Overview of the CLEF-2024 CheckThat! Lab Task 1 on Check-Worthiness Estimation of Multigenre Content

Maram Hasanain[1,†], Reem Suwaileh[2], Sanne Weering[3], Chengkai Li[4], Tommaso Caselli[3], Wajdi Zaghouani[5], Alberto Barrón-Cedeño[6], Preslav Nakov[7] and Firoj Alam[1,*,†]

[1] *Qatar Computing Research Institute, HBKU, Qatar*
[2] *Hamad Bin Khalifa University, Qatar*
[3] *University of Groningen, Netherlands*
[4] *University of Texas at Arlington, USA*
[5] *Northwestern University in Qatar, Qatar*
[6] *DIT, Università di Bologna, Forlì, Italy*
[7] *Mohamed bin Zayed University of Artificial Intelligence, UAE*

Abstract

We present an overview of the CheckThat! Lab 2024 Task 1, part of CLEF 2024. Task 1 involves determining whether a text item is check-worthy, with a special emphasis on COVID-19, political news, and political debates and speeches. It is conducted in three languages: Arabic, Dutch, and English. Additionally, Spanish was offered for extra training data during the development phase. A total of 75 teams registered, with 37 teams submitting 236 runs and 17 teams submitting system description papers. Out of these, 13, 15 and 26 teams participated for Arabic, Dutch and English, respectively. Among these teams, the use of transformer pre-trained language models (PLMs) was the most frequent. A few teams also employed Large Language Models (LLMs). We provide a description of the dataset, the task setup, including evaluation settings, and a brief overview of the participating systems. As is customary in the CheckThat! Lab, we release all the datasets as well as the evaluation scripts to the research community. This will enable further research on identifying relevant check-worthy content that can assist various stakeholders, such as fact-checkers, journalists, and policymakers.

Keywords

Check-worthiness, fact-checking, multilinguality,

1. Introduction

Check-worthiness is a crucial component of the fact-checking pipeline. It helps to alleviate the burden on fact-checkers by reducing the need to verify every claim posted or shared across multiple online and social media platforms, which contain different types of content and modalities. This content can include news reports, citizen journalism, political debates, and posts from social media platforms. Identifying and debunking misleading claims is crucial to prevent the spread of misinformation, enabling individuals to make informed decisions where false information could lead to harmful consequences. For example, in critical areas such as health, finance, natural disasters and public policy, making well-informed decisions is especially important.

The CheckThat! 2024 lab was held in the framework of CLEF 2024 [1, 2].[1] Figure 1 shows the full CheckThat! identification and verification pipeline, highlighting the six tasks targeted in this seventh edition of the lab: Task 1 on check-worthiness estimation, Task 2 on subjectivity, Task 3 on persuasion technique detection (this paper), Task 4 on detecting hero, villain, and victim from memes, Task 5 on rumor verification using evidence from authorities, and Task 6 on robustness of credibility assessment with adversarial examples.

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

*Corresponding author.

†These authors contributed equally.

mhasanain@hbku.edu.qa (M. Hasanain); rsuwaileh@hbku.edu.qa (R. Suwaileh); s.weering@student.rug.nl (S. Weering); cli@uta.edu (C. Li); t.caselli@rug.nl (T. Caselli); wajdi.zaghouani@northwestern.edu (W. Zaghouani); a.barron@unibo.it (A. Barrón-Cedeño); preslav.nakov@mbzuai.ac.ae (P. Nakov); fialam@hbku.edu.qa (F. Alam)

[1] https://checkthat.gitlab.io

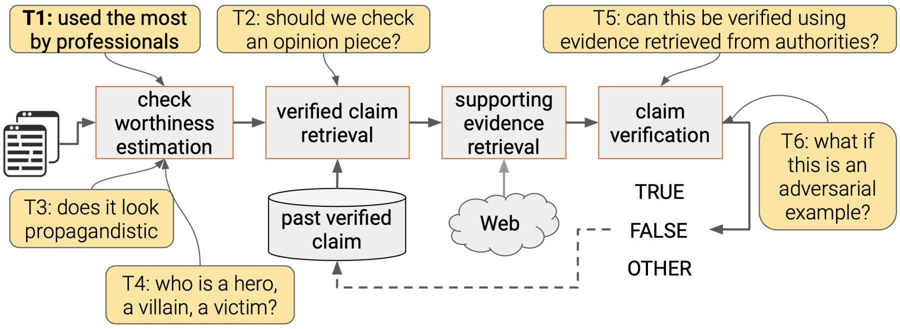
T1: used the most by professionals
T2: should we check an opinion piece?
T5: can this be verified using evidence retrieved from authorities?
check worthiness estimation
verified claim retrieval
supporting evidence retrieval
claim verification
T6: what if this is an adversarial example?
TRUE
FALSE
OTHER
T3: does it look propagandistic
past verified claim
Web
T4: who is a hero, a villain, a victim?

Figure 1: The CheckThat! lab verification pipeline. The 2024 edition of the lab covers six tasks: ***(T1)* check-worthiness estimation (this paper)**, (*T2*) subjectivity, (*T3*) persuasion technique detection, (*T4*) detecting hero, villain, and victim from memes, (*T5*) rumor verification using evidence from authorities, and (*T6*) robustness of credibility assessment with adversarial examples.

In this paper, we describe Task 1, which asks to detect whether a given text snippet from multigenre content, in a form of a tweet or a sentence from a political debate or speech, is worth fact-checking. Checkworthiness estimation simplifies and speeds up the process of fact-checking by prioritizing more important claims to be verified. In order to make that decision, one would need to consider questions such as "does it contain a verifiable factual claim?" or "is it harmful?", before deciding on the final check-worthiness label [3].

We provided manually annotated data in three languages: Arabic, Dutch, and English. Additionally, we included Spanish as an extra dataset. Among the various languages, English was the most popular target for participants. Across the submitted systems, pre-trained language models (PLMs) were widely used, with BERT, RoBERTa, and XLM-RoBERTa being the most popular models. Moreover, some teams used large language models (LLMs). The top-ranked systems also employed data augmentation and additional preprocessing steps.

The remainder of the paper is organized as follows: Section 2 describes the datasets released with the task. We present the evaluation setup in section 3. Section 4 discusses the system submissions and the official results. Section 5 presents some related work. Finally, we provide some concluding remarks in section 6.

2. Datasets

The dataset contains multigenre content in Arabic, English, Dutch, and Spanish. The Spanish subset was only offered for training purposes. The evaluation focuses on the other three languages. For all languages but English, the dataset consists of tweets collected using keywords related to a variety of topics, such as COVID-19 and vaccines, climate change, political news and the war on Gaza. The choice of topics was language-specific and was based on current events at different points of time when the dataset was being constructed. Additionally, the Spanish subset included transcriptions from Spanish politicians, and the subset was manually annotated by professional journalists who are experts in fact-checking. To annotate Arabic and Dutch data, we followed the scheme described by Alam et al. [3]. As for the English subset, it was sourced from the annotated dataset described by Arslan et al. [4], and consists of transcribed sentences from candidates during the US Presidential election debates.

We create the training, development and dev-test subsets for the 2024 edition by re-using all the data released in 2023 (or 2022 when the language was not run in the 2023 edition). Regarding the testing data, for Arabic we collected tweets using keywords relevant to the war on Gaza, that started in October 2023. For Dutch, we collected $1k$ messages between January 2021 and December 2022 on climate change and its associated debate. The English test set was constructed by manually annotating transcribed sentences that did not appear in Arslan et al. [4]. Table 1 shows statistics for all languages and partitions.

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

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

3. Evaluation Settings

We provided training, development, and dev-test subsets. The latter was intended to allow participants to validate their systems internally, while they could use the development set for hyper-parameter tuning and model selection. The test set was used for the final evaluation and ranking. The participants were allowed to submit multiple runs on the test set (without seeing the scores), and the last valid run was considered as official.

This is a binary classification task and we evaluate it on the basis of the F_1-measure on the check-worthiness class (yes) to account for class imbalance. The data and the evaluation scripts are available online.[2] The submission system was hosted on the CodaLab platform.[3]

4. Results and Overview of the Systems

A total of 13, 15 and 26 teams submitted systems for Arabic, Dutch, and English, respectively. Table 3 reports the performance results for all systems and languages. For all languages, the participating systems outperformed the baseline, except for one team in Arabic and two teams in Dutch.

Table 2 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 briefly describe the systems across all languages.
Team **Fired_from_NLP** [11] leveraged various model groups: Random Forest, SVM, and XGBoost; deep learning models such as LSTM and Bi-LSTM; and pre-trained language models (PLMs) including AraBERT for Arabic, RobBERT for Dutch, BERT-uncased for English, and Multilingual-BERT-uncased for all three languages. They trained and fine-tuned the models using the original datasets. Experiments showed that PLMs outperformed all other models.
Team **Fraunhofer SIT** [12] proposed an adapter fusion approach that combines a task adapter model with a Named Entity Recognition (NER) adapter, offering a resource-efficient alternative to fully fine-tuned PLMs. The task adapter was trained using the original training data without any preprocessing or cleaning. This method demonstrated superior performance and achieved the third place in the task.
Team **Mirela** [15] used DistilBERT-multilingual and XLM-RoBERTa-base PLMs. DistilBERT-multilingual was chosen for its lightweight and fast performance during inference, as well as its low computational training requirements. XLM-RoBERTa-base was selected due to its pre-training on 100 languages, achieving state-of-the-art performance in various NLP tasks in multilingual setups. Both models were finetuned on the original training data for English, Spanish, Arabic, and Dutch.
Team **SSN-NLP** [19] used a range of machine learning algorithms, including Support Vector Machine (SVM), Random Forest Classifier, Logistic Regression, XGBoost Classifier, CatBoost Classifier, K-Nearest Neighbors (KNN), and Passive Aggressive Classifier. Additionally, they fine-tuned several PLMs, including BERT-base-uncased, RoBERTa-base, XLM-RoBERTa-base, and DeBERTa-v3-base. Hyperparameters

[2] https://gitlab.com/checkthat_lab/clef2024-checkthat-lab/-/tree/main/task1
[3] https://codalab.lisn.upsaclay.fr/competitions/18893

Table 2
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
Checker Hacker [5]			14																				☑			
CLaC [6]			25								☑												☑			
DataBees [7]	12	10	18									☑	☑		☑	☑	☑			☑	☑			☑		
DSHacker [8]	3	2	8						☑	☑			☑		☑											
FactFinders [9]			1	☑	☑	☑																			☑	
FC_RUG [10]		6						☑																		
Fired_from_NLP [11]	7	12	10									☑	☑							☑						
Fraunhofer SIT [12]			3																							☑
HYBRINFOX [13]	10	8	12									☑	☑													☑
IAI Group [14]	1	3	9						☑	☑			☑		☑											
Mirela [15]	11	4	16												☑		☑									
OpenFact [16]	2	7	2															☑								
SemanticCuetSync [17]	5	16	6									☑	☑				☑									
SINAI [18]			7						☑				☑									☑	☑			
SSN-NLP[19]			13									☑	☑		☑			☑						☑		☑
Trio_Titans[20]			19										☑			☑	☑							☑		
TurQUaz [21]	4	1	11		☑		☑		☑	☑	☑				☑											

were optimized using GridSearchCV on the original data. Their preprocessing pipeline included text cleaning, tokenization, stopword removal, punctuation removal, URL removal, and spelling correction. For feature extraction, they used POS tagging and dependency parsing. These features were aggregated into vectors and combined with sentence embeddings generated using the Sentence-BERT PLM. The combined features were then normalized and reduced using Principal Component Analysis (PCA) to minimize computational requirements.

Team **FactFinders** [9] fine-tuned Llama2 7b on the original training data, using prompts generated by ChatGPT. A similar performance was achieved through a 2-step data pruning technique, which reduced the training data by 44% without compromising performance. The pruning involved filtering informative sentences and applying the Condensed Nearest Neighbor undersampling technique. Despite a slight performance drop (<0.5%) with the pruned dataset, results were submitted using the model fine-tuned on the original data. The models showed variability in results across different runs, so the final results were based on the majority of five iterations. Other open-source LLMs, such as Mistral, Mixtral, Llama2 13b, Llama3 8b, and CommandR, were also evaluated. Mixtral achieved the highest F1-score in the dev-test phase, followed by Llama2 7b. Due to training time considerations, Llama2 7b was used for the remainder of the study. Experiments with data expansion techniques yielded high precision but lower recall models.

Team **SemanticCuetSync** [22] fine-tuned language specific models such as RoBERTa, AraBERT, DistilBERT for English, Arabic and Dutch, respectively.

Team **Checker Hacker** [5] employed an ensemble approach integrating BERT-base-uncased and XLM-RoBERTa to improve the detection of check-worthy claims. Preprocessing steps, including tokenization and normalization, were implemented, along with data augmentation techniques to ensure the model was exposed to varied textual representations.

Team **IAI Group** [23] trained several PLMs. For English, RoBERTa-Large was fine-tuned, and for Dutch and Arabic, XLM-RoBERTa and GPT-3.5-Turbo were fine-tuned. The best models among them were selected based on their performance on the dev-test subsets. They reported that in some cases, GPT-4 in a zero-shot setting also performed well.

Table 3
Multigenre check-worthiness estimation. The F1 score is calculated with respect to the positive class.

	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	mjmanas54	0.778
5	SemanticCuetSync	0.532	5	Zamoranesis	0.601	5	ZHAW_Students	0.771
6	mjmanas54	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	mjmanas54	0.577	9	IAI Group	0.753
10	HYBRINFOX	0.519	10	DataBees	0.563	10	Fired_from_NLP	0.745
11	Mirela	0.478	11	JUNLP	0.550	11	TurQUaz	0.718
12	DataBees	0.460	12	Fired_from_NLP	0.543	12	HYBRINFOX	0.711
13	Baseline	0.418	13	Madussree	0.482	13	SSN-NLP	0.706
14	JUNLP	0.212	14	Baseline	0.438	14	Checker Hacker	0.696
			15	pandas	0.308	15	NapierNLP	0.675
			16	SemanticCuetSync	0.218	16	Mirela	0.658
						18	DataBees	0.619
						19	Trio_Titans	0.600
						20	Madussree	0.583
						21	pandas	0.579
						22	JUNLP	0.541
						23	Sinai and UG	0.517
						24	grig95	0.497
						25	CLaC	0.494
						26	Aqua_Wave	0.339
						27	Baseline	0.307

Team **OpenFact** [16] finetuned DeBERTa and mDeBERTa on multiple versions of the task dataset. This included training one model per language using the corresponding language train subset. The team also experimented with multilingual models by training over concatenated train subsets of all (or part) of the task four languages.

Team **HYBRINFOX** [13] developed a classification pipeline, consisting of three parts: a standard language model (RoBERTa for English and multilingual BERT for other languages), a component for extracting and encoding triples using OpenIE6 and Multi2OIE, and a merging neural network with a softmax layer for output. Early results indicated that including the triple encoding component improved performance over using the language model alone, especially for English. Challenges were noted in evaluating the approach for Dutch and Arabic due to limited proficiency in these languages.

Team **DSHacker** [8] conducted experiments with both monolingual and multilingual approaches. For the monolingual approach, BERT models were fine-tuned for specific languages. For the multilingual approach, XLM-RoBERTa-large was used, initially optimized and fine-tuned on the entire dataset. In a subsequent experiment, Spanish was excluded from the training data. Additionally, two LLMs, GPT-3.5-turbo and the recently released GPT-4o, were employed for each language using few-shot prompting to classify texts. A model was also fine-tuned on the DIPROMATS 2024 Task 1 dataset to predict whether the data from CheckThat! Lab 2024 Task 1 contained propaganda. This analysis aimed to indirectly determine whether check-worthy data also included propaganda. The XLM-RoBERTa-large model, fine-tuned for binary propaganda classification, was further fine-tuned for check-worthiness classification.

Team **FC_RUG** [10] tested GEITje, an LLM for Dutch based on Mistral-7B. They experimented with different prompts varying the learning settings (zero-shot vs few-shot) and the personas (helpful assistant vs fact-checker). The best model with few-shot in-context learning was selected based on the development data from the companion task of the CheckThat! 2022 Lab edition.

Team **CLaC** [6] approached the task as a binary classification task, leveraging a LLM (Google's Gemini[4])

[4] https://gemini.google.com

to classify whether a sentence is True or False, without specifying the task to classify for. The task was modeled as a multi-annotator scenario where Gemini was used to create two semantically-similar sentences to each test sentence. Then, Gemini was prompted to predict one of these labels: True, or False, for each sentence, using a single prompt. Finally, majority vote over the three annotations was used as the final label. Additionally, to improve performance, the prompt was contextualized by providing 600 randomly selected samples from the training subset.

Team **SINAI** [24] attempted two different approaches were attempted: *(i)* RoBERTa-base was fine-tuned using the original English data, and data augmentation was tried with Spanish transcription-sourced texts; *(ii)* A prompting approach with GPT-3.5-turbo was conducted, involving two experiments: one concatenating previous consecutive examples from the data (using the sentence_id) and the other using only the original text. Finally, after analyzing the results obtained from both approaches, the RoBERTa-base fine-tuning approach with the original English data was elected.

Team **Trio Titans** [20] fine-tuned different transformer models including DistilBERT, ALBERT, and RoBERTa, with the latter performing the best.

Team **DataBees** [7] fine-tuned various pre-trained models such as BERT, RoBERTa, and language-specific models like AraBERT for Arabic, along with traditional classifiers like MultinomialNB and Logistic Regression. The system was designed to work across the three languages. Their best F1 scores were achieved with DistilBERT for English, AraBERT for Arabic, and MultinomialNB for Dutch.

Team **TurQUaz** [21] developed differnet models for each language. For Arabic and English, a two-stage approach was proposed to determine check-worthy statements. This method combined a fine-tuned RoBERTa classifier with in-context learning (ICL) using multiple different instruct-tuned models. The aggregation method varied between the Arabic and English datasets. For the Dutch dataset, the fine-tuned classifier was excluded, and reliance was placed solely on in-context learning due to time constraints.

5. Related Work

5.1. Checkworiness in Fact-checking

Due to the significant surge of disinformative content online the importance of improving the capabilities of fact-checking pipeline is paramount. As depicted in Figure 1, the first part of the pipeline is finding claims that important to fact check [25]. The overall idea is to facilitate human fact-checkers to seamlessly streamline their daily fact-checking activities. To address and improve the capabilities of different components of fact-checking pipeline, there has been a considerable surge in research consisting of exploring fact-checking perspectives on fake news and associated issues [26], examining attitudes towards the detection of misinformation and disinformation [27], automating fact-checking to support human fact-checkers [28], predicting the factuality and the bias of entire news outlets [29], detecting disinformation across multiple modalities [30], and focusing on the use of abusive language on social media [31].

5.2. LLMs for Checkworthiness Task

Given that large language models (LLMs) have been demonstrating significant capabilities across various disciplines and many downstream NLP tasks, efforts have been made to utilize such models for detecting claims and their worthiness. Majer and Šnajder [32] evaluated gpt-4-turbo and demonstrated its potential for claim check-worthiness detection with minimal prompt engineering. Sawiński et al. [33] used GPT-3.5 and GPT-4 models in zero-shot and few-shot learning setups, comparing them with GPT-3, BERT, and RoBERTa-based fine-tuned models. Their findings demonstrate that the fine-tuned GPT-3 model performed the best across different models. Abdelali et al. [34] benchmarked various open and closed models for the Arabic checkworthiness task using the CT–CWT–22 dataset [35] and demonstrated that the performance of few-shot learning using GPT-4 is relatively higher; however, it is still far from state-of-the-art performance.

CT! Lab	Content Type	Modality	Language	Papers
CT-2018 [40]	Debate	Text	Ar, En	5
CT-2019 [41]	Debate, Web pages	Text	Ar, En	8
CT-2020 [42]	Tweet	Text	Ar, En	10
CT-2021 [43, 44]	Tweet, debate	Text	Ar, Bg, En, Es, Tr	10
CT-2022 [35, 45]	Tweet	Text	Ar, Bg, En, Nl, Es, Tr	13
CT-2023 [46, 47]	Tweet	Text, Image	Ar, En	12
CT-2024	Tweet, debate	Text	Ar, En, Nl	19

Table 4
Checkworthiness tasks from 2018 to 2014 offerend in different langauges and content types.

5.3. Previous Editions of Checkworthiness Shared Tasks

Since the seminal work by Hassan et al. [36], the task of check-worthiness estimation has gained broader interest. This task, proposed by Hassan et al. [36], involves assessing whether a sentence from a political debate is non-factual, trivially factual, or significantly factual enough to warrant verification. Since then, several notable studies have focused on political debates [37], tweets, and transcripts from political debates [38], as well as cross-lingual studies over tweets [39].

A major research interest has been sparked since the inception of the CLEF CheckThat! lab initiatives. The initial focus was primarily on political debates and speeches. This focus has since expanded to include social media, transcriptions, and various languages and modalities.

Significant research interest has been sparked since the inception of the CLEF CheckThat! lab initiatives. The initial focus was primarily on political debates and speeches. This focus has since expanded to include social media, transcriptions, and various languages and modalities. In Table 4, we report a summary of check-worthiness tasks over the years from 2018 to 2024. The focus has mainly been on debates and tweets, mostly in the text modality. As for languages, Arabic and English have been offered in all editions. The number of participants and system description paper submissions has increased over the years.

6. Conclusion and Future Work

We presented an overview of Task 1 of the CLEF-2024 CheckThat! lab, which focused on check-worthiness estimation of multigenre content and covering three languages: Arabic, Dutch, and English. The task attracted significant participation, with 75 registered teams and 28 teams submitting system description papers. The majority of the participating systems leveraged transformer-based models, showcasing their effectiveness in this domain. Notable approaches included the fine-tuning of language-specific models such as AraBERT for Arabic and RobBERT for Dutch, as well as the use of multilingual models like XLM-RoBERTa. Several teams experimented with large language models including GPT-3.5 and Llama2, while others implemented ensemble approaches combining multiple models. Data augmentation and preprocessing techniques were widely employed to enhance performance, and some teams incorporated named entity recognition and other linguistic features into their systems. The results show significant improvements over the baselines across all languages, highlighting the progress made in check-worthiness estimation. Future work may include covering other modalities and domains.

Acknowledgments

The work of F. Alam, M. Hasanain, R. Suwaileh and W. Zaghouani is partially supported by NPRP 14C-0916-210015 from the Qatar National Research Fund, which is a part of Qatar Research Development and Innovation Council (QRDI). The findings achieved herein are solely the responsibility of the authors.

References

[1] A. Barrón-Cedeño, F. Alam, T. Chakraborty, T. Elsayed, P. Nakov, P. Przybyła, J. M. Struß, F. Haouari, M. Hasanain, F. Ruggeri, X. Song, R. Suwaileh, The CLEF-2024 CheckThat! Lab: Check-worthiness, subjectivity, persuasion, roles, authorities, and adversarial robustness, in: N. Goharian, N. Tonellotto, Y. He, A. Lipani, G. McDonald, C. Macdonald, I. Ounis (Eds.), Advances in Information Retrieval, 2024, pp. 449–458.

[2] A. Barrón-Cedeño, F. Alam, J. M. Struß, P. Nakov, T. Chakraborty, T. Elsayed, P. Przybyła, T. Caselli, G. Da San Martino, F. Haouari, C. Li, J. Piskorski, F. Ruggeri, X. Song, R. Suwaileh, Overview of the CLEF-2024 CheckThat! Lab: Check-worthiness, subjectivity, persuasion, roles, authorities and adversarial robustness, in: L. Goeuriot, P. Mulhem, G. Quénot, D. Schwab, L. Soulier, G. M. Di Nunzio, P. Galuščáková, A. García Seco de Herrera, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024), 2024.

[3] F. Alam, S. Shaar, F. Dalvi, H. Sajjad, A. Nikolov, H. Mubarak, G. D. S. Martino, A. Abdelali, N. Durrani, K. Darwish, A. Al-Homaid, W. Zaghouani, T. Caselli, G. Danoe, F. Stolk, B. Bruntink, P. Nakov, 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, 2021, pp. 611–649.

[4] F. Arslan, N. Hassan, C. Li, M. Tremayne, A benchmark dataset of check-worthy factual claims, in: Proceedings of the International AAAI Conference on Web and Social Media, volume 14, 2020, pp. 821–829.

[5] K. Chandani, D. E. Z. Syeda, Checker Hacker at CheckThat! 2024: Ensemble models for check-worthy tweet identification, in: [48], 2024.

[6] S. Gruman, L. Kosseim, CLaC at CheckThat! 2024: A zero-shot model for check-worthiness and subjectivity classification, in: [48], 2024.

[7] T. Sriram, S. Anand, Y. Venkatesh, Databees at CheckThat! 2024: Check worthiness estimation, in: [48], 2024.

[8] P. Golik, A. Modzelewski, A. Jochym, DSHacker at CheckThat! 2024: LLMs and BERT for check-worthy claims detection with propaganda co-occurrence analysis, in: [48], 2024.

[9] Y. Li, R. Panchendrarajan, A. Zubiaga, FactFinders at CheckThat! 2024: Refining check-worthy statement detection with LLMs through data pruning, in: [48], 2024.

[10] S. Weering, T. Caselli, FC_RUG at CheckThat! 2024: Few-shot learning using GEITje for check-worthiness detection in Dutch, in: [48], 2024.

[11] M. S. A. Chowdhury, A. M. Shanto, M. M. Chowdhury, H. Murad, U. Das, Fired_from_NLP at CheckThat! 2024: Estimating the check-worthiness of tweets using a fine-tuned transformer-based approach, in: [48], 2024.

[12] I. Vogel, P. Möhle, Fraunhofer SIT at CheckThat! 2024: Adapter fusion for check-worthiness detection, in: [48], 2024.

[13] G. Faye, M. Casanova, B. Icard, J. Chanson, G. Gadek, G. Gravier, P. Égré, HYBRINFOX at CheckThat! 2024: Enhancing language models with structured information for checkworthiness estimation, in: [48], 2024.

[14] P. R. Aarnes, V. Setty, P. Galuščáková, IAI group at CheckThat! 2024: Transformer models and data augmentation for checkworthy claim detection, in: [48], 2024.

[15] M. Dryankova1, D. Dimitrov, I. Koychev, P. Nakov, Mirela at CheckThat! 2024: Check-worthiness of tweets with multilingual embeddings and adversarial training, in: [48], 2024.

[16] M. Sawinski, OpenFact at CheckThat! 2024: Optimizing training data selection through undersampling techniques, in: [48], 2024.

[17] A. I. Paran, M. S. Hossain, S. H. Shohan, J. Hossain, S. Ahsan, M. M. Hoque, SemanticCuetSync at CheckThat! 2024: Finding subjectivity in news article using Llama, in: [48], 2024.

[18] J. Valle Aguilera, A. J. Gutiérrez Megías, S. M. Jiménez Zafra, L. A. Ureña López, E. Martínez Cámara, SINAI at CheckThat! 2024: Stealthy character-level adversarial attacks using homoglyphs and

search, iterative, in: [48], 2024.

[19] S. B. K. Giridharan, S. Sounderrajan, B. Bharathi, N. R. Salim, SSN-NLP at CheckThat! 2024: Assessing the check-worthiness of tweets and debate excerpts using traditional machine learning and transformer models, in: [48], 2024.

[20] M. Prarthna, V. V. Chiranjeev Prasannaa, M. Sai Geetha, Trio Titans at CheckThat! 2024: Check worthiness estimation, in: [48], 2024.

[21] M. E. Bulut, K. E. Keleş, M. Kutlu, TurQUaz at CheckThat! 2024: A hybrid approach of fine-tuning and in-context learning for check-worthiness estimation, in: [48], 2024.

[22] S. H. Shohan, A. I. Paran, M. S. Hossain, J. Hossain, M. M. Hoque, SemanticCuetSync at CheckThat! 2024: Finetuning transformer models for checkworthy tweet identification, in: [48], 2024.

[23] M. Hasanain, R. Suwaileh, S. Weering, C. Li, T. Caselli, W. Zaghouani, A. Barrón-Cedeño, P. Nakov, F. Alam, Overview of the CLEF-2024 CheckThat! lab task 1 on check-worthiness estimation of multigenre content, in: [48], 2024.

[24] S. Stoia, J. Montañez-Collado, C. Ibáñez-Bautista, A. Montejo-Ráez, M. T. Martín-Valdivia, M. C. Díaz-Galiano, SINAI at CheckThat! 2024: Transformer-based approaches for check-worthiness classification, in: [48], 2024.

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

[26] J. Thorne, A. Vlachos, Automated fact checking: Task formulations, methods and future directions, in: Proceedings of the 27th International Conference on Computational Linguistics, COLING '18, Association for Computational Linguistics, Santa Fe, NM, USA, 2018, pp. 3346–3359.

[27] M. Hardalov, A. Arora, P. Nakov, I. Augenstein, A survey on stance detection for mis-and disinformation identification, in: Findings of the Association for Computational Linguistics: NAACL 2022, 2022, pp. 1259–1277.

[28] G. K. Shahi, Fakekg: A knowledge graph of fake claims for improving automated fact-checking (student abstract), Proceedings of the AAAI Conference on Artificial Intelligence 37 (2023) 16320–16321. doi:10.1609/aaai.v37i13.27020.

[29] P. Nakov, H. T. Sencar, J. An, H. Kwak, A survey on predicting the factuality and the bias of news media, arXiv/2103.12506 (2021).

[30] F. Alam, S. Cresci, T. Chakraborty, F. Silvestri, D. Dimitrov, G. Da San Martino, S. Shaar, H. Firooz, P. Nakov, A survey on multimodal disinformation detection, in: Proceedings of the 29th International Conference on Computational Linguistics, 2022, pp. 6625–6643.

[31] A. Arora, P. Nakov, M. Hardalov, S. M. Sarwar, V. Nayak, Y. Dinkov, D. Zlatkova, K. Dent, A. Bhatawdekar, G. Bouchard, I. Augenstein, Detecting harmful content on online platforms: What platforms need vs. where research efforts go, ACM Computing Surveys 5 (2023).

[32] L. Majer, J. Šnajder, Claim check-worthiness detection: How well do llms grasp annotation guidelines?, arXiv:2404.12174 (2024).

[33] M. Sawiński, K. Węcel, E. P. Księżniak, M. Stróżyna, W. Lewoniewski, P. Stolarski, W. Abramowicz, Openfact at checkthat! 2023: head-to-head gpt vs. bert-a comparative study of transformers language models for the detection of check-worthy claims, in: CEUR Workshop Proceedings, volume 3497, 2023.

[34] A. Abdelali, H. Mubarak, S. Chowdhury, M. Hasanain, B. Mousi, S. Boughorbel, S. Abdaljalil, Y. El Kheir, D. Izham, F. Dalvi, M. Hawasly, N. Nazar, Y. Elshahawy, A. Ali, N. Durrani, N. Milic-Frayling, F. Alam, LAraBench: Benchmarking Arabic AI with large language models, in: Y. Graham, M. Purver (Eds.), Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, St. Julian's, Malta, 2024, pp. 487–520.

[35] P. Nakov, A. Barrón-Cedeño, G. Da San Martino, F. Alam, R. Míguez, T. Caselli, M. Kutlu, W. Zaghouani, C. Li, S. Shaar, H. Mubarak, A. Nikolov, Y. S. Kartal, J. Beltrán, Overview of the CLEF-2022 CheckThat! lab task 1 on identifying relevant claims in tweets, in: N. Faggioli, Guglielmo andd Ferro, A. Hanbury, M. Potthast (Eds.), Working Notes of CLEF 2022—Conference and Labs of

the Evaluation Forum, CLEF '2022, Bologna, Italy, 2022.

[36] N. Hassan, C. Li, M. Tremayne, Detecting check-worthy factual claims in presidential debates, in: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, CIKM '15, 2015, pp. 1835–1838.

[37] S. Vasileva, P. Atanasova, L. Màrquez, A. Barrón-Cedeño, P. Nakov, It takes nine to smell a rat: Neural multi-task learning for check-worthiness prediction, in: Proceedings of the International Conference on Recent Advances in Natural Language Processing, RANLP '19, 2019, pp. 1229–1239.

[38] Y. S. Kartal, M. Kutlu, Re-think before you share: A comprehensive study on prioritizing check-worthy claims, IEEE Transactions on Computational Social Systems (2022).

[39] M. Hasanain, T. Elsayed, Cross-lingual transfer learning for check-worthy claim identification over twitter, arXiv: 2211.05087 (2022).

[40] P. Atanasova, L. Marquez, A. Barrón-Cedeño, T. Elsayed, R. Suwaileh, W. Zaghouani, S. Kyuchukov, G. Da San Martino, P. Nakov, Overview of the CLEF-2018 CheckThat! lab on automatic identification and verification of political claims. Task 1: Check-worthiness, CEUR Workshop Proceedings, 2018.

[41] P. Atanasova, P. Nakov, G. Karadzhov, M. Mohtarami, G. Da San Martino, Overview of the CLEF-2019 CheckThat! lab on automatic identification and verification of claims. Task 1: Check-worthiness, CEUR Workshop Proceedings, 2019.

[42] A. Barrón-Cedeño, T. Elsayed, P. Nakov, G. Da San Martino, M. Hasanain, R. Suwaileh, F. Haouari, N. Babulkov, B. Hamdan, A. Nikolov, S. Shaar, Z. Sheikh Ali, Overview of CheckThat! 2020: Automatic identification and verification of claims in social media, in: A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, H. Joho, C. Lioma, C. Eickhoff, A. Névéol, L. Cappellato, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Eleventh International Conference of the CLEF Association (CLEF 2020), LNCS (12260), Springer, 2020, pp. 215–236.

[43] S. Shaar, M. Hasanain, B. Hamdan, Z. S. Ali, F. Haouari, A. Nikolov, M. Kutlu, Y. S. Kartal, F. Alam, G. Da San Martino, A. Barrón-Cedeño, R. Míguez, J. Beltrán, T. Elsayed, P. Nakov, Overview of the CLEF-2021 CheckThat! lab task 1 on check-worthiness estimation in tweets and political debates, 2021.

[44] P. Nakov, G. Da San Martino, T. Elsayed, A. Barrón-Cedeño, R. Míguez, S. Shaar, F. Alam, F. Haouari, M. Hasanain, W. Mansour, B. Hamdan, Z. S. Ali, N. Babulkov, A. Nikolov, G. K. Shahi, J. M. Struß, T. Mandl, M. Kutlu, Y. S. Kartal, Overview of the CLEF-2021 CheckThat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news, in: K. Candan, B. Ionescu, L. Goeuriot, B. Larsen, H. Müller, A. Joly, M. Maistro, F. Piroi, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Twelfth International Conference of the CLEF Association, LNCS (12880), 2021.

[45] P. Nakov, A. Barrón-Cedeño, G. Da San Martino, F. Alam, J. M. Struß, T. Mandl, R. Míguez, T. Caselli, M. Kutlu, W. Zaghouani, C. Li, S. Shaar, G. K. Shahi, H. Mubarak, A. Nikolov, N. Babulkov, Y. S. Kartal, J. Beltrán, M. Wiegand, M. Siegel, J. Köhler, Overview of the CLEF-2022 CheckThat! lab on fighting the COVID-19 infodemic and fake news detection, in: A. Barrón-Cedeño, G. Da San Martino, M. Degli Esposti, F. Sebastiani, C. Macdonald, G. Pasi, A. Hanbury, M. Potthast, G. Faggioli, F. Nicola (Eds.), Proceedings of the 13th International Conference of the CLEF Association: Information Access Evaluation meets Multilinguality, Multimodality, and Visualization, CLEF '2022, Bologna, Italy, 2022.

[46] F. Alam, A. Barrón-Cedeño, G. S. Cheema, S. Hakimov, M. Hasanain, C. Li, R. Míguez, H. Mubarak, G. K. Shahi, W. Zaghouani, P. Nakov, Overview of the CLEF-2023 CheckThat! lab task 1 on check-worthiness in multimodal and multigenre content, in: M. Aliannejadi, G. Faggioli, N. Ferro, Vlachos, Michalis (Eds.), Working Notes of CLEF 2023–Conference and Labs of the Evaluation Forum, CLEF 2023, Thessaloniki, Greece, 2023.

[47] A. Barrón-Cedeño, F. Alam, T. Caselli, G. Da San Martino, T. Elsayed, A. Galassi, F. Haouari, F. Ruggeri, J. M. Struß, R. N. Nandi, G. S. Cheema, D. Azizov, P. Nakov, The CLEF-2023 CheckThat! Lab: Checkworthiness, subjectivity, political bias, factuality, and authority, in: J. Kamps, L. Goeuriot,

F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2023, pp. 506–517.

[48] G. Faggioli, N. Ferro, P. Galuščáková, A. García Seco de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CLEF 2024, 2024.