Overview of the CLEF-2024 CheckThat! Lab: Check-Worthiness, Subjectivity, Persuasion, Roles, Authorities, and Adversarial Robustness

Alberto Barrón-Cedeño[1[0000−0003−4719−3420]], Firoj Alam[2[0000−0001−7172−1997]], Julia Maria Struß[3[0000−0001−9133−4978]], Preslav Nakov[4[0000−0002−3600−1510]], Tanmoy Chakraborty[5[x]], Tamer Elsayed[6[0000−0001−5786−4668]], Piotr Przybyła,[7,8[0000−0001−9043−6817]], Tommaso Caselli[9[0000−0003−2936−0256]], Giovanni Da San Martino[10[0000−0002−2609−483X]], Fatima Haouari[6[0000−0003−4842−2467]], Maram Hasanain[2[0000−0002−7466−178X]], Chengkai Li[11[0000−0002−1724−8278]], Jakub Piskorski[12], Federico Ruggeri[1[0000−0002−1697−8586]], Xingyi Song[13[0000−0002−4188−6974]], and Reem Suwaileh[14[0000−0001−7341−1407]]

[1]Università di Bologna, Forlì, Italy [2]Qatar Computing Research Institute, Qatar [3]University of Applied Sciences Potsdam, Germany [4]Mohamed bin Zayed University of Artificial Intelligence, UAE [5]Indian Institute of Technology Delhi, New Delhi, India [6]Qatar University, Qatar [7]Univesitat Pompeu Fabra, Barcelona, Spain [8]Institute of Computer Science, Polish Academy of Sciences, Poland [9]University of Groningen, Netherlands [10]University of Padua, Padova, Italy [11]University of Texas at Arlington, USA [12]Polish Academy of Sciences, Poland [13]University of Sheffield, UK [14]Hamad bin Khalifa University, Qatar
https://checkthat.gitlab.io

Abstract. We describe the seventh edition of the `CheckThat!` lab, part of the 2024 Conference and Labs of the Evaluation Forum (CLEF). Previous editions of `CheckThat!` focused on the main tasks of the information verification pipeline: check-worthiness, identifying previously fact-checked claims, supporting evidence retrieval, and claim verification. In this edition, we introduced some new challenges, offering six tasks in fifteen languages (Arabic, Bulgarian, English, Dutch, French, Georgian, German, Greek, Italian, Polish, Portuguese, Russian, Slovene, Spanish, and code-mixed Hindi-English): Task 1 on estimation of check-worthiness (the only task that has been present in all `CheckThat!` editions), Task 2 on identification of subjectivity (a follow up of the `CheckThat!` 2023 edition), Task 3 on identification of the use of persuasion techniques (a follow up of SemEval 2023), Task 4 on detection of hero, villain, and victim from memes (a follow up of CONSTRAINT 2022), Task 5 on rumor verification using evidence from authorities (new task), and Task 6 on robustness of credibility assessment with adversarial examples (new task). These are challenging classification and retrieval problems at the document and at the span level, including multilingual and multimodal settings. This year, `CheckThat!` was one of the most popular labs at CLEF-2024 in terms of team registrations: 130 teams. More than one-third of them (a total of 46) actually participated.

Keywords: Fact-Checking · Check-Worthiness · Subjectivity · Propaganda· Rumor Verification · Credibility Assessment · Authority Finding.

1 Introduction

The aim of `CheckThat!` is to foster the development of technology to assist different tasks along the fact-checking verification pipeline, as well as auxiliary tasks supporting the process. The focus in the first five lab iterations [56,21,9,57,55] was on the core tasks of the verification pipeline (see Figure 1). From the sixth edition [8], the lab has zoomed out of the core tasks of the pipeline and opened up for auxiliary tasks helping to address the different steps of the pipeline.

This year [7], we challenged the community with six tasks in multiple mono-, multi- and cross-lingual settings covering a total of fifteen languages: Arabic, Bulgarian, Dutch, English, French, Georgian, German, Greek, Italian, Polish, Portuguese, Slovenian, Spanish, Russian, and code-mixed Hindi. Task 1 [38] focused on check-worthiness estimation and asked to identify claims that could be important to verify in social and mainstream media. This task has been organized during all editions of the lab and is the only one that was part of the core pipeline. Task 2 [82] was a follow up of the CheckThat! 2023 edition and asked to determine whether a sentence from a news article is objective or conveys subjective opinions, helping to spot text that should be processed with specific strategies [71], potentially benefiting the fact-checking pipeline [44,45,90]. Task 3 [62] was a follow up of SemEval 2023, and it addressed persuasion techniques asking participants to identify text spans in which such techniques are being issued to possibly influence the reader. Task 4 was a follow up of CONSTRAINT 2022, and it asked participants to predict the role of each entity in a meme as a *hero*, a *villain*, a *victim*, or *other*. Task 5 [35] focused on rumor verification using evidence from authorities. The participants were asked to retrieve evidence from trusted sources (authorities that have *real knowledge* on the matter) and determine whether a rumor is supported, refuted, or unverifiable according to the evidence. The aim of Task 6 [70] was to assess the *robustness* of text classifiers in the misinformation detection domain and the participants aimed at discovering examples indicating low robustness of misinformation detection models.

As in previous editions, `CheckThat!` was one of the most popular tasks at CLEF, attracting a total of 46 participating teams, using a variety of approaches to the different tasks, mostly based on encoding and decoding large language models combined with different sources of information. The only exception was Task 4, which unfortunately did not attract participants. Nevertheless, as for the other tasks, we also release all the data for Task 4.

2 Previously on the `CheckThat!` Lab

In its previous six iterations, the `CheckThat!` lab has focused on various tasks from the claim verification pipeline, in a multitude of languages and in different domains (cf. Table 1).

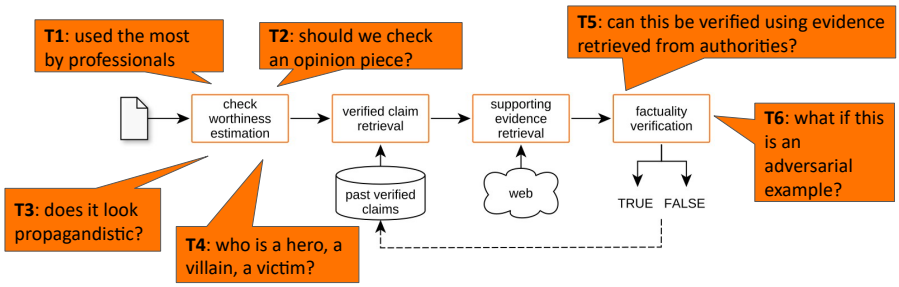

Fig. 1. The `CheckThat!` verification pipeline, featuring the four core tasks along with the `CheckThat!` 2024 tasks.

Table 1. Overview of the tasks offered in the previous editions of the lab.

tasks	years						domains				languages								
	2018	2019	2020	2021	2022	2023	debates	speeches	tweets	web pages	news articles	Arabic	Bulgarian	Dutch	English	German	Italian	Spanish	Turkish
check-worthiness estimation	■	■	■	■	■	■	■	■	■			■	■	■	■			■	■
verified claim retrieval			■	■	■		■	■	■			■			■				
supporting evidence retrieval		■	■							■		■							
claim verification	■	■	■			■						■			■				
fake news detection				■	■	■					■				■	■			■
bias						■					■								
subjectivity						■					■	■		■	■	■	■		■
topic identification				■											■				
authority finding						■			■			■							

`CheckThat!` 2018 [56] focused on check-worthiness and claim verification of political debates and speeches in Arabic and English. Both tasks continued in 2019 [21], with an additional focus on fact-checking by a task on classifying and ranking supporting evidence from the web. The 2020 edition [9] covered the full verification pipeline, with check-worthiness estimation, verified claim and supporting evidence retrieval, and claim verification. Social media data was first included in this iteration. The 2021 edition focused on multilinguality, offering tasks in five languages [57]. It also featured a fake news detection task, where the focus was on articles; this task was quite popular and it continued in 2022.

The 2023 year's edition of the `CheckThat!` lab [8] paid special attention to the various sub-aspects of check-worthiness estimation, subjectivity of news articles, factuality, bias, authority findings, again in a multitude of languages. Transformer-based models were extensively used. This edition has also introduced multimodality for check-worthiness estimation.

3 Description of the 2024 Tasks

The 2024 edition of `CheckThat!` featured a total of six tasks in a variety of languages and modalities, three of which were run for the first time (cf. Sections 3.3, 3.4 and 3.6). Moreover, two of the tasks had two subtasks each (cf. Sections 3.1 and 3.3).

3.1 Task 1: Check-Worthiness Estimation

Fact-checking is a complex process. Before assessing the truthfulness of a claim, determining whether it can be fact-checked at all is essential. Given the time-consuming nature of manual fact-checking, it is important to prioritize claims that are important to be fact-checked. Therefore, the aim of this task is to assess whether a statement sourced from a tweet, a transcript, or a political debate, requires fact-checking [8]. To make this decision, one must consider questions such as "Does it contain a verifiable factual claim?" and "Could it be harmful?" before assigning a final label for its check-worthiness. Further details about this task are discussed in [38].

3.2 Task 2: Subjectivity in News Articles

Verifiable claims are not only communicated in objective and neutral statements, but can also be found in subjectively colored ones. While objective sentences can be considered directly for verification, subjective sentences require additional processing steps, e.g., extracting an objective version of the statements or the claims they contain. Therefore, the objective of this task is to determine whether a given sentence is subjective or objective, which is set up as a binary classification task and is offered in Arabic, Bulgarian, English, German, Italian and in a multilingual setting. A more detailed description and discussion of the task can be found in [82].

3.3 Task 3: Persuasion Techniques

The goal of this task is to recognize and to classify the persuasion techniques in multilingual news at the text-span level. In particular, we used the two-tier persuasion techniques taxonomy introduced in *SemEval 2023 Shared Task 3: Detecting the Genre, the Framing, and the Persuasion Techniques in Online News in a Multi-lingual Setup* [64]. At the top level of the taxonomy, there are six coarse-grained techniques: *attack on reputation*, *justification*, *simplification*, *distraction*, *call*, and *manipulative wording*. These six types are further subdivided into 23 fine-grained techniques. The full definitions and examples are provided in [65] and [63].

3.4 Task 4: Detecting the Hero, the Villain, the Victim in Memes

Memes, characterized by their diverse multimodal nature, are frequently used to communicate intricate concepts on social media. However, this simplicity can sometimes oversimplify intricate concepts, leading to the potentially harmful content, often wrapped in humor. Identifying the narrative roles in memes is crucial for in-depth semantic analysis, especially when examining their potential connection to harmful content such as hate speech, offensive material, and cyberbullying [78]. The task aims to determine the roles of entities within memes, categorizing them as a hero, a villain, a victim, or other in a multi-class classification setting that considers systematic modeling of multimodal semiotics [79].

3.5 Task 5: Rumor Verification using Evidence from Authorities

Several studies addressed rumor verification in social media by exploiting evidence extracted from propagation networks or the Web [36,41,58]. However, finding and incorporating evidence from authorities for rumor verification in Twitter was proposed just recently [32]. In the previous edition of the lab, we offered the task of *Authority Finding in Twitter* [37]; this year, we offered a follow-up task with the objective of retrieving evidence from the timelines of authorities, and, accordingly, deciding whether the rumors are supported, refuted, or unverifiable. Task 5 is divided in two subtasks:

- **Evidence Retrieval**: Given a rumor expressed in a tweet and a set of authorities for that rumor, the system should retrieve *evidence tweets* posted by any of those authorities. An evidence tweet is a tweet that can be further used to detect the veracity of the rumor.
- **Rumor Verification**: Based solely on the evidence tweets retrieved by the above subtask, determine if the rumor is *supported* (true), *refuted* (false), or *unverifiable* (in case not enough evidence to verify it exists).

The task is offered in *Arabic* and *English*. Refer to [35] for a detailed overview.

3.6 Task 6: Robustness of Credibility Assessment with Adversarial Examples

Task 6 [70] asks to assess the robustness of text classification for misinformation detection. Automatic classifiers play an important role in many tasks in this domain, both within and outside the fact-checking pipeline explored in this lab. However, neural networks that often underpin such solutions have been shown vulnerable to *adversarial examples* (AEs) – initially for image classification [84], but later also for text classification [94] and, specifically, credibility assessment [69]. The participants were provided with a full classification setup for several domains (see 4.6), including training and attack data and three different victim models (BiLSMT, BERT and adversarially trained RoBERTa). Their goal was to find AEs by making small modifications to the text fragments in the attack set, so that the original meaning is preserved, but a victim classifier changes its decision. The quality of AEs was automatically assessed using the BODEGA framework [69] and manually through an annotation effort [70].

Table 2. Task 1: Check-worthiness in multigenre content. Statistics about the CT–CWT–24 corpus for all four languages.

	Arabic		**English**		**Spanish**		**Dutch**	
	Yes	**No**	**Yes**	**No**	**Yes**	**No**	**Yes**	**No**
Train	2,243	5,090	5,413	17,087	3,128	16,862	405	590
Dev	411	682	238	794	704	4,296	102	150
Dev-test	377	123	108	210	509	4,491	316	350
Test	218	392	88	253	-	-	397	603
Total	3,249	6,287	5,847	18,344	4,341	25,649	1,220	1,693

4 Datasets

4.1 Task 1: Check-Worthiness Estimation

The dataset contains multigenre content in Arabic, English, Dutch, and Spanish. For Arabic, it consists of tweets collected using keywords related to a variety of topics including COVID-19, following the annotation schema in [4], and political news from Arab countries. The test set includes tweets collected using keywords relevant to the war in Gaza. The dataset for English consists of transcribed sentences from candidates during the US Presidential election debates and annotated by human annotators [6]. The Dutch datasets consists of tweets collected at different moments in time and covering two topics. For training and development, we reused the datasets from the 2022 edition whose target topic was COVID-19 and vaccines, with messages spanning from January 2020 till March 2021. For testing, we collected $1k$ messages between January 2021 and December 2022 on climate change and its associated debate. The Spanish dataset consists 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 2 shows statistics for all languages and partitions.

4.2 Task 2: Subjectivity in News Articles

The dataset comprises sentences from news paper articles annotated with respect to their subjectivity. Information regarding the annotation guidelines can be found in [73]. The dataset included 2,675, 1,293, 1,776, 1,628 and 2,793 instances in Arabic (see [83] for more detail), Bulgarian, English, German, and Italian, respectively. Table 4.2 shows statistics. We provided two training sets for the multilingual scenario, one being a union of the training data for all languages offered this year and one incorporating the data for the languages offered in 2023 (Arabic, Dutch, English, Italian, German, and Turkish). The same holds for the dev and dev-test sets being compiled as balanced datasets of 50 instances per language. The test set included only data from the languages offered in 2024 consisting of 100 instances per language. The participants were free to choose from the multilingual datasets, opening room for cross-lingual approaches.

Table 3. Task 2: Subjectivity in News Articles. Dataset statistics for all five languages.

	Arabic		**Bulgarian**		**English**		**German**		**Italian**	
	obj	**subj**	**obj**	**subj**	**obj**	**subj**	**obj**	**subj**	**obj**	**subj**
Train	905	280	406	323	532	298	492	308	1,231	382
Dev	227	70	59	47	106	113	123	77	167	60
Dev-test	363	82	116	92	116	127	194	97	323	117
Test	425	323	143	107	362	122	226	111	377	136
Total	1,920	755	724	569	1,116	660	1,035	593	2,098	695

Table 4. Task 3: Persuasion Techniques. *Training*, *development* and *test* dataset statistics.

	Training		**Development**		**Test**	
language	#documents	#spans	#documents	#spans	#documents	#spans
English	536	9,002	54	1,775	98	2,599
French	211	6,831	50	1,681		
German	177	5,737	50	1,904		
Italian	303	7,961	61	2,351		
Polish	194	3,824	47	1,491		
Russian	191	4,138	72	944		
Georgian	-	-	29	218		
Greek	-	-	64	691		
Spanish	-	-	30	546		
Arabic	-	-			1,642	2,197
Bulgarian	-	-			100	1,732
Slovenian	-	-			100	4,591
Portuguese	-	-			104	1,727

4.3 Task 3: Persuasion Techniques

As training and development data, we used the corpus used in the SemEval 2023 task [64] which covers nine languages: English, German, Georgian, Greek, French, Italian, Polish, Russian, Spanish. As regards test data, we created a new dataset that covers five languages: Arabic, English, Bulgarian, Portuguese, and Slovene. English is the only language for which training, development and test data existed.

Detailed statistics about the training and development data are provided in Table 4. For more detailed characteristics of these datasets, refer to [64] and [65].

The data from the testing partition of English, Bulgarian, Portuguese and Slovene include articles about the Israeli-Palestine conflict and the Ukraine–Russia war, among others.

4.4 Task 4: Detecting the Hero, the Villain, and the Victim in Memes

We extended a previously existing dataset [80], which includes 6.9k labeled memes. Additionally, we introduced a new test dataset of 500 instances for Bulgarian, English, and code-mixed Hindi–English.

4.5 Task 5: Rumor Verification using Evidence from Authorities

The task dataset covers 160 rumors annotated with their corresponding 692 authority timelines, comprising about 34k annotated tweets in total. The rumors were randomly selected from two existing datasets namely AuFIN [33] and AuSTR [32], and the timelines were collected using the Academic Twitter search API which facilitates collecting historical user timelines.[1] Refer to [34] for more details about our data construction process.

The data was collected and annotated originally in *Arabic*, and automatically translated to *English* using GoogleTranslate.[2] A random sample of translated tweets (2,138 tweets comprising 6.3%), was manually validated to check the quality and reliability. In total, 514 (24%) tweets were edited to correct errors and inaccuracies, while 1,624 tweets (75.96%) remained unedited. More details about our data annotation process are discussed in the task overview [35]. For both *Arabic*, and *English*, we randomly split the data into 96 training, 32 development, and 32 test examples.

4.6 Task 6: Robustness of Credibility Assessment with Adversarial Examples

The task included data from five domains, each based on previously published corpora associating text with expert-assigned credibility: style-based news bias assessment (HN) [66], propaganda detection (PR) [17], fact checking (FC) [87], rumor detection (RD) [31] and COVID-19 misinformation detection (C19) [53]. These were all converted into binary classification tasks —credible vs. non-credible— and divided into training subset (for training victim classifiers) and attack subset (for preparing AEs). BiLSTM- and BERT-based classifiers were available throughout the task, while a surprise classifier (adversarially-trained RoBERTa) was only released in the testing phase. See [70] for detail.

5 Results and Overview of the Systems

5.1 Task 1: Check-Worthiness Estimation

This is a binary classification task, and we measure the performance based on the F_1-score for the check-worthiness class. The baseline is computed by randomly assigning a label from the label set to the test instance.

[1] https://developer.x.com/en/docs/twitter-api/tweets/search/api-reference/get-tweets-search-all

[2] https://py-googletrans.readthedocs.io/en/latest/

Table 5. Task 1 results on multigenre check-worthiness estimation. The F1 score is calculated with respect to the positive class. Shown are the top-10 submissions.

	Arabic			Dutch			English	
	Team	**F1**		**Team**	**F1**		**Team**	**F1**
1	IAI Group	0.569	1	TurQUaz	0.732	1	FactFinders	0.802
2	OpenFact	0.557	2	DSHacker	0.730	2	OpenFact	0.796
3	DSHacker	0.538	3	IAI Group	0.718	3	Fraunhofer SIT	0.780
4	TurQUaz	0.533	4	Mirela	0.650	4	Team_Artists	0.778
5	SemanticCUETSync	0.532	5	Zamoranesis	0.601	5	ZHAW_Students	0.771
6	Team_Artists	0.531	6	FC_RUG	0.594	6	SemanticCUETSync	0.763
7	Fired_from_NLP	0.530	7	OpenFact	0.590	7	SINAI	0.761
8	Madussree	0.530	8	HYBRINFOX	0.589	8	DSHacker	0.760
9	pandas	0.520	9	Team_Artists	0.577	9	IAI Group	0.753
10	HYBRINFOX	0.519	10	DataBees	0.563	10	Fired_from_NLP	0.745

In Table 5, we report results for the best 10 teams for each languages. A total of 13, 15 and 26 teams submitted systems for Arabic, Dutch, and English, respectively. For all languages, the participating systems outperformed the baselines, except for one team in Arabic and two teams in Dutch. Across languages, the performance was relatively higher for English, followed by Dutch.

Table 6 summarizes the approaches. Transformers were most popular. Some teams used language-specific transformers, while others opted for multilingual ones. Several teams also used large language models including variations of LLaMA, Mistral, Mixtral, and GPT. Standard preprocessing and data augmentation were also very common. Below, we discuss the top-3 systems across all languages. More details and descriptions of other systems can be found in [38].

Team **IAI Group** [1] trained several pre-trained language models (PLMs). For English, RoBERTa-Large was fine-tuned, and for Dutch and Arabic, XLM-RoBERTa and GPT-3.5-Turbo were fine-tuned.

Team **OpenFact** [77] fine-tuned DeBERTa and mDeBERTa models on multiple, curated versions of the dataset.

Team **FactFinders**[50] fine-tuned LLaMA2 7b on the training data using prompts generated by Chat-GPT. They applied a 2-step data pruning technique, including informativeness filtering and Condensed Nearest Neighbor undersampling, which did not affect performance. They further explored Mistral, Mixtral, Llama2 13b, Llama3 8b, and CommandR open-source LLMs. Mixtral achieved the highest F1-score in the dev-test phase, followed by LLaMA2 7b.

Team **Fraunhofer SIT** [91] used adapter fusion combining a task adapter with a Named Entity Recognition (NER) adapter, offering a resource-efficient alternative to fully fine-tuned PLMs. This yielded the third place in the task.

Team **DSHacker** [28] conducted monolingual and multilingual experiments. For the monolingual experiments, they fine-tuned BERT and optimized hyper-parameters per language. For the multilingual experiments, they fine-tuned XLM-RoBERTa-large and optimized hyper-parameters on the entire dataset or after

Table 6. Task 1: Overview of the approaches. The numbers in the language box refer to the position of the team in the official ranking. *Data aug:* Data augmentation.

Team	Language			Models																		Misc			
	Arabic	Dutch	English	LLama2	LLama 3	Mixtral	Mistral	GEITje	GPT-3.5	GPT-4	Gemini	BERT	RoBERTa	BERTweet	XLM-r	ALBERT	DistilBERT	DeBERTa	Electra	AraBERT	BERTje	GPT-3	Data aug	Preprocessing	Data Pruning	Info. Extraction
Aqua_Wave [10]			26									☑	☑										☑			
Checker Hacker [14]			14																				☑			
CLaC [29]			25								☑												☑			
DataBees [81]	12	10	18									☑	☑		☑	☑	☑			☑	☑			☑		
DSHacker [28]	3	2	8						☑	☑			☑		☑											
FactFinders [50]			1	☑	☑	☑																			☑	
FC_RUG [92]		6						☑																		
Fired_from_NLP [15]	7	12	10									☑	☑							☑						
Fraunhofer SIT [91]			3																							☑
HYBRINFOX [23]	10	8	12									☑	☑													☑
IAI Group [1]	1	3	9						☑	☑			☑		☑											
JUNLP [76]	14	11	22									☑														☑
Mirela [20]	11	4	16												☑		☑									
OpenFact [77]	2	7	2															☑								
pandas [85]	9	15	21									☑												☑		
SemanticCUETSync [60]	5	16	6									☑	☑				☑									
SINAI [89]			7						☑				☑									☑	☑			
SSN-NLP[27]			13									☑	☑		☑			☑						☑		☑
Team_Artists [43]	6	9	4									☑	☑	☑					☑	☑						
Trio_Titans[67]			19										☑			☑	☑							☑		
TurQUaz [12]	4	1	11		☑		☑		☑	☑	☑				☑											

excluding the Spanish data. Additionally, they leveraged GPT-3.5-turbo and GPT-4 for each language with few-shot prompting.

Team **TurQUaz** [12] developed different models for each language. For Arabic and English, they combined a fine-tuned RoBERTa model with in-context learning (ICL) using multiple different instruct-tuned models. The aggregation method varied between the Arabic and English datasets. For Dutch, they solely relied on in-context learning.

5.2 Task 2: Subjectivity in news Articles

A total of fifteen teams participated in this task, submitting 36 valid runs. Seven teams submitted valid runs for more than one language, with three teams participating in all six language settings, including the multilingual one. All teams participated in the English subtask. Table 7 shows the results achieved by the top-3 ranking teams for each language. We can see that, for most languages, at least one or two teams achieved rankings above the baseline, with the exception of Bulgarian. The best results were achieved for Italian and German, followed by English. For Arabic, none of the teams achieved a macro F1 score above 0.50. The team with the most stable results across languages was nullpointer [11]:

Table 7. Task 2: results on subjectivity classification in news articles in terms of macro F1. Shown are the top-3 submissions per language.

Rank	Team	F1	Rank	Team	F1	Rank	Team	F1
Arabic			Bulgarian			German		
1	IAI Group	0.495	1	(baseline)	0.753	1	nullpointer	0.791
2	nullpointer †	0.491	2	nullpointer	0.717	2	IAI Group	0.730
3	(baseline)	0.485	3	HYBRINFOX	0.715	3	(baseline)	0.699
English			Italian			Multilingual		
1	HYBRINFOX	0.744	1	JK_PCIC_UNAM	0.792		nullpointer*	0.712
2	ToniRodriguez	0.737	2	HYBRINFOX	0.784	1	HYBRINFOX	0.685
3	SSN-NLP	0.712	3	nullpointer	0.743	2	(baseline)	0.670
						3	IAI Group	0.629

† Team involved in the preparation of the data.
* Submitted after the official deadline.

with the exception of the English subtask, they always ranked among the top-3 teams.

All teams used neural networks, with transformer-based models being the most frequent choice. Some teams used language-specific monolingual transformer models, others chose multilingual models and some teams used English models in combination with automatic translation. An overview of the approaches is given in Table 8. More details can be found in [82].

Team **HYBRINFOX** [13] evaluated an ensemble combining a RoBERTa-based encoder, a SentenceBERT encoder, and lexical features. The RoBERTa and SentenceBERT embeddings were concatenated with subjectivity scores extracted from a rule-based expert system based on the VAGO [42] lexical database. These scores covered text aspects such as vagueness, subjectivity, detail, and objectivity. The enriched embeddings were then fed into the downstream classifier. Regarding training, only the RoBERTa model was fine-tuned, while the SentenceBERT model weights were frozen. The authors used machine translation via DeepL for all non-English sub-tasks.

Team **IAI Group** [1] experimented with the multilingual XLM-RoBERTa for all sub-tasks. They fine-tuned the model for each specific language.

Team **JK_PCIC_UNAM** [74] used a BERT-based classifier for English and Italian. They fine-tuned two distinct BERT classifiers, each tailored to a specific language. In each classification setting, they enriched BERT-based embeddings with linguistic features, including the number of nouns, adverbs, and feeling probabilities from input texts.

Team **nullpointer** [11] fine-tuned a BERT-based classifier for Arabic, Bulgarian, English, German, and Italian. They used a custom pre-processing pipeline where emojis, user mentions, and URLs were removed. The BERT model, initially pre-trained for sentiment analysis, was fine-tuned for each specific language, where the sentiment labels output by the model were mapped to subjec-

Table 8. Task 2: Overview of the approaches. The numbers in the language box refer to the position of the team in the official ranking.

Team	Language						Model												Misc			
	Multilingual	Arabic	Bulgarian	English	German	Italian	BERT	RoBERTa	DistilBERT	Gemini	mBERT	mDeBERTa	Sentence-BERT	SetFit	Mistral-7B-Instruct	XLM RoBERTa	DeBERTa	BART	Llama	Sentiment-Analysis-BERT	Data Augmentation	Translating data	Multi-lingual Training	Feature Selection
Checker Hacker [93]				4				☑													☑			
ClaC-2 [29]				14						☑											☑			
eevvgg [24]				8			☑																	☑
FactFinders				7											☑									
HYBRINFOX [13]	1	6	3	1	4	2		☑	☑		☑											☑		☑
IAI Group [1]	3	1	4	15	2	5		☑								☑								
Indigo [75]				10									☑	☑										
JK_PCIC_UNAM [74]				5		1	☑																	☑
JUNLP [76]		7	5	13			☑				☑													
nullpointer [11]	-	2	2	1	9	3														☑		☑		
SemanticCUETSync [60]		4		12								☑							☑					
SINAI				6				☑																
SSN-NLP [68]				3				☑																☑
ToniRodriguez [88]		5		2								☑				☑	☑	☑					☑	
Vigilantes				8			☑																	

- The run was submitted after the official deadline, therefore not part of the official ranking.

tivity labels. They handled class imbalance, and translated all non-English data to English.

Team **SSN-NLP** [68] compared traditional ML classifiers like K-NN and Random Forests to DL models like LSTMs, GRUs, and transformers for English. They used a custom pre-processing pipeline in which sentences are tokenized using the NLTK tool, and part-of-speech (POS) tags corresponding to retrieved tokens are added as additional features. Their best-performing model fine-tuned a RoBERTa-based classifier enriched with POS features concerning subjectivity and objectivity.

Team **ToniRodriguez** [88] fine-tuned two multilingual transformer-based classifiers, and XLM-RoBERTa, on English, German, and Italian datasets. Eventually, the mDeBERTa-v3 model was chosen as the best-performing one. Lastly, they applied zero-shot cross-lingual transfer to Arabic and Bulgarian.

5.3 Task 3: Persuasion Techniques

This was a multi-label multi-class sequence tagging task. To measure the performance of the systems, we modified the standard micro-averaged F1 to account

Table 9. Task 3: Overview of the approaches.

Team	Language					Models		Misc
	Ar	Bg	En	Pt	Sl	mBERT	DeBERTa	Data aug
Mela	1					☑		
UniBO	2	2	1	2	2		☑	☑

Table 10. Task 3: Results on persuasion techniques span identification. The team marked with * is a post competition experiment from the organizers.

Rank	Team	F1 micro	F1 macro	Rank	Team	F1 micro	F1 macro
English				**Portuguese**			
1	UniBO	0.092	0.061		PersuasionMultiSpan*	0.132	0.120
	PersuasionMultiSpan*	0.078	0.086	1	UniBO	0.107	0.073
2	Baseline	0.009	0.001	2	Baseline	0.002	
Bulgarian				**Slovenian**			
	PersuasionMultiSpan*	0.132	0.128		PersuasionMultiSpan*	0.153	0.127
1	UniBO	0.114	0.081	1	UniBO	0.123	0.075
2	Baseline	0.009	0.002	2	Baseline	0.003	0.002
Arabic							
1	Mela	0.301	0.080				
2	UniBO	0.108	0.068				
	PersuasionMultiSpan*	0.028	0.059				
3	Baseline	0.021	0.006				

for partial matching between the spans. In addition, an F1 value is computed for each persuasion technique.

Baseline. We opted for the most natural way to solve both a span identification task with a multi-label classification task: to treat it as a token classification problem, i.e., for each token, we predicted the classes with a given probability threshold, and then merged adjacent tokens with the same class in a single span.

Table 9 overviews the approaches, including the baseline. Only two teams submitted runs during the test phase (the organizers added a post competition submission), and two teams submitted system description papers. As shown in the table, the teams mostly fine-tuned transformer-based models, including data augmentation. In Table 10, we report participants results.

Team **UniBO** participated in all languages and ranked first in all but Arabic. Team **Mela** participated only in Arabic and was the top-ranked system, showing a significant improvement compared to other teams and the baseline.

In order to provide a meaningful comparison with state-of-the-art, we (the organizers) provided evaluation figures (after the competition) of a multi-lingual token-level multi-label classifier of persuasion techniques (referred to in the table with evaluation results with **PersuasionMultiSpan**) based on XML-RoBERTa [16], trained on the SemEval 2023 corpus [59,65], and whose performance on the SemEval 2023 competition [64] data oscillates around 1-3 rank across languages.

Team **UniBO** [25] proposed a system consisting of a two-part pipeline for text processing and classification. The first part was a data augmentation module using a BERT-based model fine-tuned for word alignment to project labels from source texts onto machine-translated target texts. The second part was a persuasion technique classification module, using two fine-tuned BERT-based models: a sequence classifier for detecting sentences with persuasion techniques and a set of 23 token-level classifiers for identifying specific techniques.

Team **Mela** [54] proposed a multilingual BERT-based system that incorporates both English and Arabic knowledge during its pre-training stage.

5.4 Task 4: Detecting the Hero, the Villain, the Victim in Memes

Baselines: We built a text-only system using DeBERTa (large) [40] as a baseline for this task. Due to the inherent complexity of the task, this system achieved an F1 score of 0.58, which is competitive to previous multimodal systems [80]. For evaluation, we used F_1-measure. Two role-label experts annotated each official test set, overseen by a consolidator following guidelines from previous work [80].

Unfortunately, there were no participants in this task. However, for test sets produced as part of the Lab can be obtained from the task website.

5.5 Task 5: Rumor Verification using Evidence from Authorities

In this section, we present our adopted baselines, and give an overview of the participating systems. Finally, we discuss the evaluation results.

Baselines: We adopted KGAT [51], a SOTA model for fact-checking. We fine-tuned both its evidence retrieval and rumor verification models on the FEVER English fact-checking dataset [86] following the authors setup but using multilingual BERT (mBERT) [19]. We then tested it on our *Arabic* and *English* test data as baselines for *Arabic* and *English*, respectively.

Evaluation Measures: To measure the ability of the system to retrieve evidence tweets higher in the list, we adopted the standard information retrieval rank-based measure Mean Average Precision (MAP) as the official evaluation measure, and we report Recall@5 (R@5). For rumor verification, we used the Macro-F1 to evaluate the classification of the rumors. Additionally, we considered a Strict Macro-F1 where the rumor label is considered correct only if at least one retrieved authority evidence was correct.

Systems Overview: A total of 3 and 5 teams submitted 5 and 11 runs[3] for Arabic and for English, respectively, out of which 2 teams made submissions for both languages. For *Arabic*, the participating teams either fine-tuned existing SOTA models for fact-checking on the task shared data (**bigIR**), or adopted a

[3] Each team was allowed to submit up to three runs per language.

Table 11. Task 5: Evidence retrieval (**Arabic**) official results in terms of MAP and Recall@5. The teams are ranked by the official evaluation measure MAP. Submissions with a + sign indicate submissions by task organizers.

Rank	Team (run ID)	MAP	Recall@5
1	bigIR$^+$ (bigIR-MLA-Ar)	0.618	0.673
2	IAI Group (IAI-Arabic-COLBERT)	0.564	0.581
	Baseline	0.345	0.423
3	SCUoL (SCUoL)	-	-

Table 12. Task 5: Evidence retrieval (**English**) official results in terms of MAP and Recall@5. The teams are ranked by the official evaluation measure MAP. Submissions with a + sign indicate submissions by task organizers.

Rank	Team (run ID)	MAP	Recall@5
1	bigIR$^+$ (bigIR-MLA-En)	0.604	0.677
2	Axolotl (run_rr=llama_sp=llama_rewrite=3_boundary=0)	0.566	0.617
3	DEFAULT (DEFAULT-Colbert1)	0.559	0.634
4	IAI Group (IAI-English-COLBERT)	0.557	0.590
5	AuthEv-LKolb (AuthEv-LKolb-oai)	0.549	0.587
	Baseline	0.335	0.445

zero-shot setup using existing models (**IAI Group** and **SCUoL**). **bigIR** fine-tuned KGAT [51] and MLA [47] but used MARBERTv2 [2] as the backbone model. **IAI Group** used ColBERT-XM [52] or cross-encoders for evidence retrieval, then leveraged the xlm-roberta-nli, a RoBERTa model pre-trained with a combination of Natural Language Inference (NLI) data in multiple languages [16] for rumor verification. Differently, **SCUoL** focused solely on the rumor verification subtask. They leveraged an Arabic content-based fact checking system [5], where they passed the rumor tweet to the system to get the veracity label.

For *English*, multiple approaches were adopted by the participating teams. **AuthEv-LKolb [46]** and **Axolotl [61]** used a lexical model for evidence retrieval, and used LLMs for rumor verification where they adopted OpenAI's GPT-4 assistant and Llama3 8B, respectively. **bigIR** fine-tuned two SOTA BERT-based models for fact-checking [47,51] for both subtasks. Differently, **DEFAULT [3]** formulated the task as retrieval-augmented classification and jointly trained the rumor verification classifier and the evidence retriever. A zero-shot setup was adopted by **IAI Group**, who used either ColBERT or cross-encoders for evidence retrieval and then exploited a RoBERTa pre-trained to NLI task data for rumor verification.

Evidence Retrieval Evaluation: For *Arabic*, as presented in Table 11, 2 teams outperformed the baseline by a margin. The bigIR team's primary model fine-tuned on the task data outperformed all models in terms of all evaluation mea-

Table 13. Task 5: Rumor verification (**Arabic**) official results in terms of Macro F1, and Strict Macro F1. The teams are ranked by the official evaluation measure Macro F1. Submissions with a + sign indicate submissions by task organizers.

Rank	Team (run ID)	m-F1	Strict m-F1
1	IAI Group (IAI-Arabic-COLBERT)	0.600	0.581
2	bigIR^{+} (bigIR-MLA-Ar)	0.368	0.300
3	SCUoL (SCUoL)	0.355	-
	Baseline	0.347	0.347

sures. We observe that although IAI Group adopted a zero-shot approach, it outperformed the baseline by a margin. As shown in Table 12, for *English* all the submitted runs outperformed our baseline. We observe that the models fine-tuned on our task data, bigIR-MLA-En and DEFAULT-Colbert1 runs, got the 1^{st} and 3^{rd} places respectively. The results also highlight that although Axolotl's run achieved a 2^{nd} position, bigIR outperforms it by a big margin.

Rumor Verification Evaluation: As presented in Table 13, for *Arabic* IAI Group's primary run outperformed all others significantly, although adopting a zero-shot approach. The results highlighted that even the bigIR model fine-tuned on the task data could not achieve comparable results to the best-performing model. Moreover, the bigIR model outperformed the baseline on Macro F1 only, but could not beat it in terms of Strict Macro F1. This could be attributed to the small number of training examples: 96 rumors only. Finally, the run submitted by the SCUol team performed better than the baseline, although not considering the authority evidence.

For *English*, as presented in Table 14, only 2 teams were able to outperform the baseline, AuthEv-LKolb and Axolotl, who adopted LLMs: GPT4 and Llama respectively. The results highlight that the models adopting a fine-tuning setup (bigIR and DEFAULT models), or zero-shot setup using pre-trained language models (IAI group model) could not outperform the baseline. We can conclude that, adopting LLMs can perform well on the verification task with Macro F1 of 0.895. However, further investigation is required to compare their performance against models fine-tuned on the task data but with a large number of rumors.

5.6 Task 6: Robustness of Credibility Assessment with Adversarial Examples

Task 6 received six submissions from the following teams: OpenFact [48], Text-Trojaners [30], TurQUaz [18], Palöri [39], MMU_NLP [72], and SINAI [89]. Table 15 shows the results of automatic evaluation: the teams are ranked according to BODEGA score [69], averaged over all victims and domains. It also includes two previous solutions, evaluated in the same scenario: DeepWordBug [26] and BERT-ATTACK [49], each delivering good AEs in some misinformation

Table 14. Task 5: Rumor verification (**English**) official results in terms of Macro F1, and Strict Macro F1. The teams are ranked by the official evaluation measure Macro F1. Submissions with a + sign indicate submissions by task organizers.

Rank	Team (run ID)	m-F1	Strict m-F1
1	AuthEv-LKolb (AuthEv-LKolb-oai)	0.879	0.861
2	Axolotl (run_rr=llama_sp=llama_rewrite=3_boundary=0)	0.687	0.687
	Baseline	0.495	0.495
3	DEFAULT (DEFAULT-Colbert1)	0.482	0.454
4	bigIR^{+} (bigIR-MLA-En)	0.458	0.428
5	IAI Group (IAI-English-COLBERT)	0.373	0.373

Table 15. Task 6: Results including the participating teams, BERT-ATTACK (B-A) and DeepWordBug (DWG), ranked according to average BODEGA score, as well as features of specific techniques.

			Change level				
# Team	Score	Char.	Word	Other	Word targeting	Tuning	Other
1. OpenFact	0.7458	✓	✓	✓	victim/features	✓	custom rules
2. TextTrojaners	0.7074		✓		victim/features	✓	beam search
3. TurQUaz	0.4859	✓		✓	genetic		
4. Palöri	0.4776		✓		victim		
5. MMU_NLP	0.3848	✓			none		homoghlyphs
6. SINAI	0.3507	✓	✓	✓	SHAP+KeyBERT		
- *B-A*	0.4261						
- *DWG*	0.2682						

scenarios [69]. However, here the former is easily outperformed by all submitted solutions, and the latter by most.

The table also includes information about the submitted solutions. Virtually all approaches target specific words that are likely to matter for the outcome, usually by probing the victim or relying on their features. The search methods used in this task include the BERT-ATTACK search method (MMU, Palöri, TextTrojaners, OpenFact), feature importance methods such as LIME (TextTrojaners), Genetic Algorithm (TurQUaz), brute force (SINAI), and using LLMs to suggest words to attack (TurQUaz).

Next, the candidate tokens are changed at the character- or word-level, but other modifications are also present. The best solutions are also tuned for the specific victim and/or domain.

The methods of replacement used include homoglyphs (MMU, TurQUaz, SINAI), generating words using a masked language model (TextTrojaners, OpenFact), the BERT-ATTACK replacement method (OpenFact, Palöri), word embedding similarity (OpenFact, Palöri), and LLM paraphrasing (TurQUaz).

An experimental manual evaluation was conducted to identify attack samples where the meaning was preserved from a human perspective. We selected 100 fact-checking task samples that successfully flipped the prediction of the victim classifier from each team. All samples were annotated anonymously.

During the process, two annotators evaluated each sample (the average pairwise annotator agreement was 0.59 in Cohen's Kappa), and a third annotator was introduced to resolve conflicts. The fully annotated dataset will be available soon after removing all personal identifiers. The results, showing the percentage of attack samples with preserved meaning, are as follows: SINAI: 99%, MMU_NLP: 96%, TurQUaz: 62%, Palöri: 14%, OpenFact: 11%, TextTrojaners: 7%. Based on manual evaluation results, the most successful method that preserves the meaning in this task is the homoglyphs method.

The manual evaluation of the FC results showed some discrepancies compared to the automatic evaluation of the whole task. This discrepancy might have been partly due to the manual evaluation not considering the attack's success rate. We plan to explore ways to combine both scores in the evaluation process.

6 Conclusion and Future Work

We presented the 2024 edition of the `CheckThat!` lab, which was once again one of the most popular CLEF labs, attracting a total of 46 active participating teams. This year, `CheckThat!` offered six tasks in fifteen languages (Arabic, Bulgarian, English, Dutch, French, Georgian, German, Greek, Italian, Polish, Portuguese, Russian, Slovene, Spanish, and code-mixed Hindi-English).

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 identification of the use of persuasion techniques. Task 4 detection of hero, villain, and victim from memes. Task 5 Rumor Verification using Evidence from Authorities (a first), and Task 6 robustness of credibility assessment with adversarial examples (a first).

For Task 1, most teams used pre-trained models (PLMs) and Large Language Models (LLMs). For Task 2, most teams relied on transformers, and some experimented with data augmentation or features like emojis and part-of-speech tags for classifying subjective sentences. For Task 3, the most successful team fine-tuned a multilingual transformer model. For Task 5, the results showed that the evidence retrieval models fine-tuned on the task data is the best performing models, while only the models adopting LLMs managed to outperform the rumor verification baseline. The results of Task 6 highlight the challenges of automatic evaluation, where established approaches obtain the highest quality score, but human annotators preferred homoglyph-based solutions.

7 Acknowledgments

The work of F. Haouari was supported by GSRA grant #GSRA6-1-0611-19074 from the Qatar National Research Fund (a member of Qatar Foundation). The

work of T. Elsayed was made possible by NPRP grant #NPRP-11S-1204-170060 from the Qatar National Research Fund. The statements made herein are solely the responsibility of the authors.

The work of M. Hasanain, R. Suwaileh, G. Da San Martino, and F. Alam is partially supported by NPRP 14C-0916-210015 from the Qatar National Research Fund, part of Qatar Research Development and Innovation Council (QRDI). The work of J. Struß is partially supported by the BMBF (German Federal Ministry of Education and Research) under the grant no. 01FP20031J.

The work of P. Przybyła is part of the ERINIA project, which has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101060930. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the funders. Neither the European Union nor the granting authority can be held responsible for them.

References

1. Aarnes, P.R., Setty, V., Galuščáková, P.: IAI group at CheckThat! 2024: Transformer models and data augmentation for checkworthy claim detection. In: Faggioli et al. [22]
2. 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. pp. 7088–7105 (2021)
3. Adhikari, S., Sharma, H., Kumari, R., Satapara, S., Desarkar, M.: DEFAULT at CheckThat! 2024: Retrieval Augmented Classification using Differentiable Top-k Operator for Rumor Verification based on Evidence from Authorities. In: Faggioli et al. [22]
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. pp. 611–649 (2021)
5. Althabiti, S., Alsalka, M.A., Atwell, E.: Ta'keed: The first generative fact-checking system for Arabic claims. arXiv preprint arXiv:2401.14067 (2024)
6. 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)
7. Barrón-Cedeño, A., Alam, F., Chakraborty, T., Elsayed, T., Nakov, P., Przybyła, P., Struß, J.M., Haouari, F., Hasanain, M., Ruggeri, F., Song, X., Suwaileh, R.: The CLEF-2024 CheckThat! Lab: Check-worthiness, subjectivity, persuasion, roles, authorities, and adversarial robustness. In: Goharian, N., Tonellotto, N., He, Y., Lipani, A., McDonald, G., Macdonald, C., Ounis, I. (eds.) Advances in Information Retrieval. pp. 449–458 (2024)
8. 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.: Overview of the CLEF–2023

CheckThat! Lab checkworthiness, subjectivity, political bias, factuality, and authority of news articles and their source. In: Arampatzis, A., Kanoulas, E., Tsikrika, T., Vrochidis, S., Giachanou, A., Li, D., Aliannejadi, M., Vlachos, M., Faggioli, G., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2023) (2023)

9. 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)
10. Bharathi, B., Dilsha Singh, D., Harinishree, K.: Aqua Wave at CheckThat! 2024: Check-worthiness estimation. In: Faggioli et al. [22]
11. Biswas, M.R., Tasneem Abir, A., Zaghouani, W.: Nullpointer at CheckThat! 2024: Identifying subjectivity from multilingual text sequence. In: Faggioli et al. [22]
12. Bulut, M.E., Keleş, K.E., Kutlu, M.: Turquaz at CheckThat! 2024: A hybrid approach of fine-tuning and in-context learning for check-worthiness estimation. In: Faggioli et al. [22]
13. Casanova, M., Chanson, J., Icard, B., Faye, G., Gadek, G., Gravier, G., Égré, P.: Hybrinfox at CheckThat! 2024 - task 2: Enriching BERT models with the expert system VAGO for subjectivity detection. In: Faggioli et al. [22]
14. Chandani, K., Syeda, D.E.Z.: Checker Hacker at CheckThat! 2024: Ensemble models for check-worthy tweet identification. In: Faggioli et al. [22]
15. Chowdhury, M.S.A., Shanto, A.M., Chowdhury, M.M., Murad, H., Das, U.: Fired_from_NLP at CheckThat! 2024: Estimating the check-worthiness of tweets using a fine-tuned transformer-based approach. In: Faggioli et al. [22]
16. 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. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 8440–8451 (2020)
17. 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. pp. 1377–1414. SemEval '20 (2020)
18. Demirok, B., Kutlu, M., Mergen, S., Oz, B.: Turquaz at CheckThat! 2024: Creating adversarial examples using genetic algorithm. In: Faggioli et al. [22]
19. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 4171–4186 (2019)
20. Dryankova1, M., Dimitrov, D., Koychev, I., Nakov, P.: Mirela at CheckThat! 2024: Check-worthiness of tweets with multilingual embeddings and adversarial training. In: Faggioli et al. [22]
21. 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)
22. Faggioli, G., Ferro, N., Galuščáková, P., García Seco de Herrera, A. (eds.): Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum. CLEF 2024 (2024)
23. Faye, G., Casanova, M., Icard, B., Chanson, J., Gadek, G., Gravier, G., Égré, P.: HYBRINFOX at CheckThat! 2024: Enhancing language models with structured information for checkworthiness estimation. In: Faggioli et al. [22]
24. Gajewska, E.: Eevvgg at CheckThat! 2024: Evaluative terms, pronouns and modal verbs as markers of subjectivity in text. In: Faggioli et al. [22]
25. Gajo, P., Giordano, L., Barron-Cedeño, A.: UniBO at CheckThat! 2024: Multilingual and multi-label persuasion technique detection in news with data augmentation and sequence-token classifiers
26. Gao, J., Lanchantin, J., Soffa, M.L., Qi, Y.: Black-box generation of adversarial text sequences to evade deep learning classifiers. In: Proceedings - 2018 IEEE Symposium on Security and Privacy Workshops, SPW 2018. pp. 50–56 (2018)
27. Giridharan, S.B.K., Sounderrajan, S., Bharathi, B., Salim, N.R.: SSN-NLP at CheckThat! 2024: Assessing the check-worthiness of tweets and debate excerpts using traditional machine learning and transformer models. In: Faggioli et al. [22]
28. Golik, P., Modzelewski, A., Jochym, A.: DSHacker at CheckThat! 2024: LLMs and BERT for check-worthy claims detection with propaganda co-occurrence analysis. In: Faggioli et al. [22]
29. Gruman, S., Kosseim, L.: CLaC at CheckThat! 2024: A zero-shot model for check-worthiness and subjectivity classification. In: Faggioli et al. [22]
30. Guzman Piedrahita, D., Fazla, A., Krauter, L.: TextTrojaners at CheckThat! 2024: Robustness of credibility assessment with adversarial examples through BeamAttack. In: Faggioli et al. [22]
31. Han, S., Gao, J., Ciravegna, F.: Neural language model based training data augmentation for weakly supervised early rumor detection. In: Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2019. pp. 105–112 (2019)
32. Haouari, F., Elsayed, T.: Are authorities denying or supporting? Detecting stance of authorities towards rumors in Twitter. Social Network Analysis and Mining **14**(1), 34 (2024)
33. Haouari, F., Elsayed, T., Mansour, W.: Who can verify this? Finding authorities for rumor verification in Twitter. Information Processing & Management **60**(4), 103366 (2023)
34. Haouari, F., Elsayed, T., Suwaileh, R.: AuRED: Enabling Arabic Rumor Verification using Evidence from Authorities over Twitter. In: Proceedings of ArabicNLP 2024 (2024)
35. Haouari, F., Elsayed, T., Suwaileh, R.: Overview of the CLEF-2024 CheckThat! Lab Task 5 on Rumor Verification using Evidence from Authorities. In: Faggioli et al. [22]
36. Haouari, F., Hasanain, M., Suwaileh, R., Elsayed, T.: ArCOV19-Rumors: Arabic COVID-19 Twitter dataset for misinformation detection. In: Proceedings of the Arabic Natural Language Processing Workshop. pp. 72–81. WANLP '21 (2021)
37. Haouari, F., Sheikh Ali, Z., Elsayed, T.: Overview of the CLEF-2023 CheckThat! Lab Task 5 on Authority Finding in Twitter. In: 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)

38. Hasanain, M., Suwaileh, R., Weering, S., Li, C., Caselli, T., Zaghouani, W., Barrón-Cedeño, A., Nakov, P., Alam, F.: Overview of the CLEF-2024 CheckThat! lab task 1 on check-worthiness estimation of multigenre content. In: Faggioli et al. [22]
39. He, H., Song, Y., Massey, D.: Palöri at CheckThat! 2024 shared task 6: Glota - combining GloVe embeddings with RoBERTa for adversarial attack. In: Faggioli et al. [22]
40. He, P., Liu, X., Gao, J., Chen, W.: DeBERTa: Decoding-enhanced BERT with disentangled attention. In: Proceedings of the International Conference on Learning Representations (2021)
41. Hu, X., Guo, Z., Chen, J., Wen, L., Yu, P.S.: MR2: A benchmark for multimodal retrieval-augmented rumor detection in social media. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 2901–2912. SIGIR '23 (2023)
42. Icard, B., Claveau, V., Atemezing, G., Égré, P.: Measuring vagueness and subjectivity in texts: From symbolic to neural VAGO. In: Proceedings of the IEEE International Conference on Web Intelligence and Intelligent Agent Technology. pp. 395–401. IEEE (2023)
43. Jayaswal, M., Rai, K.: Team_Artists at CheckThat! 2024: Text-based binary classification for check-worthiness detection. In: Faggioli et al. [22]
44. Jerônimo, C.L.M., Marinho, L.B., Campelo, C.E.C., Veloso, A., da Costa Melo, A.S.: Fake news classification based on subjective language. In: Proceedings of the 21st International Conference on Information Integration and Web-based Applications & Services. pp. 15–24 (2019)
45. Kasnesis, P., Toumanidis, L., Patrikakis, C.Z.: Combating fake news with transformers: A comparative analysis of stance detection and subjectivity analysis. Inf. **12**(10), 409 (2021)
46. Kolb, L., Hanbury, A.: AuthEv-LKolb at CheckThat! 2024: A Two-Stage Approach To Evidence-Based Social Media Claim Verification. In: Faggioli et al. [22]
47. Kruengkrai, C., Yamagishi, J., Wang, X.: A multi-level attention model for evidence-based fact checking. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP. pp. 2447–2460 (2021)
48. Lewoniewski, W., Stolarski, P., Stróżyna, M., Lewańska, E., Wojewoda, A., Księżniak, E., Sawiński, M.: OpenFact at CheckThat! 2024: Combining multiple attack methods for effective adversarial text generation. In: Faggioli et al. [22]
49. Li, L., Ma, R., Guo, Q., Xue, X., Qiu, X.: BERT-ATTACK: Adversarial Attack Against BERT Using BERT. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 6193–6202 (2020)
50. Li, Y., Panchendrarajan, R., Zubiaga, A.: FactFinders at CheckThat! 2024: Refining check-worthy statement detection with LLMs through data pruning. In: Faggioli et al. [22]
51. Liu, Z., Xiong, C., Sun, M., Liu, Z.: Fine-grained fact verification with kernel graph attention network. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 7342–7351 (2020)
52. Louis, A., Saxena, V., van Dijck, G., Spanakis, G.: ColBERT-XM: A modular multi-vector representation model for zero-shot multilingual information retrieval. arXiv preprint arXiv:2402.15059 (2024)
53. Mu, Y., Jiang, Y., Heppell, F., Singh, I., Scarton, C., Bontcheva, K., Song, X.: A large-scale comparative study of accurate COVID-19 information versus misinformation. In: TrueHealth 2023: Workshop on Combating Health Misinformation for Social Wellbeing (2023)

54. Nabhani, S., Riyadh, M.A.R.: Mela at CheckThat! 2024: Transferring persuasion detection from English to Arabic - a multilingual BERT approach. In: Faggioli et al. [22]
55. 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: Proceedings of the 13th International Conference of the CLEF Association: Information Access Evaluation meets Multilinguality, Multimodality, and Visualization. CLEF '2022 (2022)
56. 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)
57. 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) (2021)
58. Nielsen, D.S., McConville, R.: Mumin: A large-scale multilingual multimodal fact-checked misinformation social network dataset. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 3141–3153. SIGIR '22 (2022)
59. Nikolaidis, N., Piskorski, J., Stefanovitch, N.: Exploring the usability of persuasion techniques for downstream misinformation-related classification tasks. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). pp. 6992–7006 (2024)
60. Paran, A.I., Hossain, M.S., Shohan, S.H., Hossain, J., Ahsan, S., Hoque, M.M.: SemanticCuetSync at CheckThat! 2024: Finding subjectivity in news article using Llama. In: Faggioli et al. [22]
61. Pasin, A., Ferro, N.: SEUPD@CLEF: Team Axolotl on Rumor Verification using Evidence from Authorities. In: Faggioli et al. [22]
62. Piskorski, J., Stefanovitch, N., Alam, F., Campos, R., Dimitrov, D., Jorge, A., Pollak, S., Ribin, N., Fijavž, Z., Hasanain, M., Guimarães, N., Pacheco, A.F., Sartori, E., Silvano, P., Zwitter, A.V., Koychev, I., Yu, N., Nakov, P., Da San Martino, G.: Overview of the CLEF-2024 CheckThat! lab task 3 on persuasion techniques. In: Faggioli et al. [22]
63. Piskorski, J., Stefanovitch, N., Bausier, V.A., Faggiani, N., Linge, J., Kharazi, S., Nikolaidis, N., Teodori, G., De Longueville, B., Doherty, B., Gonin, J., Ignat, C., Kotseva, B., Mantica, E., Marcaletti, L., Rossi, E., Spadaro, A., Verile, M., Da San Martino, G., Alam, F., Nakov, P.: News categorization, framing and persuasion techniques: Annotation guidelines. Tech. rep., European Commission Joint Research Centre (2023)
64. 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 (2023)
65. Piskorski, J., Stefanovitch, N., Nikolaidis, N., Da San Martino, G., Nakov, P.: Multilingual multifaceted understanding of online news in terms of genre, framing, and persuasion techniques. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. pp. 3001–3022 (2023)
66. Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J., Stein, B.: A stylometric inquiry into hyperpartisan and fake news. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. pp. 231–240 (2018)
67. Prarthna, M., Chiranjeev Prasannaa, V.V., Sai Geetha, M.: Trio Titans at CheckThat! 2024: Check worthiness estimation. In: Faggioli et al. [22]
68. Premnath, P., Vaithiya Subramani, P., B, B., Salim, N.R.: SSN-NLP at CheckThat! 2024: From classic algorithms to transformers: A study on detecting subjectivity. In: Faggioli et al. [22]
69. Przybyła, P., Shvets, A., Saggion, H.: Verifying the robustness of automatic credibility assessment. arXiv:2303.08032 (2023)
70. Przybyła, P., Wu, B., Shvets, A., Mu, Y., Sheang, K.C., Song, X., Saggion, H.: Overview of the CLEF-2024 CheckThat! lab task 6 on robustness of credibility assessment with adversarial examples (InCrediblAE). In: Faggioli et al. [22]
71. Riloff, E., Wiebe, J.: Learning extraction patterns for subjective expressions. In: Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing. pp. 105–112. EMNLP '03 (2003)
72. Roadhouse, C., Shardlow, M., Williams, A.: MMU NLP at CheckThat! 2024: Homoglyphs are adversarial attacks. In: Faggioli et al. [22]
73. Ruggeri, F., Antici, F., Galassi, A., Korre, K., Muti, A., Barrón-Cedeño, A.: On the definition of prescriptive annotation guidelines for language-agnostic subjectivity detection. In: Proceedings the Sixth Workshop on Narrative Extraction From Texts (at ECIR). pp. 103–111 (2023)
74. Salas-Jimenez, K., Díaz, I., Gómez-Adorno, H.: JK_PCIC_UNAM at CheckThat! 2024: Analysis of subjectivity in news sentences using transformers based models. In: Faggioli et al. [22]
75. Sar, S., Roy, D.: Indigo at CheckThat! 2024: Using Setfit: A resource efficient technique for subjectivity detection in news article. In: Faggioli et al. [22]
76. Sardar, A.A.M., Fatema, K., Islam, M.A.: JUNLP at CheckThat! 2024: Enhancing check-worthiness and subjectivity detection through model optimization. In: Faggioli et al. [22]
77. Sawinski, M.: OpenFact at CheckThat! 2024: Optimizing training data selection through undersampling techniques. In: Faggioli et al. [22]
78. Sharma, S., Alam, F., Akhtar, M.S., Dimitrov, D., Da San Martino, G., Firooz, H., Halevy, A., Silvestri, F., Nakov, P., Chakraborty, T.: Detecting and understanding harmful memes: A survey. In: Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence. pp. 5597–5606 (7 2022)
79. Sharma, S., Kulkarni, A., Suresh, T., Mathur, H., Nakov, P., Akhtar, M.S., Chakraborty, T.: Characterizing the entities in harmful memes: Who is the hero, the villain, the victim? In: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. pp. 2149–2163 (2023)
80. 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 (2022)

81. Sriram, T., Anand, S., Venkatesh, Y.: Databees at CheckThat! 2024: Check worthiness estimation. In: Faggioli et al. [22]
82. Struß, J.M., Ruggeri, F., Barrón-Cedeño, A., Alam, F., Dimitrov, D., Galassi, A., Pachov, G., Koychev, I., Nakov, P., Siegel, M., Wiegand, M., Hasanain, M., Suwaileh, R., Zaghouani, W.: Overview of the CLEF-2024 CheckThat! lab task 2 on subjectivity in news articles. In: Faggioli et al. [22]
83. Suwaileh, R., Hasanain, M., Hubail, F., Zaghouani, W., Alam, F.: ThatiAR: Subjectivity detection in Arabic news sentences. arXiv: 2406.05559 (2024)
84. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. In: Proceedings of the International Conference on Learning Representations (2014)
85. Thirumurugan, R., Manimaran, M., Thota, S., Durairaj, T.: pandas at CheckThat! 2024: Ensemble models for checkworthy tweet identification. In: Faggioli et al. [22]
86. 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. pp. 809–819 (2018)
87. Thorne, J., Vlachos, A., Cocarascu, O., Christodoulopoulos, C., Mittal, A.: The Fact Extraction and VERification (FEVER) Shared Task. In: Proceedings of the First Workshop on Fact Extraction and VERification (FEVER) (2018)
88. Rodríguez de la Torre, A., Golobardes Ribé, E., Suau Martínez, J.: Tonirodriguez at CheckThat!2024: Is it possible to use zero-shot cross-lingual for subjectivity detection in low-resources languages? In: Faggioli et al. [22]
89. Valle Aguilera, J., Gutiérrez Megías, A.J., Jiménez Zafra, S.M., Ureña López, L.A., Martínez Cámara, E.: SINAI at CheckThat! 2024: Stealthy character-level adversarial attacks using homoglyphs and search, iterative. In: Faggioli et al. [22]
90. Vieira, L.L., Jerônimo, C.L.M., Campelo, C.E.C., Marinho, L.B.: Analysis of the subjectivity level in fake news fragments. In: Proceedings of the Brazillian Symposium on Multimedia and the Web. pp. 233–240. WebMedia '20 (2020)
91. Vogel, I., Möhle, P.: Fraunhofer SIT at CheckThat! 2024: Adapter fusion for checkworthiness detection. In: Faggioli et al. [22]
92. Weering, S., Caselli, T.: FC_RUG at CheckThat! 2024: Few-shot learning using GEITje for check-worthiness detection in Dutch. In: Faggioli et al. [22]
93. Zehra, S.D., Chandani, K., Khubaib, M., Aun Muhammed, A.A., Alvi, F., Samad, A.: Checker Hacker at CheckThat! 2024: Detecting check-worthy claims and analyzing subjectivity with transformers. In: Faggioli et al. [22]
94. Zhang, W.E., Sheng, Q.Z., Alhazmi, A., Li, C.: Adversarial attacks on deep-learning models in natural language processing. ACM Transactions on Intelligent Systems and Technology (TIST) **11**(3) (2020)