Overview of CheckThat! 2020 English: Automatic Identification and Verification of Claims in Social Media

Shaden Shaar[1], Alex Nikolov[2], Nikolay Babulkov[2], Firoj Alam[1], Alberto Barrón-Cedeño[3], Tamer Elsayed[4], Maram Hasanain[4], Reem Suwaileh[4], Fatima Haouari[4], Giovanni Da San Martino[1], and Preslav Nakov[1]

[1] Qatar Computing Research Institute, HBKU, Doha, Qatar
{sshaar, firoj, pnakov, gmartino}@hbku.edu.qa
[2] FMI, Sofia University "St Kliment Ohridski", Bulgaria
{nbabulkov, alexnickolow}@gmail.com
[3] DIT, Università di Bologna, Forlì, Italy
a.barron@unibo.it
[4] Computer Science and Engineering Department, Qatar University, Doha, Qatar
{telsayed, maram.hasanain, rs081123, 200159617}@qu.edu.qa

Abstract. We present an overview of the third edition of the CheckThat! Lab at CLEF 2020. The lab featured five tasks in Arabic and English, and here we focus on the three English tasks. Task 1 challenged the participants to predict which tweets from a stream of tweets about COVID-19 are worth fact-checking. Task 2 asked to retrieve verified claims from a set of previously fact-checked claims, which could help fact-check the claims made in an input tweet. Task 5 asked to propose which claims in a political debate or a speech should be prioritized for fact-checking. A total of 18 teams participated in the English tasks, and most submissions managed to achieve sizable improvements over the baselines using models based on BERT, LSTMs, and CNNs. In this paper, we describe the process of data collection and the task setup, including the evaluation measures used, and we give a brief overview of the participating systems. Last but not least, we release to the research community all datasets from the lab as well as the evaluation scripts, which should enable further research in the important tasks of check-worthiness estimation and detecting previously fact-checked claims.

Keywords: Check-Worthiness Estimation · Fact-Checking · Veracity · Verified Claims Retrieval · Detecting Previously Fact-Checked Claims · Social Media Verification · Computational Journalism · COVID-19

1 Introduction

Recent years have seen growing concerns, both in academia and in industry, in the face of the threats posed by disinformation online, commonly known as "fake news". To address the issue, a number of initiatives were launched to perform manual claim verification, with over 200 fact-checking organizations worldwide,[5] such as PolitiFact, FactCheck, Snopes, and Full Fact. Unfortunately, these efforts do not scale and they are clearly insufficient, given the scale of disinformation, which, according to the World Health Organization, has grown into the First Global Infodemic in the times of COVID-19. With this in mind, we have launched the CheckThat! Lab, which features a number of tasks aiming to help automate the fact-checking process.

The `CheckThat!` lab[6] was run for the third time in the framework of CLEF 2020. The purpose of the 2020 edition of the lab was to foster the development of technology that would enable the (semi-)automatic verification of claims posted in social media, in particular in Twitter. In this paper, we focus on the three `CheckThat!` tasks that were offered in English.[7] Figure 1 shows the full `CheckThat!` identification and verification pipeline, including four tasks on Twitter and one on debates/speeches. This year, we ran three of the five tasks in English:

Task 1 Check-worthiness estimation for tweets. Given a topic and a stream of potentially related tweets, rank the tweets according to their check-worthiness for the topic.

Task 2 Verified claim retrieval. Given a check-worthy input claim and a set of verified claims, rank those verified claims, so that the claims that can help verify the input claim, or a sub-claim in it, are ranked above any claim that is not helpful to verify the input claim.

If the model for Task 2 fails to return relevant tweets, the verification steps are triggered, i.e., Task 3 on supporting evidence retrieval and Task 4 on claim verification.[8] While Tasks 1 and 2 are offered for the first time and they focus on tweets, Task 5 is a legacy task from the two previous editions of `CheckThat!` [32, 60]. It is similar to Task 1, but it is from a different genre:

Task 5 Check-worthiness estimation on debates/speeches. Given a transcript, rank the sentences in the transcript according to the priority to fact-check them.

---

[5] `http://tiny.cc/zd1fnz`

[6] `https://sites.google.com/view/clef2020-checkthat/`

[7] Refer to [14] for an overview of the full CheckThat! 2020 lab, but with less details for the English tasks.

[8] We did not offer Tasks 3 and 4 in English this year; they were run for Arabic only. Refer to [40] for further details.

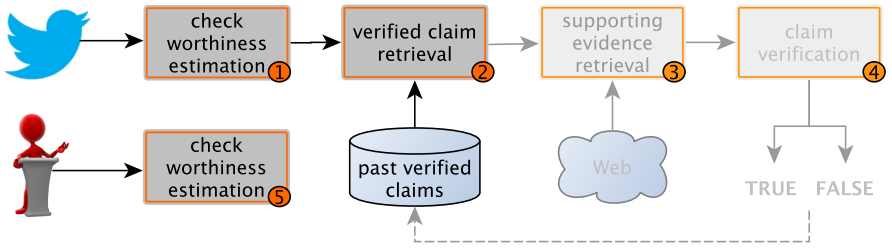

Fig. 1: The `CheckThat!` claim verification pipeline. We offer Tasks 1 and 2 on Twitter and Task 5 on political debates and speeches, all in English. Tasks 1, 3, and 4 were offered in Arabic [40].

For Task 1, we focused on COVID-19 as a topic: we crawled and manually annotated tweets from March 2020. Task 1 attracted 12 teams, and the most successful approaches used Transformers or a combination of embeddings, manually-engineered features, and neural networks. Section 3 offers more details.

For Task 2, we used claims from Snopes and corresponding tweets, where the claim originated. The task attracted 8 teams, and the most successful approaches relied on Transformers and data augmentaton. Section 4 gives more details.

For Task 5, we used PolitiFact as the main data source. The task attracted three teams, and Bi-LSTMs with word embeddings performed the best. Section 5 gives more details.

As for the rest of the paper, Section 2 discusses some related work, and Section 6 concludes with final remarks.

2 Related Work

Automatic claim fact-checking is a growing research area, covering a number of subtasks: from automatic identification and verification of claims [6, 8, 13, 31, 32, 41, 59], to identifying check-worthy claims [35, 42, 44, 78], detecting whether a target claim has been previously fact-checked [70], retrieving evidence to accept or reject these claims [10, 45], checking whether the evidence supports or denies the claim [56, 57], and inferring the veracity of the claim, e.g., using linguistic analysis [9, 21, 48, 67] or external sources [11, 12, 45, 61, 66, 76].

Check-worthiness estimation on debates/speeches. The ClaimBuster system [42] was a pioneering work on check-worthiness estimation. Given a sentence in the context of a political debate, it classified it into one of the following, manually annotated categories: *non-factual*, *unimportant factual*, or *check-worthy factual*. In later work, Gencheva & al. [35] also focused on the 2016 US Presidential debates, for which they obtained binary (*check-worthy* vs. *non-check-worthy*) annotations from different fact-checking organizations. An extension of this work resulted in the development of the ClaimRank system, which was trained on more data and also included Arabic content [44].

Other related work, also focused on political debates and speeches. For example, Patwari & al. [64] predicted whether a sentence would be selected by a fact-checking organization using a boosting-like model. Similarly, Vasileva & al. [78] used a multi-task learning neural network that predicts whether a sentence would be selected for fact-checking by each individual fact-checking organization (from a set of nine such organizations). Last but not least, the task was the topic of CLEF in 2018 and 2019, 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 Task 5 in 2020), with systems based on word embeddings and RNNs [1, 36, 38, 87]. 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 [5, 26, 29, 33, 34, 39, 55, 74].

Check-worthiness estimation for tweets. Unlike political debates, there has been less effort in identifying check-worthy claims in social media, which is Task 1 in the 2020 edition of the lab. The only directly related previous work we are aware of is [3], where they developed a multi-question annotation schema of tweets about COVID-19, organized around seven questions that model the perspective of journalists, fact-checkers, social media platforms, policy makers, and the society. The first question in the schema is influenced by [47], but overall it is much more comprehensive, and some of its questions are particularly tailored for COVID-19. For the 2020 Task 1, we use the setup and the annotations for one of the questions in their schema, as well as their data for that question, which we further extend with additional data, following their annotation instructions and using their annotation tools [2]. An indirectly related research line is on credibility assessment of tweets [37], including the CREDBANK tweet corpus [54], which has credibility annotations, as well as work on fake news [71] and on rumor detection in social media [85]; unlike that work, here we focus on detecting check-worthiness rather than predicting the credibility/factuality of the claims in the tweets. Another less relevant research line is on the development of datasets of tweets about COVID-19 [25, 73, 86]; however, none of these datasets focuses on check-worthiness estimation.

Verified claims retrieval. Task 2 in this 2020 edition of the lab focuses on retrieving and ranking verified claims. This is an underexplored task and the only directly relevant work is [70]; here we use their annotation setup and one of their datasets: Snopes (they also have experiments with claims from PolitiFact). Previous work has mentioned the task as an integral step of an end-to-end automated fact-checking pipeline, but there was very little detail provided about this component and it was not evaluated [43].

In an industrial setting, Google has developed *Fact Check Explorer*,[9] which allows users to search a number of fact-checking websites. However, the tool cannot handle a complex claim, as it uses the standard Google search funcionality, which is not optimized for semantic matching of long claims.

[9] `http://toolbox.google.com/factcheck/explorer`

Another related work is the *ClaimsKG* dataset and system [75], which includes 28K claims from multiple sources, organized into a knowledge graph (KG). The system can perform data exploration, e.g., it can find all claims that contain a certain named entity or keyphrase. In contrast, we are interested in detecting whether a claim was previously fact-checked.

Finally, the task is related to semantic relatedness tasks, e.g., from the GLUE benchmark [80], such as natural language inference (NLI) [82], recognizing textual entailment (RTE) [15], paraphrase detection [30], and semantic textual similarity (STS-B) [22]. However, it differs from them in a number of aspects; see [70] for more details and discussion.

3 Task 1_{en}. Check-Worthiness on Tweets

Task 1 (English) *Given a topic and a stream of potentially related tweets, rank the tweets according to their check-worthiness for the topic.*

Previous work on check-worthiness focused primarily on political debates and speeches, while here we focus on tweets instead.

3.1 Dataset

We focused on a single topic, namely *COVID-19*, and we collected tweets that matched one of the following keywords and hashtags: *#covid19, #CoronavirusOutbreak, #Coronavirus, #Corona, #CoronaAlert, #CoronaOutbreak, Corona*, and *covid-19*. We ran all the data collection in March 2020, and we selected the most retweeted tweets for manual annotation.

For the annotation, we considered a number of factors. These include 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 the following five questions:[10]

- **Q1: Does the tweet contain a verifiable factual claim?** This is an objective question. Positive examples include tweets that state a definition, mention a quantity in the present or the past, make a verifiable prediction about the future, reference laws, procedures, and rules of operation, discuss images or videos, and state correlation or causation, among others.[11]
- **Q2: To what extent does the tweet appear to contain false information?** This question asks for a subjective judgment; it does not ask for annotating the actual factuality of the claim in the tweet, but rather whether the claim appears to be false.

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

[11] This is influenced by [47].

– **Q3: Will the tweet have an effect on or be of interest to the general public?** This question asks for an objective judgment. Generally, claims that contain information related to potential cures, updates on number of cases, on measures taken by governments, or discussing rumors and spreading conspiracy theories should be of general public interest.
– **Q4: To what extent is the tweet harmful to the society, person(s), company(s) or product(s)?** This question also asks for an objective judgment: to identify tweets that can negatively affect society as a whole, but also specific person(s), company(s), product(s).
– **Q5: 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 informed by the answer to questions Q2, Q3 and Q4, as a check-worthy factual claim is probably one that is likely to be false, is of public interest, and/or appears to be harmful. Notice that we are stressing the fact that a professional fact-checker should verify the claim, which rules out claims easy to fact-check by the layman.

For the purpose of the task, we consider as check-worthy the tweets that received a positive answer both to Q1 and to Q5; if there was a negative answer to either Q1 or Q5, the tweet was considered not worth fact-checking. The answers to Q2, Q3, and Q4 were not considered directly, but they helped the annotators make a better decision for Q5.

The annotations were performed by 2–5 annotators independently, and then consolidated after a discussion for the cases of disagreement. The annotation setup was part of a broader COVID-19 annotation initiative; see [3] for more details about the annotation instructions and setup.

Table 1: **Task 1, English:** tweets with their check-worthiness marked.

Breaking: Congress prepares to shutter Capitol Hill for coronavirus, opens telework center	✓
China has 24 times more people than Italy...	✗
Everyone coming out of corona as barista	✗
Lord, please protect my family & the Philippines from the corona virus	✗

Examples of annotated tweets are shown in Table 1. The first example, *'Breaking: Congress prepares to shutter Capitol Hill for coronavirus, opens telework center'*, containing a verifiable factual claim on a topic of high interest to society, and thus it is labeled as check-worthy. The following tweet *'China has 24 times more people than Italy...'*, contains a verifiable factual claim, but it is trivial to fact-check, and thus it is annotated as not check-worthy. The third example, *'Everyone coming out of corona as barista'*, is a joke, and thus it is considered not check-worthy. The fourth example, *'Lord, please protect my family & the Philippines from the corona virus'* does not contain a verifiable factual claim, and it is thus not check-worthy.

Table 2 shows some statistics about the data, which is split into training, development, and testing datasets. We can see that the datasets are fairly balanced with the check-worthy claims making 34-43% of the examples.

Table 2: **Task 1, English:** Statistics about the tweets in the dataset.

Dataset	Total	Check-worthy
Train	672	231
Dev	150	59
Test	140	60

3.2 Evaluation

This is a ranking task, where a tweet has to be ranked according to its check-worthiness. Therefore, we consider mean average precision (MAP) as the official evaluation measure, which we complement with reciprocal rank (RR), R-precision (R-P), and P@k for $k \in \{1, 3, 5, 10, 20, 30\}$. The data and the evaluation scripts are available online.[12]

3.3 Overview of the Systems

A total of twelve teams took part in Task 1, using models based on state-of-the-art pre-trained Transformers such as BERT [28] and RoBERTa [50], but there were also systems that used more traditional machine learning models, such as SVMs and Logistic Regression. Table 3 shows a summary of the approaches used by the primary submissions of the participating teams. We can see that BERT and RoBERTa models were by far the most popular among the participants.

The top-ranked team **Accenture** [83] used a model based on RoBERTa, with an extra mean pooling and dropout layer on top of the basic RoBERTa network. The mean pooling layer averages the outputs from the last two RoBERTa layers in order to prevent overfitting, after which the result is passed to a dropout layer and a classification head.

The second-best **Team_Alex** [62] trained a logistic regression classifier using RoBERTa's predictions plus additional features, modeling the context of the tweet, e.g., whether the tweet comes from a verified account, the number of likes for the target tweet, whether the tweet includes a URL, whether the tweet contains a link to a news outlet that is known to be factual/questionable in its reporting, etc. Apart from some standard tweet preprocessing, such as replacing URLs and user mentions with special tokens, they further replaced the term *COVID-19* with *Ebola*, since the former is not in the RoBERTa vocabulary.

[12] `https://github.com/sshaar/clef2020-factchecking-task1/`

Table 3: **Task 1, English:** Summary of the approaches used in the primary system submissions.

Team		Models								Representation								Other	
		BERT	RoBERTa	BiLSTM	CNN	Random forest	Linear regression	Logistic regression	SVM	FastText	GloVe	PCA	Topic models	TF.IDF	Dependencies	Part of speech	Named entities	External data	Graph relations
Accenture	[83]		●																
BustingMisinformation	–				●				●		●		●	●					
check_square	[24]	●							●			●			●	●	●		
Factify	–	●																●	
NLP&IR@UNED	[51]			●							●								●
QMUL-SDS	[4]	●			●														
Team_Alex	[62]		●																
TheUofSheffield	[52]					●				●				●					
TOBB ETU	[46]	●						●								●			
UAICS	[27]	●																	
SSN_NLP	[49]		●																
ZHAW	–						●									●	●		

Team **Check_square** [24] used a variety of features such as part of speech tags, named entities, and dependency relations, in addition to a variety of word embeddings such as GloVe [65], Word2Vec [53], and FastText [16]. They also experimented with a number of custom embeddings generated with different pooling strategies from the last four layers of BERT. They further used PCA for dimensionality reduction. The remaining features were used to train an SVM model.

Team **QMUL-SDS** [4] used additional data, containing tweets annotated for rumor detection. They further used the uncased COVID-Twitter-BERT architecture [58], which is pre-trained on COVID-19 Twitter stream data, and then they passed the computed tweet representations to a 3-layer CNN model.

Team **TOBB ETU** [46] used multilingual BERT and word embeddings as features in a logistic regression model. Moreover, for each tweet they added part of speech tags, features modeling the presence of 66 special words (e.g., *unemployment*), cosine similarities between the tweet and the averaged word embedding vector of different terms describing a specific topic, e.g., *employment*, etc. They further added tweet metadata features, such as whether the account is verified, whether the tweet contains a quote/URL/hashtag/user mention, as well as the number of times it was retweeted.

Table 4: **Task 1, English:** evaluation results.

	Team	MAP	RR	R-P	P@1	P@3	P@5	P@10	P@20	P@30
[83]	**Accenture**	**0.806**$_1$	**1.000**$_1$	**0.717**$_1$	**1.000**$_1$	**1.000**$_1$	**1.000**$_1$	**1.000**$_1$	**0.950**$_1$	**0.740**$_1$
[62]	**Team_Alex**	0.803$_2$	**1.000**$_1$	0.650$_4$	**1.000**$_1$	**1.000**$_1$	**1.000**$_1$	**1.000**$_1$	**0.950**$_1$	**0.740**$_1$
	contr.-1	0.799	1.000	0.650	1.000	1.000	1.000	1.000	0.950	0.740
	contr.-2	0.781	1.000	0.667	1.000	1.000	1.000	1.000	0.850	0.680
[24]	**check_square**	0.722$_3$	**1.000**$_1$	0.667$_3$	**1.000**$_1$	0.667$_9$	0.800$_6$	0.800$_5$	0.800$_3$	0.700$_3$
	contr.-1	0.625	0.500	0.600	0.000	0.667	0.800	0.800	0.650	0.580
	contr.-2	0.714	0.500	0.683	0.000	0.667	0.600	0.800	0.850	0.700
[4]	**QMUL–SDS**	0.714$_4$	**1.000**$_1$	0.633$_5$	**1.000**$_1$	**1.000**$_1$	**1.000**$_1$	0.900$_3$	0.800$_3$	0.640$_6$
	contr.-1	0.782	1.000	0.700	1.000	1.000	1.000	1.000	0.850	0.700
	contr.-2	0.729	1.000	0.633	1.000	1.000	1.000	0.900	0.850	0.680
[46]	**TOBB ETU**	0.706$_5$	**1.000**$_1$	0.600$_6$	**1.000**$_1$	**1.000**$_1$	**1.000**$_1$	0.900$_3$	0.800$_3$	0.660$_5$
	contr.-1	0.563	0.200	0.600	0.000	0.000	0.200	0.300	0.600	0.660
	contr.-2	0.710	1.000	0.633	1.000	1.000	1.000	1.000	0.750	0.680
[49]	**SSN_NLP**	0.674$_6$	**1.000**$_1$	0.600$_6$	**1.000**$_1$	**1.000**$_1$	0.800$_6$	0.800$_5$	0.800$_3$	0.620$_7$
	contr.-1	0.674	1.000	0.600	1.000	1.000	0.800	0.800	0.800	0.620
	Factify	0.656$_7$	0.500$_{10}$	0.683$_2$	0.000$_{10}$	0.333$_3$	0.600$_9$	0.700$_7$	0.750$_7$	0.700$_3$
	contr.-1	0.696	1.000	0.683	1.000	0.333	0.600	0.800	0.800	0.740
	BustMisinfo	0.617$_8$	**1.000**$_1$	0.583$_8$	**1.000**$_1$	**1.000**$_1$	0.800$_6$	0.700$_7$	0.600$_8$	0.600$_8$
[51]	**NLP&IR@UNED**	0.607$_9$	**1.000**$_1$	0.567$_9$	**1.000**$_1$	**1.000**$_1$	**1.000**$_1$	0.700$_7$	0.600$_8$	0.580$_9$
	contr.-1	0.555	0.250	0.550	0.000	0.000	0.400	0.700	0.750	0.520
	contr.-2	0.519	0.500	0.450	0.000	0.667	0.400	0.500	0.600	0.480
	Baseline (n-gram)	0.579	1.000	0.500	1.000	0.667	0.800	0.800	0.700	0.600
	ZHAW	0.505$_{10}$	0.333$_{11}$	0.533$_{10}$	0.000$_{10}$	0.333$_{10}$	0.400$_{10}$	0.600$_{10}$	0.500$_{11}$	0.520$_{10}$
	contr.-1	0.665	1.000	0.633	1.000	1.000	0.800	0.900	0.700	0.660
[27]	**UAICS**	0.495$_{11}$	**1.000**$_1$	0.467$_{12}$	**1.000**$_1$	0.333$_{10}$	0.400$_{10}$	0.600$_{10}$	0.600$_8$	0.460$_{12}$
[52]	**TheUofSheffield**	0.475$_{12}$	0.250$_{12}$	0.533$_{10}$	0.000$_{10}$	0.000$_{12}$	0.400$_{10}$	0.200$_{12}$	0.350$_{12}$	0.480$_{11}$
	contr.-1	0.646	1.000	0.583	1.000	1.000	1.000	0.800	0.600	0.580

Team **SSN_NLP** [49] explored different approaches, such as a 5-layer CNN trained on Word2Vec representations, BERT and XLNet [84], and an SVM using TF.IDF features. Eventually, they submitted one RoBERTa and one CNN model.

Team **NLP&IR@UNED** [51] used a bidirectional LSTM with GloVe embedding representations. They used a graph model to search for up to three additional tweets that are most similar to the target tweet based on the inclusion of common hashtags, user mentions, and URLs. The text of these tweets was concatenated to the text of the original tweet and fed into the neural network. They further experimented with feed-forward neural networks and CNNs.

Team **Factify** submitted a BERT-based classifier.

Team **BustMisinfo** used an SVM with TF.IDF features and GloVe embeddings, along with topic modelling using NMF.

Team **ZHAW** used logistic regression with part-of-speech tags and named entities along with features about the location and the time of posting, etc.

Team **UAICS** [27] used a model derived from BERT, applying standard pre-processing.

Team **TheUofSheffield** [52] submitted a Random Forest model with TF.IDF features. Their pre-processing included lowercasing, lemmatization, as well as URL, emoji, stopwords, and punctuation removal. They also experimented with other models, such as a Naïve Bayes Classifier, K-Means, SVM, LSTM and FastText. They further tried Word2Vec representation as features, but this yielded worse results.

Table 4 shows the performance of the submissions to Task 1, English. We can see that *Accenture* and *Team_Alex* were almost tied and achieved very high performance on all evaluation measures and outperformed the other teams by a wide margin, e.g., by about eight points absolute in terms of MAP. We can further see that most systems managed to outperform an n-gram baseline by a very sizeable margin.

4 Task 2_{en}. Verified Claim Retrieval

Task 2 (English) *Given a check-worthy input claim and a set of verified claims, rank those verified claims, so that the claims that can help verify the input claim, or a sub-claim in it, are ranked above any claim that is not helpful to verify the input claim.*

Task 2 is a new task for the CLEF 2020 CheckThat! lab. A system solving that task could provide support to fact-checkers in their routine work: they do not need to spend hours fact-checking a claim, only to discover afterwards that it has been fact-checked already. Such a system could also help journalists during political debates and live interviews, by providing them trusted real-time information about known false claims that a politician makes, thus making it possible to put that person on the spot right away.

Table 5 shows examples of tweets (input claims), as well as the top-3 corresponding previously fact-checked claims from Snopes ranked by their relevance with respect to the input claim. The examples are ranked by our baseline model: BM25. In example (a), the model places the correct verified claim at rank 1; this can be considered as a trivial case because the same result can be achieved by using simple word overlap as a similarity score. Example (b) shows a harder case, and BM25 fails to retrieve the corresponding verified claim among the top-3 results. The claim in (c) is somewhere in between: the system assigns it a high enough score for it to be near the top, but the model is not confident enough to put the correct verified claim at rank 1, and it goes to rank 2 instead. Note that, even when expressing the same concepts, the input and the most relevant verified claim can be phrased quite differently, which makes the task difficult.

Table 5: **Task 2, English:** example input tweets and the top-3 most similar verified claims from Snopes retrieved by our baseline BM25 system. The correct previously fact-checked matching claim to be retrieved is marked with a ✓.

input tweet:	**(a)**	Former Facebook Worker: We Suppressed Conservative News — Sean Hannity (@seanhannity) May 9, 2016	
verified claims:	**(1)**	Facebook routinely suppresses conservative news in favor of liberal content.	✓
	(2)	Sarah Palin told Sean Hannity Alaska has "all sorts of Eskimos and other foreigners."	✗
	(3)	Fox News host was about to be fired in June 2016 over comments he made about Muslims.	✗
input tweet:	**(b)**	@BernieSanders Makes Epic Comeback To Win Nevada — The Young Turks (@TheYoungTurks) April 5, 2016	
verified claims:	**(1)**	Professor makes snappy comeback to female students who protest his chauvinism.	✗
	(2)	The Nevada Athletic Commission has voided Floyd Mayweather's boxing victory over Manny Pacquiao.	✗
	(3)	Article explains the difference between http and https protocols.	✗
input tweet:	**(c)**	Email from Obuma to Hilary outlines plans to take guns only from Republican voters. — wolverine7217 (@wolverine7217) May 20, 2016	
verified claims:	**(1)**	Changes coming to Social Security on 1 May 2016 'threaten the financial security' of millions of Americans.	✗
	(2)	Recovered e-mails belonging to former secretary of state Hillary Clinton have revealed plans to seize guns from Republicans on 8 November 2016.	✓
	(3)	Voters across Texas have witnessed their votes switch from 'straight Republican' to Democrat on compromised voting machines.	✗

4.1 Dataset

Each input claim was retrieved from the Snopes fact-checking website,[13] which dedicates an article to assessing the truthfulness of each claim they have analyzed. Snopes articles often list different tweets that contain (a paraphrase of) the verified claim. Together with the title of the article page and the rating of the claim, as assigned by Snopes, we collect all those tweets and we use them as input claims. The task is, given such a tweet, to find the corresponding (verified) target claim. The set of target claims consists of the claims we collected from Snopes, with additional claims added from ClaimsKG [75] that were also gathered from Snopes. Note that we have just one list of verified claims, and we match each input tweet against that list.

[13] `http://www.snopes.com`

Table 6: **Task 2, English:** Statistics about the input tweets and the matching and non-matching verified claims from Snopes.

Dataset	Tweets	Verified Claims	
Training	800	Matching	1,197
Dev	197	Non-matching	9,178
Test	200		
Total	1,197	Total	10,375

As Table 6 shows, the dataset consists of 1,197 input tweets, split into a training, a development, and a test dataset. These input tweets are to be compared to a set of 10,375 verified claims, among which only 1,197 actually match some of the input tweets.

4.2 Overview of the Systems

Eight teams participated in Task 2, using a variety of scoring functions, based on fine-tuned pre-trained Transformers such as BERT or supervised models such as SVMs, or unsupervised approaches such as simple cosine similarity and scores produced by Terrier and Elastic Search. Two teams also did data cleaning by removing URLs, hashtags, usernames and emojis from the tweets. Table 7 summarizes the approaches used by the primary submissions.

The winning team —**Buster.ai** [17]— first cleaned the tweets, and then used a pre-trained and fine-tuned version of RoBERTa. Before training on the dataset, they fine-tuned their model on external datasets such as FEVER [76], SciFact [79], and Liar [81]. While training, they used indexed search to retrieve adversarial negative examples, forcing the model to learn the proper semantics to distinguish between syntactically and lexically close sentences.

Team **UNIPI-NLE** [63] performed two cascade fine-tunings of a sentence-BERT model [68]. Initially, they fine-tuned on the task of predicting the cosine similarity between a tweet and a claim. For each tweet, they trained on the gold verified claim and on twenty negative verified claims selected randomly from a list of candidate pairs with a non-empty overlap with the input claim in terms of keywords. In the second step, they fine-tuned the model on a classification task for which sentence-BERT has to output 1 if the pair is a correct match, and 0 otherwise. They randomly selected two negative examples and used them with the gold to fine-tune the model. Before inference, they pruned the verified claim list, top-2500, using Elastic Search and simple word matching techniques.

Team **UB_ET** [77] trained a model on a limited number of tweet–claim pairs per tweet. They retrieved the top-1000 tweet–claim pairs per tweet using parameter-free DPH divergence from the randomness term weighting model in Terrier, and computed several features from weighting models (BM25, PL2 and TF-IDF) and then built a LambdaMart model on top for reranking. Moreover, the texts were pre-processed using tokenization and Porter stemming.

Table 7: **Task 2, English:** summary of the approaches used in the primary system submissions. We report which systems used search engine scores, scoring functions (supervised or not), representations (other than Transformers), and removal of tokens. We further indicate whether external data was used.

Team		**Engine**		**Scoring**							**Repr.**			**Removal**					
		Terrier	**ElasticSearch**	**LambdaMART**	**BERT**	**RoBERTa**	**Unspecified Transf.**	**KD search**	**SVM**	**Cosine**	**tf–idf**	**BM25**	**Term dependencies**	**URL removal**	**Emoji removal**	**Time removal**	**Username removal**	**Hashtag removal**	**External data**
Buster.ai	[17]					●													●
check_square	[24]						●	●											
elec-dlnlp	–				●							●							
iit	–				●					●				●	●	●			
TheUofSheffield	[52]								●		●	●		●			●	●	
trueman	[72]						●												
UB_ET	[77]	●		●							●	●	●						
UNIPI-NLE	[63]		●		●					●									

They also made submissions using the Sequential Dependence (SD) variant of the Markov Random Field for term dependence to rerank the top–1000 pairs per tweet. However, the best results were obtained by the DPH divergence from randomness term that was used for the initial claim retrieval, without the final step of reranking.

Team **NLP&IR@UNED** [51] used the Universal Sentence Encoder [23] to obtain embeddings for the tweets and for the verified claims. As features, they used sentence embeddings, the type-token ratio, the average word length, the number of verbs/nouns, the ratio of content words, and the ratio of content tags. Then, they trained a feed-forward neural network (FFNN), based on ELUs. In their `Primary` submission, they used the above seven features, without the sentence embedding, along with the FFNN, and achieved a MAP@5 score of 0.856. In their *Contrastive-1* submission, they used all features, including sentence embeddings. Finally, their *Constrastive-2* submission was identical to their `Primary` one, but with a different random initialization.

Team **TheUniversityofSheffield** [52] pre-processed the input tweets, e.g., they removed all hashtags, and then they trained a number of machine learning models, such as Logistic Regression, Random Forest, Gradient Boosted Trees, Linear SVM and Linear Regression, and features such as TF.IDF-weighted cosine similarity, BM25 score, and simple Euclidean distance between the input tweet and a candidate verified claim.

Table 8: **Task 2, English:** performance of the submissions and of the Elastic Search (ES) baseline.

Team	MAP @1	MAP @3	MAP @5	MAP –	Precision @1	Precision @3	Precision @5	RR @1	RR @3	RR @5
[17] **Buster.ai**	**0.897**$_1$	**0.926**$_1$	**0.929**$_1$	**0.929**$_1$	**0.895**$_1$	**0.320**$_1$	**0.195**$_1$	**0.895**$_1$	**0.923**$_1$	**0.927**$_1$
contr.-1	0.818	0.865	0.871	0.871	0.815	0.308	0.190	0.815	0.863	0.868
contr.-2	0.907	0.937	0.938	0.938	0.905	0.325	0.196	0.905	0.934	0.935
[63] **UNIPI–NLE**	0.877$_2$	0.907$_2$	0.912$_2$	0.913$_2$	0.875$_2$	0.315$_2$	0.193$_2$	0.875$_2$	0.904$_2$	0.909$_2$
contr.-1	0.877	0.913	0.916	0.917	0.875	0.320	0.194	0.875	0.911	0.913
[77] **UB_ET**	0.818$_3$	0.862$_3$	0.864$_3$	0.867$_3$	0.815$_3$	0.307$_3$	0.186$_3$	0.815$_3$	0.859$_3$	0.862$_3$
contr.-1	0.838	0.865	0.869	0.874	0.835	0.300	0.184	0.835	0.863	0.867
contr.-2	0.843	0.868	0.873	0.877	0.840	0.300	0.185	0.840	0.865	0.870
[51] **NLP&IR@UNED**	0.807$_4$	0.851$_4$	0.856$_4$	0.861$_4$	0.805$_4$	0.300$_4$	0.185$_4$	0.805$_4$	0.848$_4$	0.854$_4$
contr.-1	0.787	0.832	0.839	0.845	0.785	0.297	0.184	0.785	0.829	0.836
contr.-2	0.807	0.850	0.855	0.861	0.805	0.300	0.185	0.805	0.848	0.853
[52] **UofSheffield**	0.807$_4$	0.807$_5$	0.807$_5$	0.807$_5$	0.805$_4$	0.270$_5$	0.162$_7$	0.805$_5$	0.805$_5$	0.805$_5$
contr.-1	0.772	0.772	0.772	0.772	0.770	0.258	0.155	0.770	0.770	0.770
contr.-2	0.767	0.767	0.767	0.767	0.765	0.257	0.154	0.765	0.765	0.765
[72] **trueman**	0.743$_6$	0.768$_6$	0.773$_6$	0.782$_6$	0.740$_6$	0.267$_6$	0.164$_6$	0.740$_6$	0.766$_6$	0.771$_6$
elec-dlnlp	0.723$_7$	0.749$_7$	0.760$_7$	0.767$_7$	0.720$_7$	0.262$_7$	0.166$_5$	0.720$_7$	0.747$_7$	0.757$_7$
[24] **check_square**	0.652$_8$	0.690$_8$	0.695$_8$	0.706$_8$	0.650$_8$	0.247$_8$	0.152$_8$	0.650$_8$	0.688$_8$	0.692$_8$
contr.-1	0.828	0.868	0.873	0.875	0.825	0.307	0.189	0.825	0.865	0.871
contr.-2	0.718	0.746	0.754	0.763	0.715	0.260	0.163	0.715	0.743	0.751
baseline (ES)	0.470	0.601	0.609	0.619	0.472	0.249	0.156	0.472	0.603	0.611
iit	0.263$_9$	0.293$_9$	0.298$_9$	0.311$_9$	0.260$_9$	0.112$_9$	0.071$_9$	0.260$_9$	0.291$_9$	0.295$_9$

Team **trueman** [72] retrieved the top 1,000 matching claims for an input tweet along with the corresponding BM25 scores. Then, they calculated the cosine between the Sentence-BERT embedding representations for the input tweet and for a candidate verified claim, and they used these cosines to update the BM25 scores.

Team **elec-dlnlp** removed all hashtags and then used Transformer-based similarities between the input tweets and the candidate claims along with Elastic Search scores.

Team **check_square** [24] fine-tuned sentence-BERT with mined triplets and used the resulting sentence embedding to construct a KD-tree, which they used to extract the top-1000 candidate verified claims. Their tweet pre-processing included removing URLs, emails, phone numbers, and user mentions. Their `Primary` and *Constrastive-2* submissions used BERT-base and BERT-large, respectively, as well as Sentence-BERT. These models were then fine-tuned using triplet loss. Their *Contrastive-1* model was Sentence-BERT and multilingual DistilBERT [69], which was not fine-tuned. Their results show that DistilBERT performed better than the two fine-tuned BERT models.

Team **iit** used cosine similarities based on the embeddings from a pre-trained BERT model between the input tweet and the candidate verified claims.

4.3 Evaluation

The official evaluation measure for Task 2 was MAP@5. However, we further report MAP at $k \in \{1, 3, 10, 20\}$, overall MAP, R-Precision, Average Precision, Reciprocal Rank, and Precision@k.

Table 8 shows the evaluation results in terms of some of the performance measures for the primary and for the contrastive submissions for Task 2. The best and the second-best submissions —by Buster.ai and by UNIPI-NLE— are well ahead of the remaining teams by several points absolute on all evaluation measures. Most systems managed to outperform an Elastic Search baseline by a huge margin.

The data and the evaluation scripts are available online.[14]

5 Task 5_{en}. Check-Worthiness on Debates

Task 5 is a legacy task that has evolved from the first edition of the `CheckThat!` lab, and it was carried over in 2018 and 2019 [7, 8]. In each edition, more training data from more diverse sources have been added. However, all speeches and all debates are still about politics. The task focuses on mimicking the selection strategy that fact-checking organizations such as *PolitiFact* use to select the sentences and the claims to fact-check. The task is defined as follows:

Task 5 (English) *Given a transcript, rank the sentences in the transcript according to the priority they should be fact-checked.*

5.1 Dataset

Often after a major political event, such as a public debate or a speech by a government official, a professional fact-checker would go through the event transcript and would select a few claims to fact-check. Since those claims were selected for verification, we consider them as check-worthy. This is what we used to collect our data, focusing on *PolitiFact* as a fact-checking source. For a political event (debate/speech), we collected the article from *PolitiFact* and we obtained its official transcript, e.g., from ABC, Washington Post, CSPAN, etc. We would then manually match the sentences from the *PolitiFact* articles to the exact statement that was made in the debate/speech.

We collected a total of 70 transcripts and we annotated them based on overview articles from *PolitiFact* . The transcripts belonged to one of four types of political events: debates, speeches, interviews, and town-halls. We used the older 50 transcripts for training, and the more recent 20 transcripts for testing. Table 9 shows some annotated examples, and Table 10 shows the total number of sentences in the training and in the testing transcripts as well as the number of sentences that were fact-checked.

[14] `https://github.com/sshaar/clef2020-factchecking-task2/`

Table 9: **Task 5, English:** Debate fragments: the check-worthy sentences are marked with ⊙.

C. Booker:	We have systemic racism that is eroding our nation from health care to the criminal justice system.	
C. Booker:	And it's nice to go all the way back to slavery, but dear God, we have a criminal justice system that is so racially biased, we have more African-Americans under criminal supervision today than all the slaves in 1850.	⊙

(a) Fragment from the 2019 Democratic Debate in Detroit

L. Stahl:	Do you still think that climate change is a hoax?	⊙
D. Trump:	I think something's happening.	
D. Trump:	Something's changing and it'll change back again.	
D. Trump:	I don't think it's a hoax, I think there's probably a difference.	⊙
D. Trump:	But I don't know that it's manmade.	⊙

(b) Fragment from the 2018 CBS' 60 Minutes interview with President Trump

D. Trump:	We have no country if we have no border.	
D. Trump:	Hillary wants to give amnesty.	⊙
D. Trump:	She wants to have open borders.	⊙

(c) Fragment from the 2016 third presidential debate

5.2 Overview of the Systems

Three teams submitted a total of eight runs. A variety of embedding models were tried, and the best results were obtained using GloVe embeddings.

Team **NLP&IR@UNED** [51] experimented with various sampling techniques, embeddings, and models. For each of their submitted runs, they trained a Bi-LSTM model on the 6B-100D GloVE embeddings of the input sentences from the debates. Their *primary* and *Contrastive-1* runs used the training data as it was provided, while their *Constrastive-2* run used oversampling techniques. The difference between their `Primary` and *Contrastive-1* runs was in the weight initialization.

Team **UAICS** [52] used TF.IDF representation and different models: multinomial naïve Bayes for their `Primary` run, logistic regression for their *Contrastive-1* run, and decision tree for their *Constrastive-2* run.

Team **TOBB ETU** [46] used logistic regression with two main features: BERT prediction score and word2vec embeddings. They obtained the BERT prediction score by fine-tuning the base multi-lingual BERT on the classification task, and then added an additional classification layer to predict check-worthiness. They also obtained an embedding for the input sentence by averaging the word2vec embedding of the words in the sentence.

Table 10: **Task 5, English:** total number of sentences and number of sentences containing claims that are worth fact-checking, organized by type of text.

Type	Dataset	Transcripts	Sentences	Check-worthy
Debates	Train	18	25,688	254
	Test	7	11,218	56
Speeches	Train	18	7,402	163
	Test	8	7,759	50
Interviews	Train	11	7,044	62
	Test	4	2,220	23
Town-halls	Train	3	2,642	8
	Test	1	317	7
Total	**Train**	**50**	**42,776**	**487**
	Test	**20**	**21,514**	**136**

5.3 Evaluation

As this task was very similar to Task 1, but on a different genre, we used the same evaluation measures: namely, MAP as the official measure, and we also report P@k for various values of k.

Table 11 shows the performance of the primary submissions of the participating teams. The overall results are low, and only one team managed to beat our n-gram baseline.

Once again, the data and the evaluation scripts are available online.[15]

6 Conclusion and Future Work

We have presented an overview of the third edition of the CheckThat! Lab at CLEF 2020. The lab featured five tasks, which were offered in Arabic and English, and here we focus on the three English tasks. Task 1 was about check-worthiness of claims in tweets about COVID-19. Task 2 asked to rank a set of previously fact-checked claims, such that the ones that could help fact-check an input claim would be ranked higher. Task 5 asked to propose which claims in a political debate or a speech should be prioritized for fact-checking. A total of 18 teams participated in the English tasks, and most submissions managed to achieved sizable improvements over the baselines using models based on BERT, LSTMs, and CNNs.

We plan a new iteration of the CLEF CheckThat! lab, where we would offer new larger training sets, additional languages, as well as with some new tasks.

[15] `https://github.com/sshaar/clef2020-factchecking-task5/`

Table 11: **Task 5, English:** Performance of the primary submissions.

	Team	MAP	RR	R-P	P@1	P@3	P@5	P@10	P@20	P@30
[51]	**NLP&IR@UNED**	**0.087**$_1$	**0.277**$_1$	**0.093**$_1$	**0.150**$_1$	**0.117**$_1$	**0.130**$_1$	**0.095**$_1$	**0.073**$_1$	**0.039**$_1$
	contr.-1	0.085	0.259	0.092	0.150	0.100	0.120	0.090	0.068	0.037
	contr.-2	0.041	0.117	0.039	0.050	0.033	0.070	0.045	0.028	0.018
	Baseline	0.053	0.151	0.053	0.050	0.033	0.040	0.055	0.043	0.038
[52]	**UAICS**	0.052$_2$	0.225$_2$	0.053$_2$	**0.150**$_1$	0.100$_2$	0.070$_2$	0.050$_2$	0.038$_2$	0.027$_2$
	contr.-1	0.043	0.174	0.058	0.100	0.050	0.050	0.055	0.045	0.025
	contr.-2	0.033	0.114	0.028	0.050	0.050	0.030	0.035	0.018	0.019
[46]	**TOBB ETU**	0.018$_3$	0.033$_3$	0.014$_3$	0.000$_3$	0.017$_3$	0.020$_3$	0.010$_3$	0.010$_3$	0.006$_3$

Acknowledgments

This research is part of the Tanbih project, developed by the Qatar Computing Research Institute, HBKU and MIT-CSAIL, which aims to limit the effect of "fake news", propaganda, and media bias.

The work of Tamer Elsayed and Maram Hasanain was made possible by NPRP grant# NPRP 11S-1204-170060 from the Qatar National Research Fund (a member of Qatar Foundation). The work of Reem Suwaileh was supported by GSRA grant# GSRA5-1-0527-18082 from the Qatar National Research Fund and the work of Fatima Haouari was supported by GSRA grant# GSRA6-1-0611-19074 from the Qatar National Research Fund.

The statements made herein are solely the responsibility of the authors.

References

1. Agez, R., Bosc, C., Lespagnol, C., Mothe, J., Petitcol, N.: IRIT at CheckThat! 2018. In: Cappellato et al. [20]
2. Alam, F., Dalvi, F., Shaar, S., Durrani, N., Mubarak, H., Nikolov, A., Da San Martino, G., Abdelali, A., Sajjad, H., Darwish, K., Nakov, P.: Fighting the COVID-19 infodemic in social media: A holistic perspective and a call to arms. ArXiv preprint 2007.07996 (2020)
3. Alam, F., Shaar, S., Dalvi, F., Sajjad, H., Nikolov, A., Mubarak, H., Da San Martino, G., Abdelali, A., Durrani, N., Darwish, K., Nakov, P.: Fighting the COVID-19 infodemic: Modeling the perspective of journalists, fact-checkers, social media platforms, policy makers, and the society. ArXiv preprint 2005.00033 (2020)
4. Alkhalifa, R., Yoong, T., Kochkina, E., Zubiaga, A., Liakata, M.: QMUL-SDS at CheckThat! 2020: Determining COVID-19 tweet check-worthiness using an enhanced CT-BERT with numeric expressions. In: Cappellato et al. [18]
5. Altun, B., Kutlu, M.: 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, CEUR-WS.org, Lugano, Switzerland (2019)
6. Atanasova, P., Màrquez, L., Barrón-Cedeño, A., Elsayed, T., Suwaileh, R., Zaghouani, W., Kyuchukov, S., Da San Martino, G., Nakov, P.: Overview of the

CLEF-2018 CheckThat! lab on automatic identification and verification of political claims, task 1: Check-worthiness. In: 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)
7. Atanasova, P., Màrquez, L., Barrón-Cedeño, A., Elsayed, T., Suwaileh, R., Zaghouani, W., Kyuchukov, S., Da San Martino, G., Nakov, P.: Overview of the CLEF-2018 CheckThat! lab on automatic identification and verification of political claims. Task 1: Check-worthiness. In: Cappellato et al. [20]
8. Atanasova, P., Nakov, P., Karadzhov, G., Mohtarami, M., Da San Martino, G.: Overview of the CLEF-2019 CheckThat! lab on automatic identification and verification of claims. Task 1: Check-worthiness. In: Cappellato et al. [19]
9. Atanasova, P., Nakov, P., Màrquez, L., Barrón-Cedeño, A., Karadzhov, G., Mihaylova, T., Mohtarami, M., Glass, J.: Automatic fact-checking using context and discourse information. Journal of Data and Information Quality (JDIQ) **11**(3), 12 (2019)
10. Augenstein, I., Lioma, C., Wang, D., Chaves Lima, L., Hansen, C., Hansen, C., Simonsen, J.G.: 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. pp. 4685–4697. EMNLP-IJCNLP '19, Hong Kong, China (2019)
11. Baly, R., Karadzhov, G., Alexandrov, D., Glass, J., Nakov, P.: Predicting factuality of reporting and bias of news media sources. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 3528–3539. EMNLP '18, Brussels, Belgium (2018)
12. Baly, R., Mohtarami, M., Glass, J., Màrquez, L., Moschitti, A., Nakov, P.: 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. pp. 21–27. NAACL-HLT '18, New Orleans, Louisiana, USA (2018)
13. Barrón-Cedeño, A., Elsayed, T., Suwaileh, R., Màrquez, L., Atanasova, P., Zaghouani, W., Kyuchukov, S., Da San Martino, G., Nakov, P.: Overview of the CLEF-2018 CheckThat! lab on automatic identification and verification of political claims, task 2: Factuality. In: 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)
14. Barrón-Cedeño, A., Elsayed, T., Nakov, P., Da San Martino, G., Hasanain, M., Suwaileh, R., Haouari, F., Babulkov, N., Hamdan, B., Nikolov, A., Shaar, S., Sheikh Ali, Z.: Overview of CheckThat! 2020: Automatic identification and verification of claims in social media. In: Arampatzis, A., Kanoulas, E., Tsikrika, T., Vrochidis, S., Joho, H., Lioma, C., Eickhoff, C., Névéol, A., Cappellato, L., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction Proceedings of the Eleventh International Conference of the CLEF Association (CLEF 2020). LNCS (12260), Springer (2020)
15. Bentivogli, L., Dagan, I., Dang, H.T., Giampiccolo, D., Magnini, B.: The fifth PASCAL recognizing textual entailment challenge. In: Proceedings of the Text Analysis Conference. TAC '09, Gaithersburg, MD, USA (2009)
16. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics **5**, 135–146 (2017)

17. Bouziane, M., Perrin, H., Cluzeau, A., Mardas, J., Sadeq, A.: Buster.AI at Check-That! 2020: Insights and recommendations to improve fact-checking. In: Cappellato et al. [18]
18. Cappellato, L., Eickhoff, C., Ferro, N., Névéol, A. (eds.): CLEF 2020 Working Notes (2020)
19. Cappellato, L., Ferro, N., Losada, D., Müller, H. (eds.): Working Notes of CLEF 2019 Conference and Labs of the Evaluation Forum. CEUR Workshop Proceedings, CEUR-WS.org (2019)
20. Cappellato, L., Ferro, N., Nie, J.Y., Soulier, L. (eds.): Working Notes of CLEF 2018–Conference and Labs of the Evaluation Forum. CEUR Workshop Proceedings, CEUR-WS.org (2018)
21. Castillo, C., Mendoza, M., Poblete, B.: Information credibility on Twitter. In: Proceedings of the 20th International Conference on World Wide Web. pp. 675–684. WWW '11, Hyderabad, India (2011)
22. Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., Specia, L.: SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In: Proceedings of the 11th International Workshop on Semantic Evaluation. pp. 1–14. SemEval '17, Vancouver, Canada (2017)
23. Cer, D., Yang, Y., Kong, S., Hua, N., Limtiaco, N., John, R.S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y.H., Strope, B., Kurzweil, R.: Universal sentence encoder. ArXiv preprint 1803.11175 (2018)
24. Cheema, G.S., Hakimov, S., Ewerth, R.: Check_square at CheckThat! 2020: Claim detection in social media via fusion of transformer and syntactic features. In: Cappellato et al. [18]
25. Cinelli, M., Quattrociocchi, W., Galeazzi, A., Valensise, C.M., Brugnoli, E., Schmidt, A.L., Zola, P., Zollo, F., Scala, A.: The COVID-19 social media infodemic. arXiv:2003.05004 (2020)
26. Coca, L., Cusmuliuc, C.G., Iftene, A.: CheckThat! 2019 UAICS. In: CLEF 2019 Working Notes. Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum. CEUR Workshop Proceedings, CEUR-WS.org, Lugano, Switzerland (2019)
27. Cusmuliuc, C.G., Coca, L.G., Iftene, A.: UAICS at CheckThat! 2020: Fact-checking claim prioritization. In: Cappellato et al. [18]
28. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 4171–4186. NAACL-HLT '19, Minneapolis, Minnesota, USA (2019)
29. Dhar, R., Dutta, S., Das, D.: 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, CEUR-WS.org, Lugano, Switzerland (2019)
30. Dolan, W.B., Brockett, C.: Automatically constructing a corpus of sentential paraphrases. In: Proceedings of the Third International Workshop on Paraphrasing. IWP '05, Jeju Island, Korea (2005)
31. Elsayed, T., Nakov, P., Barrón-Cedeño, A., Hasanain, M., Suwaileh, R., Da San Martino, G., Atanasova, P.: CheckThat! at CLEF 2019: Automatic identification and verification of claims. In: Azzopardi, L., Stein, B., Fuhr, N., Mayr, P., Hauff, C., Hiemstra, D. (eds.) Advances in Information Retrieval. pp. 309–315. ECIR '19 (2019)

32. Elsayed, T., Nakov, P., Barrón-Cedeño, A., Hasanain, M., Suwaileh, R., Da San Martino, G., Atanasova, P.: Overview of the CLEF-2019 CheckThat!: Automatic identification and verification of claims. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. pp. 301–321. LNCS, Springer (2019)
33. Favano, L., Carman, M., Lanzi, P.: 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, CEUR-WS.org, Lugano, Switzerland (2019)
34. Gasior, J., Przybyła, P.: 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, CEUR-WS.org, Lugano, Switzerland (2019)
35. Gencheva, P., Nakov, P., Màrquez, L., Barrón-Cedeño, A., Koychev, I.: A context-aware approach for detecting worth-checking claims in political debates. In: Proceedings of the International Conference Recent Advances in Natural Language Processing. pp. 267–276. RANLP '17, Varna, Bulgaria (2017)
36. Ghanem, B., Montes-y Gómez, M., Rangel, F., Rosso, P.: UPV-INAOE-Autoritas - Check That: Preliminary approach for checking worthiness of claims. In: Cappellato et al. [20]
37. Gupta, A., Kumaraguru, P., Castillo, C., Meier, P.: TweetCred: Real-time credibility assessment of content on Twitter. In: Proceeding of the 6th International Social Informatics Conference. pp. 228–243. SocInfo'142 (2014)
38. Hansen, C., Hansen, C., Simonsen, J., Lioma, C.: 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: Cappellato et al. [20]
39. Hansen, C., Hansen, C., Simonsen, J., Lioma, C.: 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, CEUR-WS.org, Lugano, Switzerland (2019)
40. Hasanain, M., Haouari, F., Suwaileh, R., Ali, Z., Hamdan, B., Elsayed, T., Barrón-Cedeño, A., Da San Martino, G., Nakov, P.: Overview of CheckThat! 2020 Arabic: Automatic identification and verification of claims in social media. In: Cappellato et al. [18]
41. Hasanain, M., Suwaileh, R., Elsayed, T., Barrón-Cedeño, A., Nakov, P.: Overview of the CLEF-2019 CheckThat! lab on automatic identification and verification of claims. Task 2: Evidence and factuality. In: Cappellato et al. [19]
42. Hassan, N., Li, C., Tremayne, M.: Detecting check-worthy factual claims in presidential debates. In: Proceedings of the 24th ACM International Conference on Information and Knowledge Management. pp. 1835–1838. CIKM '15 (2015)
43. Hassan, N., Zhang, G., Arslan, F., Caraballo, J., Jimenez, D., Gawsane, S., Hasan, S., Joseph, M., Kulkarni, A., Nayak, A.K., Sable, V., Li, C., Tremayne, M.: ClaimBuster: The first-ever end-to-end fact-checking system. Proc. VLDB Endow. **10**(12), 1945–1948 (Aug 2017)
44. Jaradat, I., Gencheva, P., Barrón-Cedeño, A., Màrquez, L., Nakov, P.: ClaimRank: Detecting check-worthy claims in Arabic and English. In: Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics. pp. 26–30. NAACL-HLT '18, New Orleans, Louisiana, USA (2018)

45. Karadzhov, G., Nakov, P., Màrquez, L., Barrón-Cedeño, A., Koychev, I.: Fully automated fact checking using external sources. In: Proceedings of the Conference on Recent Advances in Natural Language Processing. pp. 344–353. RANLP '17, Varna, Bulgaria (2017)
46. Kartal, Y.S., Kutlu, M.: TOBB ETU at CheckThat! 2020: Prioritizing English and Arabic claims based on check-worthiness. In: Cappellato et al. [18]
47. Konstantinovskiy, L., Price, O., Babakar, M., Zubiaga, A.: Towards automated factchecking: Developing an annotation schema and benchmark for consistent automated claim detection. arXiv:1809.08193 (2018)
48. Kopev, D., Ali, A., Koychev, I., Nakov, P.: Detecting deception in political debates using acoustic and textual features. In: Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop. pp. 652–659. ASRU '19, Singapore (2019)
49. Krishan T, S., S, K., D, T., Vardhan K, R., Chandrabose, A.: Tweet check worthiness using transformers, CNN and SVM. In: Cappellato et al. [18]
50. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A robustly optimized BERT pretraining approach. Arxiv preprint 1907.11692 (2019)
51. Martinez-Rico, J., Araujo, L., Martinez-Romo, J.: NLP&IR@UNED at CheckThat! 2020: A preliminary approach for check-worthiness and claim retrieval tasks using neural networks and graphs. In: Cappellato et al. [18]
52. McDonald, T., Dong, Z., Zhang, Y., Hampson, R., Young, J., Cao, Q., Leidner, J., Stevenson, M.: The University of Sheffield at CheckThat! 2020: Claim identification and verification on Twitter. In: Cappellato et al. [18]
53. Mikolov, T., Yih, W., Zweig, G.: Linguistic regularities in continuous space word representations. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 746–751. NAACL-HLT '13, Atlanta, Georgia, USA (2013)
54. Mitra, T., Gilbert, E.: Credbank: A large-scale social media corpus with associated credibility annotations. In: Proceedings of the Ninth International AAAI Conference on Web and Social Media. pp. 258–267. ICWSM '15, Oxford, UK (2015)
55. Mohtaj, S., Himmelsbach, T., Woloszyn, V., Möller, S.: 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, CEUR-WS.org, Lugano, Switzerland (2019)
56. Mohtarami, M., Baly, R., Glass, J., Nakov, P., Màrquez, L., Moschitti, A.: Automatic stance detection using end-to-end memory networks. In: Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 767–776. NAACL-HLT '18, New Orleans, Louisiana, USA (2018)
57. Mohtarami, M., Glass, J., Nakov, P.: Contrastive language adaptation for cross-lingual stance detection. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. pp. 4442–4452. EMNLP '19, Hong Kong, China (2019)
58. Müller, M., Salathé, M., Kummervold, P.E.: COVID-Twitter-BERT: A natural language processing model to analyse COVID-19 content on Twitter. arXiv preprint arXiv:2005.07503 (2020)
59. Nakov, P., Barrón-Cedeño, A., Elsayed, T., Suwaileh, R., Màrquez, L., Zaghouani, W., Atanasova, P., Kyuchukov, S., Da San Martino, G.: 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. pp. 372–387. Lecture Notes in Computer Science, Springer, Avignon, France (2018)
60. Nakov, P., Barrón-Cedeño, A., Elsayed, T., Suwaileh, R., Màrquez, L., Zaghouani, W., Gencheva, P., Kyuchukov, S., Da San Martino, G.: Overview of the CLEF-2018 lab on automatic identification and verification of claims in political debates. In: Working Notes of CLEF 2018 – Conference and Labs of the Evaluation Forum. CLEF '18, Avignon, France (2018)
61. Nguyen, V.H., Sugiyama, K., Nakov, P., Kan, M.Y.: FANG: Leveraging social context for fake news detection using graph representation. In: Proceedings of the 29th ACM International Conference on Information and Knowledge Management. CIKM '20 (2020)
62. Nikolov, A., Da San Martino, G., Koychev, I., Nakov, P.: Team_Alex at CheckThat! 2020: Identifying check-worthy tweets with transformer models. In: Cappellato et al. [18]
63. Passaro, L., Bondielli, A., Lenci, A., Marcelloni, F.: UNIPI-NLE at CheckThat! 2020: Approaching fact checking from a sentence similarity perspective through the lens of transformers. In: Cappellato et al. [18]
64. Patwari, A., Goldwasser, D., Bagchi, S.: TATHYA: a multi-classifier system for detecting check-worthy statements in political debates. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. pp. 2259–2262. CIKM '17, Singapore (2017)
65. Pennington, J., Socher, R., Manning, C.: GloVe: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. pp. 1532–1543. EMNLP '14, Doha, Qatar (2014)
66. Popat, K., Mukherjee, S., Strötgen, J., Weikum, G.: 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. pp. 1003–1012. WWW '17, Perth, Australia (2017)
67. Rashkin, H., Choi, E., Jang, J.Y., Volkova, S., Choi, Y.: Truth of varying shades: Analyzing language in fake news and political fact-checking. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 2931–2937. EMNLP '17, Copenhagen, Denmark (2017)
68. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. pp. 3982–3992. EMNLP '19, Hong Kong, China (2019)
69. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. ArXiv preprint 1910.01108 (2019)
70. Shaar, S., Babulkov, N., Da San Martino, G., Nakov, P.: That is a known lie: Detecting previously fact-checked claims. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 3607–3618. ACL '20 (2020)
71. Shu, K., Sliva, A., Wang, S., Tang, J., Liu, H.: Fake news detection on social media: A data mining perspective. SIGKDD Explor. Newsl. **19**(1), 22–36 (2017)
72. Shukla, U., Sharma, A.: Verified claim retrieval: TIET at CLEF CheckThat! 2020. In: Cappellato et al. [18]
73. Song, X., Petrak, J., Jiang, Y., Singh, I., Maynard, D., Bontcheva, K.: Classification aware neural topic model and its application on a new COVID-19 disinformation corpus. arXiv:2006.03354 (2020)

74. Su, T., Macdonald, C., Ounis, I.: 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, CEUR-WS.org, Lugano, Switzerland (2019)
75. Tchechmedjiev, A., Fafalios, P., Boland, K., Gasquet, M., Zloch, M., Zapilko, B., Dietze, S., Todorov, K.: ClaimsKG: A knowledge graph of fact-checked claims. In: Proceedings of the 18th International Semantic Web Conference. pp. 309–324. ISWC '19, Auckland, New Zealand (2019)
76. Thorne, J., Vlachos, A., Christodoulopoulos, C., Mittal, A.: FEVER: a large-scale dataset for fact extraction and verification. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 809–819. NAACL-HLT '18, New Orleans, Louisiana, USA (2018)
77. Thuma, E., Motlogelwa, N.P., Leburu-Dingalo, T., Mudongo, M.: UB_ET at CheckThat! 2020: Exploring ad hoc retrieval approaches in verified claims retrieval. In: Cappellato et al. [18]
78. Vasileva, S., Atanasova, P., Màrquez, L., Barrón-Cedeño, A., Nakov, P.: 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. pp. 1229–1239. RANLP '19 (2019)
79. Wadden, D., Lin, S., Lo, K., Wang, L.L., van Zuylen, M., Cohan, A., Hajishirzi, H.: Fact or fiction: Verifying scientific claims. ArXiv preprint 2004.14974 (2020)
80. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.: GLUE: A multi-task benchmark and analysis platform for natural language understanding. In: Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. pp. 353–355. Brussels, Belgium (2018)
81. Wang, W.Y.: "Liar, liar pants on fire": A new benchmark dataset for fake news detection. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. pp. 422–426. ACL '17, Vancouver, Canada (2017)
82. Williams, A., Nangia, N., Bowman, S.: A broad-coverage challenge corpus for sentence understanding through inference. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 1112–1122. NAACL-HLT '18, New Orleans, Louisiana, USA (2018)
83. Williams, E., Rodrigues, P., Novak, V.: Accenture at CheckThat! 2020: If you say so: Post-hoc fact-checking of claims using transformer-based models. In: Cappellato et al. [18]
84. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R., Le, Q.V.: XLNet: Generalized autoregressive pretraining for language understanding. In: Wallach, H., Larochelle, H., Beygelzimer, A., d' Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems 32, pp. 5753–5763. Curran Associates, Inc. (2019)
85. Zhao, Z., Resnick, P., Mei, Q.: Enquiring minds: Early detection of rumors in social media from enquiry posts. In: Proceedings of the 24th International Conference on World Wide Web. pp. 1395–1405. WWW'15, Florence, Italy (2015)
86. Zhou, X., Mulay, A., Ferrara, E., Zafarani, R.: ReCOVery: A multimodal repository for COVID-19 news credibility research. arXiv:2006.05557 (2020)
87. Zuo, C., Karakas, A., Banerjee, R.: A hybrid recognition system for check-worthy claims using heuristics and supervised learning. In: Cappellato et al. [20]