Overview of the CLEF-2019 CheckThat! Lab: Automatic Identification and Verification of Claims. Task 1: Check-Worthiness*

Pepa Atanasova[1], Preslav Nakov[2], Georgi Karadzhov[3], Mitra Mohtarami[4], and Giovanni Da San Martino[2]

[1] Department of Computer Science, University of Copenhagen, Denmark
pepa@di.ku.dk
[2] Qatar Computing Research Institute, HBKU, Doha, Qatar
{pnakov, gmartino}@hbku.edu.qa
[3] SiteGround Hosting EOOD, Sofia, Bulgaria
georgi.m.karadjov@gmail.com
[4] MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA
mitram@mit.edu

Abstract. We present an overview of the 2nd edition of the CheckThat! Lab, part of CLEF 2019, with focus on Task 1: Check-Worthiness in political debates. The task asks to predict which claims in a political debate should be prioritized for fact-checking. In particular, given a debate or a political speech, the goal is to produce a ranked list of its sentences based on their worthiness for fact-checking. This year, we extended the 2018 dataset with 16 more debates and speeches. A total of 47 teams registered to participate in the lab, and eleven of them actually submitted runs for Task 1 (compared to seven last year). The evaluation results show that the most successful approaches to Task 1 used various neural networks and logistic regression. The best system achieved mean average precision of 0.166 (0.250 on the speeches, and 0.054 on the debates). This leaves large room for improvement, and thus we release all datasets and scoring scripts, which should enable further research in check-worthiness estimation.

Keywords: Computational journalism · Check-worthiness estimation · Fact-checking · Veracity

* This paper only focuses on Task 1 (Check-Worthiness). For an overview of Task 2 (Factuality), see [18].

1 Introduction

The current coverage of the political landscape in both the press and in social media has led to an unprecedented situation. Like never before, a statement in an interview, a press release, a blog note, or a tweet can spread almost instantaneously across the globe. This proliferation speed has left little time for double-checking claims against the facts, which has proven critical.

The problem caught the public attention in connection to the 2016 US Presidential Campaign, which was influenced by fake news in social media and by false claims. Indeed, some politicians were fast to notice that when it comes to shaping public opinion, facts are secondary, and that appealing to emotions and beliefs works better, especially in social media. It has been even proposed that this marked the dawn of a Post-Truth Age.

As the problem became evident, a number of fact-checking initiatives have started, led by organizations such as FactCheck and Snopes, among many others. Yet, this proved to be a very demanding manual effort, which means that only a relatively small number of claims could be fact-checked.[5] This makes it important to prioritize the claims that fact-checkers should consider first. Task 1 of the CheckThat! Lab at CLEF-2019 [10,11] aims to help in that respect, asking participants to build systems that can mimic the selection strategies of a particular fact-checking organization: `factcheck.org`. It is defined as follows:

> *Given a political debate, interview, or speech, transcribed and segmented into sentences, rank the sentences concerning the priority with which they should be fact-checked.*

This is a ranking task and the participating systems are asked to produce one score per sentence, according to which the sentences are to be ranked. This year, Task 1 was offered for English only (it was also offered in Arabic in the 2018 edition of the lab [2]).

The dataset for this task is an extension of the CT-CWC-18 dataset [2]. We added annotations from three press-conferences, six public speeches, six debates, and one post, all fact-checked by experts from `factcheck.org`.

Figure 1 shows examples of annotated debate fragments. In Figure 1a, Hillary Clinton discusses the performance of her husband, Bill Clinton, as US president. Donald Trump fires back with a claim that is worth fact-checking, namely that Bill Clinton approved NAFTA. In Figure 1b, Donald Trump is accused of having filed for bankruptcy six times, which is also a claim that is worth fact-checking. In Figure 1c, Donald Trump claims that border walls work. In a real-world scenario, the intervention by Donald Trump in Figure 1a, the second one by Hillary Clinton in Figure 1b, and the first two by Donald Trump in Figure 1c should be ranked at the top of the list in order to get the attention of the fact-checker.

[5] Full automation is not yet a viable alternative, partly because of limitations of the existing technology, and partly due to low trust in such methods by the users.

H. Clinton:	I think my husband did a pretty good job in the 1990s.	
H. Clinton:	I think a lot about what worked and how we can make it work again...	
D. Trump:	Well, he approved NAFTA...	✓

(a) Fragment from the First 2016 US Presidential Debate.

H. Clinton:	He provided a good middle-class life for us, but the people he worked for, he expected the bargain to be kept on both sides.	
H. Clinton:	And when we talk about your business, you've taken business bankruptcy six times.	✓

(b) Another fragment from the First 2016 US Presidential Debate.

D. Trump:	It's a lot of murders, but it's not close to 2,000 murders right on the other side of the wall, in Mexico.	✓
D. Trump:	So everyone knows that walls work.	✓
D. Trump:	And there are better examples than El Paso, frankly.	

(c) Fragment from Trump's National Emergency Remarks in February 2019.

Fig. 1: English debate fragments: check-worthy sentences are marked with ✓.

The rest of this paper is organized as follows. Section 2 discusses related work. Section 3 describes the evaluation framework and the task setup. Section 4 provides an overview of the participating systems, followed by the official results in Section 5, and discussion in Section 6, before we conclude in Section 7.

2 Related Work

Automatic fact-checking is envisioned in [38] as a multi-step process that includes (*i*) identifying check-worthy statements [14,19,22], (*ii*) generating questions to be asked about these statements [23], (*iii*) retrieving relevant information to create a knowledge base [28,35], and (*iv*) inferring the veracity of the statements, e.g., using text analysis [3,6,34] or external sources [4,5,23,33].

The first work to target check-worthiness was the ClaimBuster system [19]. It was trained on data that was manually annotated by students, professors, and journalists, where each sentence was annotated as *non-factual*, *unimportant factual*, or *check-worthy factual*. The data consisted of transcripts of 30 historical US election debates covering the period from 1960 until 2012 for a total of 28,029 transcribed sentences. The ClaimBuster used an SVM classifier and features such as sentiment, TF.IDF word representations, part-of-speech (POS) tags, and named entities. It did not try to mimic the check-worthiness decisions for any specific fact-checking organization; yet, it was later evaluated against CNN and PolitiFact [20]. In contrast, our dataset is based on actual annotations by a fact-checking organization, and we release freely all data and associated scripts.

More relevant to the setup of Task 1 of this Lab is the work of [14], who focused on debates from the US 2016 Presidential Campaign and used pre-existing annotations from nine respected fact-checking organizations (PolitiFact, FactCheck, ABC, CNN, NPR, NYT, Chicago Tribune, The Guardian, and Washington Post): a total of four debates and 5,415 sentences. Besides many of the features borrowed from ClaimBuster —together with the sentiment, tense, and some other features—, their model pays special attention to the context of each sentence. This includes whether it is part of a long intervention by one of the actors and even its position within such an intervention. The authors predicted both (*i*) whether any of the fact-checking organizations would select the target sentence, and also (*ii*) whether a specific organization would select it.

In follow-up work, [22] developed ClaimRank, which can mimic the claim selection strategies for each and any of the nine fact-checking organizations, as well as for the union of them all. Even though trained on English, it further supports Arabic, which is achieved via cross-language English-Arabic embeddings.

In yet another follow-up work, [37] proposed a multi-task learning neural network that learns from nine fact-checking organizations simultaneously and tries to predict for each sentence whether it would be selected for fact-checking by each of these organizations.

The work of [32] also focused on the 2016 US Election campaign, and they also used data from nine fact-checking organizations (but a slightly different dataset). They used 3 Presidential, one Vice-Presidential, and several primary debates (7 Republican and 8 Democratic) for a total of 21,700 sentences. Their setup asked to predict whether any of the fact-checking sources would select the target sentence. They used a boosting-like model that takes SVMs focusing on different clusters of the dataset, and the final outcome was considered as that coming from the most confident classifier. The features considered ranged from LDA-based topic-modeling to POS tuples and bag-of-words representations.

In the 2018 edition of the task [2,30], we followed a setup similar to that of [14,22,32], but we manually verified the selected sentences, e.g., to adjust the boundaries of the check-worthy claim, and also to include all instances of a selected check-worthy claim (as fact-checkers would only comment on one instance of a claim). We further had an Arabic version of the dataset. Finally, we chose to focus on a single fact-checking organization. The interested reader can also check the system description papers from last year's lab. Notably, the best primary submission last year was that of the *Prise de Fer* team [41], which used a multilayer perceptron and a feature-rich representation.

The present CLEF'2019 Lab is an extension of the CLEF'2018 CheckThat Lab [29], and the subtask of identifying check-worthy claims is an extension of last year's subtask [2]. This year, we provide more training and testing data, and we focus on English only.

Finally, there have been some recent related shared tasks. For example, at SemEval'2019 there was a task focusing on fact-checking in community question answering fora [24,25,31] and another one that targeted rumor detection [15].

Table 1: Total number of sentences and number of check-worthy ones in the CT19-T1 corpus.

Type	Partition	Sentences	Check-worthy
Debates	Train	10,648	256
	Test	4,584	46
Speeches	Train	2,718	282
	Test	1,883	50
Press Conferences	Train	3,011	36
	Test	612	14
Posts	Train	44	8
Total	Train	**16,421**	**433**
	Test	**7,079**	**110**

3 Evaluation Framework

In this section, we describe the evaluation framework, which includes the dataset and the evaluation measure used.

3.1 Data

The dataset for Task 1 is an extension of the CT-CWC-18 dataset [2]. The full English part of CT-CWC-18 (training and test) has become the training data this year. For the new test set, we produced labelled data from three press conferences, six public speeches, six debates, and one post.

As last year, the annotations for the new instances were derived from the publicly available analysis carried out by `factcheck.org`. We considered as check-worthy those claims whose factuality was challenged by the fact-checkers, and we made them positive instances in our CT19-T1 dataset. Note that our annotation is at the sentence level. Therefore, if only part of a sentence was fact-checked, we annotated the entire sentence as a positive instance. If a claim spanned more than one sentence, we annotated all these sentences as positive. Moreover, in some cases, the same claim was made multiple times in a debate/speech, and thus we annotated all these sentences that referred to it rather than only the one that was fact-checked. Finally, we manually refined the annotations by moving them to a neighbouring sentence (e.g., in case of an argument) or by adding/excluding some annotations. Table 1 shows some statistics about the CT19-T1 corpus.

Note that the participating systems were allowed to use external datasets with fact-checking related annotations as well as to extract information from the Web, from social media, etc.

3.2 Evaluation Measures

Recall that we defined Task 1 as an information retrieval problem, where we asked the participating systems to rank the sentences in the input document, so that the check-worthy sentences are ranked at the top of the list. Hence, we use mean average precision (MAP) as the official evaluation measure, which is defined as follows:

$$MAP = \frac{\sum_{d=1}^{D} AveP(d)}{D} \tag{1}$$

where $d \in D$ is one of the debates/speeches, and $AveP$ is the average precision, which in turn is defined as follows:

$$AveP = \frac{\sum_{k=1}^{K}(P(k) \times \delta(k))}{\#\text{ check-worthy claims}} \tag{2}$$

where $P(k)$ refers to the value of precision at rank k and $\delta(k) = 1$ iff the claim at that position is actually check-worthy.

As in the 2018 edition of the task [2], following [14] we further report some other measures: (*i*) mean reciprocal rank (MRR), (*ii*) mean R-Precision (MR-P), and (*iii*) mean precision@k (P@k). Here *mean* refers to macro-averaging over the testing debates/speeches.

4 Overview of Participants' Approaches

Eleven teams took part in Task 1. The most successful approaches relied on training supervised classification models to assign a check-worthiness score to each of the sentences. Some participants tried to model the context of each sentence, e.g., by considering the neighbouring sentences to represent an instance [12,16]. Yet, the most successful systems analyzed each sentence in isolation, ignoring the rest of the input text.

Table 2 shows an overview of the approaches used by the participating systems. While many systems relied on embedding representations, feature engineering was also popular this year. The most popular features were bag-of-words representations, part-of-speech (PoS) tags, named entities (NEs), sentiment analysis, and statistics about word use. Two systems also made use of co-reference resolution. The most popular classifiers included SVM, linear regression, Naïve Bayes, decision trees, and neural networks.

The best performing system achieved a score of 0.166 in terms of MAP. This is a clear improvement over the best score from last year's edition of the task, 0.1332 MAP, but of course the results are not directly comparable as we have a new test set. Apart from the improvement in participants' solutions, the increased performance of the systems can also be attributed to the fact that we provided twice as much data as we did last year, both for training and for testing purposes.

Table 2: Overview of the approaches to Task 1: check-worthiness. The learning model and the representations for the best system [17] are highlighted.

Learning Models	[1]	[8]	[9]	[12]	[13]	[17]	[27]	[36]
Neural Networks								
LSTM		☑				☑		
Feed-forward				☑				
SVM								☑
Naïve Bayes		☑						
Logistic regressor			☑					
Regression trees	☑							

Represent.	[1]	[8]	[9]	[12]	[13]	[17]	[27]	[36]
Embeddings								
PoS			☑					
word			☑			☑	☑	
syntactic dep.							☑	
SUSE				☑				
Bag of ...								
words		☑				☑		☑
n-grams	☑							
NEs	☑					☑		☑
PoS	☑					☑		
Readability				☑				
Synt. n-grams				☑				
Sentiment				☑		☑		
Subjectivity				☑				
Sent. context				☑				
Topics	☑							

Teams

[1] TOBB ETU
[8] UAICS
[9] JUNLP
[12] TheEarthIsFlat
[13] IPIPAN
[17] Copenhagen
[27] é proibido cochilar
[36] Terrier
[–] IIT (ISM) Dhanbad
[–] Factify
[–] nlpir01

Team **Copenhagen** achieved the best overall performance by building upon their approach from 2018 [16,17]. Their system learned dual token embeddings — domain-specific word embeddings and syntactic dependencies—, and used them in an LSTM recurrent neural network. They further pre-trained this network with previous Trump and Clinton debates, and then supervised it weakly with the ClaimBuster system.[6] In their primary submission, they used a contrastive ranking loss (excluded in their contrastive1). For their contrastive2 submission, they concatenated representations for the current and for the previous sentence.

Team **TheEarthIsFlat** [12] trained a feed-forward neural network with two hidden layers, which takes as input Standard Universal Sentence Encoder (SUSE) embeddings [7] for the current sentence as well as for the two previous sentences as a context. In their contrastive1 run, they replaced the embeddings with the Large Universal Sentence Encoder's ones, and in their constrastive2 run, they trained the model for 1,350 epochs rather than for 1,500 epochs.

Team **IPPAN** first extracted various features about the claims, including bag-of-words n-grams, word2vec vector representations [26], named entity types, part-of-speech tags, sentiment scores, and features from statistical analysis of the sentences [13]. Then, they used these features in an L1-regularized logistic regression to predict the check-worthiness of the sentences.

Team **Terrier** represented the sentences using bag-of-words and named entities [36]. They used co-reference resolution to substitute the pronouns by the referring entity/person name. They further computed entity similarity [39] and entity relatedness [40]. For prediction, they used an SVM classifier.

[6] `http://idir.uta.edu/claimbuster/`

Team **UAICS** used a Naïve Bayes classifier with bag-of-words features [8]. In their contrastive submissions, they used other models, e.g., logistic regression.

Team **Factify** used the pre-trained ULMFiT model [21] and fine-tuned it on the training set. They further over-sampled the minority class by replacing words randomly with similar words based on word2vec similarity. They also used data augmentation based on back-translation, where each sentence was translated to French, Arabic and Japanese and then back to English.

Team **JUNLP** extracted various features, including syntactic n-grams, sentiment polarity, text subjectivity, and LIX readability score, and used them to train a logistic regression classifier with high recall [9]. Then, they trained an LSTM model fed with word representations from GloVe and part-of-speech tags. The sentence representations from the LSTM model were concatenated with the extracted features and used for prediction by a fully connected layer, which had high precision. Finally, they averaged the posterior probabilities from both models in order to come up with the final check-worthiness score for the sentence.

Team **nlpir01** extracted features such as tf-idf word vectors, tf-idf PoS vectors, word, character, and PoS tag counts. Then, they used these features in a multilayer perceptron regressor with two hidden layers, each of size 2,000. For their contrastive1 run, they oversampled the minority class, and for their contrastive2 run, they reduced the number of units in each layer to 480.

Team **TOBB ETU** used linguistic features such as named entities, topics extracted with IBM Watson's NLP tools, PoS tags, bigram counts and indicators of the type of sentence to train a multiple additive regression tree [1]. They further decreased the ranks of some sentences using hand-crafted rules. In their contrastive1 run, they added the speaker as a feature, while in their contrastive2 run they used logistic regression.

Team **IIT (ISM) Dhanbad** trained an LSTM-based recurrent neural network. They fed the network with word2vec embeddings and features extracted from constituency parse trees as well as features based on named entities and sentiment analysis.

Team **é proibido cochilar** trained an SVM model on bag-of-words representations of the sentences, after performing co-reference resolution and removing all digits [27]. They further used an additional corpus of labelled claims, which they extracted from fact-checking websites, aiming at having a more balanced training corpus and potentially better generalizations.[7]

Compared to last year, there have been a number of new features introduced such as context features, readability scores, topics, and subjectivity. In terms or representation, we see not only word embeddings (as last year), but also some new representations based on PoS, syntactic dependency, and SUSE embeddings. Moreover, the participants this year actively explored ways of using external data and re-ranking techniques, with systems going beyond simple classification and introducing specialized ranking losses and regression models.

[7] Their claim crawling tool: `http://github.comx/vwoloszyn/fake_news_extractor`

Table 3: Results for Task 1: Check-worthiness. The results for the primary submission appear next to the team's identifier, followed by the contrastive submissions (if any). The subscript numbers indicate the rank of each primary submission with respect to the corresponding evaluation measure.

	Team	MAP	RR	R-P	P@1	P@3	P@5	P@10	P@20	P@50
[17]	**Copenhagen**	**.1660**$_1$	.4176$_3$	.1387$_4$	.2857$_2$	**.2381**$_1$	**.2571**$_1$	.2286$_2$	.1571$_2$	.1229$_2$
	contr.-1	.1496	.3098	.1297	.1429	.2381	.2000	.2000	.1429	.1143
	contr.-2	.1580	.2740	.1622	.1429	.1905	.2286	.2429	.1786	.1200
[12]	**TheEarthIsFlat**	.1597$_2$	.1953$_{11}$	**.2052**$_1$	.0000$_4$	.0952$_3$	.2286$_2$	.2143$_3$	**.1857**$_1$	**.1457**$_1$
	contr.-1	.1453	.3158	.1101	.2857	.2381	.1429	.1429	.1357	.1171
	contr.-2	.1821	.4187	.1937	.2857	.2381	.2286	.2286	.2143	.1400
[13]	**IPIPAN**	.1332$_3$	.2864$_6$	.1481$_2$	.1429$_3$	.0952$_3$	.1429$_5$	.1714$_5$	.1500$_3$	.1171$_3$
[36]	**Terrier**	.1263$_4$	.3253$_5$	.1088$_8$	.2857$_2$	**.2381**$_1$	.2000$_3$	.2000$_4$	.1286$_6$	.0914$_7$
[8]	**UAICS**	.1234$_5$	**.4650**$_1$	.1460$_3$	**.4286**$_1$	**.2381**$_1$	.2286$_2$	**.2429**$_1$	.1429$_4$	.0943$_6$
	contr.-1	.0649	.2817	.0655	.1429	.2381	.1429	.1143	.0786	.0343
	contr.-2	.0726	.4492	.0547	.4286	.2857	.1714	.1143	.0643	.0257
	Factify	.1210$_6$	.2285$_8$	.1292$_5$	.1429$_3$	.0952$_3$	.1143$_6$	.1429$_6$	.1429$_4$	.1086$_4$
[9]	**JUNLP**	.1162$_7$	.4419$_2$	.1128$_7$	.2857$_2$	.1905$_2$	.1714$_4$	.1714$_5$	.1286$_6$	.1000$_5$
	contr.-1	.0976	.3054	.0814	.1429	.2381	.1429	.0857	.0786	.0771
	contr.-2	.1226	.4465	.1357	.2857	.2381	.2000	.1571	.1286	.0886
	nlpir01	.1000$_8$	.2840$_7$	.1063$_9$	.1429$_3$	**.2381**$_1$	.1714$_4$	.1000$_8$	.1214$_7$	.0943$_6$
	contr.-1	.0966	.3797	.0849	.2857	.1905	.2286	.1429	.1071	.0886
	contr.-2	.0965	.3391	.1129	.1429	.2381	.2286	.1571	.1286	.0943
[1]	**TOBB ETU**	.0884$_9$	.2028$_{10}$	.1150$_6$	.0000$_4$	.0952$_3$	.1429$_5$	.1286$_7$	.1357$_5$	.0829$_8$
	contr.-1	.0898	.2013	.1150	.0000	.1429	.1143	.1286	.1429	.0829
	contr.-2	.0913	.3427	.1007	.1429	.1429	.1143	.0714	.1214	.0829
	IIT (ISM) Dhanbad	.0835$_{10}$	.2238$_9$	.0714$_{11}$	.0000$_4$	.1905$_2$	.1143$_6$	.0857$_9$	.0857$_9$	.0771$_9$
[27]	**é proibido cochilar**	.0796$_{11}$	.3514$_4$	.0886$_{10}$	.1429$_3$	**.2381**$_1$	.1429$_5$	.1286$_7$	.1071$_8$	.0714$_{10}$
	contr.-1	.1357	.5414	.1595	.4286	.2381	.2571	.2714	.1643	.1200

5 Evaluation

The participants were allowed to submit one primary and up to two contrastive runs in order to test variations of their primary models or entirely different alternative models. Only the primary runs were considered for the official ranking. A total of eleven teams submitted 21 runs. Table 3 shows the results.

The best-performing system was the one by team **Copenhagen**. They achieved a strong MAP score using a ranking loss based on contrastive sampling. Indeed, this is the only team that modelled the task as a ranking one and the decrease in the performance without the ranking loss (see their contrastive1 run) shows the importance of using this loss.

Two teams made use of external datasets: team **Copenhagen** used a weakly supervised dataset for pretraining, and team **é proibido cochilar** included claims scraped from several fact-checking Web sites.

In order to address the class imbalance in the training dataset, team **nlpir01** used oversampling in their contrastive1 run, but could not gain any improvements. Oversampling and augmenting with additional data points did not help team **é proibido cochilar** either.

Many systems used pretrained sentence or word embedding models. Team **TheEarthIsFlat**, which has the second-best performing system, used the Standard Universal Sentence Embeddings, which performed well on the task. The best MAP score overall was achieved by the contrastive2 run by this team: the only difference with respect to their primary submission was the number of training epochs. Some teams also used fine-tuning, e.g., team **Factify** fine-tuned the ULMFiT model on the training dataset.

Interestingly, the top-performing run was an unofficial one, namely the contrastive2 run by the *TheEarthIsFlat* team [12]. As described in Section 4, this model consisted of a feed-forward neural network fed with Standard Universal Sentence Encoder embeddings. The only difference with their primary run is the number of epochs they trained the network for.

6 Discussion

Debates vs. Speeches While the training data included debates only, the test data also contained speeches. Thus, it is interesting to see how the systems perform on debates vs. speeches. Table 4 shows the MAP for the primary submissions. This year again, the performance on speeches was better than on debates. The best MAP on speeches last year was .1460, while on debates it was .1011. We can see that thus year the performance on speeches improved by more than 10% absolute, while the performance on debates decreased by almost 5%. We are not sure why this should be the case, but the speeches in our test dataset contain about twice as many check-worthy claims as there are in the debates (see Table 1).

Table 4: MAP for the primary submissions for debates vs. speeches.

Team	Debates	Speeches
[17] **Copenhagen**	$.0538_{2}$	$\underline{\mathbf{.2502}}_{1}$
[12] **TheEarthIsFlat**	$.0487_{3}$	$.2430_{2}$
[13] **IPIPAN**	$\underline{\mathbf{.0632}}_{1}$	$.1858_{5}$
[36] **Terrier**	$.0210_{11}$	$.2053_{3}$
[8] **UAICS**	$.0235_{10}$	$.1983_{4}$
Factify	$.0437_{4}$	$.1790_{6}$
[9] **JUNLP**	$.0387_{5}$	$.1743_{7}$
nlpir01	$.0329_{8}$	$.1504_{8}$
[1] **TOBB ETU**	$.0314_{9}$	$.1311_{9}$
IIT (ISM) Dhanbad, India	$.0364_{6}$	$.1188_{10}$
[27] **é proibido cochilar**	$.0351_{7}$	$.1130_{11}$

Ensembles We further experimented with constructing an ensemble using the scores by the individual systems. In particular, we first performed min-max normalization of the predictions of the individual systems, and then we summed these normalized scores.[8] The overall results are shown in Table 5. We can see that there is a small improvement for the ensemble over the best individual system in terms of MAP (compare the first to the last line). The results for the other evaluation measures are somewhat mixed, which might be due to contradictions or duplication of the information coming from the different systems.

Ablation Table 5 further shows the results for ablation experiments, where we add one system to the ensemble at a time. We can see that the ensemble of the top-4 systems works best. However, as we keep adding systems, the score does not always improve. This can be due to various reasons, e.g., the noise and the wrong signal we get from the systems can harm instead of improve the final score. The best combination of teams in terms of MAP is shown on the last line of the table, and it includes the teams Copenhagen, TheEarthIsFlat and Terrier.

Table 5: Ablation results from adding one team at a time to an ensemble. The ensemble is the sum of the score by each of the teams in the ensemble. The order of adding the teams in the ensemble is based on their official rank.

Team	**MAP**	RR	R-P	P@1	P@3	P@5	P@10	P@20	P@50
Copenhagen (best team)	.1660	.4176	.1387	**.2857**	.2381	**.2571**	**.2286**	.1571	.1229
+ TheEarthIsFlat	.1694	.3959	.1613	**.2857**	.1905	.2000	**.2286**	.1714	.1286
+ IPIPAN	.1647	.3681	.1548	.1429	.1905	**.2571**	.2000	.1714	**.1314**
+ Terrier	**.1707**	**.4474**	.1694	**.2857**	**.2857**	**.2571**	.2143	.1929	.1286
+ UAICS	.1601	.4025	.1557	**.2857**	.2381	.2286	.2143	.1714	.1257
+ Factify	.1605	.3983	.1655	**.2857**	.2381	.2286	.2143	.1571	.1286
+ JUNLP	.1547	.3211	.1873	.1429	.2381	**.2571**	.2000	.1929	.1229
+ nlpir01	.1538	.3376	.1930	.1429	.2381	.2286	.1714	**.2071**	.1257
+ TOBB ETU	.1530	.3386	**.1918**	.1429	.2381	.2286	.1714	.2000	.1257
+ IIT (ISM) Dhanbad, India	.1476	.3148	**.1918**	.1429	.1905	.2000	.1714	.1857	.1257
+ é proibido cochilar	.1514	.4585	.1750	**.2857**	.2381	**.2571**	.2143	.1786	.1200
Copenhagen + TheEarthIsFlat + Terrier	**.1747**	.3771	.1877	**.2857**	.1905	.2000	**.2429**	.1929	.1257

7 Conclusion and Future Work

We have presented an overview of the CLEF–2019 CheckThat! Lab Task 1 on Automatic Identification of Claims. The task asked the participating teams to predict which claims in a political debate should be prioritized for fact-checking. As part of the CheckThat! lab, we release the dataset and the evaluation tools in order to enable further research in check-worthiness estimation.[9]

[8] We also tried summing the reciprocal ranks of the rankings that the systems assigned to each sentence, but this yielded much worse results.

[9] `http://github.com/apepa/clef2019-factchecking-task1`

A total of 11 teams participated in task 1 (compared to 7 in 2018). The evaluation results show that the most successful approaches used various neural networks and logistic regression.

In future work, we want to expand the dataset with more annotations, which should enable more interesting neural network architectures. We also plan to include information from multiple fact-checking organizations. As noted in [14], the agreement between the selection choices for different fact-checking organizations is low, meaning that there is a certain bias in the selection of claims by each of the fact-checking organizations, and aggregating the annotations from multiple sources could potentially help in that respect. It would further enable multi-task learning [37]

Acknowledgments

We want to thank Spas Kyuchukov and Oliver Ren for taking part in the annotation process of the new speeches and debates for this year's extended dataset.

This work is part of the Tanbih project,[10] which aims to limit the effect of "fake news", propaganda and media bias by making users aware of what they are reading. The project is developed in collaboration between the Qatar Computing Research Institute (QCRI), HBKU and the MIT Computer Science and Artificial Intelligence Laboratory (CSAIL).

References

1. 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)
2. 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)
3. 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)
4. 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)
5. 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

[10] `http://tanbih.qcri.org/`

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)
6. 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)
7. Cer, D., Yang, Y., Kong, S.y., Hua, N., Limtiaco, N., John, R.S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., et al.: Universal sentence encoder. arXiv preprint arXiv:1803.11175 (2018)
8. Coca, L., Cusmuliuc, C.G., Iftene, A.: 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)
9. 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)
10. 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, Springer International Publishing (2019)
11. 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. LNCS, Lugano, Switzerland (2019)
12. 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)
13. 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)
14. 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)
15. Gorrell, G., Aker, A., Bontcheva, K., Derczynski, L., Kochkina, E., Liakata, M., Zubiaga, A.: SemEval-2019 task 7: RumourEval, determining rumour veracity and support for rumours. In: Proceedings of the 13th International Workshop on Semantic Evaluation. pp. 845–854. Minneapolis, Minnesota, USA (2019)
16. Hansen, C., Hansen, C., Alstrup, S., Simonsen, J.G., Lioma, C.: Neural check-worthiness ranking with weak supervision: Finding sentences for fact-checking. arXiv preprint arXiv:1903.08404 (2019)
17. 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)

18. 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, L., Ferro, N., Losada, D., Müller, H. (eds.) 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)
19. Hassan, N., Li, C., Tremayne, M.: Detecting check-worthy factual claims in presidential debates. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. pp. 1835–1838. CIKM '15 (2015)
20. Hassan, N., Tremayne, M., Arslan, F., Li, C.: Comparing automated factual claim detection against judgments of journalism organizations. In: Computation + Journalism Symposium. Stanford, California, USA (2016)
21. Howard, J., Ruder, S.: Universal language model fine-tuning for text classification. arXiv preprint arXiv:1801.06146 (2018)
22. 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)
23. 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)
24. Mihaylova, T., Karadzhov, G., Atanasova, P., Baly, R., Mohtarami, M., Nakov, P.: SemEval-2019 task 8: Fact checking in community question answering forums. In: Proceedings of the 13th International Workshop on Semantic Evaluation. pp. 860–869 (2019)
25. Mihaylova, T., Nakov, P., Màrquez, L., Barrón-Cedeño, A., Mohtarami, M., Karadjov, G., Glass, J.: Fact checking in community forums. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence. pp. 5309–5316. AAAI '18, New Orleans, Louisiana, USA (2018)
26. Mikolov, T., Yih, W.t., 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)
27. 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)
28. 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)
29. 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)

30. Nakov, P., Barrón-Cedeño, A., Elsayed, T., Suwaileh, R., Màrquez, L., Zaghouani, W., Gencheva, P., Kyuchukov, S., Da San Martino, G.: 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)
31. Nakov, P., Mihaylova, T., Màrquez, L., Shiroya, Y., Koychev, I.: Do not trust the trolls: Predicting credibility in community question answering forums. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing. pp. 551–560. RANLP '17, Varna, Bulgaria (2017)
32. 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)
33. 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)
34. 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)
35. Shiralkar, P., Flammini, A., Menczer, F., Ciampaglia, G.L.: Finding streams in knowledge graphs to support fact checking. In: Proceedings of the IEEE International Conference on Data Mining. pp. 859–864. ICDM '17, New Orleans, Louisiana, USA (2017)
36. 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)
37. 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. RANLP '19, Varna, Bulgaria (2019)
38. Vlachos, A., Riedel, S.: Fact checking: Task definition and dataset construction. In: Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science. pp. 18–22. Baltimore, Maryland, USA (2014)
39. Zhu, G., Iglesias, C.A.: Computing semantic similarity of concepts in knowledge graphs. IEEE Transactions on Knowledge and Data Engineering **29**(1), 72–85 (2016)
40. Zhu, G., Iglesias, C.A.: Sematch: Semantic entity search from knowledge graph. In: Cheng, G., Gunaratna, K., Thalhammer, A., Paulheim, H., Voigt, M., García, R. (eds.) Joint Proceedings of the 1st International Workshop on Summarizing and Presenting Entities and Ontologies and the 3rd International Workshop on Human Semantic Web Interfaces. CEUR Workshop Proceedings, vol. 1556. CEUR-WS.org, Portoroz, Slovenia (2015)
41. Zuo, C., Karakas, A., Banerjee, R.: A hybrid recognition system for check-worthy claims using heuristics and supervised learning. In: Cappellato, L., Ferro, N., Nie, J., Soulier, L. (eds.) Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018. CEUR Workshop Proceedings, vol. 2125. CEUR-WS.org (2018)