Overview of the CLEF-2022 CheckThat! Lab Task 1 on Identifying Relevant Claims in Tweets

Preslav **Nakov**[1], Alberto **Barrón-Cedeño**[2], Giovanni **Da San Martino**[3], Firoj **Alam**[4], Rubén **Míguez**[5], Tommaso **Caselli**[6], Mucahid **Kutlu**[7], Wajdi **Zaghouani**[8], Chengkai **Li**[9], Shaden **Shaar**[10], Hamdy **Mubarak**[4], Alex **Nikolov**[11] and Yavuz Selim **Kartal**[12]

[1]*Mohamed bin Zayed University of Artificial Intelligence, UAE*

[2]*DIT, Università di Bologna, Italy*

[3]*University of Padova, Italy*

[4]*Qatar Computing Research Institute, HBKU, Qatar*

[5]*Newtral Media Audiovisual, Spain*

[6]*University of Groningen, Netherland*

[7]*TOBB University of Economics and Technology, Turkey*

[8]*Hamad Bin Khalifa University, Qatar*

[9]*University of Texas at Arlington, USA*

[10]*Cornell University, USA*

[11]*Sofia University, Bulgaria*

[12]*GESIS – Leibniz Institute for the Social Sciences, Germany*

Abstract

We present an overview of `CheckThat!` lab 2022 Task 1, part of the 2022 Conference and Labs of the Evaluation Forum (CLEF). Task 1 asked to predict which posts in a Twitter stream are worth fact-checking, focusing on COVID-19 and politics in six languages: Arabic, Bulgarian, Dutch, English, Spanish, and Turkish. A total of 19 teams participated and most submissions managed to achieve sizable improvements over the baselines using Transformer-based models such as BERT and GPT-3. Across the four subtasks, approaches that targetted multiple languages (be it individually or in conjunction, in general obtained the best performance. We describe the dataset and the task setup, including the evaluation settings, and we give a brief overview of the participating systems. As usual in the `CheckThat!` lab, we release to the research community all datasets from the lab as well as the evaluation scripts, which should enable further research on finding relevant tweets that can help different stakeholders such as fact-checkers, journalists, and policymakers.

Keywords
Check-Worthiness Estimation, Fact-Checking, Veracity, Social Media Verification, Computational Journalism, COVID-19.

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

1. Introduction

The CheckThat! 2022 lab was held in the framework of CLEF 2022 [1][1]. Figure 1 shows the full CheckThat! identification and verification pipeline, highlighting the three tasks targeted in this fifth edition of the lab: Task 1 on detecting relevant claims in tweets (this paper), Task 2 on retrieving relevant previously fact-checked tweets [2], and Task 3 on predicting the veracity of news [3]. Task 1 asks to detect relevant tweets on the basis of different complementary criteria: check-worthiness, verifiability, harmfulness, and attention-worthiness. We provided manually annotated data for these four subtasks in five languages: Arabic, Bulgarian, Dutch, English, and Turkish. For a sixth language, Spanish, we provided a larger-scale dataset annotated by investigative journalists, but only for the subtask of check-worthiness identification.

The CheckThat! 2022 Task 1 framework enabled the experimentation on identifying relevant tweets with different technologies, most of them built with transformer-based monolingual and multilingual approaches, by 18 teams from around the world. Some of the most successful approaches produced multilingual sequence-to-sequence models to take advantage of languages with large amounts of training material to help processing tweets in less-resourced languages [4]. Another team explored training a feed-forward neural network with BERT embeddings and Manifold Mixup regularization [5] to have smoother decision boundaries.

Among the different subtasks for finding relevant tweets, the check-worthiness one was the most popular. English was the most popular target language for the participants. Across the different submitted systems, transformer-based models were widely used. The top-ranked systems also used data augmentation and additional prepossessing steps.

The remainder of the paper is organized as follows. Section 2 presents the different subtasks offered this year. Section 3 describes the datasets and the evaluation measures. Section 5 discusses the system submissions and the evaluation results. Section 6 presents some related work. Section 7 offers final remarks.

---

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

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

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

2. Identifying Relevant Claims in Tweets

The aim of Task 1 is to determine whether a claim in a tweet is worth fact-checking. In order to do that, we either resort to the judgments of professional fact-checkers (for Spanish) or we ask non-expert annotators to answer several auxiliary questions [8, 9], such as "*does the tweet contain a verifiable factual claim?*", "*is it harmful?*", and "*is it of general interest?*", before deciding on the final check-worthiness label. These questions opened the door to setting up four subtasks:

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

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

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

Subtask 1D: Attention-worthy tweet detection.
　　　Given a tweet, predict whether it should get the attention of policymakers and why.

　　Subtasks 1A, 1B, and 1C are all binary problems and the models are expected to establish whether a tweet is relevant according to different criteria. Task 1D is a multi-class problem. Table 1 shows the class labels for each task. Regarding languages, Arabic, Bulgarian, Dutch, English, and Turkish are present in all four subtasks, whereas Spanish is included only in Subtask 1A. We created and released an independently labeled dataset per language as explained in the following section. The participants were free to work on any language(s) of their interest, and they could also use multilingual approaches that make use of all datasets for training.

3. Datasets

For all subtasks (**1A, 1B, 1C**, and **1D**) and for Arabic, Bulgarian, Dutch, and English languages we used the datasets described in [9]. The datasets were developed based on a multi-question annotation schema and the annotated tweets for the languages mentioned above [8]. Following the same annotation schema, we further annotated a Turkish dataset.

Table 2
Statistics about the CT–CWT–22 corpus for all six languages and four subtasks. The bottom part of the table shows the main topics covered.

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

For Spanish, we used a different approach. The Spanish tweets for subtask 1A were manually annotated by journalists from Newtral —a Spanish fact-checking organization— and came from the Twitter accounts of 300 Spanish politicians. Moreover, the Spanish dataset is the largest one across the six languages, by a margin. Table 2 shows some statistics about the datasets, which are split into training, development, and testing partitions.

Although all languages tackled major topics such as COVID-19 and politics, the crawling and the annotation were done differently across the languages due to different resources available. Below, we provide more detail about the crawling and the annotation for each language.

3.1. Arabic, Bulgarian, Dutch and English Datasets

We collected tweets by specifying the target language, a set of COVID-19 keywords (shown on Figure 2), and time frames from January 2020 till March 2021. We removed retweets, replies, duplicates (using a similarity-based approach) [10], as well as tweets with less than five words. Finally, we selected the most frequently liked and retweeted tweets for annotation.

Arabic: #كورونا, كورونا, #كورونا (Corona), #فيروس_كورونا_المستجد, #فيروس_كورونا_الجديد, فيروس_كورونا (novel Coronavirus), #كورونا_فيروس, #كورنا, فيروس_كورن(Coronavirus), #كورونا_الجديد and كورونا_الجديد (new Corona), لقاحات لقاح, تطعيم, مطعوم, لقاح

Bulgarian: #корона, #коронавирус, коронавирус, корона

Dutch: #coronavirus, #COVID19, #coronaviruschina, #coronavirusNederland, #Italië, #RIVM, #coronavirusnederland, #CoronavirusOutbreak, #COVID-19, #CoronaVirusUpdates, #persconferentie, #Vindicat, #COVID_19, #hamsteren, #coronagekte, #coronapocalypse, #coronadebat, #scholendicht, #COVID19NL, #samentegencorona, #StayHomeSaveLives, #thuisonderwijs, #thuisblijven, #thuiswerken, #ikleesthuis, #groepimmuniteit, #LockdownNow, #blijfthuis, #houdafstand, #anderhalvemeter, #testen

English: #covid19, #CoronavirusOutbreak, #Coronavirus, #Corona, #CoronaAlert, #CoronaOutbreak, Corona, covid-19, COVID vaccine, Covid-19 vaccine, #covidvaccine, corona vaccine, #vaccinate, #vaccine, vaccine

Figure 2: The keywords used to collect the Arabic, the Bulgarian, the Dutch, and the English tweets.

For the data, we considered a number of factors, including tweet popularity in terms of retweets, which is already taken into account as part of the data collection process. We further asked the annotators to answer four questions for each instance:[2]

- **Verifiable factual claims: Does the tweet contain a verifiable factual claim?** This is an objective question, and it proved easy to annotate. Influenced by [11], positive examples include tweets that state a definition, mention a quantity in the present or in the past, make a verifiable prediction of the future, reference laws, procedures, and rules of operation, discuss images or videos, and state correlation or causation, among others.

- **Check-worthiness: Do you think that a professional fact-checker should verify the claim in the tweet?** This question asks for a subjective judgment. Yet, its answer should be based on whether the claim is likely to be false, is of public interest, and/or appears to be harmful. Note that we stress the fact that a professional fact-checker should verify the claim, ruling out claims that are easy to fact-check by a layperson.

- **Harmful tweet detection: Is the tweet harmful to the society and why?** This is an objective question. It further asks to categorize the nature of the harm if any. (even if we do not ask this to the participating models).

- **Attention-worthy tweet: Do you think that this tweet should get the attention of a government entity?** This question asks for a subjective judgment about whether the target tweet should get the attention of a government entity or of policymakers in general. The answers to this question are categorical and are not on an ordinal scale.

The annotations were performed by 2–5 people independently, and consolidated for the cases of disagreement. The annotation setup was part of a broader initiative; see [9] for details.

---

[2]We used the following MicroMappers setup for the annotations:
http://micromappers.qcri.org/project/covid19-tweet-labelling/.

3.2. Spanish Dataset

The Spanish dataset is an extended version from the one used in the 2021 edition of the lab [12]. A total of 5,000 new tweets have been added as test set, sampled from recent tweets from 350 well-known Spanish politicians. As in the previous year, professional journalists with expertise in fact-checking determined the level of check-worthiness of the tweets on the basis of diverse editorial criteria, such as actuality, public relevance of the character behind the claim, and potential impact on the general audience. All tweets in Spanish were annotated independently by three experts and the final decision was made by majority voting.

3.3. Turkish Dataset

We crawled Turkish tweets tracking keywords related to COVID-19 using Twitter API from 24 June, 2021 to September 29, 2021, yielding 10.87 M tweets in total. Keywords we used include "covid", "corona", "kovid", "korona", "aşı", "asi", "pfizer", "biontech", "sinovac", "astrazeneca', "moderna", "turkovac", "salgin", "pandemi", and "salgın". Subsequently, we deduplicated our crawl and eliminated tweets which (*i*) quote another tweet, (*ii*) are shorter than five words, (*iii*) start with a URL, or (*iv*) are posted as a reply to other messages. Subsequently, we sorted tweets based on their *popularity* in terms of number of retweets and *likes*. We picked the most popular 4K tweets whose cosine similarity score was less than 0.75 for the annotation.

Each tweet was annotated by three people independently using the same platform utilized for creating datasets for Arabic, Bulgarian, Dutch, and English. Disagreements on the annotations were solved by discussions among the annotators. However, if disagreements for a particular tweet continued even after the discussion, that tweet is discarded.

4. Evaluation Settings

For the lab, we provided the training, development dev-test set to enable the participants to validate their systems internally, while they could use the dev set for parameter tuning. For each language and subtask, we have annotated new instances, using three or four annotators per instance. Class label has been assigned by majority voting and disagreements have been solved by a consolidator or discussions among annotators. The test set has been used for final evaluation and system ranking. Participants were allowed to submit as many runs as they wanted on the test set, but only the last one was considered as official.

The evaluation framework is completed by the evaluation metrics for each subtask. For **subtasks 1A and 1C**, we used the F_1-measure with respect to the positive class (yes), to account for class imbalance. For **subtask 1B**, we used accuracy, as the data is fairly balanced. For **subtask 1D**, we used weighted-F_1, as there are multiple classes and we wanted them appropriately weighted. The data and the evaluation scripts are available online.[3]

---

[3]https://gitlab.com/checkthat_lab/clef2022-checkthat-lab/

5. Overview of the Systems and Evaluation Results

Eighteen teams took part in this task, with English being the most popular language. Across the different subtasks, some teams paid special attention to multiple languages, mostly through three different strategies. **MT-based data augmentation**. Team TOBB-ETU [13] (the only team that targeted all four subtasks in almost all languages available) applied both translation and back-translation to increase the amount of training data in the different languages. Language-specific models were then trained. **Multilingual transformer**. Team NUS-IDS [4] adopted a multilingual mT5 transformer to train one single model and apply it to multiple languages. **Zero-shot**. Team PoliMi-FlatEarthers [14] fine-tuned a GPT-3 model for each of the four subtasks feeding only instances in English and applied them to other languages during testing.

In the rest of the section we zoom into each of the subtasks, by looking into the models and resources explored and the performance of the official runs. Tables 3, 5, 7 and 9 offer a bird's eye overview of the participants' approaches to each of the four subtasks. Tables 4, 6, 8 and 10 show the official evaluation scores for each of the four subtasks. Appendix A includes a brief description of the approaches for every participating team.

5.1. Subtask 1A. Check-Worthiness Estimation

A total of 18 teams took part in this task, with English, Bulgarian, and Dutch being the most popular languages. Three teams —NUS-IDS [4]), PoliMi-FlatEarthers [14] and TOBB ETU [13]— participated in five out of the six languages offered. Table 3 shows an overview of the approaches, whereas Table 4 shows the performance of the official submissions, ranked on the basis of F_1 with respect to the positive class. The baseline consists of a random system. We now zoom into the different languages.

Arabic Four teams participated. As observed for other languages, the multilingual approach of team **NUS-IDS** [4] allowed them to effectively take advantage of the supervised data in other languages, resulting in the top-performing approach. It is based on mT5, a multilingual sequence-to-sequence transformer pretrained on the mC4 corpus, which covers 101 languages. The second best system, **TOBB ETU** [13], used fine-tuned AraBERT.

Bulgarian Five teams took part. Once again **NUS-IDS** [4] was the top-ranked team, followed by Team **TOBB ETU** [13], using the same approaches as for Arabic.

Dutch Five teams participated. The multilingual approach by team **NUS-IDS** [4] outperformed the other systems again. This time, team **AI Rational** [15] arrived second with a system built with RoBERTa, after a data augmentation process based on back-translation.

English A total of 13 teams took part. The top-ranked team was **AI Rational** [15], with a similar approach to the one they applied for Dutch. Team **Zorros** [23] submitted the second-best system, based on an ensemble approach combining BERT and RoBERTa. Two other aspects are worth highlighting. On the one hand, **PoliMi-FlatEarthers** [14] fine-tuned a unique GPT-3 model with all instances in English. When applying it on the English test set, they obtained

Table 3

Overview of the approaches to **subtask 1A**. The numbers in the language box refer to the position of the team in the official ranking; marks under multilingual flag models that have been trained on more than one language at once; ☑=part of the official submission; ✔=considered in internal experiments.

Team		Languages							Transformers								Repr.			Classifiers					Miscellaneous					
		Arabic	Bulgarian	Dutch	English	Spanish	Turkish	Multilingual	BERT	DistilBERT	Electra	GPT-3	mT5-XL	RoBERTa	XLM RoBERTa	XLNet	ELMo	LIWC	word n-grams	CNN	Random forest	RNN	SVM	XGBoost	Data augmentation	Data normalization	Ensemble	Multi-task learn.	Quantum NLP	Semi-supervised
AI Rational	[15]		3	2	1		2		✔	✔				☑	☑										☑	☑				
ARC-NLP	[16]				3				✔					☑					☑					☑	☑	☑				
Asatya	[17]				10				☑																☑	☑	✔			
Fraunhofer SIT	[18]				5				☑					☑											☑	☑	☑		✔	☑
iCompass	[19]	3							☑											☑		☑					✔			
NUS-IDS	[4]	1	1	1	8	1		☑					☑	✔			✔											☑		
PoliMi-FlatE.	[14]	5	5	4	3	2		☑					☑																	
RUB-DFL	[20]				6		1		☑	✔					✔		☑	☑							☑	☑				
TOBB ETU	[13]	2	2	3	4		4		☑	✔		☑													☑					
VTU BGM	[21]				11										☑								☑		☑					
Z-Index	[22]		5	12	3			☑	☑					✔				☑			✔		✔		☑	☑				
Zorros	[23]			2					☑					☑											☑	☑				

the third best performance. The zero-shot application of the model to other languages was not as successful. On the other hand, teams that trained actual multilingual models, that is **NUS-IDS** [4] and **Z-index** [22], struggled when dealing with English.

Spanish   Three teams took part. The multilingual approach by team **NUS-IDS** [4] was the most successful by a margin over the top runner —the zero-shot approach by **PoliMi-FlatEarthers** [14]. A first sight might suggest that such zero-shot approach hints to be working for Spanish (their performance goes below the baseline in the other languages), but the distance to the top model is still bigger than two points. It is worth observing that the performances over the Spanish datasets are the lowest. Whether the reason behind is that this is the only one annotated by expert journalists rather than by crowdsourcing remains an open topic.

Turkish   Four teams participated. All participants used BERT-based models and GPT-3. Team **RUB-DFL** [20] is the top-performing one. The system uses BERT with a combination of both ELMo and LIWC features. The runner up team **AI Rational** applied standard pre-processing and data augmentation with back translation.

5.2. Subtask 1B: Verifiable Factual Claims Detection

Thirteen teams took part in Subtask 1B, with English, Bulgarian and Arabic being the most popular languages. Team **TOBB ETU** [13] participated in all five languages, whereas team

Table 4
Subtask 1A: Check-Worthiness estimation, results for the official submissions in all six languages. F1 with respect to the positive class. Baseline is the random baseline.

Team	F1	Team	F1	Team	F1
Arabic		**English**		**Spanish**	
1. NUS-IDS [4]	0.628	1. AI Rational [15]	0.698	1. NUS-IDS [4]	0.571
2. TOBB ETU [13]	0.495	2. Zorros [23]	0.667	2. PoliMi-FlatEarthers [14]	0.323
3. iCompass [19]	0.462	3. PoliMi-FlatEarthers [14]	0.626	3. Z-Index [22]	0.303
4. Baseline	0.347	4. TOBB ETU [13]	0.561	4. Baseline	0.139
5. PoliMi-FlatEarthers [14]	0.321	5. Fraunhofer SIT [18]	0.552	**Turkish**	
Bulgarian		6. RUB-DFL [20]	0.525	1. RUB-DFL [20]	0.801
1. NUS-IDS [4]	0.617	7. hinokicrum*	0.522	2. AI Rational [15]	0.789
2. TOBB ETU [13]	0.542	8. NUS-IDS [4]	0.519	3. ARC-NLP [16]	0.760
3. AI Rational [15]	0.483	9. TonyTTTTT*	0.500	4. TOBB ETU [13]	0.729
4. Baseline	0.434	10. Asatya [17]	0.500	5. Baseline	0.496
5. PoliMi-FlatEarthers [14]	0.341	11. VTU_BGM [21]	0.482		
6. pogs2022*	0.000	12. Z-Index [22]	0.478		
Dutch		13. NLP&IR@UNED*	0.469		
1. NUS-IDS [4]	0.642	14. Baseline	0.253		
2. AI Rational [15]	0.620				
3. TOBB ETU [13]	0.534				
4. PoliMi-FlatEarthers [14]	0.532				
5. Z-Index [22]	0.497				
6. Baseline	0.451				

*No working note submitted.

AI Rational participated in four. Table 5, shows an overview of the explored approaches and Table 6 shows the performance of the official submissions on the test set, including the random baseline. The table shows the runs ranked on the basis of the official *accuracy* measure.

Arabic   Three teams participated. Team **TOBB ETU** [13] ranked best, being the only one that surpassed the baseline. They used a four-layer feedforward network with Manifold Mixup regularization and BERT embeddings. Arabic showed to be the most challenging language for Task 1B, with a top performance shorter than 0.60 and ony one team beating the baseline.

Bulgarian   Two teams stook part. Team **AI Rational** [15] topped using XLM-RoBERTa with data augmentation, followed by Team **TOBB ETU** [13], which fine-tuned RoBERTa.

Dutch   Two teams participated. Team **AI Rational** [15] and **TOBB ETU** [13] used similar approaches as for Bulgarian.

English   Nine teams participated. Team **PoliMi-FlatEarthers** [14] ranked as the best system thanks to their GPT-3 fine-tuning. Team **Asatya** [17] arrived second by fine-tuning BERT.

Turkish   Four teams participated. The top-ranked team is **RUB-DFL** [20], which used again a combination of RoBERTa, Electra, and BERTurk. The second-best team is **AI Rational** [15],

Table 5
Overview of the approaches to **subtask 1B**. The numbers in the language box refer to the position of the team in the official ranking; ☑=part of the official submission; ✔=considered in internal experiments.

Team		Languages					Transformers								Repr.		Misc	
		Arabic	Bulgarian	Dutch	English	Turkish	DistilBERT	BERT	RoBERTa	XLM RoBERTa	GPT-3	ELMo	Electra	XLNet	LIWC	word n-grams	Data augmentation	Preprocessing
AI Rational	[15]		1	1	4	2	✔	✔	☑	☑							☑	☑
ARC-NLP	[16]				3			☑								✔		
Asatya	[17]			2	3		☑										☑	☑
PoliMi-FlatEarthers	[14]				1					☑								
RUB-DFL	[20]				6	1	☑	☑				☑	☑		✔	✔		
TOBB ETU	[13]	1	2	2	9	4	✔	☑	☑									☑
VTU BGM	[21]				7			☑						☑				☑
Zorros	[23]				5			☑	☑									☑

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

Team	Acc	Team	Acc	Team	Acc
Arabic		**English**		**Turkish**	
1. TOBB ETU [13]	0.570	1. PoliMi-FlatEarthers [14]	0.761	1. RUB-DFL [20]	0.801
2. Baseline	0.531	2. Asatya [17]	0.749	3. AI Rational [15]	0.789
3. claeser*	0.454	3. NLP&IR@UNED*	0.725	3. ARC-NLP [16]	0.760
4. pogs2022*	0.454	4. AI Rational [15]	0.713	4. TOBB ETU [13]	0.729
Bulgarian		5. Zorros [23]	0.709	5. Baseline	0.496
1. AI Rational [15]	0.839	6. RUB-DFL [20]	0.709		
2. TOBB ETU [13]	0.742	7. VTU_BGM [21]	0.709		
3. Baseline	0.535	8. hinokicrum*	0.665		
Dutch		9. TOBB ETU [13]	0.641		
1. AI Rational [15]	0.736	10. Baseline	0.494		
2. TOBB ETU [13]	0.658				
3. Baseline	0.521				

*No working note submitted.

which used BERT, RoBERTa, and DistilBERT.

5.3. Subtask 1C: Harmful Tweets Detection

Thirteen teams participated in subtask 1C, with English and Turkish being the most popular languages. Teams TOBB ETU [13] and AI Rational [15] participated in five and four languages,

Table 7

Overview of the approaches to **subtask 1C**. The numbers in the language box refer to the position of the team in the official ranking; ☑=part of the official submission; ✔=considered in internal experiments.

Team		Languages					Transformers									Repr.		Misc	
		Arabic	Bulgarian	Dutch	English	Turkish	DistilBERT	BERT	RoBERTa	XLM RoBERTa	GPT-3	ELMo	Electra	XLNet	LIWC	word n-grams	sentiment	Data augmentation	Preprocessing
AI Rational	[15]		1	2	2	3	✔	✔	☑	☑								☑	☑
ARC-NLP	[16]				6	1		☑								☑			
Asatya	[17]				3			☑										☑	☑
COURAGE	[24]				8								☑						☑
iCompass	[19]	1						☑									☑		☑
PoliMi-FlatEarthers	[14]				10					☑									
RUB-DFL	[20]				9	2			☑			☑	☑	✔		✔			✔
TOBB ETU	[13]	2	2	1	5	4	✔	☑	☑									☑	
VTU BGM	[21]				7			☑					☑						☑
Zorros	[23]				1		☑	☑											☑

respectively. Table 7 overviews the teams' approaches and Table 8 shows the performance of the official submissions on the test set, together with the random baseline. The table shows the runs ranked based on the official *F1 with respect to the positive class*.

Arabic   Two teams participated. Team **iCompass** [19] obtained the top performance by fine-tuning AraBERT and ARBERT. Team **TOBB ETU** [13] opted for a feedforward network trained with Manifold Mixup regularization and represented tweets with AraBERT embeddings. Notice that the top model doubles the performance of the second one.

Bulgarian   Two teams participated. Team **AI Rational** [15] topped using XLM-RoBERTa. **TOBB ETU** [13] arrived second with a fine-tuned RoBERTa. Both applied data augmentation via back-translation. Bulgarian showed to be the most challenging language in task 1C.

Dutch   Two teams participated. Team **TOBB ETU** [13] ranked on top by fine-tuning BERTje [25] after data-augmentation. Team **AI Rational** [15] used XLM-RoBERTa.

English   Eleven teams participated. Team **Zorros** [23] ranked as the best system, using an ensemble of five transformers. Team **ARC-NLP** [16] ranked second. Besides transformer-based models across all approaches, some teams have also used data augmentation.

Turkish   Four teams participated. Team **ARC-NLP** [16] ranked on top by approaching harm as a contradiction detection problem. They extracted facts related to COVID-19 from reliable

Table 8
Subtask 1C: Harmful Tweet Detection, results for the official submissions in all five languages.

Team	F1	Team	F1	Team	F1
Arabic		**English**		**Turkish**	
1. iCompass [19]	0.557	1. Zorros [23]	0.397	1. ARC-NLP [16]	0.366
2. TOBB ETU [13]	0.268	2. AI Rational [15]	0.361	2. RUB-DFL [20]	0.353
3. Baseline	0.118	3. Asatya [17]	0.361	3. AI Rational [15]	0.346
Bulgarian		4. NLP&IR@UNED*	0.347	4. TOBB ETU [13]	0.262
1. AI Rational [15]	0.286	5. TOBB ETU [13]	0.329	5. Baseline	0.061
2. TOBB ETU [13]	0.054	6. ARC-NLP [16]	0.300		
3. Baseline	0.000	7. hinokicrum*	0.281		
Dutch		8. COURAGE [24]	0.280		
1. TOBB ETU [13]	0.358	9. RUB-DFL [20]	0.273		
2. AI Rational [15]	0.147	10. PoliMi-FlatEarthers [14]	0.270		
3. Baseline	0.114	11. Baseline	0.200		
		12 VTU_BGM [21]	0.000		

*No working note submitted.

Table 9
Overview of the approaches to **subtask 1D**. The numbers in the language box refer to the position of the team in the official ranking; ☑=part of the official submission; ✔=considered in internal experiments.

Team	Languages					Transformers						Misc	
	Arabic	Bulgarian	Dutch	English	Turkish	DistilBERT	BERT	RoBERTa	XLM RoBERTa	GPT-3		Data augmentation	Preprocessing
AI Rational [15]		1	1	3	1	✔	✔	☑	☑			☑	☑
PoliMi-FlatEarthers [14]				7						☑			
TOBB ETU [13]	2	2	2	4	3	☑	☑	☑					✔
Zorros [23]				1				☑	☑				☑

sources and associated them with tweets based on similarity. The pairs were used to fine-tune BERTurk. Team **RUB-DFL** [20] arrived second with a fine-tuned ConvBert.

5.4. Subtask 1D: Attention-Worthy Tweet Detection

Seven teams participated in subtask 1D, with English being the most popular language. Again, teams TOBB ETU [13] and AI Rational [15] participated in five and four languages. Table 9 overviews the approaches whereas Table 10 shows the performance of the official submissions on the test, together with the random baseline. The ranking is based on the official *weighted F1*.

Table 10
Subtask 1D: Attention-Worthy Tweet Detection, results for the official submissions in all five languages. Performance is reported as weighted F1.

Team	F1	Team	F1	Team	F1
Arabic		**English**		**Turkish**	
1. Baseline	0.206	1. Zorros [23]	0.725	1. AI Rational [15]	0.895
2. TOBB ETU [13]	0.184	2. Baseline	0.695	2. Baseline	0.853
Bulgarian		3. AI Rational [15]	0.684	3. TOBB ETU [13]	0.806
1. AI Rational [15]	0.915	4. TOBB ETU [13]	0.670	**Dutch**	
2. TOBB ETU [13]	0.877	5. NLP&IR@UNED*	0.650	1. AI Rational [15]	0.715
3. Baseline	0.875	6. hinokicrum*	0.643	2. TOBB ETU [13]	0.694
		7. PoliMi-FlatEarthers [14]	0.636	3. Baseline	0.641

*No working note submitted.

Arabic   Only one team participated. Team **TOBB ETU** [13] ran short and did not manage to pass the random baseline. They tried with a dual approach: the combination of a binary model attention-worthy vs not and a multi-class model with the original classes.

Bulgarian   Two teams participated. Both Teams **AI Rational** [15] and **TOBB ETU** [13] used similar models as the ones for subtask 1C.

Dutch   Two teams participated, resulting in a similar picture as for Bulgarian.

English   Six teams participated. Team **Zorros** [23] ranked first, by fine-tuning a COVID Twitter BERT pre-trained model. The random baseline ranked second.

Turkish   Two teams participated. Team **AI Rational** [15] was the only team to beat the baseline with a fine-tuned XLM-RoBERTa model.

6. Related Work

There has been a significant research interest in recent years in identifying disinformation, misinformation, and "fake news", which thrive in social media, political debates and speeches. Several recent works highlighted how information is disseminated and consumed in social media [26], fact-checking perspective on "fake news" and related problems [27], truth discovery [28], stance towards misinformation and disinformation detection [29], automatic fact-checking to assist human fact-checkers [30], predicting the factuality and the bias of entire news outlets [31], multimodal disinformation detection [32], and on abusive language in social media [33].

Within the scope in identifying disinformation, misinformation and the "fake news" in general, the research interests have focused on more specific problems such as automatic identification and verification of claims [34, 35, 36, 37, 7, 38, 39], to identifying check-worthy claims [40, 41,

42, 43, 44], detecting whether a claim has been previously fact-checked [45, 46, 47], retrieving evidence to accept or to reject a claim [48, 49], checking whether the evidence supports or denies the claim [50, 51], and inferring the veracity of the claim [52, 53, 54, 55, 56, 57, 49, 58, 59, 60]. Such specific tasks can help fact-checkers and/or journalists.

Among these tasks check-worthiness estimation received an wider attention since the pioneering work proposed by [41], where the idea is to detect whether a sentence in a political debate is *non-factual*, *unimportant factual*, or *check-worthy factual*. The proposed system later extended with more data and to cover Arabic content [42]. Most of the earlier work on check-worthiness estimation was mainly focused political debates [61, 43] and lately attention has been focused on social media [9, 8, 62].

Major research attention emerged due to the `CheckThat!` lab initiatives in CLEF 2018, 2019, 2020, and 2021 where the focus was once again on political debates and speeches, from a single fact-checking organization. In the 2018 edition of the task, a total of seven teams submitted runs for Task 1 (which corresponds to Subtask 1B in 2021), with systems based on word embeddings and RNNs [63, 64, 65, 66]. In the 2019 edition of the task, eleven teams submitted runs for the corresponding Task 1, again using word embeddings and RNNs, and further trying a number of interesting representations [67, 68, 69, 70, 71, 72, 73, 74]. In the 2020 edition of the task, three teams submitted runs for the corresponding Task 5 with systems based on word embeddings and BiLSTM [75], TF.IDF representation with Naïve Bayes, logistic regression, decision trees [76], BERT prediction scores, and word embeddings with logistic regression [77]. Several teams fine-tuned pre-trained models such as AraBERT and multilingual BERT [78, 79, 77]. Other approaches relied on pre-trained models such as GloVe and Word2vec [80, 75] to obtain embeddings for the tweets, which were fed into a neural network or an SVM. In addition to text representations, some teams used other features, namely morphological and syntactic, part-of-speech (POS) tags, named entities, and sentiment features [81, 82]. As for the English task, we also observed the popularity of pre-trained Transformers, namely BERT and RoBERTa [78, 80, 83, 84, 77]. In the 2021 edition [12], check-worthiness estimation has been offered for political debates/speeches and tweets. Top ranked system used transformer based models [85, 86].

A large body of work has been devoted to identifying the factuality of claims, which are often expressed and disseminated through social networks [87]. The studies include fact-checking on news media [88], fact-checking such as fact-checked URL recommendation model [89] , fact-checking with stance detection [57], factuality of media outlets [90], generating justifications for verdicts on claims [91], and fact-checking claims from Wikipedia [92].

For harmfulness detection, research mainly focused on offensive and hateful content on social media that can harm an individual, organization, and society [93, 94, 95].

Attention-worthiness is a relatively new area, which has recently been proposed in [9, 8] and has not been explored yet in the current literature, and the `CheckThat!` lab 2022 initiative opened up a new research interest on this topic.

7. Conclusions

We have presented an overview of task 1 of the CLEF-2022 `CheckThat!` lab. The lab featured tasks that span the full verification pipeline: from spotting check-worthy claims to checking

whether they have been fact-checked before. Task 1 asked to identify relevant claims in tweets in terms of check-worthiness, verifiability, harmfulness, and attention-worthiness.

Inline with the general mission of CLEF, we promoted multilinguality by offering the task in six different languages. This edition of the task has attracted diverse approaches in terms of model (e.g., kind of transformer), representations (e.g., embeddings, n-grams) and data augmentation (e.g., back-translation). Among the most innovative ones, we highlight the use of quantum NLP [4], GPT-3 [14], and mT5 [4]. Teams targeting multiple languages tended to rank at the top positions across tasks and languages. The exception is English. Teams excelling in multiple languages tended to rank relatively low in this language.

The general problem of check-worthiness estimation remains open in general and further efforts could be paid on considering external evidence when assessing whether a tweet is calling for the attention of a verifier. In this edition of the lab, we have observed already efforts in this direction and the results are promising.

Acknowledgments

References

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

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

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

[4] S. K. N. Mingzhe Du, Sujatha Das Gollapalli, NUS-IDS at CheckThat! 2022: identifying check-worthiness of tweets using CheckthaT5, in: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, CLEF '2022, Bologna, Italy, 2022.

---

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

[5] V. Verma, A. Lamb, C. Beckham, A. Najafi, I. Mitliagkas, D. Lopez-Paz, Y. Bengio, Manifold mixup: Better representations by interpolating hidden states, in: International Conference on Machine Learning, PMLR, 2019, pp. 6438–6447.

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

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

[8] F. Alam, F. Dalvi, S. Shaar, N. Durrani, H. Mubarak, A. Nikolov, G. Da San Martino, A. Abdelali, H. Sajjad, K. Darwish, P. Nakov, Fighting the COVID-19 infodemic in social media: A holistic perspective and a call to arms, in: Proceedings of the International AAAI Conference on Web and Social Media, ICWSM '21, 2021, pp. 913–922.

[9] 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.

[10] F. Alam, H. Sajjad, M. Imran, F. Ofli, CrisisBench: Benchmarking crisis-related social media datasets for humanitarian information processing, in: Proceedings of the International AAAI Conference on Web and Social Media, ICWSM '21, 2021, pp. 923–932. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/18115.

[11] L. Konstantinovskiy, O. Price, M. Babakar, A. Zubiaga, Towards automated factchecking: Developing an annotation schema and benchmark for consistent automated claim detection, arXiv:1809.08193 (2018).

[12] 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, in: [96], 2021.

[13] A. B. Eyuboglu, M. B. Arslan, E. Sonmezer, M. Kutlu, TOBB ETU at CheckThat! 2022: detecting attention-worthy and harmful tweets and check-worthy claims, in: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, CLEF '2022, Bologna, Italy, 2022.

[14] S. Agrestia, A. S. Hashemianb, M. J. Carmanc, PoliMi-FlatEarthers at CheckThat! 2022: GPT-3 applied to claim detection, in: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, CLEF '2022, Bologna, Italy, 2022.

[15] A. Savchev, AI Rational at CheckThat! 2022: using transformer models for tweet classification, in: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, CLEF '2022, Bologna, Italy, 2022.

[16] C. Toraman, O. Ozcelik, F. Şahinuç, U. Sahin, ARC-NLP at CheckThat! 2022: contradiction for harmful tweet detection, in: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, CLEF '2022, Bologna, Italy, 2022.

[17] P. K. Manan Suri, S. Dudeja, Asatya at CheckThat! 2022: multimodal BERT for identifying claims in tweets, in: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, CLEF '2022, Bologna, Italy, 2022.

[18] R. A. Frick, I. Vogel, I. Nunes Grieser, Fraunhofer SIT at CheckThat! 2022: semi-supervised ensemble classification for detecting check-worthy tweets, in: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, CLEF '2022, Bologna, Italy, 2022.

[19] T. Bilel, B. N. Mohamed Aziz, H. Haddad, iCompass at CheckThat! 2022: ARBERT and AraBERT for Arabic checkworthy tweet identification, in: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, CLEF '2022, Bologna, Italy, 2022.

[20] Z. M. Hüsünbeyi, O. Deck, T. Scheffler, RUB-DFL at CheckThat! 2022: transformer models and linguistic features for identifying relevant claims, in: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, CLEF '2022, Bologna, Italy, 2022.

[21] S. Kavatagi, R. Rachh, M. Mulimani, VTU_BGM at Check That! 2022: an autoregressive encoding model for verifying check-worthy claims, in: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, CLEF '2022, Bologna, Italy, 2022.

[22] P. Tarannum, H. Md. Arid, F. Alam, S. R. H. Noori, Z-Index at CheckThat! Lab 2022: check-worthiness identification on tweet text, in: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, CLEF '2022, Bologna, Italy, 2022.

[23] R. M. Buliga Nicu, Zorros at CheckThat! 2022: ensemble model for identifying relevant claims in tweets, in: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, CLEF '2022, Bologna, Italy, 2022.

[24] F. Lomonaco, G. Donabauer, M. Siino, COURAGE at CheckThat! 2022: harmful tweet detection using graph neural networks and ELECTRA, in: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, CLEF '2022, Bologna, Italy, 2022.

[25] W. de Vries, A. van Cranenburgh, A. Bisazza, T. Caselli, G. van Noord, M. Nissim, Bertje: A dutch bert model, arXiv preprint arXiv:1912.09582 (2019).

[26] K. Shu, A. Sliva, S. Wang, J. Tang, H. Liu, Fake news detection on social media: A data mining perspective, SIGKDD Explor. Newsl. 19 (2017) 22–36. URL: http://doi.acm.org/10.1145/3137597.3137600. doi:10.1145/3137597.3137600.

[27] 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. URL: http://www.aclweb.org/anthology/C18-1283.

[28] Y. Li, J. Gao, C. Meng, Q. Li, L. Su, B. Zhao, W. Fan, J. Han, A survey on truth discovery, SIGKDD Explor. Newsl. 17 (2016) 1–16. URL: http://doi.acm.org/10.1145/2897350.2897352. doi:10.1145/2897350.2897352.

[29] M. Hardalov, A. Arora, P. Nakov, I. Augenstein, A survey on stance detection for mis- and disinformation identification, arXiv/2103.00242 (2021).

[30] 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.

[31] 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).

[32] F. Alam, S. Cresci, T. Chakraborty, F. Silvestri, D. Dimitrov, G. D. S. Martino, S. Shaar, H. Firooz, P. Nakov, A survey on multimodal disinformation detection, arXiv/2103.12541 (2021).

[33] P. Nakov, V. Nayak, K. Dent, A. Bhatawdekar, S. M. Sarwar, M. Hardalov, Y. Dinkov, D. Zlatkova, G. Bouchard, I. Augenstein, Detecting abusive language on online platforms: A critical analysis, arXiv/2103.00153 (2021).

[34] P. Atanasova, L. Màrquez, 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, in: CLEF 2018 Working Notes. Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2018.

[35] 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, in: [97], 2019.

[36] A. Barrón-Cedeño, T. Elsayed, R. Suwaileh, L. Marquez, P. Atanasova, 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 2: Factuality, in: [98], 2018.

[37] T. Elsayed, P. Nakov, A. Barrón-Cedeño, M. Hasanain, R. Suwaileh, G. Da San Martino, P. Atanasova, CheckThat! at CLEF 2019: Automatic identification and verification of claims, in: L. Azzopardi, B. Stein, N. Fuhr, P. Mayr, C. Hauff, D. Hiemstra (Eds.), Advances in Information Retrieval, ECIR '19, 2019, pp. 309–315.

[38] M. Hasanain, R. Suwaileh, T. Elsayed, A. Barrón-Cedeño, P. Nakov, Overview of the CLEF-2019 CheckThat! lab on automatic identification and verification of claims. Task 2: Evidence and factuality, in: [97], 2019.

[39] P. Nakov, A. Barrón-Cedeño, T. Elsayed, R. Suwaileh, L. Màrquez, W. Zaghouani, P. Atanasova, S. Kyuchukov, G. Da San Martino, Overview of the CLEF-2018 CheckThat! lab on automatic identification and verification of political claims, in: Proceedings of the Ninth International Conference of the CLEF Association: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Lecture Notes in Computer Science, 2018, pp. 372–387.

[40] P. Gencheva, P. Nakov, L. Màrquez, A. Barrón-Cedeño, I. Koychev, A context-aware approach for detecting worth-checking claims in political debates, in: Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, 2017, pp. 267–276.

[41] 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.

[42] I. Jaradat, P. Gencheva, A. Barrón-Cedeño, L. Màrquez, P. Nakov, ClaimRank: Detecting check-worthy claims in Arabic and English, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, 2018, pp. 26–30.

[43] 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.

[44] 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).

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

[46] P. Nakov, D. Corney, M. Hasanain, F. Alam, T. Elsayed, A. Barrón-Cedeño, P. Papotti, S. Shaar, G. D. S. Martino, Automated fact-checking for assisting human fact-checkers (2021).

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

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

[49] G. Karadzhov, P. Nakov, L. Màrquez, A. Barrón-Cedeño, I. Koychev, Fully automated fact checking using external sources, in: Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, 2017, pp. 344–353.

[50] M. Mohtarami, R. Baly, J. Glass, P. Nakov, L. Màrquez, A. Moschitti, Automatic stance detection using end-to-end memory networks, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, 2018, pp. 767–776.

[51] M. Mohtarami, J. Glass, P. Nakov, Contrastive language adaptation for cross-lingual stance detection, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, 2019, pp. 4442–4452.

[52] P. Atanasova, P. Nakov, L. Màrquez, A. Barrón-Cedeño, G. Karadzhov, T. Mihaylova, M. Mohtarami, J. Glass, Automatic fact-checking using context and discourse information, Journal of Data and Information Quality (JDIQ) 11 (2019) 12.

[53] C. Castillo, M. Mendoza, B. Poblete, Information credibility on twitter, in: S. Srinivasan, K. Ramamritham, A. Kumar, M. P. Ravindra, E. Bertino, R. Kumar (Eds.), Proceedings of the 20th International Conference on World Wide Web, WWW 2011, 2011, pp. 675–684.

[54] D. Kopev, A. Ali, I. Koychev, P. Nakov, Detecting deception in political debates using acoustic and textual features, in: Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop, ASRU '19, 2019, pp. 652–659.

[55] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, Y. Choi, Truth of varying shades: Analyzing language in fake news and political fact-checking, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 2931–2937.

[56] R. Baly, G. Karadzhov, D. Alexandrov, J. Glass, P. Nakov, Predicting factuality of reporting and bias of news media sources, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 3528–3539.

[57] R. Baly, M. Mohtarami, J. Glass, L. Màrquez, A. Moschitti, P. Nakov, Integrating stance

detection and fact checking in a unified corpus, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), 2018, pp. 21–27.

[58] V. Nguyen, K. Sugiyama, P. Nakov, M. Kan, FANG: leveraging social context for fake news detection using graph representation, in: M. d'Aquin, S. Dietze, C. Hauff, E. Curry, P. Cudré-Mauroux (Eds.), CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020, 2020, pp. 1165–1174.

[59] K. Popat, S. Mukherjee, J. Strötgen, G. Weikum, Where the truth lies: Explaining the credibility of emerging claims on the web and social media, in: Proceedings of the 26th International Conference on World Wide Web Companion, WWW '17, 2017, pp. 1003–1012.

[60] J. Thorne, A. Vlachos, C. Christodoulopoulos, A. Mittal, FEVER: a large-scale dataset for fact extraction and VERification, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 2018, pp. 809–819.

[61] A. Patwari, D. Goldwasser, S. Bagchi, TATHYA: A multi-classifier system for detecting check-worthy statements in political debates, in: E. Lim, M. Winslett, M. Sanderson, A. W. Fu, J. Sun, J. S. Culpepper, E. Lo, J. C. Ho, D. Donato, R. Agrawal, Y. Zheng, C. Castillo, A. Sun, V. S. Tseng, C. Li (Eds.), Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM, 2017, pp. 2259–2262.

[62] S. Shaar, F. Alam, G. Da San Martino, A. Nikolov, W. Zaghouani, P. Nakov, A. Feldman, Findings of the nlp4if-2021 shared tasks on fighting the covid-19 infodemic and censorship detection, in: Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda, 2021, pp. 82–92.

[63] R. Agez, C. Bosc, C. Lespagnol, J. Mothe, N. Petitcol, IRIT at CheckThat! 2018, in: [98], 2018.

[64] B. Ghanem, M. Montes-y Gómez, F. Rangel, P. Rosso, UPV-INAOE-Autoritas - Check That: Preliminary approach for checking worthiness of claims, in: [98], 2018.

[65] C. Hansen, C. Hansen, J. Simonsen, C. Lioma, The Copenhagen team participation in the check-worthiness task of the competition of automatic identification and verification of claims in political debates of the CLEF-2018 fact checking lab, in: [98], 2018.

[66] C. Zuo, A. Karakas, R. Banerjee, A hybrid recognition system for check-worthy claims using heuristics and supervised learning, in: [98], 2018.

[67] B. Altun, M. Kutlu, TOBB-ETU at CLEF 2019: Prioritizing claims based on check-worthiness, in: CLEF 2019 Working Notes. Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2019.

[68] L. Coca, C.-G. Cusmuliuc, A. Iftene, CheckThat! 2019 UAICS, in: CLEF 2019 Working Notes. Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2019.

[69] R. Dhar, S. Dutta, D. Das, A hybrid model to rank sentences for check-worthiness, in: CLEF 2019 Working Notes. Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2019.

[70] L. Favano, M. Carman, P. Lanzi, TheEarthIsFlat's submission to CLEF'19 CheckThat!

challenge, in: CLEF 2019 Working Notes. Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2019.

[71] J. Gasior, P. Przybyła, The IPIPAN team participation in the check-worthiness task of the CLEF2019 CheckThat! lab, in: CLEF 2019 Working Notes. Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2019.

[72] C. Hansen, C. Hansen, J. Simonsen, C. Lioma, Neural weakly supervised fact check-worthiness detection with contrastive sampling-based ranking loss, in: CLEF 2019 Working Notes. Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2019.

[73] S. Mohtaj, T. Himmelsbach, V. Woloszyn, S. Möller, The TU-Berlin team participation in the check-worthiness task of the CLEF-2019 CheckThat! lab, in: CLEF 2019 Working Notes. Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2019.

[74] T. Su, C. Macdonald, I. Ounis, Entity detection for check-worthiness prediction: Glasgow Terrier at CLEF CheckThat! 2019, in: CLEF 2019 Working Notes. Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2019.

[75] J. Martinez-Rico, L. Araujo, J. Martinez-Romo, NLP&IR@UNED at CheckThat! 2020: A preliminary approach for check-worthiness and claim retrieval tasks using neural networks and graphs, in: [99], 2020.

[76] T. McDonald, Z. Dong, Y. Zhang, R. Hampson, J. Young, Q. Cao, J. Leidner, M. Stevenson, The University of Sheffield at CheckThat! 2020: Claim identification and verification on Twitter, in: [99], 2020.

[77] Y. S. Kartal, M. Kutlu, TOBB ETU at CheckThat! 2020: Prioritizing English and Arabic claims based on check-worthiness, in: [99], 2020.

[78] E. Williams, P. Rodrigues, V. Novak, Accenture at CheckThat! 2020: If you say so: Post-hoc fact-checking of claims using transformer-based models, in: [99], 2020.

[79] M. Hasanain, T. Elsayed, bigIR at CheckThat! 2020: Multilingual BERT for ranking Arabic tweets by check-worthiness, in: [99], 2020.

[80] G. S. Cheema, S. Hakimov, R. Ewerth, Check_square at CheckThat! 2020: Claim detection in social media via fusion of transformer and syntactic features, in: [99], 2020.

[81] A. Hussein, A. Hussein, N. Ghneim, A. Joukhadar, DamascusTeam at CheckThat! 2020: Check worthiness on Twitter with hybrid CNN and RNN models, in: [99], 2020.

[82] I. Touahri, A. Mazroui, EvolutionTeam at CheckThat! 2020: Integration of linguistic and sentimental features in a fake news detection approach, in: [99], 2020.

[83] R. Alkhalifa, T. Yoong, E. Kochkina, A. Zubiaga, M. Liakata, QMUL-SDS at CheckThat! 2020: Determining COVID-19 tweet check-worthiness using an enhanced CT-BERT with numeric expressions, in: [99], 2020.

[84] A. Nikolov, G. Da San Martino, I. Koychev, P. Nakov, Team_Alex at CheckThat! 2020: Identifying check-worthy tweets with transformer models, in: [99], 2020.

[85] E. Williams, P. Rodrigues, S. Tran, Accenture at CheckThat! 2021: Interesting claim identification and ranking with contextually sensitive lexical training data augmentation, in: [96], 2021.

[86] X. Zhou, B. Wu, P. Fung, Fight for 4230 at CLEF CheckThat! 2021: Domain-specific preprocessing and pretrained model for ranking claims by check-worthiness, in: [96],

2021.

[87] F. Alam, S. Cresci, T. Chakraborty, F. Silvestri, D. Dimitrov, G. D. S. Martino, S. Shaar, H. Firooz, P. Nakov, A survey on multimodal disinformation detection, arXiv:2103.12541 (2021).

[88] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, Y. Choi, Truth of varying shades: Analyzing language in fake news and political fact-checking, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 2931–2937. URL: https://www.aclweb.org/anthology/D17-1317. doi:`10.18653/v1/D17-1317`.

[89] N. Vo, K. Lee, The rise of guardians: Fact-checking URL recommendation to combat fake news, in: K. Collins-Thompson, Q. Mei, B. D. Davison, Y. Liu, E. Yilmaz (Eds.), The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018, ACM, 2018, pp. 275–284. URL: https://doi.org/10.1145/3209978.3210037. doi:`10.1145/3209978.3210037`.

[90] R. Baly, G. Karadzhov, J. An, H. Kwak, Y. Dinkov, A. Ali, J. Glass, P. Nakov, What was written vs. who read it: News media profiling using text analysis and social media context, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL '20, Association for Computational Linguistics, Online, 2020, pp. 3364–3374. URL: https://aclanthology.org/2020.acl-main.308. doi:`10.18653/v1/2020.acl-main.308`.

[91] P. Atanasova, J. G. Simonsen, C. Lioma, I. Augenstein, Generating fact checking explanations, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL '20, Association for Computational Linguistics, Online, 2020, pp. 7352–7364. URL: https://aclanthology.org/2020.acl-main.656. doi:`10.18653/v1/2020.acl-main.656`.

[92] A. Sathe, S. Ather, T. M. Le, N. Perry, J. Park, Automated fact-checking of claims from Wikipedia, in: Proceedings of the 12th Language Resources and Evaluation Conference, ACL '20, European Language Resources Association, Marseille, France, 2020, pp. 6874–6882. URL: https://aclanthology.org/2020.lrec-1.849.

[93] S. Sharma, F. Alam, M. Akhtar, D. Dimitrov, G. D. S. Martino, H. Firooz, A. Halevy, F. Silvestri, P. Nakov, T. Chakraborty, et al., Detecting and understanding harmful memes: A survey, arXiv preprint arXiv:2205.04274 (2022).

[94] B. Haidar, M. Chamoun, F. Yamout, Cyberbullying detection: A survey on multilingual techniques, in: EMS, 2016, pp. 165–171. doi:`10.1109/EMS.2016.037`.

[95] F. Husain, O. Uzuner, A survey of offensive language detection for the Arabic language, TALLIP 20 (2021) 1–44.

[96] G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), CLEF 2021 Working Notes. Working Notes of CLEF 2021–Conference and Labs of the Evaluation Forum, 2021.

[97] L. Cappellato, N. Ferro, D. Losada, H. Müller (Eds.), Working Notes of CLEF 2019 Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2019.

[98] L. Cappellato, N. Ferro, J.-Y. Nie, L. Soulier (Eds.), Working Notes of CLEF 2018–Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2018.

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

[100] M. Müller, M. Salathé, P. E. Kummervold, Covid-twitter-bert: A natural language pro-

cessing model to analyse covid-19 content on twitter, arXiv preprint arXiv:2005.07503 (2020).

A. Summary of the Approaches

Here we give a brief description of the approaches explored by the different teams for all four subtasks.

Team AI Rational [15] (1A-bg:3 1B-bg:1 1C-bg:1 1D-bg:1 1A-nl:2 1B-nl:1 1C-nl:2 1D-nl:1 1A-en:1 1B-en:4 1B-en:2 1C-en:2 1D-en:3 1B-tr:2 1C-tr:3 1D-tr:1) experimented with different transformer models: DistilBERT, BERT, RoBERTa. It used DistilBERT to find the best parameters for the model and in the final the model they used RoBERTa large for English and XLM-RoBERTa large for Bulgarian, Dutch and Turkish. For data transformation, all links from tweets were replaced with "@link". The proposed system used data augmentation using back translation for English and Bulgarian. For the English tweets texts were translated to French then back translated to English and combined to the training dataset. The Bulgarian text are back translated with English, then combined with the training set.

Team ARC-NLP [16] (1C-en:6 1A-tr:3 1B-tr:3 1C-tr:1) proposed an approach called *ARC-NLP-contra*, in which the idea is that harmful tweets contradict with the real-life facts in the scope of COVID-19 pandemic. In addition, authors explore two other models. The first model, called *ARC-NLP-hc*, which utilizes hand-crafted tweet and user features. The second model, called *ARC-NLP-pretrain*, which pretrains a transformer-based model by using COVID-related Turkish tweets.

Team Asatya [17] (1A-en:10 1B-en:2 1C-en:3) proposes a methodology for the subtasks 1A, 1B and 1C, which includes data augmentation to increase the size of our training dataset, followed by preprocessing of the tweets and feature extraction for the tweet. Authors used a multimodal model, which uses numerical and categorical features in addition to the textual data from tweets. The multimodal network is combined with BERT for training the model.

Team COURAGE [24] (1C-en:8) proposed a deep learning model based on graph machine learning (i.e. Graph Attention Convolution) and a pretrained transformer-based model (i.e. ELECTRA). The representation of each tweet into a graph starts with text preprocessing (i.e., lowercasing, removing url, user mention and hashtag symbol), and Part Of Speech (POS) tagging. The POS-tagged tweet is then transformed into an undirected and attributed graph using a window equal to 3 to populate the adjacency matrix and using ELECTRA word embedding as nodes' attribute. The model reaches average performance on the English test set, but it improves F1-Score for positive class with respect to both the baseline and the simple ELECTRA embedding.

Team iCompass [19] (1A-ar:3) finetuned pre-trained language models such as AraBERT and ARBERT. The first model consisted of adding stacked gated recurrent units and one-dimensional convolutional neural networks to ARBERT and finding the optimal configuration, dropout rates,

and training strategy to classify the tweet as harmful or normal. The second model was composed of a gated recurrent network layer, a dense layer, and a dropout layer on the top of the pre-trained AraBERT (V1) model to predict whether a tweet is worth fact-checking.

Fraunhofer SIT [18] (**1A-en:5**) used an ensemble classification method that took advantage of state-of-the-art transformer networks and semi-supervised learning using GAN-BERT as well as data augmentation and data preprocessing. The ensemble classifier consisted of fine-tuned BERT-base-cased, BERTweet and RoBERTa-base[5] models that were trained using cross-validation on the training and validation split of the released dataset. Similarly, the BERT-base-cased and RoBERTa-base models were fine-tuned using GAN-BERT, while include additional unlabeled training data. Using a meta-classifier, the classification system was able to rank fifth best in the competition. Early experiments with quantum natural language processing (QNLP)[6] were used. However, the current state of the technique (i.e., QNLP) posed some problems and was therefore not included in the final model.

Team NUS-IDS [4] (**1A-ar:1 1A-bg:1 1A-nl:1 1A-en:8 1A-es:1**) describes the system CheckthaT5, which was designed in the context of the CheckThat! lab 2022 competition at CLEF. CheckthaT5 explores the feasibility of adapting sequence-to-sequence models for detecting check-worthy social media content in a multilingual texts (Arabic, Bulgarian, Dutch, English, Spanish and Turkish) provided in the competition. CheckThaT5 system takes all languages as input uniformly, thus enabling knowledge transfer from high-resource languages to low-resource languages. Empirically, CheckthaT5 outperforms strong baselines in all low-resource languages. In addition, the system incorporates tasks based on non-textual features that complement tweets and other related CheckThat! 2022 tasks through multitask learning further improving the average classification performance by 3%.

Team PoliMi-FlatEarthers [14] (**1A-ar:5 1A-bg:5 1A-nl:4 1A-en:3 1B-en:1 1C-en:10 1D-en:7 1A-es:2**) propose a system which is based on GPT-3, which outperforms previous language models on the task of finding relevant tweets. Though GPT-3 model is originally trained on English, however, it shows competitive performances on other languages as well.

Team RUB-DFL [20] (**1A-en:6 1B-en:6 1C-en:9 1A-tr:1 1B-tr:1 1C-tr:2**) used transformer-based pre-trained language models, as well as ELMo embeddings, which are combined with a range of linguistic features in attention networks. They tried to include some forms of pre-processing with LIWC features and URL resolution. In the end, the best results were achieved with fine-tuned transformer-based models for both English and Turkish in subtasks 1A, 1B, and 1C.

Team TOBB ETU [13] (**1A-ar:2 1B-ar:1 1C-ar:2 1D-ar:2 1A-bg:2 1B-bg:2 1C-bg:2 1D-bg:2 1A-nl:3 1B-nl:2 1C-nl:1 1D-nl:2 1A-en:4 1B-en:9 1C-en:5 1D-en:4 1A-tr:3 1B-tr:4 1C-tr:4 1D-tr:3**) participated in all subtasks for Arabic, Bulgarian, Dutch, English, and Turkish, yielding 20 submissions in total. They investigated fine-tuning transformer models pre-trained on various texts. In

---

[5]https://huggingface.co/cardiffnlp/twitter-roberta-base-2021-124m
[6]https://github.com/CQCL/lambeq

addition, they explored two data augmentation methods including i) machine translating labeled datasets in other languages to the corresponding language and ii) back-translation. Furthermore, as another approach, they represent tweets using BERT embeddings and train a feedforward neural network with Manifold Mixup regularization [5]. Based on their experiments on the test development dataset for each subtask and language, they submitted one of the three methods accordingly: i) a transformer model fine-tuned with the original data, ii) a transformer model fine-tuned with the dataset augmented by back-translation, iii) a feedforward neural network trained with Manifold Mixup regularization.

Team VTU BGM [21] (1A-en:11 1B-en:7 1C-en:12) used autoregressive model XLNet for feature extraction and SVM for the classification of tweets.

Team Z-Index [22] (1A-nl:5 1A-en:12 1A-es:3) preprocessed the claims data by removing username, URLs, non-ASCII characters, and stopwords. They experimented with both deep learning and traditional approaches. For deep learning approach, transformer based models BERT multilingual and XLM-RoBERTa base were used for training. For Dutch and English they obtained better results using transformer based model and for Spanish they obtained better results using SVM and Random Forest.

Team Zorros [23] (1A-en:2) used fine-tuned and pre-trained transformer-based models like BERT and RoBERTa. They built ensemble models which combine the fine-tuned transformer models. The preprocessing step in their system include removing URLs, hashtags, numbers and other symbols. For subtask 1A, check-worthiness of tweets, authors used an ensemble of ten transformer-based models, pre-trained on tweets about COVID-19. A classification header and a dropout layer is used to avoid over-fitting. For subtask 1B (verifiable factual claims detection) and 1C (Harmful tweet detection) they used an ensemble of five transformer-base models. The ensemble models obtained a better performance than simple fine-tuned transformer models. For subtask 1D (attention-worthy tweet detection) they used COVID Twitter BERT v2[100].