Overview of the CLEF–2023 CheckThat! Lab on Checkworthiness, Subjectivity, Political Bias, Factuality, and Authority of News Articles and Their Source

Overview of the CLEF–2023 CheckThat! Lab on Checkworthiness, Subjectivity, Political Bias, Factuality, and Authority of News Articles and Their Source

Alberto Barrón-Cedeño[1][0000−0003−4719−3420], Firoj Alam[2][0000−0001−7172−1997], Andrea Galassi[1][0000−0001−9711−7042], Giovanni Da San Martino[3][0000−0002−2609−483X], Preslav Nakov[4][0000−0002−3600−1510], Tamer Elsayed[5][0000−0001−5786−4668], Dilshod Azizov[5][0000−0002−0572−8813], Tommaso Caselli[3][0000−0003−2936−0256], Gullal S. Cheema[7][0000−0003−4354−9629], Fatima Haouari[5][0000−0003−4842−2467], Maram Hasanain[2][0000−0002−7466−178X], Mucahid Kutlu[8][0000−0002−5660−4992], Chengkai Li[9][0000−0002−1724−8278], Federico Ruggeri[1][0000−0002−1697−8586], Julia Maria Struß[10][0000−0001−9133−4978], Wajdi Zaghouani[11][0000−0003−1521−5568]

[1]Università di Bologna, Italy [2]Qatar Computing Research Institute, Qatar [3]University of Padua, Italy [4]Mohamed bin Zayed University of Artificial Intelligence, UAE [5]Qatar University, Qatar [6]University of Groningen, Netherlands [7]L3S Research Center, Leibniz University of Hannover, Germany [8]TOBB University of Economics and Technology, Türkiye [9]University of Texas at Arlington, USA [10]University of Applied Sciences Potsdam, Germany [11]Hamad bin Khalifa University, Qatar
https://checkthat.gitlab.io

Abstract. We describe the sixth edition of the `CheckThat!` lab, part of the 2023 Conference and Labs of the Evaluation Forum (CLEF). The five previous editions of `CheckThat!` focused on the main tasks of the information verification pipeline: check-worthiness, verifying whether a claim was fact-checked before, supporting evidence retrieval, and claim verification. In this sixth edition, we zoom into some new problems and for the first time we offer five tasks in seven languages: Arabic, Dutch, English, German, Italian, Spanish, and Turkish. Task 1 asks to determine whether an item —text or text plus image— is check-worthy. Task 2 aims to predict whether a sentence from a news article is subjective or not. Task 3 asks to assess the political bias of the news at the article and at the media outlet level. Task 4 focuses on the factuality of reporting of news media. Finally, Task 5 looks at identifying authorities in Twitter that could help verify a given target claim. For a second year, `CheckThat!` was the most popular lab at CLEF-2023 in terms of team registrations: 127 teams. About one-third of them (a total of 37) actually participated.

Keywords: Fact Checking · Check-Worthiness · Subjectivity · Political Bias · Factuality of Reporting · Authority Finding.

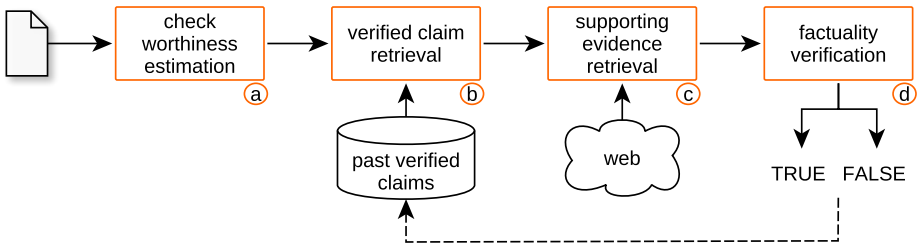

Fig. 1. The `CheckThat!` verification pipeline, featuring the four core tasks. Task 1 on check-worthiness this year is the only one that belongs to these core tasks.

1 Introduction

From its conception, the `CheckThat!` lab has been dedicated to promoting and fostering the development of technology to assist investigative journalists to perform fact-checking, focusing on political debates, social media posts, and news articles. The five previous editions of the lab have been held annually from 2018 to 2022, targeting diverse Natural Language Processing (NLP) and Information Retrieval (IR) tasks, part of the `CheckThat!` pipeline [69,37,36,19,20,72,73,68,67] is shown in Figure 1.

For the first time, `CheckThat!` 2023 [18] zooms out of the core pipeline and focuses on *auxiliary* tasks that help in addressing the different steps of the pipeline. For that, it challenged the research community with five tasks in seven languages: Arabic, Dutch, English, German, Italian, Spanish, and Turkish. Task 1 [2], the only one that follows up from previous editions and the only one that is part of the core pipeline, asks systems to find out whether a given claim in a tweet is worth fact-checking. This year, for the first time, Task 1 offers a multimodal track. Task 2 [41] requires to determine whether a sentence from a news article is objective or conveys a subjective point of view, influenced by personal feelings, tastes, or opinions. Task 3 [29] asks systems to measure the level of political bias of news reporting at the article and at the media level. Task 4 [65] asks to assess the factuality of reporting at the news media level. Task 5 [46] challenges models to retrieve a set of authority Twitter accounts for a given rumor propagating in Twitter.

Task 1 is what professionals take the most advantage of, since the amount of information online is impossible for one to keep up with. Task 2 helps check-worthiness, by spotting opinionated snippets that might not be relevant for fact-checking. Task 3 could help factuality verification by contributing information about both the stance of a claim and a piece of evidence. Task 4 could help to determine whether the information from a news outlet can be trusted a priori. Finally, Task 5 could help in identifying people/institutions that can challenge a claim. Table 1 showcases the language coverage and the type of documents included in this year's tasks.

Table 1. Overview of the 2023 tasks: language coverage and type of document.

task	ar	de	du	en	es	it	tr	documents
Task 1	■			■	■			tweets (incl. multimodal), debates and speeches
Task 2	■	■	■	■		■	■	news articles and tweets
Task 3				■				news articles
Task 4				■				news articles
Task 5	■							tweets

Table 2. Overview of the tasks offered in the previous editions of the lab.

tasks	years					domains					languages						
	2018	2019	2020	2021	2022	debates	speeches	tweets	web pages	news articles	English	Arabic	Bulgarian	Spanish	Turkish	German	Dutch
check-worthiness estimation	■	■	■	■	■	■	■	■			■	■	■	■	■		■
verified claim retrieval			■	■	■	■	■	■			■	■					
supporting evidence retrieval		■	■						■			■					
claim verification	■	■	■								■	■					
fake news detection				■	■					■	■					■	
topic identification				■							■						

2 Previously on the CheckThat! Lab

During the previous five iterations of the CheckThat! lab, it has focused on various tasks from the claim verification pipeline, in a multitude of languages and in different domains (cf. Table 2).

The first iteration of CheckThat! in 2018 [10,21] focused on check-worthiness and claim verification of political debates and speeches in Arabic and English. Both tasks were then continued in the following iteration, with an additional focus on fact-checking by a task on classifying and ranking supporting evidence from the web [37,11,48]. The 2020 edition [20] of the CheckThat! lab covered the full claim verification pipeline, with check-worthiness estimation, verified claim and supporting evidence retrieval, and claim verification; social media data was first included in that iteration of the lab [47,91]. The fourth edition of the lab in 2021 put focus on multilinguality by offering tasks in five languages [73,90,89]. The edition also featured a new fake news detection task [92], where the focus was not on a claim, but on an article; this task was quite popular and it was continued in the 2022 edition of the CheckThat! lab.

The 2022 year's edition of the CheckThat! lab [67] paid special attention to the various sub-aspects of check-worthiness estimation, namely factuality, harmfulness, and attention-worthiness estimation, again in a multitude of languages. Transformer-based models were extensively used.

The highest-ranking systems additionally implemented data augmentation and supplementary preprocessing measures [66]. The second task in the 2022 edition of the lab asked to detect previously fact-checked claims from tweets, political debates, and speeches [71]. The best system used the Sentence T5 transformer and GPT-Neo models. The third task in the 2022 edition of the `CheckThat!` lab asked to predict the veracity of the main claim in an English news article, with English or German training data. The most successful approaches fine-tuned a BERT-based model. The cross-language nature of the task has mainly been addressed using machine translation [54].

3 Description of the 2023 Tasks

The 2023 edition of the `CheckThat!` lab is organized around five tasks, four of which are run for the first time (cf. Sections 3.2 to 3.5). Moreover, two tasks have two subtasks (cf. Sections 3.1 and 3.3).

3.1 Task 1: Check-Worthiness in Multimodal and Multigenre Content

The goal of this task is to assess whether a given statement, in a tweet or from a political debate, is worth fact-checking [2]. In order to make that decision, one would need to ponder about questions, such as "does it contain a verifiable factual claim?" or "is it harmful?", before deciding on the final check-worthiness label [4]. Task 1 is divided into two subtasks. Subtask 1A is offered in Arabic and English, whereas subtask 1B is offered in Arabic, English and Spanish.

Subtask 1A: Multimodality Given a tweet with the text and its corresponding image, predict whether it is worth fact-checking. Here, answers to the questions relevant for deriving a label are based on both the image and the text. The image plays two roles for check-worthiness estimation: (*i*) there is a piece of evidence (e.g., an event, an action, a situation, a person's identity, etc.) or illustration of certain aspects from the textual claim, and/or (*ii*) the image contains overlaid text that contains a claim (e.g. misrepresented facts and figures) in a textual form.

Subtask 1B: Multigenre A text snippet alone, either from a tweet or from a political debate or speech, has to be assessed for check-worthiness.

3.2 Task 2: Subjectivity in News Articles

Given a sentence from a news article or a tweet (in the case of Turkish), Task 2 asks systems to determine whether the sentence is subjective or objective [41]. A sentence is **subjective** if its contents are based on or influenced by personal feelings, tastes, or opinions; otherwise, it is considered **objective**. This task pays the most attention to multilinguality this year. It is offered in Arabic, Dutch, English, Italian, German, and Turkish, with an additional multilingual setting.

3.3 Task 3: Political Bias of News Articles and News Media

The goal of the task is to detect political bias of news reporting at the article and at the media level. This is an ordinal classification task and it is offered in English [29]. It includes two subtasks:

Subtask 3A: Political Bias of News Articles Given an article, classify its leaning as left, center, or right.

Subtask 3B: Political Bias of News Media Given a news outlet, predict its overall political bias as left, center, or right.

3.4 Task 4: Factuality of Reporting of News Media

In this task, we specifically target media credibility. The goal is to predict the factuality of reporting at the media level, given a set of articles from the target news outlet: low, mixed, and high. This is another ordinal classification task, and it is offered in English only [65].

3.5 Task 5: Authority Finding in Twitter

The task asks systems to retrieve authority Twitter accounts for a given rumor that propagates in Twitter [46]. Given a tweet spreading a rumor, the participating systems need to retrieve a ranked list of authority Twitter accounts that can help verify that rumor, as such accounts may tweet evidence that supports or denies the rumor [43]. This task is offered in Arabic.

4 Datasets

4.1 Task 1: Check-Worthiness of Multimodal and Multigenre Content

Subtask 1A: Multimodality The dataset used for subtask 1A was derived from [25], with the existing data repurposed for training and development purposes. We followed the schema from [25] to produce a new testing set.

The dataset focused on three topics: COVID-19, climate change and technology. Each tweet was labeled using both the image and the text, with Optical Character Recognition (OCR) performed using the Google Vision API to extract the text from the images. We provided 3,175 annotated examples and around 110k unlabeled tweets of text–image pairs and OCR output. Two annotators, one of them an expert, annotated the new test data. The non-expert went through a dry run of 50 examples, where disagreements were discussed and resolved. For the final test set of 736 examples, the Cohen's Kappa inter-annotator agreement [27] was 0.49 for the check-worthy label, indicating moderate agreement. The expert annotator resolved any remaining disagreements.

Table 3. Task 1: Check-worthiness in multimodal and multigenre content. Statistics about the CT–CWT–23 corpus for all three languages.

Subtask	Class	Train	Dev	Dev-Test	Test	Total
1A Arabic	No	1,421	207	402	792	2,822
	Yes	776	113	220	203	1,312
	Total	2,197	320	622	995	**4,134**
1A English	No	1,536	184	374	459	2,553
	Yes	820	87	174	277	1,358
	Total	2,356	271	548	736	**3,911**
1B Arabic	No	4,301	789	682	123	5,895
	Yes	1,758	485	411	377	3,031
	Total	6,059	1,274	1,093	500	**8,926**
1B English	No	12,818	4,270	794	210	18,092
	Yes	4,058	1,355	238	108	5,759
	Total	16,876	5,625	1,032	318	**23,851**
1B Spanish	No	14,805	2,157	4,190	4,491	25,643
	Yes	2,682	391	759	509	4,341
	Total	17,487	2,548	4,949	5,000	**29,984**

For subtask 1A Arabic, we followed several steps for training, development, dev-test, and test datasets. For the former three partitions, we used the CT-CWT-21 [90] and the CT-CWT-22 [66] datasets annotated for check-worthiness with topics focusing on COVID-19 and politics. The labelling of the datasets follows the annotation schema discussed in [3,4]. To develop multimodal datasets based on these datasets, we crawled images associated with tweets. For tweets with multiple images, we retrieved only the first one. For the former three partitions, we derived the label for multimodality from the textual modality, and thus these can be seen as weakly labeled annotations. For the test set, we crawled tweets using similar keywords to those reported in [3,4]. For the annotation, three annotators followed the same annotation schema, but for multimodality We used majority voting to select the final labels.

Subtask 1B: Multigenre The dataset for Subtask 1B consists of tweets in Arabic and Spanish as well as statements from political debates in English. The Arabic tweets are collected using keywords related to COVID-19 and vaccines, using the annotation schema in [4]. The training, the development, and the dev-test partitions of the dataset come from CT-CWT-21 [90] and CT-CWT-22 [66]. For the test set, we used the same approach as for subtask 1A. The dataset for English consists of transcribed sentences from candidates during the US presidential election debates and annotated by human annotators [9]. For the first three partitions, we used essentially the same dataset reported in [9], with some updates that reflect improved annotation accuracy. The test set contains sentences that were not included in [9].

Table 4. Task 2: Subjectivity in news articles. Statistics about the datasets for all six languages and the multilingual setting, and the distribution of objective (Obj) and subjective (Subj) examples.

	Training		Dev		Test		
Language	**Obj**	**Subj**	**Obj**	**Subj**	**Obj**	**Subj**	**Total**
Arabic	905	280	227	70	363	82	**1,927**
Dutch	489	311	107	93	263	237	**1,500**
English	532	298	106	113	116	127	**1,292**
German	492	308	123	77	194	97	**1,291**
Italian	1,231	382	167	60	323	117	**2,280**
Turkish	422	378	100	100	129	111	**1,140**
Multilingual	4,371	2,257	300	300	300	300	**7,828**

The Spanish dataset is also a combination of CT-CWT-21 [90], CT-CWT-22 [66] and newly collected content. It is composed of tweets collected from Twitter accounts and transcriptions from Spanish politicians, which are manually annotated by professional journalists who are experts in fact-checking.

Table 3 shows statistics about the datasets for Task 1. Across the different subtasks, dataset sizes range from 3,911 to 29,984, which are the largest so far across different languages over the years for the check-worthiness task.

4.2 Task 2: Subjectivity in News Articles

The datasets for all languages in Task 2 were produced on the basis of the subjectivity identification guidelines outlined in [7]. The sentences were extracted from news articles, with the exception of Turkish, in which each sentence is manually extracted from tweets about politics. Table 4 shows the label distribution. The training set of the multilingual dataset is the union of the training material from all languages. The development and the testing sets, on the other hand, are formed by randomly selecting 50 subjective and 50 objective sentences from the respective development and testing sets in all languages. In total, we annotated 9,351 sentences covering six languages.

4.3 Task 3: Political Bias of News Articles and News Media

Table 5 reports the label distribution of the datasets for Task 3.

For subtask 3A we release a collection of 55k articles from 1,023 media sources annotated for bias at the article level. The articles were crawled from AllSides.[1] To make sure the data is up to date and pertinent to the present political environment, the dataset includes news articles published from the end of 2022 to the beginning of 2023.

[1] https://www.allsides.com

Table 5. Task 3: Political bias of news articles and news media. Statistics about the CT–Bias–23 datasets [29].

Class	Train	Dev	Test	Total
left	12,073	1,342	2,589	**16,004**
center	15,449	1,717	1,959	**19,125**
right	17,544	1,949	650	**20,143**
	19,125	**16,044**	**20,143**	**55,272**

Task 3A: Bias of Articles

Class	Train	Dev	Test	Total
left	216	31	25	**272**
center	296	34	29	**359**
right	305	39	48	**392**
	817	**104**	**102**	**1,023**

Task 3B: Bias of News Media

Table 6. Task 4: Factuality of reporting of news media. Statistics about the CT–Factuality–23 dataset [65].

Class	Train	Dev	Test	Total
High	593	72	72	**737**
Mixed	233	32	31	**296**
Low	121	16	19	**156**
	947	**120**	**122**	**1,189**

AllSides curates articles from a variety of reputable national and international news sources to ensure a balanced representation across different political spectra. The articles are annotated following a rigorous scheme that involves expert reviewers. For the data split, we divide them into 80%, 10% and 10% for the training, development, and test, respectively.

For subtask 3B our dataset is sourced from Media Bias/Fact Check,[2] which follows a meticulous approach to characterize media sources, conducted by experts. This dataset contains a subset of data used in previous research [15]. Note that we remap the bias from a 7-point scale (extreme-left, left, center-left, center, center-right, right, and extreme-right) to a 3-point scale: left, center, and right (for this, we exclude center-left and center-right). For the data split, we divide them by news media as the same splits as subtask 3A, but we randomly select up to 11 news articles from each news medium. Finally, we release annotated articles for each medium, which are to be used for the classification.

4.4 Task 4: Factuality of Reporting of News Media

We use the same kind of data as for Task 3, but with labels for factuality (again on an ordinal scale). We obtain the annotations and the analysis of the factuality of reporting from Media Bias/Fact Check, where they are manually labeled by professional fact-checkers. We use a 3-point scale: low, mixed, and high factuality. The dataset consists of 1,189 news media: see Table 6 for detailed statistics. For each new medium, we include approximately ten articles.

[2] https://www.mediabiasfactcheck.com

Table 7. Task 5: Authority finding. Rumor collection and relevance judgments statistics.

Data split	Rumors	Authorities
Training	120	849
Development	30	195
Testing	30	172
Total	**180**	**1,216**

4.5 Task 5: Authority Finding in Twitter

For training, we adopted the AuFIN [44] collection, which comprises 150 rumors expressed in tweets, associated with 1,044 authority Twitter accounts, and a user collection of 395,231 accounts along with their Twitter lists (1,192,284 unique lists). Each authority is graded as *highly relevant* or *relevant* to the rumor, i.e., having a higher priority to be contacted for verification or not. The rumors cover three categories: politics, sports, and health; 50 from each category. We split the rumors into 120 for training and 30 for development.

For testing, we collected 30 new rumors from AraFacts [5], where we focused on the ones collected from Misbar[3] and Fatabayyano,[4] which were used recently to construct Arabic rumor verification [45] and fake news detection [53] datasets. We selected 10 rumors from each of the three categories. For each rumor, two annotators separately identified all authority Twitter accounts that can help support or debunk the rumor following the same annotation guidelines as for AuFIN. The Cohen's Kappa inter-annotator agreement [27] was 0.91 and 0.42 for the authority label and the graded relevance, which correspond to almost perfect and moderate agreements, respectively [56]. Table 7 presents the overall statistics about the rumor collection and the relevance judgments for each data split. For an overall summary of the user collection, we refer the reader to [44].

5 Results

In this section, we present the top-performing submissions for each of the five tasks. For details about all participating approaches, refer to the corresponding task paper: Task 1 [2], Task 2 [41], Task 3 [29], Task 4 [65], and Task 5 [46].

5.1 Task 1: Check-Worthiness in Multimodal and Unimodal Contents

A total of 14 teams participated in Task 1 and submitted 35 runs.

[3] https://misbar.com/
[4] https://fatabyyano.net/

Table 8. Overview of the approaches for **subtask 1A**. The numbers in the language box show the position of the team in the official ranking; ☑=part of the official submission; ✔=considered in internal experiments.

Team		Lang		Transformers									CNN			Repr.			Misc			
		Arabic	English	BERT	RoBERTa	mBERT	XLM RoBERTa	AraBERT	BERTweet	GPT-3	Electra	Vision Transformer	ConvNext	ResNet	VGG	EfficientNet	Text	Image	OCR Output	CLIP	LSTM	CatBoost	Preprocessing
CSECU-DSG	[12]	1	4				☑	☑	☑				☑				☑	☑			☑		
ES-VRAI	[81]		-	☑								✔		☑			☑	☑		✔			
Fraunhofer SIT	[40]		1	☑								✔				✔	☑	✔	☑				☑
Mtop*		2	6														☑	☑				☑	
Z-Index	[97]	3	5	☑		☑	✔							☑			☑	☑					☑
ZHAW-CAI	[105]		2		☑					☑	☑	☑					☑	☑					☑

- Run submitted after the deadline. *No working note submitted.

Table 9. Subtask 1A: Multimodal check-worthiness estimation. Shown are the top-3 submissions for Arabic and English. The F1 score is computed with respect to the positive class.

Team		F1	Team		F1
Arabic			**English**		
1	CSECU-DSG [12]	0.399	1	Fraunhofer SIT [40]	0.712
2	Mtop*	0.312	2	ZHAW-CAI [105]	0.708
3	Z-Index [97]	0.301	-	ES-VRAI [81]	0.704

- Run submitted after the deadline. *No working note submitted.

Subtask-1A A total of 7 and 4 teams submitted their runs for English and for Arabic, respectively, out of which four made submissions for both languages. Table 8 gives an overview of the submitted models per language. This was a binary classification task, and we used the F1 score for the positive class as the official evaluation measure. Table 9 shows the performance of the top official submissions on the test set.

Starting with the best-performing system for English: team *Fraunhofer SIT* [40] tackled the problem by fine-tuning individual text classifiers on the tweet text and on the OCR text, respectively. They further used pre-processing for the tweet text and extracted the text from images using *easyOCR*.[5] Two BERT [33] models were fine-tuned on each input, and the final label for each example in the test set was a re-weighted combination of the two predictions based on the validation loss.

[5] https://github.com/JaidedAI/EasyOCR

Team *ZHAW-CAI* [105] submitted official runs for the English track only. They trained different unimodal and multimodal systems and then combined them using a kernel-based ensemble. This ensemble was trained using an SVM for classification. For the text-based model, n-gram features were extracted separately from the tweet text, and prompt response from GPT-3 (Open AI's text-davinci-003), and SVMs were trained on these features. In addition, an Electra [26] model was fine-tuned over the tweet text for classification. For the multimodal model, features from Twitter-based RoBERTa [59] and ViT were extracted, fused via pooling, and passed through a dense layer for classification. The submission model is an ensemble of the four features described earlier with their individual kernels and combined with an average kernel to be used in an SVM for classification.

Team *ES-VRAI* [81] comprehensively evaluated several pre-trained vision and text models, different classifiers, and several early and late fusion strategies to select the best model for the English data. Their submitted model combined BERT and ResNet50 [49] features in an early fusion mode.

Team *CSECU-DSG* [12] participated in both the Arabic and the English tracks. They used a model that jointly fine-tunes two transformers. A language-specific BERT is used to represent the tweet text, and ConvNext [61] is used for image feature extraction. They uses BERTweet [74] for English data, and AraBERT [8] for Arabic. In addition, a BiLSTM was used on top of the text features to handle long-term contextual dependency. Finally, the features from the BiLSTM and the ConvNext were concatenated and followed up by a multisample dropout [51] to predict the final label.

Team *Z-Index* [97] also participated in both languages. They used BERT for English tweet text and ResNet50 for images, and a feed-forward neural network for fusion and classification. In addition, mBERT [33] was used for Arabic text. The backbone networks were fine-tuned along with the feed-forward network to train the model for the task. In their internal evaluation, they also experimented with XLM-Roberta [28], which performed better by 4% than the BERT variant for both languages.

To summarize all the contributions of the participating teams: one common theme across the methods was the use of large pre-trained models and their features for semantic information extraction. Only team *Fraunhofer SIT* used two separate classifiers, while the rest used fusion and ensemble strategies. Two of the teams further used OCR.

Subtask-1B A total of 11, 6, and 7 teams submitted their runs for English, Arabic, and Spanish, respectively, out of which 6 teams submitted runs for all languages. Table 10 gives an overview of the submitted systems per language. This was a binary classification task, and we used F1 score with respect to the positive class as the official evaluation measure. Table 11 shows the performance of the top official submissions on the test set.

Table 10. Overview of the systems for **subtask 1B**. The numbers in the language box refer to the position of the team in the official ranking; ☑=part of the official submission.

Team		Langs			Models											Misc.	
		English	Arabic	Spanish	BERT	RoBERTa	XLM RoBERTa	MarBERT	AraBERT	BERTweet	BETO	BERTIN	Spanish RoBERTa	GPT	MultinomialNB	LSTM	Data augmentation	Preprocessing
Accenture	[99]	3	2	5	☑	☑											☑	
CSECU-DSG	[12]	7	4	3			☑		☑	☑	☑					☑		
DSHacker	[64]	9	5	1			☑							☑			☑	
ES-VRAI	[81]	5	1	2			☑	☑		☑								
Fraunhofer SIT	[40]	2			☑													☑
OpenFact	[85]	1												☑				
Z-Index	[97]	6	3	6			☑											☑

Table 11. Subtask 1B: Multigenre (unimodal) check-worthiness estimation. Shown are the top-3 submissions for English debates, and for Arabic and Spanish tweets. The F1 score is calculated with respect to the positive class.

	Team	F1		Team	F1		Team	F1
	English			**Arabic**			**Spanish**	
1	OpenFact [85]	0.898	1	ES-VRAI [81]	0.809	1	DSHacker [64]	0.641
2	Fraunhofer SIT [40]	0.878	2	Accenture [99]	0.733	2	ES-VRAI [81]	0.627
3	Accenture [99]	0.860	3	Z-Index [97]	0.710	3	CSECU-DSG [12]	0.599

The best-performing team on English is *OpenFact* [85], who fine-tuned GPT-3[6] using 7.7K examples of sentences from debates and speeches annotated for check-worthiness, extracted from an already existing dataset [9]. In addition to that, during internal experiments, the team also experimented with fine-tuning a variety of BERT models and found that fine-tuning DeBERTaV3 [50] leads to near-identical performance to that of the model based on GPT-3.

Team *Fraunhofer SIT* [40] fine-tuned a BERT model [33] three times starting with a different seed for model initialization, resulting in three models. The team used ensemble learning using a model souping technique that adaptively adjusts the influence of each individual model based on its performance on the dev subset.

[6] https://platform.openai.com/docs/models/gpt-3

Team *Accenture* [99] also fine-tuned large pre-trained models: RoBERTa [60] for English and GigaBERT for Arabic [55]. They further proposed to extend the training subset with examples resulting from back-translating the same subset using AWS translation.[7]

Team *ES-VRAI* [81] achieved the best and the second best performance for Arabic and for Spanish, respectively. After comprehensive evaluation of several language-specific pre-trained models, their official submission for Arabic was based on fine-tuning MARBERT [1] using the training subset, after downsampling examples from the majority class. Fine-tuned XLM-RoBERTa model was used to produce the official submitted run for the Spanish test set.

Team *Z-Index* [97] participated in all three languages using the same system architecture. Their system includes a feed forward network, where input is represented using embeddings.[8] The network was trained using the training set released per language.

Team *DSHacker* [64] achieved the best performance for Spanish. Their system is based on fine-tuning XLM-RoBERTa [28] using the available train data, and additional datasets obtained by data augmentation. For data augmentation, they used GPT-3.5[9] to translate the input train set to English and to Arabic resulting in two additional training subsets. GPT-3.5 was also used to paraphrase the original Spanish training data, resulting in a third augmented training subset.

Team *CSECU-DSG* [12] also participated in all three languages. Their model includes jointly fine-tuning two transformers: a language-specific BERT and Twitter XLM-RoBERTa [17] to represent the input text. In addition, a BiLSTM module was used on top of the text features to handle long-term contextual dependency. Finally, the features from the BiLSTM were followed by a multisample dropout strategy [51] to produce the final prediction.

5.2 Task 2: Subjectivity in News Articles

Task 2 has seen the participation of 12 teams, with a total of 45 runs. The majority of the participants (seven out of 12) submitted runs for more than one language, with four teams participating in all languages.

Table 12 offers a snapshot of the approaches, whereas Table 13 reports the performance results for the top-three submissions per task, ranked on the basis of macro-averaged F_1 (cf. [41] for the whole ranking, including submissions after the deadline).

All systems used neural networks. Two teams, Fraunhofer SIT [39] and TOBB ETU [31], based their submissions on GPT-3* models, structuring the task via prompts in zero-shot or few-shot settings. All other participants fine-tuned encoder-based Transformers mostly using multi-lingual models (e.g., mBERTaV3, XML-R, and mBERT). Generative models based on the GPT-3* family were mostly used to augment the training data, rather than adopting standard upsampling and downsampling methods.

[7] https://aws.amazon.com/translate/
[8] No enough details were available about the source of these embeddings.
[9] https://platform.openai.com/docs/models/gpt-3-5

Table 12. Task 2 Overview of the approaches. The numbers in the language box refer to the position of the team in the official ranking.

Team		Languages							Models												Misc			
		Multilingual	Arabic	Dutch	German	English	Italian	Turkish	BERT	RoBERTa	XLM RoBERTa	GigaBERT	M-BERT	M-DeBERTa	S-BERT	SetFit	ChatGPT	GPT-3	BART	LSTM	Gradient Boosting	Multi-lingual training	Data augmentation	Feature Selection	Ensemble
Accenture	[100]		3	5	7	8	3	4	✓	✓		✓											✓		
Awakened						10													✓	✓					✓
DWReCo	[86]				4	1		2	✓	✓								✓					✓		
ES-VRAI	[82]	-											✓												
Fraunhofer SIT	[39]				5	6												✓							
Gpachov	[76]					2					✓				✓	✓									✓
KUCST						4			✓												✓		✓		
NN	[34]	1	1	2	2	5	2	3			✓											✓			
tarrekko		-	-	-	-	-	-	-																	
Thesis Titan	[57]	2	2	1	1	3	1	1						✓								✓			
TOBB ETU	[31]	4	5	3	3	9	5	6									✓								
TUDublin	[95]					11	6						✓				✓						✓		

- Run submitted after the deadline.

Only two teams, Thesis Titan [57] and NN [34] achieved consistently good results across all languages, with a ranking in the top-3 positions. As in previous editions of the `CheckThat!` lab, using all languages helped to boost the performance.

All the top systems are in the range [0.75, 0.80] in terms of macro-F1, well above the baselines. The only language outlier is Turkish, where the top approach by Thesis Titan [57] achieved an outstanding F1 score of 0.89. In the majority of the languages, the distance between the first and the second system tends to be higher than two points. In the Arabic, English, and multilingual settings, the distances range between 0.05 and 0.01 points. Rather than pointing only to differences in the annotation of the data, this may suggest that some approaches have found optimal language-specific hyper-parameters.

5.3 Task 3: Political Bias of News Articles and News Media

Table 14 shows the results for Task 3, in which four teams participated. Two teams participated in both subtasks, and all teams outperformed the baseline.

Team Accenture [102] used back-translation to augment the minority classes in the datasets that label the article and the news source bias into three categories: left, center, and right. Then, they used this augmented data to fine-tune RoBERTa transformer models. Team TOBB ETU [31] explored zero-shot and few-shot classification by using ChatGPT exclusively for subtask 3A.

Table 13. Task 2 Top-3 performing models per language.

Team	F1	Team	F1	Team	F1
Multilingual		**English**		**Italian**	
1 NN [34]	81.97	1 DWReCo [86]	78.18	1 Thesis Titan [57]	75.75
2 Thesis Titan [57]	81.00	2 Gpachov [76]	77.34	2 NN [34]	71.01
3 *baseline*	73.56	3 Thesis Titan [57]	76.78	3 Accenture [100]	65.52
Arabic		**German**		**Turkish**	
1 NN [34]	78.75	1 Thesis Titan [57]	81.52	1 Thesis Titan [57]	89.94
2 Thesis Titan [57]	77.53	2 NN [34]	74.13	2 DWReCo [86]	84.11
3 Accenture [100]	72.53	3 TOBB ETU [31]	71.19	3 NN [34]	81.21
Dutch					
† 1 Thesis Titan [57]	81.43				
2 NN [34]	75.57				
3 TOBB ETU [31]	73.01				

† Team involved in the preparation of the data.

Table 14. Task 3: Top-3 performing models when identifying political bias of news articles and news media (MAE score).

Subtask 3A		Subtask 3B	
Team	bf MAE	**Team**	**MAE**
1 Accenture [102]	0.473	1 Accenture [102]	0.549
2 TOBB ETU [31]	0.646	2 Awakened [103]	0.765
3 KUCST	0.736	3 *baseline*	0.902

5.4 Task 4: Factuality of Reporting of News Media

Five teams participated in this task, with participants proposing three distinct approaches to predict the veracity of the news outlets.

In an effort to reduce the influence of redundant data and to enhance the model resilience, team CUCPLUS [58] used RoBERTa coupled with regularized adversarial training. Team Accenture [101] aimed at maximizing the amount of training data and developed a RoBERTa model that learns the factual reporting patterns of news articles and news sources.

5.5 Task 5: Authority Finding in Twitter

Two teams participated in this task, submitting four runs. Both teams adopted the Twitter profile name and descriptions, and the Twitter lists as a user presentation. Moreover, both teams incorporated the lexical matching between the rumor and the users in addition to the users network features to retrieve the corresponding authorities.

Table 15. Task 4: Top-3 models on the factuality of reporting of news media outlets task (MAE score).

Team	MAE
1 CUCPLUS [58]	0.295
2 NLPIR-UNED	0.344
3 Accenture [101]	0.467

Table 16. Task 5 Evaluation results, in terms of P@1, P@5, and nDCG@5, ranked by P@5. Teams with a + sign include task organisers.

Team (run ID)	P@5	P@1	nDCG@5
1 +bigIR (Hybrid3)	0.260	0.367	0.297
2 +bigIR (Hybrid1)	0.247	0.367	0.282
3 +bigIR (Hybrid2)	0.227	0.333	0.247
BM25 baseline	0.087	0.133	0.104
4 ES-VRAI [83] (Model1)	0.067	0.067	0.071

Team bigIR further used semantic matching by adopting Arabic BERT [84] fine-tuned on the full training data and deployed in the Tahaqqaq real-time system [94]. As shown in Table 16, all runs by the bigIR team managed to outperform the baseline by a sizable margin.

6 Related Work

Related work has focused on detecting misinformation and fact-checking across a variety of sources: news articles, forums, and social media [13,15,75,108]. This has given rise to variety of tasks, such as claim extraction [80], check-worthiness estimation [52,68], relevant document retrieval [70,107], detecting previously fact-checked claims [62,87,88], profiling articles and news outlets for their bias [14,96] and factuality [15,16,77], and recommendation systems to encourage people to engage in fact-checking [104]. There have been also a number of related shared tasks, which focused on rumour veracity [32], fact-checking in community question answering forums [63], propaganda techniques and framing in text and images [30,35,78,93], and fact verification and evidence finding for tabular data in scientific documents [106]. Other initiatives include FEVER [98], the Fake News Challenge [42], and the multimodal task at MediaEval [79].

7 Conclusions and Future Work

We presented the 2023 edition of the `CheckThat!` lab, which was once again one of the most popular CLEF labs, attracting a total of 37 active participating teams. This year, `CheckThat!` offered five tasks in seven languages: Arabic, Dutch, English, German, Italian, Spanish, and Turkish.

Task 1 focused on determining the check-worthiness of an item, whether it is a text or a combination of a text and image. Task 2 asked to predict the subjectivity or the objectivity of sentences. Task 3 aimed at detecting the political bias both at the level of a news article and of a news medium. Task 4 asked to measure the level of factuality of reporting of a news medium. Finally, Task 5 tasked the participants to identify authoritative sources on Twitter that could assist in verifying a given input claim. Tasks 2, 3, 4, and 5 were organized this year for the first time. For Task 1, most teams used large pre-trained models, OCR and data augmentation. In Task 2, most teams relied on transformers, and some used generative models (GPT*) to augment the training data or to flag subjective sentences. In Task 3, the most successful participants used RoBERTa and ChatGPT. In Task 4, most participants used RoBERTa, and some used stylistic features. In Task 5, the best team used lexical and semantic matching.

In the future, we plan to continue this year's trend of considering tasks that could play a relevant role in the analysis of journalistic and social media posts, and that go beyond factuality.

8 Acknowledgments

The work of F. Alam, M. Hasanain and W. Zaghouani is partially supported by NPRP 13S-0206-200281 and NPRP 14C-0916-210015 from the Qatar National Research Fund (a member of Qatar Foundation). The work of A. Galassi is supported by the European Commission NextGeneration EU programme, PNRR-M4C2-Investimento 1.3, PE00000013-"FAIR" Spoke 8. The work of Fatima Haouari was supported by GSRA grant #GSRA6-1-0611-19074 from the Qatar National Research Fund. The work of Tamer Elsayed was made possible by NPRP grant #NPRP-11S-1204-170060 from the Qatar National Research Fund.

The findings achieved herein are solely the responsibility of the authors.

References

1. Abdul-Mageed, M., Elmadany, A., et al.: Arbert & marbert: Deep bidirectional transformers for arabic. 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). pp. 7088–7105 (2021)
2. Alam, F., Barrón-Cedeño, A., Cheema, G.S., Hakimov, S., Hasanain, M., Li, C., Míguez, R., Mubarak, H., Shahi, G.K., Zaghouani, W., Nakov, P.: Overview of the CLEF-2023 CheckThat! lab task 1 on check-worthiness in multimodal and multigenre content. In: Aliannejadi et al. [6]
3. Alam, F., Dalvi, F., Shaar, S., Durrani, N., Mubarak, H., Nikolov, A., Da San Martino, G., Abdelali, A., Sajjad, H., Darwish, K., Nakov, P.: Fighting the COVID-19 infodemic in social media: A holistic perspective and a call to arms. In: Proceedings of the Conference on Web and Social Media. pp. 913–922. ICWSM '21 (2021)

4. Alam, F., Shaar, S., Dalvi, F., Sajjad, H., Nikolov, A., Mubarak, H., Da San Martino, G., Abdelali, A., Durrani, N., Darwish, K., Al-Homaid, A., Zaghouani, W., Caselli, T., Danoe, G., Stolk, F., Bruntink, B., Nakov, P.: 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. Aliannejadi, M., Faggioli, G., Ferro, N., Vlachos, Michalis (eds.): Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum. CLEF 2023, Thessaloniki, Greece (2023)
7. Antici, F., Galassi, A., Ruggeri, F., Korre, K., Muti, A., Bardi, A., Fedotova, A., Barrón-Cedeño, A.: A corpus for sentence-level subjectivity detection on english news articles (2023)
8. Antoun, W., Baly, F., Hajj, H.: AraBERT: Transformer-based model for Arabic language understanding. In: Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools. pp. 9–15. OSAC '20, Marseille, France (2020)
9. Arslan, F., Hassan, N., Li, C., Tremayne, M.: A benchmark dataset of check-worthy factual claims. In: Proceedings of the International AAAI Conference on Web and Social Media. vol. 14, pp. 821–829 (2020)
10. Atanasova, P., Marquez, L., Barrón-Cedeño, A., Elsayed, T., Suwaileh, R., Zaghouani, W., Kyuchukov, S., Da San Martino, G., Nakov, P.: Overview of the CLEF-2018 CheckThat! lab on automatic identification and verification of political claims. Task 1: Check-worthiness. In: Cappellato et al. [24]
11. Atanasova, P., Nakov, P., Karadzhov, G., Mohtarami, M., Da San Martino, G.: Overview of the CLEF-2019 CheckThat! lab on automatic identification and verification of claims. Task 1: Check-worthiness. In: Cappellato et al. [23]
12. Aziz, A., Hossain, M., Chy, A.: CSECU-DSG at CheckThat! 2023: Transformer-based fusion approach for multimodal and multigenre check-worthiness. In: Aliannejadi et al. [6]
13. 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 companion on world wide web. pp. 159–162 (2016)
14. Baly, R., Da San Martino, G., Glass, J., Nakov, P.: We can detect your bias: Predicting the political ideology of news articles. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. pp. 4982–4991. EMNLP '20 (2020)
15. Baly, R., Karadzhov, G., An, J., Kwak, H., Dinkov, Y., Ali, A., Glass, J., Nakov, P.: 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. pp. 3364–3374 (2020)
16. Baly, R., Karadzhov, G., Saleh, A., Glass, J., Nakov, P.: Multi-task ordinal regression for jointly predicting the trustworthiness and the leading political ideology of news media. In: Proceedings of the 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 2109–2116. NAACL-HLT '19, Minneapolis, MN, USA (2019)
17. Barbieri, F., Espinosa Anke, L., Camacho-Collados, J.: Xlm-t: A multilingual language model toolkit for twitter. arXiv e-prints pp. arXiv–2104 (2021)

18. Barrón-Cedeño, A., Alam, F., Galassi, A., Da San Martino, G., Nakov, P., , Elsayed, T., Azizov, D., Caselli, T., Cheema, G., Haouari, F., Hasanain, M., Kutlu, M., Li, C., Ruggeri, F., Struß, J.M., Zaghouani, W.: The CLEF-2023 CheckThat! Lab: Checkworthiness, subjectivity, political bias, factuality, and authority. In: Kamps, J., Goeuriot, L., Crestani, F., Maistro, M., Joho, H., Davis, B., Gurrin, C., Kruschwitz, U., Caputo, A. (eds.) Advances in Information Retrieval. pp. 506–517. Springer Nature Switzerland, Cham (2023)
19. Barrón-Cedeño, A., Elsayed, T., Nakov, P., Da San Martino, G., Hasanain, M., Suwaileh, R., Haouari, F.: CheckThat! at CLEF 2020: Enabling the automatic identification and verification of claims in social media. In: Advances in Information Retrieval. pp. 499–507. Springer International Publishing, Cham (2020)
20. Barrón-Cedeño, A., Elsayed, T., Nakov, P., Da San Martino, G., Hasanain, M., Suwaileh, R., Haouari, F., Babulkov, N., Hamdan, B., Nikolov, A., Shaar, S., Sheikh Ali, Z.: Overview of CheckThat! 2020: Automatic identification and verification of claims in social media. In: Arampatzis, A., Kanoulas, E., Tsikrika, T., Vrochidis, S., Joho, H., Lioma, C., Eickhoff, C., Névéol, A., Cappellato, L., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Eleventh International Conference of the CLEF Association (CLEF 2020). pp. 215–236. LNCS (12260), Springer (2020)
21. Barrón-Cedeño, A., Elsayed, T., Suwaileh, R., Marquez, L., Atanasova, P., Zaghouani, W., Kyuchukov, S., Da San Martino, G., Nakov, P.: Overview of the CLEF-2018 CheckThat! lab on automatic identification and verification of political claims. Task 2: Factuality. In: Cappellato et al. [24]
22. Cappellato, L., Eickhoff, C., Ferro, N., Névéol, A. (eds.): CLEF 2020 Working Notes. CEUR Workshop Proceedings (2020)
23. Cappellato, L., Ferro, N., Losada, D., Müller, H. (eds.): Working Notes of CLEF 2019 Conference and Labs of the Evaluation Forum. CEUR Workshop Proceedings (2019)
24. Cappellato, L., Ferro, N., Nie, J.Y., Soulier, L. (eds.): Working Notes of CLEF 2018–Conference and Labs of the Evaluation Forum. CEUR Workshop Proceedings (2018)
25. Cheema, G.S., Hakimov, S., Sittar, A., Müller-Budack, E., Otto, C., Ewerth, R.: MM-Claims: A dataset for multimodal claim detection in social media. In: Findings of the Association for Computational Linguistics: NAACL 2022. pp. 962–979. Seattle, Washington (2022)
26. Clark, K., Luong, M., Le, Q.V., Manning, C.D.: ELECTRA: pre-training text encoders as discriminators rather than generators. In: Proceedings of the 8th International Conference on Learning Representations. ICLR '20 (2020)
27. Cohen, J.: A coefficient of agreement for nominal scales. Educational and psychological measurement **20**(1), 37–46 (1960)
28. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. CoRR **abs/1911.02116** (2019)
29. Da San Martino, G., Alam, F., Hasanain, M., Nandi, R.N., Azizov, D., Nakov, P.: Overview of the CLEF-2023 CheckThat! lab task 3 on political bias of news articles and news media. In: Aliannejadi et al. [6]
30. Da San Martino, G., Barrón-Cedeño, A., Wachsmuth, H., Petrov, R., Nakov, P.: Semeval-2020 task 11: Detection of propaganda techniques in news articles. In: Proceedings of the Workshop on Semantic Evaluation. pp. 1377–1414 (2020)

31. Deniz Türkmen, M., Coşgun, G., Kutlu, M.: TOBB ETU at CheckThat! 2023: Utilizing chatgpt to detect subjective statements and political bias. In: Aliannejadi et al. [6]
32. 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)
33. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 4171–4186. NAACL-HLT '19, Minneapolis, MN, USA (2019)
34. Dey, K., Tarannum, P., Hasan, M.A., Noori, S.R.H.: Nn at CheckThat! 2023: Subjectivity in news articles classification with transformer based models. In: Aliannejadi et al. [6]
35. Dimitrov, D., Bin Ali, B., Shaar, S., Alam, F., Silvestri, F., Firooz, H., Nakov, P., Da San Martino, G.: SemEval-2021 Task 6: Detection of persuasion techniques in texts and images. In: Proceedings of the International Workshop on Semantic Evaluation. pp. 70–98. SemEval '21 (2021)
36. Elsayed, T., Nakov, P., Barrón-Cedeño, A., Hasanain, M., Suwaileh, R., Da San Martino, G., Atanasova, P.: 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)
37. Elsayed, T., Nakov, P., Barrón-Cedeño, A., Hasanain, M., Suwaileh, R., Da San Martino, G., Atanasova, P.: 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)
38. 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)
39. Frick, R.A.: Fraunhofer sit at CheckThat! 2023: Can llms be used for data augmentation & few-shot classification? detecting subjectivity in text using chatgpt. In: Aliannejadi et al. [6]
40. Frick, R.A., Vogel, I., Choi, J.E.: Fraunhofer SIT at CheckThat! 2023: Enhancing the detection of multimodal and multigenre check-worthiness using optical character recognition and model souping. In: Aliannejadi et al. [6]
41. Galassi, A., Ruggeri, F., Barrón-Cedeño, A., Alam, F., Caselli, T., Kutlu, M., Struss, J., Antici, F., Hasanain, M., Köhler, J., Korre, K., Leistra, F., Muti, A., Siegel, M., Mehmet Deniz, T., Wiegand, M., Zaghouani, W.: Overview of the CLEF-2023 CheckThat! lab task 2 on subjectivity in news articles. In: Aliannejadi et al. [6]
42. Hanselowski, A., Avinesh, P., Schiller, B., Caspelherr, F., Chaudhuri, D., Meyer, C.M., Gurevych, I.: A retrospective analysis of the fake news challenge stance-detection task. In: Proceedings of the 27th International Conference on Computational Linguistics. pp. 1859–1874 (2018)
43. Haouari, F., Elsayed, T.: Detecting Stance of Authorities Towards Rumors in Arabic Tweets: A Preliminary Study. In: Advances in Information Retrieval. pp. 430–438. Springer Nature Switzerland, Cham (2023)

44. Haouari, F., Elsayed, T., Mansour, W.: Who can verify this? finding authorities for rumor verification in Twitter. Information Processing & Management **60**(4), 103366 (2023)
45. Haouari, F., Hasanain, M., Suwaileh, R., Elsayed, T.: ArCOV19-Rumors: Arabic COVID-19 Twitter Dataset for Misinformation Detection. In: Proceedings of the Sixth Arabic Natural Language Processing Workshop. pp. 72–81 (2021)
46. Haouari, F., Sheikh Ali, Z., Elsayed, T.: Overview of the CLEF-2023 CheckThat! lab task 5 on authority finding in twitter. In: Aliannejadi et al. [6]
47. Hasanain, M., Haouari, F., Suwaileh, R., Ali, Z., Hamdan, B., Elsayed, T., Barrón-Cedeño, A., Da San Martino, G., Nakov, P.: Overview of CheckThat! 2020 Arabic: Automatic identification and verification of claims in social media. In: Cappellato et al. [22]
48. Hasanain, M., Suwaileh, R., Elsayed, T., Barrón-Cedeño, A., Nakov, P.: Overview of the CLEF-2019 CheckThat! lab on automatic identification and verification of claims. Task 2: Evidence and factuality. In: Cappellato et al. [23]
49. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778. CVPR '16, Las Vegas, Nevada, USA (2016)
50. He, P., Gao, J., Chen, W.: Debertav3: improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. In: The Eleventh International Conference on Learning Representations (2023)
51. Inoue, H.: Multi-sample dropout for accelerated training and better generalization. CoRR **abs/1905.09788** (2019)
52. Kartal, Y.S., Kutlu, M.: Re-think before you share: A comprehensive study on prioritizing check-worthy claims. IEEE Transactions on Computational Social Systems (2022)
53. Khalil, A., Jarrah, M., Aldwairi, M., Jararweh, Y.: Detecting Arabic fake news using machine learning. In: Proceedings of the International Conference on Intelligent Data Science Technologies and Applications. pp. 171–177 (2021)
54. Köhler, J., Shahi, G.K., Struß, J.M., Wiegand, M., Siegel, M., Mandl, T., Schütz, M.: 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)
55. Lan, W., Chen, Y., Xu, W., Ritter, A.: An empirical study of pre-trained transformers for arabic information extraction. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 4727–4734 (2020)
56. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics pp. 159–174 (1977)
57. Leistra, F., Caselli, T.: Thesis titan at CheckThat! 2023: Language-specific fine-tuning of mdebertav3 for subjectivity detection. In: Aliannejadi et al. [6]
58. Li, C., Xue, R., Lin, C., Fan, W.: CUCPLUS at CheckThat! 2023: Text combination and regularized adversarial training for news media factuality evaluation. In: Aliannejadi et al. [6]
59. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A robustly optimized BERT pretraining approach. CoRR **abs/1907.11692** (2019)
60. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)

61. Liu, Z., Mao, H., Wu, C., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11966–11976. CVPR '22, New Orleans, LA, USA (2022)
62. Mansour, W., Elsayed, T., Al-Ali, A.: This is not new! spotting previously-verified claims over Twitter. Information Processing & Management **60**(4), 103414 (2023)
63. 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 Workshop on Semantic Evaluation. pp. 860–869 (2019)
64. Modzelewski, A., Sosnowski, W., Wierzbicki, A.: DSHacker at CheckThat! 2023: Check-Worthiness in Multigenre and Multilingual Content With GPT-3.5 Data Augmentation. In: Aliannejadi et al. [6]
65. Nakov, P., Alam, F., Da San Martino, G., Hasanain, M., Nandi, R.N., Azizov, D., Panayotov, P.: Overview of the CLEF-2023 CheckThat! lab task 4 on factuality of reporting of news media. In: Aliannejadi et al. [6]
66. Nakov, P., Barrón-Cedeño, A., Da San Martino, G., Alam, F., Míguez, R., Caselli, T., Kutlu, M., Zaghouani, W., Li, C., Shaar, S., Mubarak, H., Nikolov, A., Kartal, Y.S., Beltrán, J.: 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)
67. Nakov, P., Barrón-Cedeño, A., Da San Martino, G., Alam, F., Struß, J.M., Mandl, T., Míguez, R., Caselli, T., Kutlu, M., Zaghouani, W., Li, C., Shaar, S., Shahi, G.K., Mubarak, H., Nikolov, A., Babulkov, N., Kartal, Y.S., Beltrán, J., Wiegand, M., Siegel, M., Köhler, J.: Overview of the CLEF-2022 CheckThat! lab on fighting the COVID-19 infodemic and fake news detection. In: Proc. of the Conference of the CLEF Association: Information Access Evaluation meets Multilinguality, Multimodality, and Visualization. CLEF '2022, Bologna, Italy (2022)
68. Nakov, P., Barrón-Cedeño, A., Da San Martino, G., Alam, F., Struß, J.M., Mandl, T., Míguez, R., Caselli, T., Kutlu, M., Zaghouani, W., Li, C., Shaar, S., Shahi, G.K., Mubarak, H., Nikolov, A., Babulkov, N., Kartal, Y.S., Beltrán, J.: The CLEF-2022 CheckThat! Lab on fighting the COVID-19 infodemic and fake news detection. In: Advances in Information Retrieval. pp. 416–428. Springer International Publishing, Cham (2022)
69. Nakov, P., Barrón-Cedeño, A., Elsayed, T., Suwaileh, R., Màrquez, L., Zaghouani, W., Gencheva, P., Kyuchukov, S., Da San Martino, G.: 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 '18 (2018)
70. Nakov, P., Corney, D., Hasanain, M., Alam, F., Elsayed, T., Barrón-Cedeño, A., Papotti, P., Shaar, S., Martino, G.D.S.: Automated fact-checking for assisting human fact-checkers. In: Proceedings of the 30th International Joint Conference on Artificial Intelligence. pp. 4551–4558 (2021)
71. 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)
72. Nakov, P., Da San Martino, G., Elsayed, T., Barrón-Cedeño, A., Míguez, R., Shaar, S., Alam, F., Haouari, F., Hasanain, M., Babulkov, N., Nikolov, A., Shahi, G.K., Struß, J.M., Mandl, T.: The CLEF-2021 CheckThat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news. In: Advances in Information Retrieval. pp. 639–649. Springer International Publishing, Cham (2021)

73. Nakov, P., Da San Martino, G., Elsayed, T., Barrón-Cedeño, A., Míguez, R., Shaar, S., Alam, F., Haouari, F., Hasanain, M., Mansour, W., Hamdan, B., Ali, Z.S., Babulkov, N., Nikolov, A., Shahi, G.K., Struß, J.M., Mandl, T., Kutlu, M., Kartal, Y.S.: Overview of the CLEF-2021 CheckThat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news. In: Candan, K., Ionescu, B., Goeuriot, L., Larsen, B., Müller, H., Joly, A., Maistro, M., Piroi, F., Faggioli, G., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Twelfth International Conference of the CLEF Association. LNCS (12880), Springer (2021)
74. Nguyen, D.Q., Vu, T., Nguyen, A.T.: Bertweet: A pre-trained language model for english tweets. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 9–14 (2020)
75. 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 and Knowledge Management. pp. 1165–1174. CIKM '20 (2020)
76. Pachov, G., Dimitrov, D., Koychev, I., Nakov, P.: Gpachov at CheckThat! 2023: A diverse multi-approach ensemble for subjectivity detection in news articles. In: Aliannejadi et al. [6]
77. Panayotov, P., Shukla, U., Sencar, H.T., Nabeel, M., Nakov, P.: GREENER: Graph neural networks for news media profiling. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 7470–7480. EMNLP '22, Abu Dhabi, UAE (2022)
78. Piskorski, J., Stefanovitch, N., Da San Martino, G., Nakov, P.: SemEval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multi-lingual setup. In: Proceedings of the 17th International Workshop on Semantic Evaluation. SemEval'23, Toronto, Canada (July 2023)
79. Pogorelov, K., Schroeder, D.T., Burchard, L., Moe, J., Brenner, S., Filkukova, P., Langguth, J.: FakeNews: Corona virus and 5G conspiracy task at MediaEval 2020. In: MediaEval (2020)
80. Reddy, R.G., Chinthakindi, S.C., Wang, Z., Yi Fung, K.C., Elsayed, A., Palmer, M., Nakov, P., Hovy, E., Small, K., Ji, H.: NewsClaims: A new benchmark for claim detection from news with attribute knowledge. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 6002–6018. EMNLP '22, Abu Dhabi, UAE (December 2022)
81. Sadouk, H.T., Sebbak, F., Zekiri, H.E.: Es-vrai at CheckThat! 2023: Analyzing checkworthiness in multimodal and multigenre contents through fusion and sampling approaches. In: Aliannejadi et al. [6]
82. Sadouk, H.T., Sebbak, F., Zekiri, H.E.: Es-vrai at CheckThat! 2023: Enhancing model performance for subjectivity detection through multilingual data augmentation. In: Aliannejadi et al. [6]
83. Sadouk, H.T., Sebbak, F., Zekiri, H.E.: Es-vrai at CheckThat! 2023: Leveraging bio and lists information for enhanced rumor verification in twitter. In: Aliannejadi et al. [6]
84. Safaya, A., Abdullatif, M., Yuret, D.: KUISAIL at SemEval-2020 task 12: BERT-CNN for offensive speech identification in social media. In: Proceedings of the Fourteenth Workshop on Semantic Evaluation. pp. 2054–2059. International Committee for Computational Linguistics, Barcelona (online) (Dec 2020)
85. Sawiński, M., Wecel, K., Księżniak, E., Stróżyna, M., Lewoniewski, W., Stolarski, P., Abramowicz, W.: Openfact at CheckThat! 2023: Head-to-head gpt vs. bert -

a comparative study of transformers language models for the detection of check-worthy claims. In: Aliannejadi et al. [6]
86. Schlicht, I.B., Khellaf, L., Altiok, D.: Dwreco at CheckThat! 2023: Enhancing subjectivity detection through style-based data sampling. In: Aliannejadi et al. [6]
87. Shaar, S., Alam, F., Da San Martino, G., Nakov, P.: The role of context in detecting previously fact-checked claims. In: Findings of the Association for Computational Linguistics: NAACL 2022. pp. 1619–1631 (2022)
88. 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. pp. 3607–3618 (2020)
89. Shaar, S., Haouari, F., Mansour, W., Hasanain, M., Babulkov, N., Alam, F., Da San Martino, G., Elsayed, T., Nakov, P.: Overview of the CLEF-2021 CheckThat! lab task 2 on detecting previously fact-checked claims in tweets and political debates. In: Faggioli et al. [38]
90. Shaar, S., Hasanain, M., Hamdan, B., Ali, Z.S., Haouari, F., Nikolov, A., Kutlu, M., Kartal, Y.S., Alam, F., Da San Martino, G., Barrón-Cedeño, A., Míguez, R., Beltrán, J., Elsayed, T., Nakov, P.: Overview of the CLEF-2021 CheckThat! lab task 1 on check-worthiness estimation in tweets and political debates. In: Faggioli et al. [38]
91. Shaar, S., Nikolov, A., Babulkov, N., Alam, F., Barrón-Cedeño, A., Elsayed, T., Hasanain, M., Suwaileh, R., Haouari, F., Da San Martino, G., Nakov, P.: Overview of CheckThat! 2020 English: Automatic identification and verification of claims in social media. In: Cappellato et al. [22]
92. Shahi, G.K., Struß, J.M., Mandl, T.: Overview of the CLEF-2021 CheckThat! lab: Task 3 on fake news detection. In: Faggioli et al. [38]
93. Sharma, S., Suresh, T., Kulkarni, A., Mathur, H., Nakov, P., Akhtar, M.S., Chakraborty, T.: Findings of the CONSTRAINT 2022 shared task on detecting the hero, the villain, and the victim in memes. In: Proceedings of the Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situations. pp. 1–11. Dublin, Ireland (2022)
94. Sheikh Ali, Z., Mansour, W., Haouari, F., Hasanain, M., Elsayed, T., Al-Ali, A.: Tahaqqaq: A Real-Time System for Assisting Twitter Users in Arabic Claim Verification. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (2023)
95. Shushkevich, E., Cardiff, J.: Tudublin at CheckThat! 2023: Chatgpt for data augmentation. In: Aliannejadi et al. [6]
96. Stefanov, P., Darwish, K., Atanasov, A., Nakov, P.: Predicting the topical stance and political leaning of media using tweets. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics. pp. 527–537. ACL '20 (2020)
97. Tarannum, P., Hasan, M.A., Alam, F., Noori, S.R.H.: Z-Index at CheckThat! 2023: Unimodal and multimodal checkworthiness classification. In: Aliannejadi et al. [6]
98. Thorne, J., Vlachos, A., Christodoulopoulos, C., Mittal, A.: FEVER: a large-scale dataset for fact extraction and VERification. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 809–819. NAACL-HLT '18, New Orleans, Louisiana, USA (2018)
99. Tran, S., Rodrigues, P., Strauss, B., Williams, E.: Accenture at CheckThat! 2023: Identifying claims with societal impact using nlp data augmentation. In: Aliannejadi et al. [6]

100. Tran, S., Rodrigues, P., Strauss, B., Williams, E.: Accenture at CheckThat! 2023: Impacts of back-translation on subjectivity detection. In: Aliannejadi et al. [6]
101. Tran, S., Rodrigues, P., Strauss, B., Williams, E.: Accenture at CheckThat! 2023: Learning to detect factuality levels of news sources. In: Aliannejadi et al. [6]
102. Tran, S., Rodrigues, P., Strauss, B., Williams, E.: Accenture at CheckThat! 2023: Learning to detect political bias of news articles and sources. In: Aliannejadi et al. [6]
103. 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)
104. Vo, N., Lee, K.: The rise of guardians: Fact-checking url recommendation to combat fake news. In: Proceedings of the International ACM SIGIR Conference on Research & Development in Information Retrieval. pp. 275–284. SIGIR '18 (2018)
105. von Däniken, P., Deriu, J., Cieliebak, M.: Zhaw-cai at CheckThat! 2023: Ensembling using kernel averaging. In: Aliannejadi et al. [6]
106. 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. pp. 317–326. SemEval '21 (2021)
107. 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)
108. Zubiaga, A., Liakata, M., Procter, R., Wong Sak Hoi, G., Tolmie, P.: Analysing how people orient to and spread rumours in social media by looking at conversational threads. PloS one **11**(3), e0150989 (2016)