with components using BERT [44] and LSTMs [20,44].

Previous editions of the lab had targeted other tasks of the verification pipeline (cf. Fig. 1) on different kinds of texts. **The 2020 edition** featured three main tasks: detecting previously fact-checked claims, evidence retrieval, and actual fact-checking of claims [11,12]. The major focus was on Twitter. **The 2019 edition** covered the various modules necessary to verify a claim: from check-worthiness, to ranking and classification of evidence in the form of Web pages, to actual fact-checking of claims against specific text snippets [24,25]. **The 2018 edition** of the lab focused on check-worthiness and fact-checking of claims in political debates [62].

3 Description of the Tasks

The 2022 edition of the `CheckThat!` lab is organized around three tasks, each of which in turn has several subtasks. Figure 1 shows the full `CheckThat!` verification pipeline, and the three tasks we target this year are highlighted.

Table 1. Class labels for Subtasks 1A, 1B, 1C, and 1D.

Subtask 1A	Subtask 1C	Subtask 1D	
1. No	1. No	1. No	6. Yes, contains advice
2. Yes	2. Yes	2. Yes, asks question	7. Yes, discusses action taken
Subtask 1B		3. Yes, blame authorities	8. Yes, discusses cure
1. No		4. Yes, calls for action	9. Yes, other
2. Yes		5. Yes, Harmful	

3.1 Task 1: Identifying Relevant Claims in Tweets

The aim of Task 1 is to determine whether a claim in a tweet is worth fact-checking. In order to do that, we either resort to the judgments of professional fact-checkers or we ask human annotators to answer several auxiliary questions [3,4], such as "does it contain a verifiable factual claim?", "is it harmful?" and "is it of general interest?", before deciding on the final check-worthiness label. Tasks 1A to 1C are all binary problems and the models are expected to establish whether a tweet is relevant according to different criteria. Task 1D is a multi-class problem. Table 1 shows the class labels for all four subtasks. Regarding languages, Arabic, Bulgarian, Dutch, English, and Turkish are present in all four subtasks, whereas Spanish is included only in Subtask 1A. The participants were free to work on any language(s) of their choice, and they could also use multilingual approaches that make use of all datasets for training.

Subtask 1A: Check-worthiness of tweets. Given a tweet, predict whether it is worth fact-checking.

Subtask 1B: Verifiable factual claims detection. Given a tweet, predict whether it contains a verifiable claim or not.

Subtask 1C: Harmful tweet detection. Given a tweet, predict whether it is harmful to the society.

Subtask 1D: Attention-worthy tweet detection. Given a tweet, predict whether it should get the attention of policy makers and why. Table 1 shows the nine classes. More details of the label definitions can be found in [4].

3.2 Task 2: Detecting Previously Fact-Checked Claims

Given a check-worthy claim, and a set of previously-checked claims, determine whether the claim has been previously fact-checked with respect to a collection of fact-checked claims. Both subtasks are ranking problems, where systems are asked to produce a list of top-n candidates. Subtask 2A focuses on tweets and was offered in both Arabic and English. Subtask 2B focuses on political debates and speeches and was given only in English.

Subtask 2A: Detect previously fact-checked claims from tweets. Given a tweet, detect whether the claim it makes was previously fact-checked with respect to a collection of previously fact-checked claims. This is a ranking task, where the systems were asked to produce a list of top-n candidates.

Subtask 2B: Detect previously fact-checked claims in political debates or speeches. Within the context of a political debate or a speech, detect whether a claim has been previously fact-checked with respect to a collection of previously fact-checked claims.

3.3 Task 3: Fake News Detection

Task 3 asks to predict the veracity news articles and is designed as a multi-class classification problem. This task was offered as a monolingual task in English and as a cross-lingual task for English and German (English training data, German test data). The idea of the latter is to use the English data and cross-language representation learning (e.g., [19,71]) to build a classification model for the German language as well.

Task 3: Multi-class fake news detection of news articles. Given the text of a news article, determine whether the claims made in the article are *true*, *partially true*, *false*, or *other* (e.g., claims in dispute).

4 Datasets

Here, we briefly describe the datasets for each of the three tasks. For more details, refer to the task description papers for Task 1 [60], Task 2 [64], and Task 3 [43].

4.1 Task 1: Identifying Relevant Claims in Tweets

For all **1A, 1B, 1C**, and **1D** subtasks and all languages, but Spanish, we used the dataset reported in [4]. The dataset is developed based on a multi-question annotation schema and annotated tweets for Arabic, Bulgarian, Dutch, and English [3]. Following the same annotation schema, a Turkish dataset has also been produced. The dataset reported in [4] comes with a training, development, and test set. For the shared task, we provided the test set as a dev-test set to enable the participants to validate their systems internally, while they can use the dev set for parameter tuning. For each language and subtask, we have annotated new instances, using three or four annotators per instance. Class labels have been assigned by majority voting and disagreements have been solved by a consolidator.

For Spanish, the tweets were manually annotated by journalists from Newtral—a Spanish fact-checking organization—and came from the Twitter accounts of 300 Spanish politicians. The Spanish collection is the largest one compared to the other languages ; more than three times the second largest dataset: 14,990 vs 4,121 for Arabic. However, Spanish is only available for Subtask **1A**.

Table 2 summarizes the data available for each subtask and each language.

Table 2. Task 1 (Identifying Relevant Claims in Tweets): Statistics about the CT–CWT–22 corpus for all six languages. The bottom part of the table shows the main topics covered.

Subtask	Partition	AR	BG	NL	EN	ES	TR	Total
1A	Train	2,513	1,871	923	2,122	4,990	2,417	14,836
	Dev	235	177	72	195	2,500	222	3,401
	Dev-Test	691	519	252	574	2,500	660	5,196
	Test	682	130	666	149	5,000	303	6,930
	Total	4,121	2,697	1,913	3,040	14,990	3,602	
1B	Train	3,631	2,710	1,950	3,324		2,417	14,032
	Dev	339	251	181	307		222	1,300
	Dev-Test	996	736	534	911		660	3,837
	Test	1,248	329	1,358	251		512	3,698
	Total	6,214	4,026	4,023	4,793		3,811	
1C	Train	3,624	2,708	1,946	3,323		2,417	14,018
	Dev	336	250	179	307		222	1,294
	Dev-Test	994	735	531	910		660	3,830
	Test	1,201	325	1,360	251		512	3,649
	Total	6,155	4,018	4,016	4,791		3,811	
1D	Train	3,621	2,710	1,949	3,321		1,904	13,505
	Dev	338	251	179	306		178	1,252
	Dev-Test	995	736	533	909		533	3,706
	Test	1,186	329	1,356	251		465	3,587
	Total	6,140	4,026	4,017	4,787		3,080	
Main topics								
COVID-19		■	■	■	■	■	■	
Politics						■	■	

4.2 Task 2: Detecting Previously Fact-Checked Claims

Subtask 2A: Detecting previously fact-checked claims from tweets. For English, we have 1,610 annotated tweets, each matching a single claim in a set of 13,835 verified claims from Snopes. For Arabic, we have 858 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 include 5,921 Arabic claims from AraFacts [5] and 24,408 English claims from ClaimsKG [91], translated to Arabic using the Google Translate API.[1]

[1] http://cloud.google.com/translate.

Table 3. 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**

Subtask 2B: Detecting previously fact-checked claims in political debates/speeches. We have 752 claims from political debates [83], matched against 869 verified claims (some input claims match more than one verified claim) in a collection of 20,771 verified claims in PolitiFact.

Table 3 shows statistics about the CT–VCR–22 corpus for Task 2, including both subtasks and languages. CT–VCR–22 stands for `CheckThat!` verified claim retrieval 2022. *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 [84].

4.3 Task 3: Fake News Detection

For the creation of the data for Task 3 the AMUSED framework [86] was followed. The starting point for the data collection was finding suitable fact-checking organizations and their websites. On those websites, the authors of the individual articles discuss and rate the truthfulness of claims that are made in different sources. We scraped the links to those sources as well as the judgment about the claim made on the fact-checking sites. To ensure that only news articles are in the corpus, automatic filtering was applied. Thus, all links leading to a social media platform or a non-textual document (e.g., image, video) were deleted. Furthermore, the remaining links were manually checked. During this step, in addition to deleting non-relevant URLs, we examined, if the links actually (still) lead to the claim source and if the document still existed in its original form. Following those quality evaluations, we scraped the title and the full text for each of the remaining articles.

Table 4. Task 3: Statistics about the number of documents and class distribution for the CT-FAN-22 corpus for English and German fake news detection.

Class	EN Training	EN Dev.	EN Test	DE Test
False	465	113	315	191
True	142	69	210	243
Partially false	217	141	56	97
Other	76	41	31	55
Total	900	364	612	586

Task 3 was offered in English and as a cross-language task in German. As training material, we only provided the English data of last year's iteration. Thus, 900 English news articles for training and 364 articles as development set were given to the participants. Those documents were collected from a total of 15 fact-checking websites (e.g., PolitiFact) [88]. Because the German task was intended as a cross-language classification problem, no German training data was necessary. The training data contained an ID for each article as unique identifier, the title of the given target article as well as its full-text, and finally, a label stating the truthfulness of the article. We took the labels from the judgment on the fact-checking sites. Yet, each fact-checking site had their own label inventory if any at all, such as *incorrect*, *inaccurate*, or *misinformation* for *false*. Therefore, we merged the labels with a similar meaning according to [87], leading to the following four classes: *true*, *false*, *partially false* (meaning any mix of false and true information, such as mostly true or mostly false), and *other*.

As test data, we collected 612 English and 586 German articles from a total of 20 fact-checking websites (14 for the English data and 7 for the German data; the AFP website was consulted for both languages). We did not provide any other information (e.g., a link to the article, a publication date, eventual tags, authors, location of publication, etc.). An overview of the different datsets can be found in Table 4. Both training and test data set are available on Zenodo[2].

5 Evaluation

We used different official evaluation metrics, depending on the nature of the tasks at hand and the involved datasets.

Task 1 and Task 3 included both binary and multi-class classification subtasks. For **Subtasks 1A and 1C**, we used the F_1-measure with respect to the positive class (yes), to account for class imbalance. For **Subtask 1B**, we used accuracy, as the data is fairly balanced. For **Subtask 1D**, we used weighted-F_1, as there are multiple classes and we wanted them appropriately weighted.

Task 2 included ranking subtasks. The official measure for both **Subtasks 2A and 2B** was mean-average precision at 5 (MAP@5); these are the same evaluation measures as in the 2021 edition of the `CheckThat!` lab.

For **Task 3**, we used macro F_1-measure, as in the previous iteration of the task.

[2] https://zenodo.org/record/6555293.

6 Results for Task 1: Identifying Relevant Claims in Tweets

Below, we report the evaluation results for Task 1 and its four subtasks for all six languages.

6.1 Task 1A. Check-Worthiness Estimation

A total of 20 teams took part in this task, with English, Bulgarian, and Dutch being the most popular languages. Two teams (TOBB ETU [26] and NUS-IDS [57]) participated in five languages out of six. For all six languages, we had a monolingual random baseline. Table 5 shows the performance of the official submissions—the last valid blind submission by each team—on the test set, in addition to the random baseline. The table shows the runs ranked on the basis of the official F_1 with respect to the positive class and includes all six languages.

Table 5. Task 1A: Check-Worthiness estimation, results for the official submissions in all six languages. F1 with respect to the positive class. Baseline is the random baseline.

Team	F1
Arabic	
1. NUS-IDS [57]	0.628
2. TOBB ETU [26]	0.495
3. iCompass [13]	0.462
4. Baseline	0.347
5. PoliMi-FlatEarthers [2]	0.321
Bulgarian	
1. NUS-IDS [57]	0.617
2. TOBB ETU [26]	0.542
3. AI Rational [77]	0.483
4. Baseline	0.434
5. PoliMi-FlatEarthers [2]	0.341
6. pogs2022*	0.000
Dutch	
1. NUS-IDS [57]	0.642
2. AI Rational [77]	0.620
3. TOBB ETU [26]	0.534
4. PoliMi-FlatEarthers [2]	0.532
5. Z-Index [90]	0.497
6. Baseline	0.451
English	
1. AI Rational [77]	0.698
2. Zorros [16]	0.667
3. PoliMi-FlatEarthers [2]	0.626
4. TOBB ETU [26]	0.561
5. Fraunhofer SIT [29]	0.552
6. RUB-DFL [39]	0.525
7. hinokicrum*	0.522
8. NUS-IDS [57]	0.519
9. TonyTTTTT	0.500
10. Asatya [50]	0.500
11. VTU_BGM [41]	0.482
12. Z-Index [90]	0.478
13. NLP&IR@UNED*	0.469
14. Baseline	0.253
Spanish	
1. NUS-IDS [57]	0.571
2. PoliMi-FlatEarthers [2]	0.323
3. Z-Index [90]	0.303
4. Baseline	0.139
Turkish	
1. RUB-DFL [39]	0.801
2. AI Rational [77]	0.789
3. ARC-NLP [93]	0.760
4. TOBB ETU [26]	0.729
5. Baseline	0.496

*No working note submitted.

Arabic. Four teams participated for Arabic, submitting a total of 12 runs. All participating teams fine-tuned existing pre-trained models, such as BERT, AraBERT, GPT-3 and mT5 models. The top performing system, NUS-IDS [57], used mT5 model, which is a multilingual sequence-to-sequence transformer pre-trained on the mC4 corpus covering 101 languages. They performed both data augmentation and preprocessing. The second best system, TOBB ETU [26], used fine-tuned AraBERT.

Bulgarian. Five teams took part for Bulgarian, submitting a total of six runs. Once again **NUS-IDS** [57] was the top-ranked team, followed by Team **TOBB ETU** [26]. BERT, RoBERTa, DistilBERT and the common pretrained models have been used by all participating systems. Several systems also used data augmentation and standard preprocessing.

Dutch. Five teams participated for Dutch, submitting a total of 11 runs. Team **NUS-IDS** [57] also ranked first, followed by Team **AI Rational** [77] is the second-best system. Across different teams, BERT is the most commonly used pre-trained model. Other pre-trained models include RoBERTa, DistilBERT, and GPT-3. Data augmentation and standard preprocessing have also been used for Dutch.

English. A total of 13 teams took part in task 1A for English, with a total of 59 runs. The top-ranked team was **AI Rational** [77], and they fine-tuned several pre-trained transformers models such as DistilBERT, BERT, RoBERTa. For the system submission they used RoBERTa-large. The second best system—Team **Zorros** [16]—also used BERT and RoBERTa with an ensemble approach.

Spanish. Three teams took part for Spanish, with a total of eight runs. Team **NUS-IDS** [57] is the top-ranked team. Team **PoliMi-FlatEarthers** [2] is second, with a system based on a GPT-3 pretrained model.

Turkish. Four teams participated for Turkish, submitting a total of five runs. All participants used BERT-based models and GPT-3. The top ranked team is **RUB-DFL** [39], which used BERT-based models and LIWC features. The runner up team **AI Rational** applied standard pre-processing and data augmentation with back translation.

6.2 Subtask 1B: Verifiable Factual Claims Detection

Thirteen teams took part in Subtask 1B, with English, Bulgarian and Arabic being the most popular languages. Team **TOBB ETU** [26] participated in all five languages. Team **AI Rational** participated in four languages. Table 6 shows the performance of the official submissions on the test set including the random baseline. The table shows the runs ranked on the basis of the official *accuracy* measure in all five languages for this subtask.

Arabic. Three teams participated in the Arabic factual claim detection subtask, submitting a total of seven runs. The system of team **TOBB ETU** [26] ranked best for this subtask, which uses a four-layer feed-forward network with Manifold Mixup regularization and BERT embeddings.

Bulgarian. Two teams submitted three runs: Team **AI Rational** [77] tops the ranking, followed by Team **TOBB ETU** [26]. AI Rational used XLM-RoBERTa with data augmentation while TOBB ETU used fine-tuned RoBERTa.

Dutch. As for Bulgarian, two teams submitted three runs. Team **AI Rational** [77] and **TOBB ETU** [26] ranked as the first and second systems. Similar approaches (i.e., BERT, RoBERTa, and DistilBERT) have been used.

Table 6. Task 1B: Verifiable Factual Claims Detection, results for the official submissions in all five languages.

Team	Acc	Team	Acc	Team	Acc
Arabic		**English**		**Turkish**	
1. TOBB ETU [26]	0.570	1. PoliMi-FlatEarthers [2]	0.761	1. RUB-DFL [39]	0.801
2. Baseline	0.531	2. Asatya [50]	0.749	3. AI Rational [77]	0.789
3. claeser*	0.454	3. NLP&IR@UNED*	0.725	3. ARC-NLP [93]	0.760
4. pogs2022*	0.454	4. AI Rational [77]	0.713	4. TOBB ETU [26]	0.729
Bulgarian		5. Zorros [16]	0.709	5. Baseline	0.496
1. AI Rational [77]	0.839	6. RUB-DFL [39]	0.709		
2. TOBB ETU [26]	0.742	7. VTU_BGM [41]	0.709		
3. Baseline	0.535	8. hinokicrum*	0.665		
Dutch		9. TOBB ETU [26]	0.641		
1. AI Rational [77]	0.736	10. Baseline	0.494		
2. TOBB ETU [26]	0.658				
3. Baseline	0.521				

*No working note submitted.

English. Nine teams participated with 21 runs. Team **PoliMi-FlatEarthers** [2] ranked as the best system and **Asatya** [50] as the second. The top-performing system used GPT-3, whereas other teams used BERT, RoBERTa, and DistilBERT as pretrained models for fine-tuning.

Turkish. Four teams participated, submitting five runs. The top-ranked team is **RUB-DFL** [39], which used RoBERTa, Electra, and BERTurk pre-trained models. The second-best team is **AI Rational** [77], which used BERT, RoBERTa, and DistilBERT.

6.3 Subtask 1C: Harmful Tweet Detection

Thirteen teams participated in Subtask 1C, with English and Turkish being the most popular languages. Teams TOBB ETU [26] and AI Rational [77] participated in five and four languages, respectively. Table 7 shows the performance of the official submissions on the test set, together with the random baseline. The table shows the runs ranked based on the official *F1 with respect to positive class* for five languages.

Table 7. Task 1C: Harmful Tweet Detection, results for the official submissions in all five languages.

Team	F1	Team	F1	Team	F1
Arabic		**English**		**Turkish**	
1. iCompass [13]	0.557	1. Zorros [16]	0.397	1. ARC-NLP [93]	0.366
2. TOBB ETU [26]	0.268	2. AI Rational [77]	0.361	2. RUB-DFL [39]	0.353
3. Baseline	0.118	3. Asatya [50]	0.361	3. AI Rational [77]	0.346
Bulgarian		4. NLP&IR@UNED*	0.347	4. TOBB ETU [26]	0.262
1. AI Rational [77]	0.286	5. TOBB ETU [26]	0.329	5. Baseline	0.061
2. TOBB ETU [26]	0.054	6. ARC-NLP [93]	0.300		
3. Baseline	0.000	7. hinokicrum*	0.281		
Dutch		8. COURAGE [47]	0.280		
1. TOBB ETU [26]	0.358	9. RUB-DFL [39]	0.273		
2. AI Rational [77]	0.147	10. PoliMi-FlatEarthers [2]	0.270		
3. Baseline	0.114	11. Baseline	0.200		
		12 VTU_BGM [41]	0.000		

*No working note submitted.

Arabic. Two teams participated, submitting a total of 12 runs. Team **iCompass** [13] is the best system, followed by Team **TOBB ETU** [26]. iCompass finetuned the AraBERT and ARBERT pre-trained language models.

Bulgarian. Two teams participated, submitting 4 runs. Team **AI Rational** [77] ranked as the best system using XLM-RoBERTa while the second best system **TOBB ETU** [26] fine-tuned RoBERTa. Both teams applied data augmentation via back-translation.

Dutch. Two teams participated with 3 runs. Team **TOBB ETU** [26] ranked on top and **AI Rational** [77] ranked second. For this subtask, AI Rational used XLM-RoBERTa without data-augmentation while TOBB ETU fine-tuned BERT and applied data-augmentation via back-translation.

English. A total of 11 teams participated with 17 submissions. Team **Zorros** [16] ranked as the best system, using an ensemble of five transformer-based models. Team **ARC-NLP** [93] ranked second. Besides transformer-based models across all approaches, some teams have also used data augmentation.

Turkish. Four teams participated with five runs submitted. Team **ARC-NLP** [93] ranked as the best system by approaching harmful detection as a contradiction detection problem. They first extracted facts related to the COVID-19 pandemic from reliable sources, and then associated tweets with facts based on their textual similarity. Next, they fine-tuned BERTurk using fact and tweet pairs as data instances. The second best system is by Team **RUB-DFL** [39], which fine-tuned ConvBert with standard pre-processing.

6.4 Subtask 1D: Attention-Worthy Tweet Detection

Seven teams participated in subtask 1D, with English being the most popular language. As for subtask 1C, teams TOBB ETU [26] and AI Rational [77] participated in five and four languages, respectively. Table 8 shows the performance of the official submissions on the test, together with the random baseline. The ranking is based on the official *weighted F1*.

Arabic. Only one team participated. The random baseline outperformed feed-forward network with BERT embeddings and Manifold Mixup regularization proposed by team **TOBB ETU** [26].

Bulgarian. Two teams participated, submitting 4 runs. Team **AI Rational** [77] ranked on top whereas **TOBB ETU** [26] arrived second. While AI Rational used the same transformer-based model in Subtask 1C, TOBB ETU utilized a manifold mixup approach.

Dutch. Two teams participated, making three runs. As for Bulgarian, teams **AI Rational** [77] and **TOBB ETU** [26] ranked first and second.

English. Six teams participated with a total of 14 runs. Team **Zorros** [16] ranked first, by fine-tuning a COVID Twitter BERT pre-trained model. The random baseline ranked second.

Table 8. Task 1D: Attention-Worthy Tweet Detection, results for the official submissions in all five languages. Performance is reported as weighted F1.

Team	F1	Team	F1	Team	F1
Arabic		**English**		**Turkish**	
1. Baseline	0.206	1. Zorros [16]	0.725	1. AI Rational [77]	0.895
2. TOBB ETU [26]	0.184	2. Baseline	0.695	2. Baseline	0.853
Bulgarian		3. AI Rational [77]	0.684	3. TOBB ETU [26]	0.806
1. AI Rational [77]	0.915	4. TOBB ETU [26]	0.670	**Dutch**	
2. TOBB ETU [26]	0.877	5. NLP&IR@UNED*	0.650	1. AI Rational [77]	0.715
3. Baseline	0.875	6. hinokicrum*	0.643	2. TOBB ETU [26]	0.694
		7. PoliMi-FlatEarthers [2]	0.636	3. Baseline	0.641

*No working note submitted.

Turkish. Two teams participated, with three runs. Team **AI Rational** [77] ranked on top, followed by a random baseline.

7 Results for Task 2: Verified Claim Retrieval

Six teams took part in Task 2. Subtask 2A was more popular than subtask 2B. Only team **SimBa** took part in both subtasks, whereas team **BigIR** was the only one that participated in both languages.

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

Table 9 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. 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** [54] submitted the top run, based on a sentence transformer (sentence-t5) for candidate selection and a generative model (gpt-neo [14]) for re-ranking. Team **AI Rational** ranked second, using a pretrained SBERT, ElasticSearch, and an SVM.

Table 9. 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 [54]	0.957	0.943	0.955	0.956	0.956	0.322	0.194	0.098
2. AI Rational	0.922	0.904	0.919	0.922	0.922	0.313	0.190	0.095
3. BigIR [51]	0.923	0.900	0.921	0.921	0.921	0.316	0.189	0.095
4. SimBa [38]	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 [28]	0.624	0.557	0.601	0.610	0.617	0.221	0.141	0.075
Task 2B: English								
SimBa [38]	0.475	0.408	0.446	0.459	0.459	0.190	0.126	0.063

7.2 Subtask 2B: Detecting Previously Fact-Checked Claims in Political Debates and Speeches

Table 9 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.

Team **SimBa** [38] submitted a total of four runs. They computed different kinds of similarities between input claims and verified claims, including the cosine between sentence embeddings and different lexical similarity metrics. 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.

8 Results for Task 3: Fake News Detection

In this section, we present the results of the evaluation for Task 3 and for each of the two languages, English (monolingual subtask) and German (cross-language subtask). Each team could submit up to 200 runs. Yet, only the last submission was taken into account for the evaluation. In total, there were 32 teams submitting runs for the English and 14 for the German task. Runs which were either incorrectly formatted or consisted of incomplete files were rejected, resulting in 25 and 8 runs for the English and German subtasks, respectively.

As in the 2021 edition [88], most experiments involved deep learning models (16 teams), especially applications of BERT (12 teams), RoBERTa (6 teams) or other BERT variations (8 teams) and one of the publicly available BERT language models. However, almost as many teams (14 teams) experimented with feature-based supervised-learning approaches as well. Examples are SVMs (10 teams), logistic regression (9 teams), random forests (8 teams) and naïve bayes (7 teams). Yet, the majority merely fine-tuned a pre-trained language model and only very few experimented with other approaches.

English. Last year, the best submission made extensive use of external data resources [88]. This year, in total, 12 teams worked with additional English, and one team with additional German training data that was not provided by the organizers of this task. The best submission for the monolingual subtask was by team **iCompass** (macro-averaged F_1: 0.339). They applied *bert-base-uncased* and fine-tuned their model. They also experimented with RoBERTa for which they got worse results. No additional external resources were employed in the final classifier.

The second-best submission, by team **NLP&IR@UNED** (macro-averaged F_1: 0.332), made use of an ensemble of classifiers. It was built out of a Funnel Transformer and a Feed Forward Neural Network. The features were extracted by the *LIWC* text analysis tool.

Overall, all teams had a macro-averaged F_1 score lower than 0.5. Table 10 shows a complete overview of the teams and their results. The baseline system [79], a standard bert-base-cased model from HuggingFace, was made available to the participants at the beginning of the lab cycle.

German. Eight teams attempted to solve the second subtask, which was the English–German cross-language setting. Team **ur-iw-hn** was the team with the most successful submission (macro-averaged F_1: 0.290). They translated the first 5,000 tokens of an article from the German test data using the service

Table 10. Task 3 English: Official evaluation results for English Fake News Detection ranked by the macro-F_1 score, including the F_1 scores for individual classes and the overall accuracy

	Team	True	False	Partially False	Other	Accuracy	Macro-F1
1	iCompass [89]	0.383	0.721	0.173	0.080	0.547	0.339
2	NLP&IR@UNED [52]	0.446	0.729	0.097	0.057	0.541	0.332
3	Awakened [95]	0.328	0.744	0.185	0.035	0.531	0.323
4	UNED	0.346	0.725	0.191	0.000	0.544	0.315
	Baseline	0.244	0.701	0.157	0.144	0.480	0.312
5	NLytics [75]	0.339	0.707	0.184	0.000	0.513	0.308
6	SCUoL [6]	0.377	0.709	0.133	0.000	0.526	0.305
7	NITK-IT_NLP [34]	0.325	0.734	0.133	0.000	0.536	0.298
8	CIC [7]	0.111	0.682	0.215	0.136	0.475	0.286
9	ur-iw-hnt [94]	0.290	0.733	0.110	0.000	0.533	0.283
10	BUM [46]	0.207	0.694	0.140	0.063	0.472	0.276
11	boby232	0.255	0.676	0.126	0.045	0.475	0.275
12	HBDCI [17]	0.177	0.708	0.209	0.000	0.508	0.273
13	DIU_SpeedOut	0.195	0.706	0.182	0.000	0.521	0.271
14	DIU_Carbine	0.192	0.626	0.157	0.056	0.472	0.258
15	CODE [15]	0.126	0.662	0.203	0.029	0.444	0.255
16	MNB	0.160	0.701	0.142	0.000	0.507	0.251
17	subMNB	0.160	0.701	0.142	0.000	0.507	0.251
18	FoSIL [48]	0.141	0.670	0.169	0.022	0.462	0.251
19	TextMinor [45]	0.250	0.555	0.086	0.048	0.377	0.235
20	DLRG	0.009	0.694	0.092	0.000	0.513	0.199
21	DIU_Phoenix	0.420	0.040	0.092	0.000	0.278	0.159
22	AIT_FHSTP [78]	0.280	0.146	0.154	0.039	0.199	0.155
23	DIU_SilentKillers	0.407	0.070	0.135	0.000	0.260	0.153
24	DIU_Fire71	0.430	0.006	0.094	0.000	0.275	0.133
25	AI Rational	0.296	0.000	0.196	0.090	0.098	0.117

of Google Translate. They applied an extractive summarization techniques and a $BERT_{Large}$ model for the multi-class classification.

Team **NITK-IT_NLP**, which was the first runner up, divided the news article into windows of 500 tokens. Those windows were shifted over the text to avoid losing context. They experimented with different transformer models, with an *mDeBERTa* model yielding the best results. Table 11 shows the individual results of all eight submissions. Again, the baseline [79] (macro-averaged F_1 score 0.242) results are listed in the table as well. The baseline translated the German articles into English to classify them in accordance to the monolingual subtask.

Table 11. Task 3 German: Official evaluation results for German Fake News Detection ranked by the macro-F_1 score, including the F_1 scores for individual classes and the overall accuracy

	Team	True	False	Partially False	Other	Accuracy	Macro-F1
1	ur-iw-hnt [94]	0.401	0.536	0.189	0.033	0.427	0.290
	Baseline	0.405	0.328	0.029	0.204	0.280	0.242
2	NITK-IT_NLP [34]	0.268	0.490	0.077	0.063	0.362	0.225
3	UNED	0.298	0.166	0.210	0.162	0.213	0.209
4	AIT_FHSTP [78]	0.378	0.168	0.151	0.081	0.254	0.195
5	Awakened [95]	0.098	0.452	0.194	0.000	0.283	0.186
6	CIC [7]	0.000	0.449	0.240	0.000	0.282	0.172
7	NoFake	0.000	0.492	0.000	0.000	0.326	0.123
8	AI Rational	0.268	0.000	0.166	0.122	0.114	0.111

9 Related Work

There has been a significant number of work on detecting fake news, identifying factuality/credibility of a claim appearing in different sources [8,10,40,49,68,69,73,102]. Typical sources include news article, social media (e.g., Facebook status, tweets, WhatsApp messages, posts in different forums). Major research attention has been paid to the social media [63,81]. Within the realm of misinformation and disinformation there are a number of research areas such as identifying the checkworthiness of a claim [74,85], claim detection [30,35–37], fact-checked claims [32,83,96] etc.

Shared tasks has also been organized in the last few years, which are similar to `CheckThat!` . Such initiatives include SemEval on determining rumour veracity and support for rumours [22,31], on stance detection [58], on fact-checking in community question answering forums [56], on propaganda detection [21,23], and on semantic textual similarity [1,67]. It is also related to the FEVER task [92] on fact extraction and verification, Fake News Challenge [33], and the FakeNews task at MediaEval [72], fact verification and evidence finding for tabular data [98], detecting and rating humor and offense [53], toxic span detection [70], and multimodal fake news detection [59].

10 Conclusion and Future Work

We have presented the 2022 edition of the `CheckThat!` lab, which was again the most popular lab regarding the number of registrations, with a total of 137 registered teams.

Task 1 asked to identify relevant claims in tweets in terms of check-worthiness, verifiability, harmfulness, and attention-worthiness. Task 2 asked to detect previously fact-checked claims that could be relevant to fact-check a new claim. Task 3 asked to predict the veracity of the main claim in a news article. As in `CheckThat!` 2021, BERT and BERT-derived transformers were at the core of the majority of the explored approaches (other transformers explored were GPT-3 and sentence-t5). Back-translation was a popular data augmentation strategy. Regarding Task 1, the use of the mT5 transformer outperformed all other participants in four out of six languages for subtask 1A. The most successful model for subtask 1B approached harmful detection as a contradiction problem. Addressing the retrieval Task 2 with the sentence-t5 transformer and gpt-neo resulted in the best performance, whereas search engines ran short. As for Task 3, the most successful approaches fine-tuned a BERT-based model (which also represented the baseline) and feature-based approaches ran short. The cross-language nature of this task was addressed by machine translating German instances into English.

The approaches to all `CheckThat!` 2022 tasks reflect convergence toward the fine-tuning of transformers. 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. Part of this research is carried out under the Tanbih mega-project, developed at the Qatar Computing Research Institute, HBKU, which aims to limit the impact of "fake news", propaganda, and media bias, thus promoting digital literacy and critical thinking.

Part of this work has been funded by the German Federal Ministry of Education and Research (BMBF) under the grant no. 01FP20031J. The responsibility for the contents of this publication lies with the authors.

References

1. Agirre, E., et al.: SemEval-2016 task 1: semantic textual similarity, monolingual and cross-lingual evaluation. In: Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval 2016. pp. 497–511 (2016)
2. Agrestia, S., Hashemianb, A.S., Carmanc, M.J.: PoliMi-FlatEarthers at CheckThat! 2022: GPT-3 applied to claim detection. In: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, CLEF 202022, Bologna, Italy (2022)
3. Alam, F., et al.: Fighting the COVID-19 infodemic in social media: a holistic perspective and a call to arms. In: Proceedings of the International AAAI Conference on Web and Social Media, ICWSM 2021, pp. 913–922 (2021)
4. Alam, F., et al.: Fighting the COVID-19 infodemic: modeling the perspective of journalists, fact-checkers, social media platforms, policy makers, and the society. In: Findings of EMNLP 2021, pp. 611–649 (2021)
5. Ali, Z.S., Mansour, W., Elsayed, T., Al-Ali, A.: Arafacts: the first large Arabic dataset of naturally occurring claims. In: Proceedings of the Sixth Arabic Natural Language Processing Workshop, pp. 231–236 (2021)

6. Althabiti, S., Alsalka, M.A., Atwell, E.: SCUoL at CheckThat! 2022: fake news detection using transformer-based models. In: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, CLEF 2022, Bologna, Italy (2022)
7. Arif, M., et al.: CIC at CheckThat! 2022: multi-class and cross-lingual fake news detection. In: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, CLEF 2022, Bologna, Italy (2022)
8. Ba, M.L., Berti-Equille, L., Shah, K., Hammady, H.M.: VERA: a platform for veracity estimation over web data. In: Proceedings of the 25th International Conference on World Wide Web, WWW 2016, pp. 159–162 (2016)
9. Balouchzahi, F., Shashirekha, H., Sidorov, G.: MUCIC at CheckThat! 2021: FaDo-fake news detection and domain identification using transformers ensembling. In: Faggioli, G., Ferro, N., Joly, A., Maistro, M., Piroi, F. (eds.) CLEF 2021 Working Notes. Working Notes of CLEF 2021-Conference and Labs of the Evaluation Forum, pp. 455–464 (2021)
10. Baly, R., et al.: What was written vs. who read it: news media profiling using text analysis and social media context. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, pp. 3364–3374 (2020)
11. Barrón-Cedeño, A., et al.: `CheckThat!` at CLEF 2020: enabling the automatic identification and verification of claims in social media. In: Jose, J.M., Yilmaz, E., Magalhães, J., Castells, P., Ferro, N., Silva, M.J., Martins, F. (eds.) ECIR 2020. LNCS, vol. 12036, pp. 499–507. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-45442-5_65
12. Barrón-Cedeño, A., et al.: Overview of CheckThat! 2020: automatic identification and verification of claims in social media. In: Arampatzis, A., et al. (eds.) CLEF 2020. LNCS, vol. 12260, pp. 215–236. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58219-7_17
13. Bilel, T., Mohamed Aziz, B.N., Haddad, H.: iCompass at CheckThat! 2022: ARBERT and AraBERT for Arabic checkworthy tweet identification. In: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, CLEF 2022, Bologna, Italy (2022)
14. Black, S., Gao, L., Wang, P., Leahy, C., Biderman, S.: GPT-neo: large scale autoregressive language modeling with mesh-tensorflow (2021). https://doi.org/10.5281/zenodo.5297715
15. Blanc, O., Pritzkau, A., Schade, U., Geierhos, M.: CODE at CheckThat! 2022: multi-class fake news detection of news articles with BERT. In: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, CLEF 2022, Bologna, Italy (2022)
16. Buliga Nicu, R.M.: Zorros at CheckThat! 2022: ensemble model for identifying relevant claims in tweets. In: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, CLEF 2022, Bologna, Italy (2022)
17. Capetillo, C.P., Lecuona-Gómez, D., Gómez-Adorn, H., Arroyo-Fernández, I., Neri-Chávez, J.: HBDCI at CheckThat! 2022: fake news detection using a combination of stylometric features and deep learning. In: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, CLEF 2022, Bologna, Italy (2022)
18. Chernyavskiy, A., Ilvovsky, D., Nakov, P.: Aschern at CLEF CheckThat! 2021: Lambda-calculus of fact-checked claims. In: Faggioli, G., Ferro, N., Joly, A., Maistro, M., Piroi, F. (eds.) CLEF 2021 Working Notes. Working Notes of CLEF 2021-Conference and Labs of the Evaluation Forum (2021)

19. Conneau, A., et al.: Unsupervised cross-lingual representation learning at scale. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), pp. 8440–8451 (2020). https://doi.org/10.18653/v1/2020.acl-main.747
20. Cusmuliuc, C.G., Amarandei, M.A., Pelin, I., Cociorva, V.I., Iftene, A.: UAICS at CheckThat! 2021: fake news detection. In: Faggioli, G., Ferro, N., Joly, A., Maistro, M., Piroi, F. (eds.) CLEF 2021 Working Notes. Working Notes of CLEF 2021-Conference and Labs of the Evaluation Forum (2021)
21. Da San Martino, G., Barrón-Cedeno, A., Wachsmuth, H., Petrov, R., Nakov, P.: SemEval-2020 task 11: detection of propaganda techniques in news articles. In: Proceedings of the 14th Workshop on Semantic Evaluation, SemEval 2020, pp. 1377–1414 (2020)
22. Derczynski, L., Bontcheva, K., Liakata, M., Procter, R., Hoi, G.W.S., Zubiaga, A.: SemEval-2017 task 8: RumourEval: determining rumour veracity and support for rumours. In: Proceedings of the 11th International Workshop on Semantic Evaluation, SemEval 2017, pp. 69–76 (2017)
23. Dimitrov, D., et al.: SemEval-2021 task 6: detection of persuasion techniques in texts and images. In: Proceedings of the International Workshop on Semantic Evaluation, SemEval 2021, pp. 70–98 (2021)
24. Elsayed, T., et al.: CheckThat! at CLEF 2019: automatic identification and verification of claims. In: Azzopardi, L., Stein, B., Fuhr, N., Mayr, P., Hauff, C., Hiemstra, D. (eds.) Advances in Information Retrieval, pp. 309–315. Springer International Publishing, Cham (2019). https://doi.org/10.1007/978-3-030-15719-7_41
25. Elsayed, T., et al.: Overview of the CLEF-2019 CheckThat!: automatic identification and verification of claims. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction, pp. 301–321. LNCS (2019)
26. Eyuboglu, A.B., Arslan, M.B., Sonmezer, E., Kutlu, M.: TOBB ETU at CheckThat! 2022: detecting attention-worthy and harmful tweets and check-worthy claims. In: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, CLEF 2022, Bologna, Italy (2022)
27. Faggioli, G., Ferro, N., Joly, A., Maistro, M., Piroi, F. (eds.): CLEF 2021 Working Notes. Working Notes of CLEF 2021-Conference and Labs of the Evaluation Forum (2021)
28. Frick, R.A., Vogel, I.: 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)
29. Frick, R.A., Vogel, I., Nunes Grieser, I.: Fraunhofer SIT at CheckThat! 2022: semi-supervised ensemble classification for detecting check-worthy tweets. In: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, CLEF 2022, Bologna, Italy (2022)
30. Gencheva, P., Nakov, P., Màrquez, L., Barrón-Cedeño, A., Koychev, I.: 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, pp. 267–276 (2017)
31. Gorrell, G., et al.: SemEval-2019 task 7: RumourEval, determining rumour veracity and support for rumours. In: Proceedings of the 13th International Workshop on Semantic Evaluation, SemEval 2019, pp. 845–854 (2019)

32. Guo, Z., Schlichtkrull, M., Vlachos, A.: A survey on automated fact-checking. Trans. Assoc. Comput. Linguist. **10**, 178–206 (2022)
33. Hanselowski, A., et al.: A retrospective analysis of the fake news challenge stance-detection task. In: Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, pp. 1859–1874 (2018)
34. Hariharan, R.L., Anand Kumar, M.: NITK-IT_NLP at CheckThat! 2022: window based approach for fake news detection using transformers. In: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, CLEF 2022, Bologna, Italy (2022)
35. Hassan, N., Li, C., Tremayne, M.: Detecting check-worthy factual claims in presidential debates. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, CIKM 2015, pp. 1835–1838 (2015)
36. Hassan, N., Tremayne, M., Arslan, F., Li, C.: Comparing automated factual claim detection against judgments of journalism organizations. In: Computation+Journalism Symposium, pp. 1–5 (2016)
37. Hassan, N., et al.: ClaimBuster: the first-ever end-to-end fact-checking system. Proc. VLDB Endowment **10**(12), 1945–1948 (2017)
38. Hövelmeyer, A., Boland, K., Dietze, S.: 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)
39. Hüsünbeyi, Z.M., Deck, O., Scheffler, T.: RUB-DFL at CheckThat! 2022: transformer models and linguistic features for identifying relevant claims. In: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, CLEF 2022, Bologna, Italy (2022)
40. Karadzhov, G., Nakov, P., Màrquez, L., Barrón-Cedeño, A., Koychev, I.: Fully automated fact checking using external sources. In: Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, pp. 344–353 (2017)
41. Kavatagi, S., Rachh, R., Mulimani, M.: VTU_BGM at Check That! 2022: an autoregressive encoding model for verifying check-worthy claims. In: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, CLEF 2022, Bologna, Italy (2022)
42. Kazemi, A., Garimella, K., Shahi, G.K., Gaffney, D., Hale, S.A.: Research note: Tiplines to uncover misinformation on encrypted platforms: a case study of the 2019 Indian general election on whatsapp. Harvard Kennedy School Misinformation Review (2022)
43. Köhler, J., et al.: 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)
44. Kovachevich, N.: BERT fine-tuning approach to CLEF CheckThat! fake news detection. In: Faggioli, G., Ferro, N., Joly, A., Maistro, M., Piroi, F. (eds.) CLEF 2021 Working Notes. Working Notes of CLEF 2021-Conference and Labs of the Evaluation Forum (2021)
45. Kumar, S., Kumar, G., Singh, S.R.: TextMinor at CheckThat! 2022: fake news article detection using RoBERT. In: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, CLEF 2022, Bologna, Italy (2022)
46. La Barbera, D., Roitero, K., Mackenzie, J., Damiano, S., Demartini, G., Mizzaro, S.: BUM at CheckThat! 2022: a composite deep learning approach to fake news detection using evidence retrieval. In: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, CLEF 2022, Bologna, Italy (2022)

47. Lomonaco, F., Donabauer, G., Siino, M.: COURAGE at CheckThat! 2022: harmful tweet detection using graph neural networks and ELECTRA. In: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, CLEF 2022, Bologna, Italy (2022)
48. Ludwig, A., Felser, J., Xi, J., Labudde, D., Spranger, M.: FoSIL at CheckThat! 2022: using human behaviour-based optimization for text classification. In: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, CLEF 2022, Bologna, Italy (2022)
49. Ma, J., Gao, W., Mitra, P., Kwon, S., Jansen, B.J., Wong, K.F., Cha, M.: Detecting rumors from microblogs with recurrent neural networks. In: Proceedings of the International Joint Conference on Artificial Intelligence, IJCAI 2016, pp. 3818–3824 (2016)
50. Manan Suri, P.K., Dudeja, S.: Asatya at CheckThat! 2022: multimodal BERT for identifying claims in tweets. In: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, CLEF 2022, Bologna, Italy (2022)
51. Mansour, W., Elsayed, T., Al-Ali, A.: Did i see it before? detecting previously-checked claims over twitter. In: European Conference on Information Retrieval. pp. 367–381 Springer (2022)
52. Martinez-Rico, J.R., Martinez-Romo, J., Araujo, L.: NLP&IRUNED at CheckThat! 2022: ensemble of classifiers for fake news detection. In: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, CLEF 2022, Bologna, Italy (2022)
53. Meaney, J., Wilson, S., Chiruzzo, L., Lopez, A., Magdy, W.: Semeval 2021 task 7: hahackathon, detecting and rating humor and offense. In: Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pp. 105–119 (2021)
54. Michael Shliselberg, S.D.H.: 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)
55. Mihaylova, S., Borisova, I., Chemishanov, D., Hadzhitsanev, P., Hardalov, M., Nakov, P.: DIPS at CheckThat! 2021: verified claim retrieval. In: Faggioli, G., Ferro, N., Joly, A., Maistro, M., Piroi, F. (eds.) CLEF 2021 Working Notes. Working Notes of CLEF 2021-Conference and Labs of the Evaluation Forum (2021)
56. Mihaylova, T., Karadzhov, G., Atanasova, P., Baly, R., Mohtarami, M., Nakov, P.: SemEval-2019 task 8: Fact checking in community question answering forums. In: Proceedings of the 13th International Workshop on Semantic Evaluation, SemEval 2019, pp. 860–869 (2019)
57. Mingzhe, D., Sujatha Das Gollapalli, S.K.N.: NUS-IDS at CheckThat! 2022: identifying check-worthiness of tweets using CheckthaT5. In: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, CLEF 2022, Bologna, Italy (2022)
58. Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X., Cherry, C.: SemEval-2016 task 6: Detecting stance in tweets. In: Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval 2016, pp. 31–41 (2016)

59. Nakamura, K., Levy, S., Wang, W.Y.: r/fakeddit: a new multimodal benchmark dataset for fine-grained fake news detection. arXiv preprint arXiv:1911.03854 (2019)
60. Nakov, P., et al.: Overview of the CLEF-2022 CheckThat! lab task 1 on identifying relevant claims in tweets. In: Working Notes of CLEF 2022–Conference and Labs of the Evaluation Forum, CLEF 2022, Bologna, Italy (2022)
61. Nakov, P., et al.: The CLEF-2022 CheckThat! Lab on fighting the covid-19 infodemic and fake news detection. In: Hagen, M., et al. (eds.) Advances in Information Retrieval, pp. 416–428. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-99739-7_52
62. Nakov, P., et al.: Overview of the CLEF-2018 lab on automatic identification and verification of claims in political debates. In: Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, CLEF 2018 (2018)
63. Nakov, P., et al.: Automated fact-checking for assisting human fact-checkers. In: Proceedings of the 30th International Joint Conference on Artificial Intelligence, IJCAI 2021, pp. 4551–4558 (2021)
64. Nakov, P., Da San Martino, G., Alam, F., Shaar, S., Mubarak, H., Babulkov, N.: 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)
65. Nakov, P., et al.: The CLEF-2021 CheckThat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news. In: Hiemstra, D., Moens, M.-F., Mothe, J., Perego, R., Potthast, M., Sebastiani, F. (eds.) ECIR 2021. LNCS, vol. 12657, pp. 639–649. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-72240-1_75
66. Nakov, P., et al.: Overview of the CLEF–2021 CheckThat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news. In: Candan, S., et al. (eds.) CLEF 2021. LNCS, vol. 12880, pp. 264–291. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-85251-1_19
67. Nakov, P., et al.: SemEval-2016 Task 3: community question answering. In: Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval 2015, pp. 525–545 (2016)
68. Nguyen, V.H., Sugiyama, K., Nakov, P., Kan, M.Y.: FANG: leveraging social context for fake news detection using graph representation. In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, CIKM 2020, pp. 1165–1174 (2020)
69. Oshikawa, R., Qian, J., Wang, W.Y.: A survey on natural language processing for fake news detection. In: Proceedings of the 12th Language Resources and Evaluation Conference, LREC 2020, pp. 6086–6093 (2020)
70. Pavlopoulos, J., Sorensen, J., Laugier, L., Androutsopoulos, I.: Semeval-2021 task 5: toxic spans detection. In: Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pp. 59–69 (2021)
71. Pires, T., Schlinger, E., Garrette, D.: How Multilingual is Multilingual BERT. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), Florence, Italy, pp. 4996–5001 (2019). https://doi.org/10.18653/v1/P19-1493
72. Pogorelov, K., et al.: FakeNews: corona virus and 5G conspiracy task at MediaEval 2020. In: Proceedings of the MediaEval 2020 Workshop, MediaEval 2020 (2020)
73. Popat, K., Mukherjee, S., Strötgen, J., Weikum, G.: Credibility assessment of textual claims on the web. In: Proceedings of the 25th ACM International Con-

ference on Information and Knowledge Management, CIKM 2016, pp. 2173–2178 (2016)
74. Pritzkau, A.: NLytics at CheckThat! 2021: check-worthiness estimation as a regression problem on transformers. In: Faggioli, G., Ferro, N., Joly, A., Maistro, M., Piroi, F. (eds.) CLEF 2021 Working Notes. Working Notes of CLEF 2021-Conference and Labs of the Evaluation Forum (2021)
75. Pritzkau, A., Blanc, O., Geierhos, M., Schade, U.: NLytics at CheckThat! 2022: hierarchical multi-class fake news detection of news articles exploiting the topic structure. In: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, CLEF 2022, Bologna, Italy (2022)
76. Röchert, D., Shahi, G.K., Neubaum, G., Ross, B., Stieglitz, S.: The networked context of covid-19 misinformation: informational homogeneity on youtube at the beginning of the pandemic. Online Social Netw. Media **26**, 100164 (2021)
77. Savchev, A.: AI Rational at CheckThat! 2022: using transformer models for tweet classification. In: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, CLEF 2022, Bologna, Italy (2022)
78. Schütz, M., et al.: AIT FHSTP at CheckThat! 2022: cross-lingual fake news detection with a large pre-trained transformer. In: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, CLEF 2022, Bologna, Italy (2022)
79. Schütz, M., Siegel, M.: Baseline for clef2022 - checkthat! lab task 3 (2022). https://doi.org/10.5281/zenodo.6362498
80. Sepúlveda-Torres, R., Saquete, E.: GPLSI team at CLEF CheckThat! 2021: fine-tuning BETO and RoBERTa. In: Faggioli, G., Ferro, N., Joly, A., Maistro, M., Piroi, F. (eds.) CLEF 2021 Working Notes. Working Notes of CLEF 2021-Conference and Labs of the Evaluation Forum (2021)
81. Shaar, S., Alam, F., Da San Martino, G., Nakov, P.: The role of context in detecting previously fact-checked claims. Arxiv:2104.07423 (2021)
82. Shaar, S., Alam, F., Martino, G.D.S., Nakov, P.: Assisting the human fact-checkers: detecting all previously fact-checked claims in a document. arXiv preprint arXiv:2109.07410 (2021)
83. Shaar, S., Babulkov, N., Da San Martino, G., Nakov, P.: That is a known lie: Detecting previously fact-checked claims. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, pp. 3607–3618 (2020)
84. Shaar, S., et al.: Overview of the CLEF-2021 CheckThat! lab task 2 on detecting previously fact-checked claims in tweets and political debates. In: Faggioli, G., Ferro, N., Joly, A., Maistro, M., Piroi, F. (eds.) CLEF 2021 Working Notes. Working Notes of CLEF 2021-Conference and Labs of the Evaluation Forum (2021)
85. Shaar, S., et al.: Overview of the CLEF-2021 CheckThat! lab task 1 on check-worthiness estimation in tweets and political debates. In: Faggioli, G., Ferro, N., Joly, A., Maistro, M., Piroi, F. (eds.) CLEF 2021 Working Notes. Working Notes of CLEF 2021-Conference and Labs of the Evaluation Forum (2021)
86. Shahi, G.K.: AMUSED: an annotation framework of multi-modal social media data. arXiv:2010.00502 (2020)
87. Shahi, G.K., Dirkson, A., Majchrzak, T.A.: An exploratory study of COVID-19 misinformation on Twitter. Online Social Netw. Media **22**, 100104 (2021)
88. Shahi, G.K., Struß, J.M., Mandl, T.: Overview of the CLEF-2021 CheckThat! lab: Task 3 on fake news detection. In: Faggioli, G., Ferro, N., Joly, A., Maistro, M., Piroi, F. (eds.) CLEF 2021 Working Notes. Working Notes of CLEF 2021-Conference and Labs of the Evaluation Forum (2021)

89. Taboubi, B., Nessir, M.A.B., Haddad, H.: iCompass at CheckThat! 2022: combining deep language models for fake news detection. In: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, CLEF 2022, Bologna, Italy (2022)
90. Tarannum, P., Md. Arid, H., Alam, F., Noori, S.R.H.: Z-Index at CheckThat! Lab 2022: check-worthiness identification on tweet text. In: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, CLEF 2022, Bologna, Italy (2022)
91. Tchechmedjiev, A., et al.: ClaimsKG: a knowledge graph of fact-checked claims. In: Ghidini, C., et al. (eds.) ISWC 2019. LNCS, vol. 11779, pp. 309–324. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-30796-7_20
92. Thorne, J., Vlachos, A., Christodoulopoulos, C., Mittal, A.: FEVER: a large-scale dataset for fact extraction and VERification. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2018, pp. 809–819 (2018)
93. Toraman, C., Ozcelik, O., Şahinuç, F., Sahin, U.: ARC-NLP at CheckThat! 2022: contradiction for harmful tweet detection. In: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, CLEF 2022, Bologna, Italy (2022)
94. Tran, H.N., Kruschwitz, U.: ur-iw-hnt at CheckThat! 2022: cross-lingual text summarization for fake news detection. In: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, CLEF 2022, Bologna, Italy (2022)
95. Truică C.O., Apostol, E.S., Paschke, A.: Awakened at CheckThat! 2022: fake news detection using BiLSTM and sentence transformer. In: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, CLEF 2022, Bologna, Italy (2022)
96. Varma, H., Jain, A., Ratadiya, P., Rathi, A.: Attestable at semeval-2021 task 9: extending statement verification with tables for unknown class, and semantic evidence finding. In: Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pp. 1276–1282 (2021)
97. Vasileva, S., Atanasova, P., Màrquez, L., Barrón-Cedeño, A., Nakov, P.: It takes nine to smell a rat: neural multi-task learning for check-worthiness prediction. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing, RANLP 2019, pp. 1229–1239 (2019)
98. Wang, N.X., Mahajan, D., Danilevsky, M., Rosenthal, S.: Semeval-2021 task 9: fact verification and evidence finding for tabular data in scientific documents (sem-tab-facts). In: Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pp. 317–326 (2021)
99. Williams, E., Rodrigues, P., Tran, S.: Accenture at CheckThat! 2021: interesting claim identification and ranking with contextually sensitive lexical training data augmentation. In: Faggioli, G., Ferro, N., Joly, A., Maistro, M., Piroi, F. (eds.) CLEF 2021 Working Notes. Working Notes of CLEF 2021-Conference and Labs of the Evaluation Forum (2021)
100. Yasser, K., Kutlu, M., Elsayed, T.: Re-ranking web search results for better fact-checking: a preliminary study. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management, pp. 1783–1786 (2018)

101. Zhou, X., Wu, B., Fung, P.: Fight for 4230 at CLEF CheckThat! 2021: domain-specific preprocessing and pretrained model for ranking claims by check-worthiness. In: Faggioli, G., Ferro, N., Joly, A., Maistro, M., Piroi, F. (eds.) CLEF 2021 Working Notes. Working Notes of CLEF 2021-Conference and Labs of the Evaluation Forum (2021)
102. Zubiaga, A., Liakata, M., Procter, R., Hoi, G.W.S., Tolmie, P.: Analysing how people orient to and spread rumours in social media by looking at conversational threads. PLoS ONE **11**(3), e0150989 (2016)