Overview of the CLEF-2018 CheckThat! Lab on Automatic Identification and Verification of Political Claims. Task 1: Check-Worthiness*

Pepa Atanasova[1], Lluís Màrquez[2], Alberto Barrón-Cedeño[3],
Tamer Elsayed[4], Reem Suwaileh[4], Wajdi Zaghouani[5],
Spas Kyuchukov[6], Giovanni Da San Martino[3], and Preslav Nakov[3]

[1] SiteGround, Sofia, Bulgaria
pepa.gencheva@siteground.com
[2] Amazon, Barcelona, Spain
lluismv@amazon.com
[3] Qatar Computing Research Institute, HBKU, Doha, Qatar
{albarron, gmartino, pnakov}@qf.org.qa
[4] Computer Science and Engineering Department, Qatar University, Doha, Qatar
{telsayed, reem.suwaileh}@qu.edu.qa
[5] College of Humanities and Social Sciences, HBKU, Doha, Qatar
wzaghouani@hbku.edu.qa
[6] Sofia University "St Kliment Ohridski", Sofia, Bulgaria
spas.kyuchukov@gmail.com

Abstract. We present an overview of the CLEF-2018 CheckThat! Lab on Automatic Identification and Verification of Political Claims, with focus on Task 1: Check-Worthiness. 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 was to produce a ranked list of its sentences based on their worthiness for fact checking. We offered the task in both English and Arabic, based on debates from the 2016 US Presidential Campaign, as well as on some speeches during and after the campaign. A total of 30 teams registered to participate in the Lab and seven teams actually submitted systems for Task 1. The most successful approaches used by the participants relied on recurrent and multi-layer neural networks, as well as on combinations of distributional representations, on matchings claims' vocabulary against lexicons, and on measures of syntactic dependency. The best systems achieved mean average precision of 0.18 and 0.15 on the English and on the Arabic test datasets, respectively. This leaves large room for further improvement, and thus we release all datasets and the scoring scripts, which should enable further research in check-worthiness estimation.

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

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

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 in politics, e.g., during 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 were secondary, and that appealing to emotions and beliefs worked better, especially in social media. It has been even proposed that this was marking 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 has proved to be a very demanding manual effort, which means that only a relatively small number of claims could be fact-checked.[7] This makes it important to prioritize the claims that fact-checkers should consider first. Task 1 of the CheckThat! Lab at CLEF-2018 [17] 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 transcription of a political debate/speech, predict which claims should be prioritized for fact-checking.*

The goal is to produce a *ranked list* of sentences ordered by their worthiness for fact-checking. This is the first step in the pipeline of the full fact-checking process, displayed in Figure 1. Refer to [2] for details on the fact-checking task.

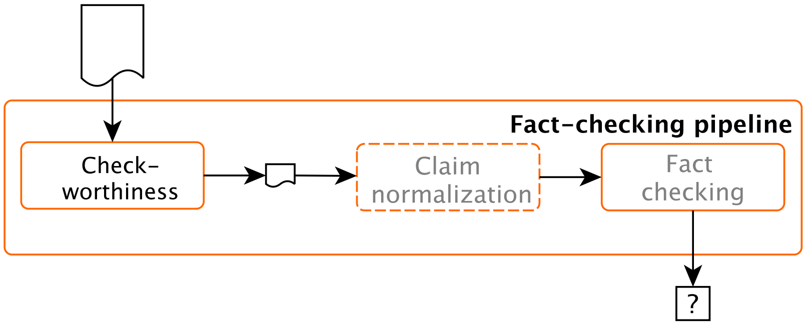

Fig. 1: The general fact-checking pipeline. First, the input document is analyzed to identify sentences containing check-worthy claims (this task), then these claims are extracted and normalized, and finally they are fact-checked.

[7] 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 human users.

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

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

Hillary Clinton:	Take clean energy	
Hillary Clinton:	Some country is going to be the clean-energy superpower of the 21st century.	
Hillary Clinton:	Donald thinks that climate change is a hoax perpetrated by the Chinese.	✓
Hillary Clinton:	I think it's real.	
Donald Trump:	I did not.	

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

Fig. 2: English debate fragments: check-worthy sentences are marked with ✓.

We offered the task in two languages, English and Arabic. Figure 2 shows examples of English debate fragments. In example 2a, Hillary Clinton discusses the performance of her husband Bill Clinton while he was US president. Donald Trump fires back with a claim that is worth fact-checking: that Bill Clinton approved NAFTA. In example 2b, whether Donald Trump thinks about climate change as charged by Hillary Clinton is also worth fact-checking.

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

Journalists, online users, and researchers are well aware of the proliferation of false information. For example, there was a 2016 special issue of the ACM Transactions on Information Systems journal on Trust and Veracity of Information in Social Media [18], and there is a Workshop on Fact Extraction and Verification at EMNLP'2018. Moreover, there have been several related shared tasks, e.g., a SemEval-2017 shared task on Rumor Detection [5], an ongoing FEVER challenge on Fact Extraction and VERification at EMNLP'2018, the present CLEF'2018 Lab on Automatic Identification and Verification of Claims in Political Debates, and an upcoming task at SemEval'2019 on Fact-Checking in Community Question Answering Forums.

Automatic fact-checking was envisioned in [23] as a multi-step process that includes (*i*) identifying check-worthy statements [8, 12, 14], (*ii*) generating questions to be asked about these statements [15], (*iii*) retrieving relevant information to create a knowledge base [22], and (*iv*) inferring the veracity of the statements, e.g., using text analysis [4, 21] or external sources [15, 20].

The first work to target check-worthiness was the ClaimBuster system [12]. 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 historical US election debates covering the period from 1960 until 2012 for a total of 30 debates and 28,029 transcribed sentences. In each sentence, the speaker was marked: candidate vs. moderator. The ClaimBuster used an SVM classifier and a manifold of features such as sentiment, TF.IDF word representations, part-of-speech (POS) tags, and named entities. It produced a check-worthiness ranking on the basis of the SVM prediction scores. The ClaimBuster system did not try to mimic the check-worthiness decisions for any specific fact-checking organization; yet, it was later evaluated against CNN and PolitiFact [13]. In contrast, our dataset is based on actual annotations by a fact-checking organization, and we release freely all data and associated scripts (while theirs is not available).

More relevant to the setup of Task 1 of this Lab is the work of [8], 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. Beside many of the features borrowed from ClaimBuster —together with 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 one would select it.

In follow-up work, [14] 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.

The work of [19] also focused on the 2016 US Election campaign, and they also used data from nine fact-checking organizations (but slightly different set from above). They used presidential (3 presidential one vice-presidential) and 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 topic-modeling to POS tuples and bag-of-words representations.

We follow a setup that is similar to that of [8, 14, 19], but we manually verify 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 have an Arabic version of the dataset. Finally, we chose to focus on a single fact-checking organization.

3 Evaluation Framework

3.1 Data

For Task 1, we produced the CT-CWC-18 dataset,[8] which stands for CheckThat! Check-Worthiness 2018 corpus. It includes transcripts from the 2016 US Presidential campaign, together with some more recent political speeches. In order to derive the annotation, we used the publicly available analysis carried out by FactCheck.org.[9] We considered those claims whose factuality was challenged by the fact-checkers as check-worthy and we made them positive instances in the 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 example. 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 the one that was annotated by the fact-checkers. Finally, we manually refined the annotations by moving them to a neighboring sentence (e.g., in case of argument) or by adding/excluding some annotations.

As shown in Table 1, the English CT-CWC-18 is comprised of five debates and five speeches. To produce Arabic data, we hired translators to translate five debates and Donald Trump's acceptance speech. We released the first three debates as training data, and we used the remaining debates/speeches for testing.

	Type	**Partition**	**Sent.**	**CW**
	Debates			
文A	1st Presidential	train	1,403	37
文A	2nd Presidential	train	1,303	25
文A	Vice-Presidential	train	1,358	28
文A	3rd Presidential	test	1,351	77
文A	9th Democratic	test	1,464	17
	Speeches			
文A	Donald Trump Acceptance	test	375	21
	Donald Trump at the World Economic Forum	test	245	11
	Donald Trump at a Tax Reform Event	test	412	16
	Donald Trump's Address to Congress	test	390	15
	Donald Trump's Miami Speech	test	645	35
Total English			**8,946**	**282**
Total Arabic			**7,254**	**205**

Table 1: Total number of sentences and those identified as check-worthy (CW) in CT-CWC-18 . The documents available in Arabic are marked with 文A.

[8] http://github.com/clef2018-factchecking/clef2018-factchecking

[9] See for example, http://transcripts.factcheck.org/presidential-debate-hofstra-university-hempstead-new-york/

Note that it was forbidden to use external datasets with fact-checking related annotations. However, it was allowed to extract information from the Web, from Twitter, etc., but the retrieved URLs had to be checked for sanity using a script that we provided to the participants. The script tried to make sure no information from fact-checking websites would be used.

3.2 Evaluation Measures

As we shaped this task as an information retrieval problem, in which check-worthy instances should be ranked at the top of the list, we opted for using mean average precision as the official evaluation measure. It 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:

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

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

Following [8], we further report the results for 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

Table 2 offers a summary of the used approaches and representations; see the system description papers for more detail.

Prise de Fer [25] normalized the texts, e.g., by unifying the speakers' names, and also created additional datasets out of the provided debates by collecting the sentences by a single participant in the debate, thus mimicking speeches. They used averaged word embeddings and bag-of-words representations, after stemming and stopword removal. They also considered the number of negations, verbal forms, as well as clauses and phrases and named entities, among other features. Their prediction model comes in the form of either a multilayer perceptron or a support vector machine. In any case, the decisions made by the model can be overridden by a number of heuristic rules that take into account the length of the intervention or the appearance of certain phrases such as "thank you" or a question mark.

Copenhagen [11] used a recurrent neural network. Their input consists of a combination of word2vec embeddings [16], part of speech tags, and syntactic dependencies. These representations are fed to a GRU neural network with attention. They further combined their approach with that proposed in [8]. This combination boosted their performance on cross-validation, but their neural network alone performed better on the test dataset.

Learning Models	[1]	[9]	[11]	[24]	[25]
Recurrent neural nets			✓		
Multilayer perceptron					✓
Support vector machines	✓				✓
Random forest	✓				
k-nearest neighbors		✓			
Gradient boosting				✓	

Teams

[1] RNCC
[9] UPV-INAOE-Autoritas
[11] Copenhagen
[24] bigIR
[25] Prise de Fer
[–] fragarach
[–] blue

Representations	[1]	[9]	[11]	[24]	[25]
Bag of words					✓
Character n-grams		✓			
Part of speech tags			✓	✓	✓
Verbal forms					✓
Negations					✓
Named entities				✓	✓
Sentiment				✓	✓
Topics				✓	
IR nutritional labels	✓				
Clauses					✓
Syntactic dependency			✓		✓
Word embeddings			✓	✓	✓

Table 2: Summary of the models and representations used by the participants.

bigIR [24]used a learning-to-rank approach based on the MART algorithm [6]. Their features are organized in five families: (i) word embeddings, and binary features expressing the presence of (ii) different types of named entities, (iii) part-of-speech tags, (iv) sentiment labels, and (v) topics. Moreover, they over-sampled the positive instances in the training set in order to alleviate the impact of class imbalance.

UPV-INAOE-Autoritas [9] used a k-nearest neighbors classifier. Their representation is based on character n-grams, after removing irrelevant contents by means of text distortion [10]. Regardless of the outcome of the distortion model, words were retained if they were part of named entities or were found in some linguistic lexicons.

RNCC [1] used support vector machines with different kernels as well as random forests. Their representations are a subset of the values included in the so-called information retrieval nutritional labels of [7], which they trained on various datasets.

Two of the participating teams did not submit system description papers, and below we describe their systems based on the limited information that they provided as a short description at system submission time:

The **fragarach** team, from the Faculty of Mathematics and Informatics, Sofia University, used a linear SVM with a variety of features including averaged word embeddings, sentence length, average length of the words, number of punctuation marks, number of stop words, positive/negative sentiment, and part of speech tags. They further performed feature selection to be able to focus on the most promising words and n-grams.

The **blue** team, from the Indian Institute of Technology Kharagpur, used an LSTM with 100-hidden dimensions with attention, taking the five sentences that preceded the target sentence as context.

	MAP	MRR	MR-P	MP@1	MP@3	MP@5	MP@10	MP@20	MP@50
Prise de Fer [25]									
primary	**.1332**$_{(1)}$	**.4965**$_{(1)}$	**.1352**$_{(1)}$	**.4286**$_{(1)}$	**.2857**$_{(1)}$	.2000$_{(2)}$	.1429$_{(3)}$	**.1571**$_{(1)}$	.1200$_{(2)}$
cont. 1	.1366	.5246	.1475	.4286	.2857	.2286	.1571	.1714	.1229
cont. 2	.1317	.4139	.1523	.2857	.1905	.1714	.1571	.1571	.1429
Copenhagen [11]									
primary	.1152$_{(2)}$	.3159$_{(5)}$	.1100$_{(5)}$	.1429$_{(3)}$	.1429$_{(4)}$	.1143$_{(3)}$	.1286$_{(4)}$	.1286$_{(2)}$	**.1257**$_{(1)}$
cont. 1	.1810	.6224	.1875	.5714	.4286	.3143	.2571	.2357	.1514
UPV–INAOE–Autoritas [9]									
primary	.1130$_{(3)}$	.4615$_{(2)}$	.1315$_{(2)}$	.2857$_{(2)}$	.2381$_{(2)}$	**.3143**$_{(1)}$	**.2286**$_{(1)}$	.1214$_{(3)}$	.0886$_{(4)}$
cont. 1	.1232	.3451	.1022	.1429	.2857	.2286	.1429	.1143	.0771
cont. 2	.1253	.5535	.0849	.4286	.4286	.2571	.1429	.1286	.0771
bigIR [24]									
primary	.1120$_{(4)}$	.2621$_{(6)}$	.1165$_{(4)}$	.0000$_{(4)}$	.1429$_{(4)}$	.1143$_{(3)}$	.1143$_{(5)}$	.1000$_{(5)}$	.1114$_{(3)}$
cont. 1	.1319	.2675	.1505	.1429	.0952	.0857	.1714	.1786	.1343
cont. 2	.1116	.2195	.1294	.0000	.1429	.1429	.1857	.1429	.0886
fragarach									
primary	.0812$_{(5)}$	.4477$_{(3)}$	.1217$_{(3)}$	.2857$_{(2)}$	.1905$_{(3)}$	.2000$_{(2)}$	.1571$_{(2)}$	.1071$_{(4)}$	.0743$_{(5)}$
blue									
primary	.0801$_{(6)}$	.2459$_{(7)}$	.0576$_{(7)}$	.1429$_{(3)}$	.0952$_{(5)}$	.0571$_{(4)}$	.0571$_{(6)}$	.0857$_{(6)}$	.0600$_{(6)}$
RNCC [1]									
primary	.0632$_{(7)}$	.3775$_{(4)}$	.0639$_{(6)}$	.2857$_{(2)}$	.1429$_{(4)}$	.1143$_{(3)}$	.0571$_{(6)}$	.0571$_{(7)}$	.0486$_{(7)}$
cont. 1	.0886	.4844	.0945	.4286	.1429	.1714	.1286	.1000	.0714
cont. 2	.0747	.2198	.0984	.0000	.0952	.1143	.1000	.1000	.0829
Baselines									
n-gram	.1201	.4087	.1280	.1429	.2857	.1714	.1571	.1357	.1143
random	.0485	.0633	.0359	.0000	.0000	.0000	.0286	.0214	.0429

Table 3: English results, ranked based on MAP, the official evaluation measure. The best score per evaluation measure is shown in bold.

5 Evaluation Results

The participants were allowed to submit one primary and up to two contrastive runs in order to test variations or alternative models. For ranking purposes, only the primary submissions were considered. A total of seven teams submitted runs for English, and two of them also did so for Arabic.

English. Table 3 shows the results for English. The best primary submission was that of the *Prise de Fer* team [25], which used a multilayer perceptron and a feature-rich representation. We can see that they had the best overall performance not only on the official MAP measure, but also on six out of nine evaluation measures (and they were 2nd or 3rd on the rest).

	MAP	MRR	MR-P	MP@1	MP@3	MP@5	MP@10	MP@20	MP@50
bigIR [24]									
primary	**.0899**$_{(1)}$	.1180$_{(2)}$	**.1105**$_{(1)}$	.0000$_{(2)}$	.0000$_{(2)}$	.0000$_{(2)}$	**.1333**$_{(1)}$	**.1000**$_{(1)}$	**.1133**$_{(1)}$
cont. 1	.1497	.2805	.1760	.0000	.3333	.3333	.2667	.2333	.1533
cont. 2	.0962	.1660	.0895	.0000	.1111	.2000	.1667	.1000	.0867
UPV–INAOE–Autoritas [9]									
primary	.0585$_{(2)}$	**.3488**$_{(1)}$	.0087$_{(2)}$	**.3333**$_{(1)}$	**.1111**$_{(1)}$	**.0667**$_{(1)}$	.0333$_{(2)}$	.0167$_{(2)}$	.0400$_{(2)}$
cont. 1	.1168	.6714	.0649	.6667	.6667	.4000	.2000	.1000	.0733
Baselines									
n-gram	.0861	.2817	.0981	.0000	.3333	.2667	.1667	.1667	.0867
random	.0460	.0658	.0375	.0000	.0000	.0000	.0333	.0167	.0333

Table 4: Arabic results, ranked based on MAP, the official evaluation measure. The best score per evaluation measure is shown in bold.

Interestingly, the top-performing run for English was an unofficial one, namely the contrastive 1 run by the *Copenhagen* team [11]. As described in Section 4, this model consisted of a recurrent neural network on three representations. They submitted a system that combined their neural network with the model of [8] as their primary submission, but their neural network alone (submitted as contrastive 1), performed better on the test set. This can be due to the model of [8] relying on structural information, which was not available for the speeches included in the test set (cf. Section 3.1).

To put these results in perspective, the bottom of Table 3 shows the results for two baselines: (*i*) a random permutation of the input sentences, and (*ii*) an n-gram based classifier. We can see that all systems managed to outperform the *random* baseline on all measures by a margin. However, only two runs managed to beat the *n-gram* baseline: the primary run of the *Prise de Fer* team, and the contrastive 1 run of the *Copenhagen* team.

Arabic. Only two teams participated in the Arabic task [9, 24], using basically the same models that they had for English. The *bigIR* [24] team translated automatically the test input to English and then ran their English system, while *UPV–INAOE–Autoritas* translated to Arabic the English lexicons their representation was based on, and then trained an Arabic system on the Arabic training data, which they finally ran on the Arabic test input. It is worth noting that for English *UPV–INAOE–Autoritas* outperformed *bigIR*, but for Arabic it was the other way around. We suspect that a possible reason might be the direction of machine translation and also the presence/lack of context. On one hand, translation into English tends to be better than into Arabic. Moreover, the translation of sentences is easier as there is context, whereas such a context is missing when translating lexicon entries in isolation.

Finally, similarly to English, all runs managed to outperform the *random* baseline by a margin, while the *n-gram* baseline was strong yet possible to beat.

English		
Team	**Debates**	**Speeches**
[25] Prise de Fer	.1011$_{(1)}$	.1460$_{(1)}$
[11] Copenhagen	.0757$_{(2)}$	.1310$_{(3)}$
[9] UPV–INAOE–Aut.	.0521$_{(4)}$	.1373$_{(2)}$
[24] bigIR	.0693$_{(3)}$	.1290$_{(4)}$
fragarach	.0512$_{(5)}$	.0932$_{(5)}$
blue	.0506$_{(6)}$	.0920$_{(6)}$
[1] RNCC	.0417$_{(7)}$	.0717$_{(7)}$

Arabic		
Team	**Debates**	**Speeches**
[24] bigIR	.0650$_{(1)}$	.1397$_{(1)}$
[9] UPV–INAOE–Aut.	.0461$_{(2)}$	.0834$_{(2)}$

Table 5: MAP for the primary submissions for debates vs. speeches.

6 Discussion

While the training data included debates only, the test data also contained speeches. Thus, it is interesting to see how systems perform on debates vs. speeches. Table 5 shows the MAP for the primary submissions for both English and Arabic. Interestingly, speeches turn out to be easier than debates. We are not sure why this should be the case, but it might be because the speeches in our test dataset have about twice as many check-worthy claims as there are in the debates (see Table 1).

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.[10] The results are shown in Table 6. We can see that there is small improvement for the ensemble over the best individual system in terms of MAP for both English and Arabic. The results for the other evaluation measures are somewhat mixed for English, but there is clear improvement for Arabic.

Table 6 further shows the results for ablation experiments, where we remove one system from the ensemble. We can see that in most cases, removing an individual system yields lower MAP. A notable exception is *blue*, removing which yields improvements in terms of MAP and some other evaluation measures. Moreover, we can see that different ablations can improve over any of the evaluation measures. This suggests that there is potential for improving the overall results by combining the approaches used by the different teams; this should be also possible at the feature/model level.

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

	ENGLISH								
	MAP	MRR	MR-P	MP@1	MP@3	MP@5	MP@10	MP@20	MP@50
Best team: ***Prise de Fer***	.1332	**.4965**	.1352	**.4286**	**.2857**	.2000	.1429	.1571	.1200
Ensemble: ***SUM scores***	**.1378**	.4479	**.1726**	.2857	.2381	.2000	**.2000**	.1571	.1200
−**blue**	.1437	.4533	.1839	.2857	.2381	.2571	.2000	.2000	.1286
−**Prise de Fer**	.1341	.3890	.1537	.1429	.2381	.2286	.2000	.1571	.1171
−**Copenhagen**	.1322	.4449	.1473	.2857	.2381	.2286	.2143	.1357	.1200
−**fragarach**	.1302	.3888	.1574	.1429	.2381	.2000	.2286	.1500	.1257
−**RNCC**	.1298	.3885	.1596	.1429	.2381	.2286	.2143	.1500	.1171
−**UPV-INAOE-Autoritas**	.1257	.4545	.1466	.2857	.2857	.1714	.1857	.1429	.1200
−**bigIR**	.1205	.5250	.1195	.4286	.2381	.2286	.1857	.1357	.1114

	ARABIC								
	MAP	MRR	MR-P	MP@1	MP@3	MP@5	MP@10	MP@20	MP@50
Best team: ***bigIR***	.0899	.1180	.1105	.0000	.0000	.0000	.1333	.1000	.1133
Ensemble: ***SUM scores***	**.0931**	**.4083**	.1105	**.3333**	**.1111**	**.0667**	.1333	**.1167**	**.1200**

Table 6: Ablation results for an ensemble summing the participating systems, as well as for ablation excluding each of the systems from the ensemble.

7 Conclusion and Future Work

We provided an overview of the CLEF-2018 CheckThat! Lab on Automatic Identification and Verification of Political Claims, with focus on Task 1: Check-Worthiness, which asked to predict which claims in a political debate should be prioritized for fact-checking. We offered the task in both English and Arabic.

Our evaluation framework consisted of a dataset of five debates and five speeches divided into training and testing set, and a MAP-based evaluation. A total of thirty teams registered to participate in the Lab and seven teams actually submitted systems for Task 1. The most successful approaches used by the participants relied on recurrent and multi-layer neural networks, as well as on combinations of distributional representations, on matchings claims' vocabulary against lexicons, and on measures of syntactic dependency. The best systems achieved mean average precision of 0.18 and 0.15 on the English and on the Arabic test datasets, respectively. This leaves large room for further improvement, and thus we release[11] all datasets and the scoring scripts, which should enable further research in check-worthiness estimation.

In future iterations of the lab, we plan to add more debates and speeches, both annotated and unannotated, which would enable semi-supervised learning. We further want to add annotations for the same debates/speeches from different fact-checking organizations, which would allow using multi-task learning [8].

[11] http://alt.qcri.org/clef2018-factcheck/

Acknowledgments

This work was made possible in part by NPRP grant# NPRP 7-1313-1-245 from the Qatar National Research Fund (a member of Qatar Foundation). 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. [3]
2. 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: Cappellato et al. [3]
3. 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, Avignon, France (2018)
4. 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)
5. Derczynski, L., Bontcheva, K., Liakata, M., Procter, R., Wong Sak Hoi, G., Zubiaga, A.: SemEval-2017 Task 8: RumourEval: Determining rumour veracity and support for rumours. In: Proceedings of the 11th International Workshop on Semantic Evaluation. pp. 60–67. SemEval '17, Vancouver, Canada (2017)
6. Friedman, J.H.: Greedy function approximation: A gradient boosting machine. The Annals of Statistics **29**(5), 1189–1232 (2001)
7. Fuhr, N., Nejdl, W., Peters, I., Stein, B., Giachanou, A., Grefenstette, G., Gurevych, I., Hanselowski, A., Jarvelin, K., Jones, R., Liu, Y., Mothe, J.: An Information Nutritional Label for Online Documents. ACM SIGIR Forum **51**, 46–66 (2018)
8. 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)
9. 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. [3]
10. Granados, A., Cebrian, M., Camacho, D., de Borja Rodriguez, F.: Reducing the loss of information through annealing text distortion. IEEE Transactions on Knowledge and Data Engineering **23**(7), 1090–1102 (2011)
11. 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. [3]
12. 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, Melbourne, Australia (2015)

13. Hassan, N., Tremayne, M., Arslan, F., Li, C.: Comparing automated factual claim detection against judgments of journalism organizations. In: Computation + Journalism Symposium. Stanford, CA, USA (2016)
14. 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. NAACL-HLT '18, New Orleans, LA, USA (2018)
15. 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)
16. 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, GA, USA (2013)
17. 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: Bellot, P., Trabelsi, C., Mothe, J., Murtagh, F., Nie, J., Soulier, L., Sanjuan, E., Cappellato, L., Ferro, N. (eds.) Proceedings of the Ninth International Conference of the CLEF Association: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Lecture Notes in Computer Science, Springer, Avignon, France (2018)
18. Papadopoulos, S., Bontcheva, K., Jaho, E., Lupu, M., Castillo, C.: Overview of the special issue on trust and veracity of information in social media. ACM Trans. Inf. Syst. **34**(3), 14:1–14:5 (Apr 2016)
19. 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)
20. 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)
21. 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 (2017)
22. 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. ICDM '17, New Orleans, LA, USA (2017)
23. 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, MD, USA (2014)
24. Yasser, K., Kutlu, M., , Elsayed, T.: bigIR at CLEF 2018: Detection and Verifi cation of Check-Worthy Political Claims. In: Cappellato et al. [3]
25. Zuo, C., Karakas, A., Banerjee, R.: A Hybrid Recognition System for Check-worthy Claims Using Heuristics and Supervised Learning. In: Cappellato et al. [3]