

MBZUAI

Digital.Commons@MBZUAI

---

Natural Language Processing Faculty  
Publications

Scholarly Works

---

9-2023

## Overview of the CLEF-2023 CheckThat! Lab Task 1 on Check-Worthiness of Multimodal and Multigenre Content

Firoj Alam

*Qatar Computing Research Institute*

Alberto Barrón-Cedeño

*Alma Mater Studiorum Università di Bologna*

Gullal S. Cheema

*Forschungszentrum L3S*

Gautam Kishore Shahi

*Universität Duisburg-Essen*

Sherzod Hakimov

*Universität Potsdam*

Follow this and additional works at: <https://dclibrary.mbzuai.ac.ae/nlpfp>



Part of the [Artificial Intelligence and Robotics Commons](#)

Archived from the original on 2024-04-03.

Open Access

License: CC by 4.0

Uploaded: April 03, 2024

---

### Recommended Citation

F. Alam and A. Barrón-Cedeño and G. Cheema and G. Shahi and S. Hakimov and M. Hasanain and C. Li and R. Míguez and H. Mubarak and W. Zaghouani and P. Nakov, "Overview of the CLEF-2023 CheckThat! Lab Task 1 on Check-Worthiness of Multimodal and Multigenre Content," *CEUR Workshop Proceedings*, vol. 3497, pp. 219 - 235, Sep 2023.

This Conference Proceeding is brought to you for free and open access by the Scholarly Works at Digital.Commons@MBZUAI. It has been accepted for inclusion in Natural Language Processing Faculty Publications by an authorized administrator of Digital.Commons@MBZUAI. For more information, please contact [libraryservices@mbzuai.ac.ae](mailto:libraryservices@mbzuai.ac.ae).

---

## Authors

Firoj Alam, Alberto Barrón-Cedeño, Gullal S. Cheema, Gautam Kishore Shahi, Sherzod Hakimov, Maram Hasanain, Chengkai Li, Rubén Míguez, Hamdy Mubarak, Wajdi Zaghoulani, and Preslav Nakov

# Overview of the CLEF-2023 CheckThat! Lab Task 1 on Check-Worthiness of Multimodal and Multigenre Content

Firoj Alam<sup>1,\*</sup>, Alberto Barrón-Cedeño<sup>2</sup>, Gullal S. Cheema<sup>3</sup>, Gautam Kishore Shahi<sup>4</sup>, Sherzod Hakimov<sup>5</sup>, Maram Hasanain<sup>1</sup>, Chengkai Li<sup>6</sup>, Rubén Míguez<sup>7</sup>, Hamdy Mubarak<sup>1</sup>, Wajdi Zaghrouani<sup>8</sup> and Preslav Nakov<sup>9</sup>

<sup>1</sup>*Qatar Computing Research Institute, HBKU, Qatar*

<sup>2</sup>*DIT, Università di Bologna, Italy*

<sup>3</sup>*L3S Research Center, Leibniz University of Hannover, Germany*

<sup>4</sup>*University of Duisburg-Essen, Germany*

<sup>5</sup>*University of Potsdam, Germany*

<sup>6</sup>*University of Texas at Arlington, USA*

<sup>7</sup>*Newtral Media Audiovisual, Spain*

<sup>8</sup>*Hamad Bin Khalifa University, Qatar*

<sup>9</sup>*Mohamed bin Zayed University of Artificial Intelligence, UAE*

## Abstract

We present an overview of CheckThat! Lab's 2023 Task 1, which is part of CLEF-2023. Task 1 asks to determine whether a text item, or a text coupled with an image, is check-worthy. This task places a special emphasis on COVID-19, political debates and transcriptions, and it is conducted in three languages: Arabic, English, and Spanish. A total of 15 teams participated, and most submissions managed to achieve significant improvements over the baselines using Transformer-based models. Out of these, seven teams participated in the multimodal subtask (1A), and 12 teams participated in the Multigenre subtask (1B), collectively submitting 155 official runs for both subtasks. Across both subtasks, approaches that targeted multiple languages, either individually or in conjunction, generally achieved the best performance. We provide a description of the dataset and the task setup, including the evaluation settings, and we briefly overview the participating systems. As is customary in the CheckThat! lab, we have released all datasets from the lab as well as the evaluation scripts to the research community. This will enable further research on finding relevant check-worthy content that can assist various stakeholders such as fact-checkers, journalists, and policymakers.

## Keywords

Check-worthiness, fact-checking, multilinguality, multimodality.

---

*CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece*

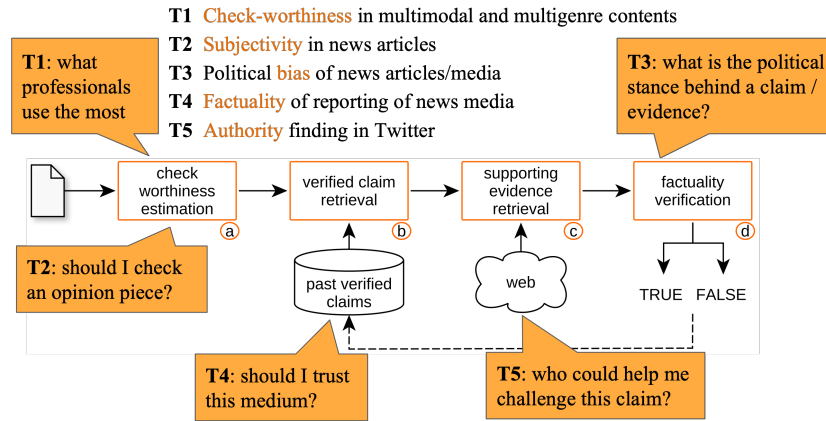
\*Corresponding author.

✉ fialam@hbku.edu.qa (F. Alam); a.barron@uniibo.it (A. Barrón-Cedeño); Gullal.Cheema@tib.eu (G. S. Cheema); gautam.shahi@uni-due.de (G. K. Shahi); sherzod.hakimov@uni-potsdam.de (S. Hakimov); mhasanain@hbku.edu.qa (M. Hasanain); cli@uta.edu (C. Li); ruben.miguez@newtral.es (R. Míguez); hmubarak@hbku.edu.qa (H. Mubarak); wzaghrouani@hbku.edu.qa (W. Zaghrouani); preslav.nakov@mbzuai.ac.ae (P. Nakov)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)



**Figure 1:** The CheckThat! lab verification pipeline. The 2023 edition of the lab covers five tasks: (T1) check-worthiness estimation in multimodal and multigenre content (the focus of this paper), (T2) subjectivity, (T3) political bias of the news articles/media, (T4) factuality of the news media, and (T5) authority finding in Twitter.

## 1. Introduction

Fact-checking of multimodal and multigenre content is crucial to ensure the accuracy and the reliability of information shared on different communication channels such as news, political debates, and social media platforms. It helps to prevent the spread of misinformation and to promote informed decision-making. By verifying the claims in such content, individuals can make well-informed judgments and contribute to a more accurate and trustworthy online discourse.

The CheckThat! 2023 lab was held in the framework of CLEF 2023 [1, 2, 3].<sup>1</sup> Figure 1 shows the full CheckThat! identification and verification pipeline, highlighting the five tasks targeted in this fifth edition of the lab: Task 1 on detecting check-worthiness (this paper), Task 2 on subjectivity in sentences [4], and Task 3 on detecting political bias of the news articles/media [5], Task 4 on detecting factuality of news media [6], and Task 5 on authority finding in twitter [7].

Task 1 ask for the detection of check-worthiness in multimodal and multigenre content. We provided manually annotated data for two subtasks in three languages: Arabic, English, and Spanish. Among the different subtasks, the check-worthiness on multigenre content was popular, with 12 teams participating. English was the most popular target language for the participants. Across the different submitted systems, transformer-based models were widely used, with XLM-RoBERTa being the most popular among the models. The top-ranked systems also employed data augmentation and additional preprocessing 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.

<sup>1</sup><https://checkthat.gitlab.io/>



**Figure 2:** Examples of tweets with their check-worthiness labels.

## 2. Task

The goal of this task is to assess whether a given statement, in a tweet or from a political debate, is worth fact-checking [3]. In order to make that decision, one would need to ponder about questions, such as “does it contain a verifiable factual claim?” or “is it harmful?”, before deciding on the final check-worthiness label [8]. Task 1 is divided into two subtasks. Subtask 1A is offered in Arabic and English, Subtask 1B is offered in Arabic, English, and Spanish.

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

**Subtask 1B: Multigenre** The task requires the assessment of a text snippet for check-worthiness. This snippet could be a standalone segment extracted from a variety of sources such as a tweet, a political debate, or a speech. The objective is to evaluate whether the information contained within the snippet is reliable and worthy of further fact-checking, contributing to the credibility and the integrity of the information ecosystem.

## 3. Datasets

### 3.1. Subtask 1A: Multimodality

In subtask 1A for English, we followed the annotation schema reported in [9]. The dataset used for the challenge was derived from [9], with the existing data repurposed for training and development purposes, and new data developed for the evaluation. The dataset focused on three topics: COVID-19, climate change, and technology. Each tweet was labeled using both the image and the text, with OCR performed using the Google Vision API to extract the text from the images. We provided 3,175 annotated examples and around 110k unlabeled tweets of text-image pairs and OCR output to all participants.

Two annotators, one expert and one new, annotated the new test set. The new annotator went through a dry run of 50 examples, where disagreements were discussed and resolved. For the final test set of 736 examples, the Cohen’s Kappa inter-annotator agreement [10] was 0.49 for the check-worthiness label, indicating a moderate agreement. The expert annotator resolved any remaining disagreements for a higher quality test set.

For Subtask 1A Arabic, our data curation involved several steps for the training, the development, the dev-test, and the test datasets. For the first three partitions, we used the CT-CWT-21 [11] and CT-CWT-22 [12] datasets, both of which had been annotated for check-worthiness and focused on topics related to COVID-19 and politics. These datasets followed the annotation schema described in [13, 8]. In order to develop multimodal datasets from these resources, we crawled images linked to the tweets. Since a tweet can be associated with multiple images, we only selected the first image for our study. The labels for multimodality in the first three partitions were derived from the textual modality, so these annotations can be considered as weakly labeled. For the test set, we collected tweets using similar keywords to those reported in [13, 8]. For the annotation of the test set, we followed the same annotation schema. Our three annotators had prior experiences in annotating datasets for similar tasks. We used majority voting (and sometimes discussion) to select the final labels in case of disagreements.

### 3.2. Subtask 1B: Multigenre

The dataset for Subtask 1B consists of tweets in Arabic, English, and Spanish, as well as statements from English political debates. The Arabic tweets for Subtask 1B were collected using keywords related to COVID-19 and vaccines, using the annotation schema described in Alam et al. [8]. The training, the development, and the dev-test partitions of the dataset were obtained from CT-CWT-21 [11] and CT-CWT-22 [12]. For the test dataset, we used the same approach as discussed in the previous section 3.1.

The English dataset comprises sentences made by presidential election candidates during the US general election debates, annotated by human annotators [14]. While the first three partitions primarily use the same dataset described in Arslan et al. [14], there have been some updates made to improve the quality of the annotations. The test set includes sentences that were not featured in [14].

The Spanish dataset, a combination of CT-CWT-21 [11], CT-CWT-22 [12], as well as newly collected content, consists of tweets from Twitter accounts and transcriptions from Spanish politicians. These were annotated by professional journalists with expertise in fact-checking.

Statistics about the datasets for Task 1 are given in Table 1. Across the different subtasks, dataset sizes range from 3,911 to 29,984, which are the largest so far across different languages over the years for the check-worthiness task. Figure 2 shows examples of checkworthy and non-checkworthy tweets.

## 4. Evaluation Settings

For the lab, we provided a training, a development, and a dev-test dataset. 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.

**Table 1**

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

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

For each language and subtask, we annotated new instances, using three annotators per instance. The final label was assigned using majority voting and disagreements were resolved by a consolidator or by discussion among the annotators. 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.

For evaluation, we used the  $F_1$ -measure with respect to the positive class (yes) to account for class imbalance. The data and the evaluation scripts are available online.<sup>2</sup> The submission system was hosted on the CodaLab platform.<sup>3</sup>

## 5. Results and Overview of the Systems

Fourteen teams participated in this task and submitted 35 final runs, with English being the most popular language. Below, we provide a summary of the systems submitted by the participants.

### 5.1. Subtask-1A

A total of 7 and 4 teams submitted their runs for English and for Arabic, respectively, out of which four made submissions for both languages. Table 2 gives an overview of the submitted systems, and Table 3 shows the performance of the official submissions on the test set. We also provide results for a random baseline.

<sup>2</sup>[https://gitlab.com/checkthat\\_lab/clef2023-checkthat-lab/-/tree/main/task1](https://gitlab.com/checkthat_lab/clef2023-checkthat-lab/-/tree/main/task1)

<sup>3</sup><https://codalab.lisn.upsaclay.fr/competitions/12936>

**Table 2**

Overview of the approaches for **Subtask 1A**. The numbers in the language box show the position of the team in the official ranking.

Team		Lang	Transformers								CNN		Repr.			Misc		
			Arabic English	BERT	RoBERTa	mBERT	AraBERT	BERTweet	GPT-3	Electra	ViT	ConvNext	ResNet	Text	Image	OCR Output	LSTM	CatBoost
CSECU-DSG	[15]	1 4																
ES-VRAI	[16]	-																
FraunhoferSIT	[17]	1																
Mtop*		2 3																
TeamX	[18]	- -																
ZHAW-CAI	[19]	2																
Z-Index	[20]	3 5																

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

**Team Fraunhofer SIT [17]** tackled the problem by fine-tuning individual text classifiers on the tweet text and on the OCR text, respectively. They further used pre-processing for the tweet text and extracted the text from images using *easyOCR*.<sup>4</sup> Two BERT [21] models were fine-tuned on each input, and the final label for each example on the test set was a re-weighted combination of the two predictions based on the validation loss.

**Team ZHAW-CAI [19]** submitted official runs for the English track only. They trained different unimodal and multimodal systems and then combined them using a kernel-based ensemble. This ensemble was trained using an SVM for classification. For the text-based model,  $n$ -gram features were extracted separately from the tweet text, and from the prompt response from GPT-3 (Open AI’s text-davinci-003), and SVMs were trained on these features. In addition, an Electra [22] model was fine-tuned on the tweet text for classification. For the multimodal model, features from Twitter-based RoBERTa [23] and ViT [24] were extracted, fused via pooling, and passed through a dense layer for classification. The submission model is an ensemble of the four features described earlier with their individual kernels and combined with an average kernel to be used in an SVM for classification.

**Team ES-VRAI [16]** comprehensively evaluated several pre-trained vision and text models, different classifiers, and several early and late fusion strategies to select the best model for the English data. Their submitted model combined BERT and ResNet50 [25] features in an early fusion mode.

**Team CSECU-DSG [15]** participated in both the Arabic and the English tracks. They jointly fine-tuned two transformers. A language-specific BERT was used to represent the tweet text, and ConvNext [26] was used for image feature extraction. They used BERTweet [27] for English data, and AraBERT [28] for Arabic. In addition, a BiLSTM was used on top of the text features to handle long-term contextual dependency. Finally, the features from BiLSTM and ConvNext were concatenated and followed up by a multi-sample dropout [29] to predict the final label.

<sup>4</sup><https://github.com/JaiedAI/EasyOCR>



**Table 3**

**Subtask 1A:** Multimodal check-worthiness estimation results. The F1 score is computed with respect to the positive class.

Team	F1	Team	F1
Arabic		English	
1 CSECU-DSG [15]	0.399	1 Fraunhofer SIT [17]	0.712
2 Mtop*	0.312	2 ZHAW-CAI [19]	0.708
3 Z-Index [20]	0.301	- ES-VRAI [16]	0.704
- TeamX [18]	0.300	3 Mtop*	0.697
4 Baseline	0.299	- TeamX [18]	0.671
		4 CSECU-DSG [15]	0.628
		5 Z-Index [20]	0.495
		6 Baseline	0.474

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

**TeamX [18]** also participated in both languages. The proposed architecture uses Vision Transformer (ViT) [24] for image feature encoding and multilingual BERT (mBERT) [21] for the textual representation of English, and AraBERT [28] for Arabic. Finally, BLOCK fusion was used to combine both modalities. For the textual representation, since the OCR text was available in the English dataset, the model was trained by merging the tweet text and the OCR text, while only the tweet text was used with the Arabic dataset.

**Team Z-Index [20]** also participated in both languages. They used BERT for the English tweet text and ResNet50 for images, and a feed-forward neural network for fusion and classification. They further used mBERT [21] for the Arabic text. The backbone networks were fine-tuned along with the feed-forward network to train the model for the task. In their internal evaluation, they also experimented with XLM-RoBERTa [30], which performed better by 4% than the BERT variant for both languages.

To summarize: one common theme was the use of large pre-trained models and their features for semantic information extraction. Three of the teams further used OCR. All teams but one included both the text and the image modality into the system architecture design.

## 5.2. Subtask-1B

A total of 11, 6, and 7 teams submitted their runs for English, Arabic, and Spanish, respectively, out of which 6 teams submitted runs for all languages. Table 4 gives an overview of the submitted systems per language, and Table 5 shows the performance of the official submissions on the test set, in addition to the performance of a random baseline.

**Team OpenFact [35]** was the best-performing team on English. They fine-tuned GPT-3<sup>5</sup> using 7.7K examples of sentences from debates and speeches annotated for check-worthiness, extracted from a pre-existing dataset [14]. Moreover, during internal experiments, they also experimented with fine-tuning a variety of BERT models and found that fine-tuning DeBERTaV3 [36] yielded near-identical performance to GPT-3.

<sup>5</sup><https://platform.openai.com/docs/models/gpt-3>

**Table 4**

Overview of the systems for **Subtask 1B**. The numbers in the language box refer to the position of the team in the official ranking.

Team		Langs			Models										Misc		
		English	Arabic	Spanish	BERT	RoBERTa	XLNet-RoBERTa	MARBERT	AraBERT	BERTweet	BETO	Spanish RoBERTa	GPT	MultinomialNB	LSTM	Data augmentation	Preprocessing
Accenture	[31]	3	2	5	⬅	⬅											⬅
CSECU-DSG	[15]	7	4	3			⬅		⬅	⬅	⬅				⬅		
DSHacker	[32]	9	5	1			⬅						⬅			⬅	
ES-VRAI	[16]	5	1	2			⬅	⬅		⬅							
FakeDTML	[33]	8	-	-			⬅							⬅			⬅
Fraunhofer SIT	[17]	2			⬅	⬅											⬅
NLPIR-UNED	[34]	4	4		⬅							⬅					
OpenFact	[35]	1											⬅				
Z-Index	[20]	6	3	6			⬅										⬅

- Run submitted after the deadline.

**Team Fraunhofer SIT** [17] fine-tuned BERT [21] three times starting with a different seed for model initialization, resulting in three models, which they combined in an ensemble using a model souping technique that adaptively adjusts the influence of each individual model based on its performance on the dev set.

**Team Accenture** [31] also fine-tuned large pre-trained models: RoBERTa [23] for English and GigaBERT for Arabic [37]. They further proposed to extend the training subset with examples resulting from back-translating the same set using AWS translation.<sup>6</sup>

**Team ES-VRAI** [16] achieved the best and the second best performance for Arabic and for Spanish, respectively. After comprehensive evaluation of several language-specific pre-trained models, their official submission for Arabic was based on fine-tuning MARBERT [38] using the training set, after downsampling examples from the majority class. A fine-tuned XLNet-RoBERTa model was used to produce the official submitted run for the Spanish test set.

**Team Z-Index** [20] participated in all three languages using the same system architecture. Their system includes a feed forward network, where the input is represented using embeddings.<sup>7</sup> The network was trained using the training set released per language.

**Team NLPIR-UNED** [34] proposed to include the context of debates/speeches for the English dataset. Their system combined three BERT models fine-tuned independently. A transformer operates on one of the following: the instance to be classified, the sentence before it or the one that follows it. The fine-tuned models are then followed by a feed-forward network (FFN) that

<sup>6</sup><https://aws.amazon.com/translate/>

<sup>7</sup>No enough details were available about the source of these embeddings.

**Table 5**

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

Team	F1	Team	F1	Team	F1
English		Arabic		Spanish	
1 OpenFact	0.898	1 ES-VRAI	0.809	1 DSHacker	0.641
2 Fraunhofer SIT	0.878	2 Accenture	0.733	2 ES-VRAI	0.627
3 Accenture	0.860	3 Z-Index	0.710	3 CSECU-DSG	0.599
4 NLPIR-UNED	0.851	4 CSECU-DSG	0.662	4 NLPIR-UNED	0.589
5 ES-VRAI	0.843	5 DSHacker	0.633	5 Accenture	0.509
6 Z-Index	0.838	6 Baseline	0.625	6 Z-Index	0.496
7 CSECU-DSG	0.834	- FakeDTML	0.530	- FakeDTML	0.440
8 FakeDTML	0.833			7 Baseline	0.172
9 DSHacker	0.819				
10 Pikachu	0.767				
- UGPLN y SINAI	0.757				
11 Baseline	0.462				

concatenates the outputs from all three transformer models and is trained on the same training set. For the Spanish tweet dataset, a similar architecture is followed using an ensemble of three classifiers: a Spanish RoBERTa [39] fine-tuned on the training dataset, a feed-forward network classifier trained using the tweets represented as TF.IDF vectors, and a second FFN classifier that has as inputs the discrete features generated by the LIWC text analysis tool [40].

**Team DSHacker** [32] achieved the best overall performance for Spanish. Their system is based on fine-tuning XLM-RoBERTa [?] using the available training data, and additional datasets obtained by data augmentation. For data augmentation, they used GPT-3.5<sup>8</sup> to translate the training set to English and to Arabic resulting in two additional training subsets. GPT-3.5 was also used to paraphrase the original Spanish training data, resulting in a third augmented training subset.

**Team CSECU-DSG** [15] also participated in all three languages. Their model includes jointly fine-tuning two transformers: a language-specific BERT and Twitter XLM-RoBERTa [30] to represent the input text. In addition, a BiLSTM module was used on top of the text features to handle long-term contextual dependencies. Finally, the features from the BiLSTM were followed by a multisample dropout strategy [29] to produce the final prediction.

**Team FakeDTML** [33] submitted runs for all three languages. For the English data, the team opted to fine-tuning XLM-RoBERTa for the task. As for the Twitter datasets, a multinomial Naïve Bayes model was used, using  $n$ -grams to represent the input.

In all participating systems, we again observe the popularity of fine-tuning pre-trained models, with XLM-RoBERTa being the most-commonly used model. We also observe GPT-3 being used by at least two teams, once fine-tuned for classification, and a second time as a tool for data augmentation.

<sup>8</sup><https://platform.openai.com/docs/models/gpt-3-5>

## 6. Related Work

There has been a considerable surge in research interest in identifying disinformation, misinformation, and fake news in recent years. These phenomena flourish on social media and within political debates and speeches. Numerous recent studies have shed light on various aspects of this problem. These include understanding the ways information is shared and received on social media platforms [41, 42], exploring fact-checking perspectives on fake news and associated issues [43], investigating truth discovery [44], examining attitudes towards the detection of misinformation and disinformation [45], automating fact-checking to support human fact-checkers [46, 47, 48], predicting the factuality and the bias of entire news outlets [49, 50], detecting disinformation across multiple modalities [51], and focusing on the use of abusive language on social media [52].

Within the wider context of identifying disinformation, misinformation, and fake news, research interest has focused on more specific issues. These include the automatic identification and verification of claims [53, 54, 55, 56, 57, 58, 59], recognizing check-worthy claims [60, 61, 62], and assessing whether a claim has been previously fact-checked [63, 64, 65, 66]. Additionally, there has been research into evidence retrieval for substantiating or refuting a claim [67], and the evaluation of whether this evidence supports or denies the claim [68]. Finally, efforts have been made to infer the veracity of a given claim [69, 70]. Such specific tasks can prove highly beneficial to fact-checkers and journalists.

Since the pioneering work of Hassan et al. [71], the task of check-worthiness estimation has garnered wider attention. The aim is to determine whether a sentence from a political debate is non-factual, unimportantly factual, or check-worthy factual. Follow-up work added more data and covered Arabic content [72]. Initially, most of the work on check-worthiness estimation was primarily concentrated on political debates [61]. However, recently, the focus has shifted towards social media [8, 13, 73, 74].

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 the 2018 edition of the task, seven teams submitted runs for Task 1. Their systems were primarily based on word embeddings and Recurrent Neural Networks (RNNs) [75].

In the 2019 edition of the task, eleven teams submitted runs for the corresponding Task 1. They continued to use word embeddings and RNNs, while also experimenting with various new representations [54].

In the 2020 edition, three teams submitted runs for the corresponding Task 5 with systems based on word embeddings and BiLSTM, TF.IDF representation with Naïve Bayes, logistic regression, decision trees, BERT prediction scores, and word embeddings with logistic regression [76].

In the 2021 edition of the task [11], fifteen teams submitted entries for the check-worthiness estimation task. The top-ranked systems used transformers such as BERT and RoBERTa [77, 11].

In the 2022 edition of the task [12, 78], nineteen teams participated. Most submissions successfully achieved considerable improvements over the baselines by using transformers such as BERT and GPT-3.

This year, for the first time, the task was offered in multiple modalities, incorporating both the tweet text and images; it was offered in both English and Arabic.

## 7. Conclusion and Future Work

We presented an overview of task 1 of the CLEF-2023 CheckThat! lab. The lab featured tasks that span the full verification pipeline: from spotting check-worthy claims to checking whether a claim has been fact-checked before. Task 1 asked to identify check-worthiness in multimodal and multigenre content. For the multimodality, notable systems used fusion of the text and the image modalities (BERT and ViT-based Vision Transformer). For the multigenre text classification, the majority of the systems fine-tuned pre-trained models, with XLM-RoBERTa being most popular. The top-performing system was based on GPT-3. In general, the current iteration of the task has encompassed a variety of strategies, involving different models such as various types of transformers.

In future work, we plan to expand the task in a variety of ways, e.g., by enlarging the dataset and by incorporating more languages.

## Acknowledgments

The work of F. Alam, M. Hasanain and W. Zaghouni is partially supported by NPRP 13S-0206-200281 and NPRP 14C-0916-210015 from the Qatar National Research Fund (a member of Qatar Foundation). The findings achieved herein are solely the responsibility of the authors.

## References

- [1] A. Barrón-Cedeño, F. Alam, A. Galassi, G. Da San Martino, P. Nakov, T. Elsayed, D. Azizov, T. Caselli, G. Cheema, F. Haouari, M. Hasanain, M. Kutlu, C. Li, F. Ruggeri, J. M. Struß, W. Zaghouni, 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.
- [2] A. Barrón-Cedeño, F. Alam, A. Galassi, G. Da San Martino, P. Nakov, T. Elsayed, D. Azizov, T. Caselli, G. Cheema, F. Haouari, M. Hasanain, M. Kutlu, C. Li, F. Ruggeri, J. M. Struß, W. Zaghouni, Overview of the CLEF-2023 CheckThat! Lab checkworthiness, subjectivity, political bias, factuality, and authority of news articles and their source, in: A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, A. Giachanou, D. Li, M. Aliannejadi, M. Vlachos, G. Faggioli, N. Ferro (Eds.), *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2023)*, 2023.
- [3] F. Alam, A. Barrón-Cedeño, G. S. Cheema, S. Hakimov, M. Hasanain, C. Li, R. Míguez, H. Mubarak, G. K. Shahi, W. Zaghouni, P. Nakov, Overview of the CLEF-2023 CheckThat! lab task 1 on check-worthiness in multimodal and multigenre content, in: [79], 2023.
- [4] A. Galassi, F. Ruggeri, A. Barrón-Cedeño, F. Alam, T. Caselli, M. Kutlu, J. Struss, F. Antici, M. Hasanain, J. Köhler, K. Korre, F. Leistra, A. Muti, M. Siegel, T. Mehmet Deniz, M. Wiegand, W. Zaghouni, Overview of the CLEF-2023 CheckThat! lab task 2 on subjectivity in news articles, in: [79], 2023.

- [5] G. Da San Martino, F. Alam, M. Hasanain, R. N. Nandi, D. Azizov, P. Nakov, Overview of the CLEF-2023 CheckThat! lab task 3 on political bias of news articles and news media, in: [79], 2023.
- [6] P. Nakov, F. Alam, G. Da San Martino, M. Hasanain, R. N. Nandi, D. Azizov, P. Panayotov, Overview of the CLEF-2023 CheckThat! lab task 4 on factuality of reporting of news media, in: [79], 2023.
- [7] F. Haouari, Z. Sheikh Ali, T. Elsayed, Overview of the CLEF-2023 CheckThat! lab task 5 on authority finding in twitter, in: [79], 2023.
- [8] F. Alam, S. Shaar, F. Dalvi, H. Sajjad, A. Nikolov, H. Mubarak, G. Da San 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.
- [9] G. S. Cheema, S. Hakimov, A. Sittar, E. Müller-Budack, C. Otto, R. Ewerth, MM-Claims: A dataset for multimodal claim detection in social media, in: Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, Washington, 2022, pp. 962–979.
- [10] J. Cohen, A coefficient of agreement for nominal scales, Educational and psychological measurement 20 (1960) 37–46.
- [11] 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.
- [12] 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: Working Notes of CLEF 2022—Conference and Labs of the Evaluation Forum, CLEF ’2022, Bologna, Italy, 2022.
- [13] 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 Conference on Web and Social Media, ICWSM ’21, 2021, pp. 913–922.
- [14] 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.
- [15] A. Aziz, M. Hossain, A. Chy, CSECU-DSG at CheckThat! 2023: Transformer-based fusion approach for multimodal and multigenre check-worthiness, in: [79], 2023.
- [16] H. T. Sadouk, F. Sebbak, H. E. Zekiri, Es-vrai at CheckThat! 2023: Analyzing checkworthiness in multimodal and multigenre contents through fusion and sampling approaches, in: [79], 2023.
- [17] R. A. Frick, I. Vogel, J.-E. Choi, Fraunhofer SIT at CheckThat! 2023: Enhancing the detection of multimodal and multigenre check-worthiness using optical character recognition and model souping, in: [79], 2023.
- [18] R. N. Nandi, F. Alam, TeamX at CheckThat! 2023: Multilingual and multimodal approach for check-worthiness detection, in: [79], 2023.



- [19] P. von Däniken, J. Deriu, M. Cieliebak, Zhaw-cai at CheckThat! 2023: Ensembling using kernel averaging, in: [79], 2023.
- [20] P. Tarannum, M. A. Hasan, F. Alam, S. R. H. Noori, Z-Index at CheckThat! 2023: Unimodal and multimodal checkworthiness classification, in: [79], 2023.
- [21] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT '19, Minneapolis, MN, USA, 2019, pp. 4171–4186.
- [22] K. Clark, M. Luong, Q. V. Le, C. D. Manning, ELECTRA: pre-training text encoders as discriminators rather than generators, in: Proceedings of the 8th International Conference on Learning Representations, ICLR '20, 2020.
- [23] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A robustly optimized BERT pretraining approach, CoRR abs/1907.11692 (2019). [arXiv:1907.11692](#).
- [24] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations, 2020.
- [25] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR '16, Las Vegas, Nevada, USA, 2016, pp. 770–778.
- [26] Z. Liu, H. Mao, C. Wu, C. Feichtenhofer, T. Darrell, S. Xie, A ConvNet for the 2020s, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR '22, New Orleans, LA, USA, 2022, pp. 11966–11976.
- [27] D. Q. Nguyen, T. Vu, A. T. Nguyen, BERTweet: A pre-trained language model for English tweets, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2020, pp. 9–14.
- [28] W. Antoun, F. Baly, H. Hajj, AraBERT: Transformer-based model for Arabic language understanding, in: Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, OSAC '20, Marseille, France, 2020, pp. 9–15.
- [29] H. Inoue, Multi-sample dropout for accelerated training and better generalization, CoRR abs/1905.09788 (2019). [arXiv:1905.09788](#).
- [30] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, É. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 8440–8451.
- [31] S. Tran, P. Rodrigues, B. Strauss, E. Williams, Accenture at CheckThat! 2023: Identifying claims with societal impact using nlp data augmentation, in: [79], 2023.
- [32] A. Modzelewski, W. Sosnowski, A. Wierzbicki, DSHacker at CheckThat! 2023: Check-Worthiness in Multigenre and Multilingual Content With GPT-3.5 Data Augmentation, in: [79], 2023.
- [33] A. A. M. Sardar, M. Z. Karim, K. Dey, M. A. Hasan, Fakedtml at CheckThat! 2023: Identifying check-worthiness of tweets and debate snippets, in: [79], 2023.
- [34] J. Martinez-Rico, L. Araujo, J. Martinez-Romo, NLPiR-UNED at CheckThat! 2023: Ensemble

- of Classifiers for Check-Worthiness Estimation, in: [79], 2023.
- [35] M. Sawiński, K. Wecel, E. Ksieźniak, 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: [79], 2023.
  - [36] P. He, J. Gao, W. Chen, DeBERTaV3: improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing, in: Proceedings of the Eleventh International Conference on Learning Representations, 2023.
  - [37] W. Lan, Y. Chen, W. Xu, A. Ritter, An empirical study of pre-trained transformers for arabic information extraction, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 4727–4734.
  - [38] M. Abdul-Mageed, A. Elmadany, et al., ARBERT & MARBERT: Deep bidirectional transformers for Arabic, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 7088–7105.
  - [39] J. D. la Rosa y Eduardo G. Ponferrada y Manu Romero y Paulo Villegas y Pablo González de Prado Salas y María Grandury, BERTIN: Efficient pre-training of a Spanish language model using perplexity sampling, *Procesamiento del Lenguaje Natural* 68 (2022) 13–23.
  - [40] J. W. Pennebaker, R. L. Boyd, K. Jordan, K. Blackburn, The development and psychometric properties of LIWC2015, Technical Report, 2015.
  - [41] K. Shu, A. Sliva, S. Wang, J. Tang, H. Liu, Fake news detection on social media: A data mining perspective, *ACM SIGKDD explorations newsletter* 19 (2017) 22–36.
  - [42] G. K. Shahi, T. A. Majchrzak, Amused: an annotation framework of multimodal social media data, in: International Conference on Intelligent Technologies and Applications, Springer, 2021, pp. 287–299.
  - [43] 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.
  - [44] 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.
  - [45] 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.
  - [46] 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.
  - [47] S. Shaar, N. Georgiev, F. Alam, G. Da San Martino, A. Mohamed, P. Nakov, Assisting the human fact-checkers: Detecting all previously fact-checked claims in a document, in: Findings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '22, Abu Dhabi, UAE, 2022, pp. 2069–2080.
  - [48] G. K. Shahi, Fakekg: A knowledge graph of fake claims for improving automated fact-checking (student abstract), in: Proceedings of the AAAI Conference on Artificial Intelli-



gence, volume 37, 2023, pp. 16320–16321.

- [49] 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).
- [50] P. Panayotov, U. Shukla, H. T. Sencar, M. Nabeel, P. Nakov, GREENER: Graph neural networks for news media profiling, in: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP '22*, Abu Dhabi, UAE, 2022, pp. 7470–7480.
- [51] 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.
- [52] 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).
- [53] P. Atanasova, L. Màrquez, A. Barrón-Cedeño, T. Elsayed, R. Suwaileh, W. Zaghoulani, 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: L. Cappellato, N. Ferro, J.-Y. Nie, L. Soulier (Eds.), *CLEF 2018 Working Notes. Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS.org, Avignon, France, 2018*.
- [54] 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: [80], 2019.
- [55] A. Barrón-Cedeño, T. Elsayed, R. Suwaileh, L. Marquez, P. Atanasova, W. Zaghoulani, 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: [81], 2018.
- [56] 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.
- [57] 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.
- [58] 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: [80], 2019.
- [59] P. Nakov, A. Barrón-Cedeño, T. Elsayed, R. Suwaileh, L. Màrquez, W. Zaghoulani, 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.

- [60] 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'17, 2017, pp. 267–276.
- [61] 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.
- [62] 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).
- [63] 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.
- [64] 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, in: Proceedings of the 30th International Joint Conference on Artificial Intelligence, 2021, pp. 4551–4558.
- [65] S. Shaar, F. Alam, G. Da San Martino, P. Nakov, The role of context in detecting previously fact-checked claims, in: Findings of the Association for Computational Linguistics: NAACL 2022, 2022, pp. 1619–1631.
- [66] G. K. Shahi, D. Nandini, FakeCovid—a multilingual cross-domain fact check news dataset for COVID-19, in: Workshop Proceedings of the 14th International AAAI Conference on Web and Social Media, ICWSM 2020, Atlanta, Georgia, USA, 2020.
- [67] 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.
- [68] 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, Association for Computational Linguistics, Hong Kong, China, 2019, pp. 4442–4452.
- [69] 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), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 21–27.
- [70] 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), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 809–819.
- [71] 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.
- [72] 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, New Orleans, Louisiana, 2018, pp. 26–30.
  - [73] 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, NLP4IF '21, 2021, pp. 82–92.
  - [74] G. K. Shahi, W. Kana Tsoplefack, Mitigating harmful content on social media using an interactive user interface, in: International Conference on Social Informatics, Springer, 2022, pp. 490–505.
  - [75] 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, in: [81], 2018.
  - [76] 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. Névél, 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.
  - [77] 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), Springer, 2021.
  - [78] 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: Proc. of the Conference of the CLEF Association: Information Access Evaluation meets Multilinguality, Multimodality, and Visualization, CLEF '2022, Bologna, Italy, 2022.
  - [79] M. Aliannejadi, G. Faggioli, N. Ferro, Vlachos, Michalis (Eds.), Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, 2023.
  - [80] 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.
  - [81] 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.