



# Overview of the CLEF-2018 CheckThat! Lab on Automatic Identification and Verification of Political Claims

Preslav Nakov<sup>1</sup>(✉), Alberto Barrón-Cedeño<sup>1</sup>, Tamer Elsayed<sup>2</sup>,  
Reem Suwaileh<sup>2</sup>, Lluís Màrquez<sup>3</sup>, Wajdi Zaghouani<sup>4</sup>, Pepa Atanasova<sup>5</sup>,  
Spas Kyuchukov<sup>6</sup>, and Giovanni Da San Martino<sup>1</sup>

- <sup>1</sup> Qatar Computing Research Institute, HBKU, Doha, Qatar  
{pnakov, albarron, gmartino}@qf.org.qa
- <sup>2</sup> Computer Science and Engineering Department, Qatar University, Doha, Qatar  
{telsayed, reem.suwaileh}@qu.edu.qa
- <sup>3</sup> Amazon, Barcelona, Spain  
lluismv@amazon.com
- <sup>4</sup> College of Humanities and Social Sciences, HBKU, Doha, Qatar  
wzaghouani@hbku.edu.qa
- <sup>5</sup> SiteGround, Sofia, Bulgaria  
pepa.gencheva@siteground.com
- <sup>6</sup> 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. In its starting year, the lab featured two tasks. Task 1 asked to predict which (potential) 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. Task 2 asked to assess whether a given check-worthy claim made by a politician in the context of a debate/speech is factually true, half-true, or false. We offered both tasks in English and in Arabic. In terms of data, for both tasks, we focused on debates from the 2016 US Presidential Campaign, as well as on some speeches during and after the campaign (we also provided translations in Arabic), and we relied on comments and factuality judgments from [factcheck.org](http://factcheck.org) and [snopes.com](http://snopes.com), which we further refined manually. A total of 30 teams registered to participate in the lab, and 9 of them actually submitted runs. The evaluation results show that the most successful approaches used various neural networks (esp. for Task 1) and evidence retrieval from the Web (esp. for Task 2). We release all datasets, the evaluation scripts, and the submissions by the participants, which should enable further research in both check-worthiness estimation and automatic claim verification.

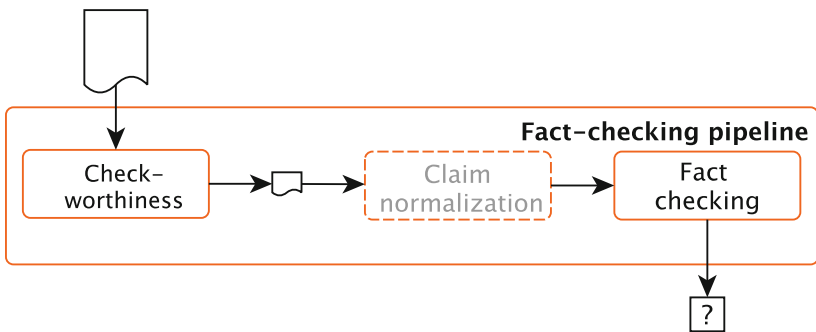
**Keywords:** Computational journalism  
Check-worthiness estimation · Fact-checking · Veracity

## 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 speed of proliferation has left little time for double-checking claims against the facts, which has proven critical in politics. For instance, the 2016 US Presidential Campaign was arguably 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. 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<sup>1</sup> and Snopes<sup>2</sup> 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.<sup>3</sup> This makes it important to prioritize the claims that fact-checkers should consider first, and then to help them discover the veracity of those claims.

The **CheckThat! Lab** at CLEF-2018 aims at helping in that respect, by promoting the development of tools for computational journalism. Figure 1 illustrates the fact-checking pipeline, which includes three steps: (i) *check-worthiness estimation*, (ii) *claim normalization*, and (iii) *fact-checking*. The CheckThat! Lab focuses on the first and the last steps, while taking for granted (and thus excluding) the intermediate claim normalization step.



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

<sup>1</sup> <http://www.factcheck.org>.

<sup>2</sup> <http://www.snopes.com>.

<sup>3</sup> Fully automating the process of fact-checking 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:	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.	
Hillary 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.		

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

**Task 1 (Check-Worthiness)** aims to help fact-checkers prioritize their efforts. In particular, it asks participants to build systems that can mimic the selection strategies of a particular fact-checking organization: [factcheck.org](http://factcheck.org). The task is defined as follows:

*Given a transcription of a political debate/speech, predict which claims should be prioritized for fact-checking.*

Figure 2 shows examples of English debate fragments with annotations for Task 1. 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, Donald Trump is accused of having filed for bankruptcy six times, which is also worth checking.

Task 1 is a *ranking* task. The goal is to produce a *ranked list* of sentences ordered by their worthiness for fact-checking. Each of the identified claims then becomes an input for the next step (after being manually normalized, i.e., edited to be self-contained with no ambiguous or unresolved references).

**Task 2 (Fact-Checking)** focuses on tools intended to verify the factuality of a check-worthy claim. The task is defined as follows:

*Given a check-worthy claim in the form of a (transcribed) sentence, determine whether the claim is likely to be true, half-true, or false.*

For example, the sentence “*Well, he approved NAFTA...*” from example 2a is normalized to “*President Bill Clinton approved NAFTA.*” and the target label is set to HALF-TRUE. Similarly, the sentence “*And when we talk about your business, you’ve taken business bankruptcy six times.*” from example 2b is normalized to “*Donald Trump has filed for bankruptcy of his business six times.*” and the target label is set to TRUE.

Task 2 is a *classification* task. The goal is to *label* each check-worthy claim with an estimated/predicted veracity. Note that we provide the participants not only with the normalized claim, but also with the original sentence it originated from, which is in turn given in the context of the entire debate/speech. Thus, this is a novel task for fact-checking claims *in context*, an aspect that has been largely ignored in previous research on fact-checking.

Note that the intermediate task of *claim normalization* is challenging and requires dealing with anaphora resolution, paraphrasing, and dialogue analysis, and thus we decided not to offer it as a separate task.

We produced data based on professional fact-checking annotations of debates and speeches from [factcheck.org](http://factcheck.org), which we modified in three ways: (i) we did some minor adjustments of which sentences were selected for fact-checking, (ii) we generated normalized versions of the claims in the selected sentences, and (iii) we generated veracity labels for each normalized claim based on the fact-checker’s free-text analysis. As a result, we created CT-C-18, the Check-That! 2018 corpus, which combines two sub-corpora: CT-CWC-18 to predict check-worthiness, and CT-FCC-18 to assess the veracity of claims. We offered each of the two tasks in two languages: English and Arabic. For Arabic, we hired professional translators to translate the English data, and we also had a separate Arabic-only part for Task 2, based on claims from [snopes.com](http://snopes.com).

Nine teams participated in the lab this year. The most successful systems relied on supervised models using a manifold of representations. We believe that there is still large room for improvement, and thus we release the corpora, the evaluation scripts, and the participants’ predictions, which should enable further research on check-worthiness estimation and automatic claim verification.<sup>4</sup>

The remainder of the paper is organized as follows. Section 2 presents an overview of related work. Section 3 describes the datasets. Section 4 discusses Task 1 (check-worthiness) in detail, including the evaluation framework and the setup, the approaches used by the participating teams, and the official results. Section 5 provides similar details for Task 2 (fact-checking). Finally, Sect. 6 draws some conclusions.

## 2 Related Work

Journalists, online users, and researchers are well aware of the proliferation of false information, and topics such as credibility and fact-checking are becoming increasingly important. 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 [24], and there is a Workshop on Fact Extraction and Verification at EMNLP’2018. Moreover, there is a SemEval-2017 shared task on Rumor Detection [7], 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.

<sup>4</sup> <https://github.com/clef2018-factchecking>.

Automatic fact-checking was envisioned in [31] as a multi-step process that includes (i) identifying check-worthy statements [9, 14, 16], (ii) generating questions to be asked about these statements [18], (iii) retrieving relevant information to create a knowledge base [29], and (iv) inferring the veracity of the statements, e.g., using text analysis [6, 28] or external sources [18, 27].

The first work to target check-worthiness was the ClaimBuster system [14]. 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 [15]. 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 [9], 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, [16] 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 [25] 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 (three presidential one vice-presidential) and primary debates (seven Republican and eight 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.

For Task 1, we follow a setup that is similar to that of [9, 16, 25], 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.

Regarding Task 2, which targets fact-checking a claim, there have been several datasets that focus on rumor detection. The gold labels are typically extracted from fact-checking websites such as Politifact with datasets ranging in size from 300 for the Emergent dataset [8] to 12.8K claims for the Liar dataset [33]. Another fact-checking source that has been used is [snopes.com](http://snopes.com), with datasets ranging in size from 1k claims [20] to 5k claims [26].



Less popular as a source has been Wikipedia with datasets ranging in size from 100 claims [26] to 185k for the FEVER dataset [30]. These datasets rely on crowdsourced annotations, which allows them to get large-scale, but risks having lower quality standards compared to the rigorous annotations by fact-checking organizations. Other crowdsourced efforts include the SemEval-2017’s shared task on Rumor Detection [7] with 5.5k annotated rumor tweets, and CREDBANK with 60M annotated tweets [22]. Finally, there have been manual annotation efforts, e.g., for fact-checking the answers in a community question answering forums with size of 250 [21]. Note that while most datasets have been targeting English, there have been also efforts focusing on other languages, e.g., Chinese [20], Arabic [3], and Bulgarian [13].








Unlike the above work, our focus in Task 2 is on claims in both their normalized and unnormalized form and in the context of a political debate or speech.

### 3 Corpora

We produced the CT-C-18 corpus, which stands for CheckThat! 2018 corpus. It is composed of CT-CWC-18 (check-worthiness corpus) and CT-FCC-18 (fact-checking corpus). CT-C-18 includes transcripts from debates, together with political speeches, and isolated claims. Table 1 gives an overview.

The training sets for both tasks come from the first and the second Presidential debates and the Vice-Presidential debate in the 2016 US campaign. The labels for both tasks were derived from manual fact-checking analysis published on [factcheck.org](http://factcheck.org). For Task 1, a claim was considered check-worthy if a journalist had fact-checked it. For Task 2 a judgment was generated based on the free-text discussion by the fact-checking journalists: true, half-true, or false. We followed the same procedure for texts in the test set: two other debates and five speeches by Donald Trump, which occurred after he took office as a US President. Note that there are cases in which the number of claims intended for predicting factuality is lower than the reported number of check-worthy claims. The reason is that claims exist which were formulated more than once in both debates and speeches and, whereas we do consider them all as positive instances for Task 1, we consider them only once for Task 2.

**Table 1.** Overview of the debates, speeches, and isolated claims in the CT-C-18 corpus. It includes the number of utterances, those identified as check-worthy (task 1), and those claims identified as factually- true, half-true, and false. The debates/speeches that are available in Arabic are marked with . Note that the claims from [snopes.com](http://snopes.com) were released in Arabic only, and are marked with .

	Set	Claims	Check- worthy	true	half-true	false	
<b>Debates</b>							
	1st Presidential	training	1,403	37	8	9	13
	2nd Presidential	training	1,303	25	4	7	14
	Vice-Presidential	training	1,358	28	7	6	14
	3rd Presidential	test	1,351	77	19	8	21
	9th Democratic	test	1,464	17	3	3	4
<b>D. Trump Speeches</b>							
	Acceptance	test	375	21	8	5	7
	World Economic Forum	test	245	11	6	2	3
	Tax Reform Event	test	412	16	4	4	4
	Address to Congress	test	390	15	6	3	4
	Miami Speech	test	645	35	4	9	12
	<b>Snopes.com claims</b>	test	–	150	30	10	110

The Arabic version of the corpus was produced manually by professional translators who translated some of the English debates/speeches to Arabic as shown in Table 1. We used this strategy for all three training debates, for the two testing debates, and for one of the five speeches that we used for testing. In order to balance the number of examples for Task 2, we included fresh Arabic-only instances by selecting 150 claims from [snopes.com](http://snopes.com) that were related to the Arab world or to Islam. As the language of [snopes.com](http://snopes.com) is English, we translated these claims to Arabic but this time using Google Translate, and then some of the task organizers (native Arabic speakers) post-edited the result in order to come up with proper Arabic versions. Further details about the construction of the CT-CWC-18 and the CT-FCC-18 corpora can be found in [2, 4].

## 4 Task 1: Check-Worthiness

### 4.1 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:

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

**Table 2.** Task 1 (check-worthiness): overview of the learning models and of the representations used by the participants.

Learning Models	[1][11][12][35][36]	Representations	[1][11][12][35][36]
Recurrent neural nets	☑	Bag of words	☑
Multilayer perceptron		Character $n$ -grams	☑
Support vector machines	☑	Part of speech tags	☑ ☑ ☑
Random forest	☑	Verbal forms	☑
$k$ -nearest neighbors	☑	Negations	☑
Gradient boosting		Named entities	☑ ☑
<b>Teams</b>		Sentiment	☑ ☑
		Topics	☑
[1] RNCC	[-] fragarach	IR nutritional labels	☑
[11] UPV-INAOE-Autoritas	[-] blue	Clauses	☑
[12] Copenhagen		Syntactic dependency	☑ ☑
[35] bigIR		Word embeddings	☑ ☑ ☑
[36] Prise de Fer			

where  $d \in D$  is one of the debates/speeches, and  $AveP$  is the average precision:

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

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 [9], 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.

## 4.2 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 4 shows the results for English. The best primary submission was that of the *Prise de Fer* team [35], 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).



Interestingly, the top-performing run for English was an unofficial one, namely the contrastive 1 run by the *Copenhagen* team [12]. This model consisted of a recurrent neural network on three representations. They submitted a system that combined their neural network with the model of [9] 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 [9] relying on structural information, which was not available for the speeches included in the test set.

To put these results in perspective, the bottom of Table 4 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 [11, 34], using basically the same models that they had for English. The *bigIR* [34] 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.

## 5 Task 2: Factuality

### 5.1 Evaluation Measures

Task 2 (factuality) the claims have to be labeled as *true*, *half-true*, or *false*. Note that, unlike standard multi-way classification problems, here we have a natural ordering between the classes and confusing one extreme with the other one is more harmful than confusing it with a neighboring class. This is known as an *ordinal classification* problem (aka *ordinal regression*), and it requires an evaluation measure that would take this ordering into account. Therefore, we opted for using mean absolute error (MAE), which is standard for such kinds of problems, as the official measure. MAE is defined as follows:

$$MAE = \frac{\sum_{c=1}^C |y_c - x_c|}{C} \quad (3)$$

where  $y_c$  and  $x_c$  are gold and predicted labels of claim  $c$  and  $|\cdot|$  is the difference between them: either zero, one, or two.

Following [23], we also compute macro-average mean absolute error, accuracy, macro-averaged  $F_1$ , and macro-averaged recall.<sup>5</sup>

## 5.2 Evaluation Results

When dealing with the factuality task, participants opted for retrieving evidence from the Web in order to assess the factuality of the claims. After retrieving a number of search engine snippets or full documents, they performed different operations, including calculating similarities or levels of contradiction and stance between the supporting document and the claim. For example, the Copenhagen team [32] concatenating the representations of claim and of the document in a neural network. Table 3 gives a brief overview. Refer to [4] and the corresponding participants' reports for further details.

**Table 3.** Task 2 (factuality): overview of the learning models and of the representations used by the participants.

	[10]	[20]	[33]	[35]	$f(\text{claim, doc})$	[10]	[20]	[33]	[35]
<b>Learning Models</b>					Similarity	✓		✓	
Logistic regression				✓	Alexa rank	✓			
Long short-term memory		✓			Stance				✓
Conv. neural network			✓		Contradiction				✓
Support vector machine			✓		NN concatenation			✓	
Random forests	✓			✓					
<b>Search Engines</b>					<b>Teams</b>				
Google	✓		✓	✓	[10] UPV-INAOE-Autoritas				
Bing	✓				[33] Copenhagen				
<b>Representations</b>					[20] Check it out				
Bag of words	✓		✓	✓	[35] bigIR				
Word embeddings	✓	✓	✓	✓	[—] FACTR				

Note that the *bigIR* team [34] tried to identify the relevant fragments in the supporting documents by considering only those with high similarity against the claim. Various approaches [32, 34] are based at some extent on [17]. Only one team, Check it out [19], did not use external supporting documents (Table 5).

**English.** Table 6 shows the results on the English dataset. Overall, the top-performing system is the one by the Copenhagen team [32]. One aspect that

<sup>5</sup> The implementation of the evaluation measures is available at <https://github.com/clef2018-factchecking/clef2018-factchecking/>.

**Table 4.** Task 1 (check-worthiness): English results, ranked based on MAP, the official evaluation measure. The best score per evaluation measure is in shown in bold.

	MAP	MRR	MR-P	MP@1	MP@3	MP@5	MP@10	MP@20	MP@50
<b>Prise de Fer [35]</b>									
primary	<b>.1332</b> <sub>(1)</sub>	<b>.4965</b> <sub>(1)</sub>	<b>.1352</b> <sub>(1)</sub>	<b>.4286</b> <sub>(1)</sub>	<b>.2857</b> <sub>(1)</sub>	.2000 <sub>(2)</sub>	.1429 <sub>(3)</sub>	<b>.1571</b> <sub>(1)</sub>	.1200 <sub>(2)</sub>
cont. 1	.1366	.5246	.1475	.4286	.2857	.2286	.1571	.1714	.1229
cont. 2	.1317	.4139	.1523	.2857	.1905	.1714	.1571	.1571	.1429
<b>Copenhagen [12]</b>									
primary	.1152 <sub>(2)</sub>	.3159 <sub>(5)</sub>	.1100 <sub>(5)</sub>	.1429 <sub>(3)</sub>	.1429 <sub>(4)</sub>	.1143 <sub>(3)</sub>	.1286 <sub>(4)</sub>	.1286 <sub>(2)</sub>	<b>.1257</b> <sub>(1)</sub>
cont. 1	.1810	.6224	.1875	.5714	.4286	.3143	.2571	.2357	.1514
<b>UPV-INAOE-Autoritas [11]</b>									
primary	.1130 <sub>(3)</sub>	.4615 <sub>(2)</sub>	.1315 <sub>(2)</sub>	.2857 <sub>(2)</sub>	.2381 <sub>(2)</sub>	<b>.3143</b> <sub>(1)</sub>	<b>.2286</b> <sub>(1)</sub>	.1214 <sub>(3)</sub>	.0886 <sub>(4)</sub>
cont. 1	.1232	.3451	.1022	.1429	.2857	.2286	.1429	.1143	.0771
cont. 2	.1253	.5535	.0849	.4286	.4286	.2571	.1429	.1286	.0771
<b>bigIR [34]</b>									
primary	.1120 <sub>(4)</sub>	.2621 <sub>(6)</sub>	.1165 <sub>(4)</sub>	.0000 <sub>(4)</sub>	.1429 <sub>(4)</sub>	.1143 <sub>(3)</sub>	.1143 <sub>(5)</sub>	.1000 <sub>(5)</sub>	.1114 <sub>(3)</sub>
cont. 1	.1319	.2675	.1505	.1429	.0952	.0857	.1714	.1786	.1343
cont. 2	.1116	.2195	.1294	.0000	.1429	.1429	.1857	.1429	.0886
<b>fragarach</b>									
primary	.0812 <sub>(5)</sub>	.4477 <sub>(3)</sub>	.1217 <sub>(3)</sub>	.2857 <sub>(2)</sub>	.1905 <sub>(3)</sub>	.2000 <sub>(2)</sub>	.1571 <sub>(2)</sub>	.1071 <sub>(4)</sub>	.0743 <sub>(5)</sub>
<b>blue</b>									
primary	.0801 <sub>(6)</sub>	.2459 <sub>(7)</sub>	.0576 <sub>(7)</sub>	.1429 <sub>(3)</sub>	.0952 <sub>(5)</sub>	.0571 <sub>(4)</sub>	.0571 <sub>(6)</sub>	.0857 <sub>(6)</sub>	.0600 <sub>(6)</sub>
<b>RNCC [1]</b>									
primary	.0632 <sub>(7)</sub>	.3775 <sub>(4)</sub>	.0639 <sub>(6)</sub>	.2857 <sub>(2)</sub>	.1429 <sub>(4)</sub>	.1143 <sub>(3)</sub>	.0571 <sub>(6)</sub>	.0571 <sub>(7)</sub>	.0486 <sub>(7)</sub>
cont. 1	.0886	.4844	.0945	.4286	.1429	.1714	.1286	.1000	.0714
cont. 2	.0747	.2198	.0984	.0000	.0952	.1143	.1000	.1000	.0829
<b>Baselines</b>									
n-gram	.1201	.4087	.1280	.1429	.2857	.1714	.1571	.1357	.1143
random	.0485	.0633	.0359	.0000	.0000	.0000	.0286	.0214	.0429

**Table 5.** Task 1 (check-worthiness): Arabic results, ranked based on MAP, the official evaluation measure. The best score per evaluation measure is in bold.

	MAP	MRR	MR-P	MP@1	MP@3	MP@5	MP@10	MP@20	MP@50
<b>bigIR [34]</b>									
primary	<b>.0899</b> <sub>(1)</sub>	.1180 <sub>(2)</sub>	<b>.1105</b> <sub>(1)</sub>	.0000 <sub>(2)</sub>	.0000 <sub>(2)</sub>	.0000 <sub>(2)</sub>	<b>.1333</b> <sub>(1)</sub>	<b>.1000</b> <sub>(1)</sub>	<b>.1133</b> <sub>(1)</sub>
cont. 1	.1497	.2805	.1760	.0000	.3333	.3333	.2667	.2333	.1533
cont. 2	.0962	.1660	.0895	.0000	.1111	.2000	.1667	.1000	.0867
<b>UPV-INAOE-Autoritas [11]</b>									
primary	.0585 <sub>(2)</sub>	<b>.3488</b> <sub>(1)</sub>	.0087 <sub>(2)</sub>	<b>.3333</b> <sub>(1)</sub>	<b>.1111</b> <sub>(1)</sub>	<b>.0667</b> <sub>(1)</sub>	.0333 <sub>(2)</sub>	.0167 <sub>(2)</sub>	.0400 <sub>(2)</sub>
cont. 1	.1168	.6714	.0649	.6667	.6667	.4000	.2000	.1000	.0733
<b>Baselines</b>									
n-gram	.0861	.2817	.0981	.0000	.3333	.2667	.1667	.1667	.0867
random	.0460	.0658	.0375	.0000	.0000	.0000	.0333	.0167	.0333

might explain the relatively large difference in performance compared to the other teams is the use of additional training material. The Copenhagen team incorporated hundreds of labeled claims from Politifact to their training set. Their model combines the claim and supporting texts to build representations. Their primary submission is an SVM, whereas their contrastive one uses a CNN.

**Table 6.** Task 2 (factuality): English results, ranked based on MAE, the official evaluation measure. The best score per evaluation measure is in bold.

	MAE	Macro MAE	Acc	Macro F1	Macro AvgR
<b>Copenhagen [32]</b>					
primary	<b>.7050</b> <sub>(1)</sub>	<b>.6746</b> <sub>(1)</sub>	<b>.4317</b> <sub>(1)</sub>	<b>.4008</b> <sub>(1)</sub>	<b>.4502</b> <sub>(1)</sub>
cont. 1	.7698	.7339	.4676	.4681	.4721
<b>FACTR</b>					
primary	.9137 <sub>(2)</sub>	.9280 <sub>(2)</sub>	.4101 <sub>(2)</sub>	.3236 <sub>(2)</sub>	.3684 <sub>(2)</sub>
cont. 1	.9209	.9358	.4029	.3063	.3611
cont. 2	.9281	.9314	.4101	.3420	.3759
<b>UPV-INAOE-Autoritas [10]</b>					
primary	.9496 <sub>(3)</sub>	.9706 <sub>(3)</sub>	.3885 <sub>(4)</sub>	.2613 <sub>(3)</sub>	.3403 <sub>(3)</sub>
<b>bigIR [34]</b>					
primary	.9640 <sub>(4)</sub>	1.0000 <sub>(4)</sub>	.3957 <sub>(3)</sub>	.1890 <sub>(4)</sub>	.3333 <sub>(4)</sub>
cont. 1	.9640	1.0000	.3957	.1890	.3333
cont. 2	.9424	.9256	.3525	.3297	.3405
<b>Check It Out [19]</b>					
primary	.9640 <sub>(4)</sub>	1.0000 <sub>(4)</sub>	.3957 <sub>(3)</sub>	.1890 <sub>(4)</sub>	.3333 <sub>(4)</sub>
<b>Baselines</b>					
n-gram	.9137	.9236	.3957	.3095	.3588
random	.8345	.8139	.3597	.3569	.3589

Unfortunately, not much information is available regarding team FACTR, as no paper was submitted to describe their model. They used a similar approach as most other teams: converting the claim into a query for a search engine, computing stance, sentiment and other features over the supporting documents, and using them in a supervised model.

**Arabic.** Table 7 shows the results of the two teams that participated in the Arabic task. In order to deal with it, FACTR translated all the claims into English and performed the rest of the process in that language. In contrast, UPV-INAOE-Autoritas [10] translated the claims into English, but only in order to query the search engines,<sup>6</sup> and then translated the retrieved evidence into Arabic in order to keep working in that language. Perhaps, the noise generated by using two imperfect translations caused their performance to decrease (the performance of the two teams in the English task was much closer).

<sup>6</sup> The reason is that the Arabic dataset was produced by translating the datasets from an English version. Hence it was difficult to find evidence in Arabic.

**Table 7.** Task 2 (factuality): Arabic results, ranked based on MAE, the official evaluation measure. The best score per evaluation measure is in bold.

	MAE	Macro MAE	Acc	Macro F1	Macro AvgR
<b>FACTR</b>					
primary	<b>.6579</b> <sub>(1)</sub>	<b>.8914</b> <sub>(1)</sub>	<b>.5921</b> <sub>(1)</sub>	<b>.3730</b> <sub>(1)</sub>	<b>.3804</b> <sub>(1)</sub>
cont. 1	.7018	.9461	.5833	.3691	.3766
cont. 2	.6623	.9153	.5965	.3657	.3804
<b>UPV–INAOE–Autoritas [10]</b>					
primary	.8202 <sub>(2)</sub>	1.0417 <sub>(2)</sub>	.5175 <sub>(2)</sub>	.2796 <sub>(2)</sub>	.3027 <sub>(2)</sub>
<b>Baselines</b>					
n-gram	.6798	.9850	.5789	.2827	.3267
random	.9868	.9141	.3070	.2733	.2945

Overall, the performance of the models in Arabic is better than in English. The reason is that the isolated claims from [snopes.com](http://snopes.com)—which were released only in Arabic (cf. Table 1)—were easier to verify.

## 6 Conclusions and Future Work

We have presented an overview of the CLEF-2018 CheckThat! Lab on Automatic Identification and Verification of Political Claims. Task 1 asked to predict which claims in a political debate or speech should be prioritized for fact-checking. Task 2 asked to assess whether a claim made by a politician is factually true, half-true, or false. We proposed both tasks in English and Arabic, relying on comments and factuality judgments from both [factcheck.org](http://factcheck.org) and [snopes.com](http://snopes.com) to obtain a further-refined gold standard and on translation for the Arabic versions of the corpus. A total of 30 teams registered to participate in the lab, and 9 of them actually submitted runs. The evaluation results showed that the most successful approaches used various neural networks (esp. for Task 1) and evidence retrieved from the Web (esp. for Task 2). The corpora and the evaluation measures we have released as a result of this lab should enable further research in check-worthiness estimation and in automatic claim verification.

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 [9].

**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. [5]
2. Atanasova, P., et al.: Overview of the CLEF-2018 CheckThat! Lab on automatic identification and verification of political claims. Task 1: Check-worthiness. In: Cappellato et al. [5]
3. 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. NAACL-HLT 2018, New Orleans, Louisiana, USA, pp. 21–27 (2018)
4. Barrón-Cedeño, A., et al.: Overview of the CLEF-2018 CheckThat! Lab on automatic identification and verification of political claims. Task 2: Factuality. In: Cappellato et al. [5]
5. 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)
6. Castillo, C., Mendoza, M., Poblete, B.: Information credibility on Twitter. In: Proceedings of the 20th International Conference on World Wide Web, WWW 2011, Hyderabad, India, pp. 675–684 (2011)
7. 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, SemEval 2017, Vancouver, Canada, pp. 60–67 (2017)
8. Ferreira, W., Vlachos, A.: Emergent: a novel data-set for stance classification. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. NAACL-HLT 2016, San Diego, California, USA, pp. 1163–1168 (2016)
9. 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. RANLP 2017, Varna, Bulgaria, pp. 267–276 (2017)
10. Ghanem, B., Montes-y Gómez, M., Rangel, F., Rosso, P.: UPV-INA OE-Autoritas - check that: an approach based on external sources to detect claims credibility. In: Cappellato et al. [5]
11. Ghanem, B., Montes-y Gómez, M., Rangel, F., Rosso, P.: UPV-INA OE-Autoritas - check that: preliminary approach for checking worthiness of claims. In: Cappellato et al. [5]
12. 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. [5]
13. Hardalov, M., Koychev, I., Nakov, P.: In search of credible news. In: Proceedings of the 17th International Conference on Artificial Intelligence: Methodology, Systems, and Applications. AIMS A 2016, Varna, Bulgaria, pp. 172–180 (2016)
14. 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, CIKM 2015, Melbourne, Australia, pp. 1835–1838 (2015)

15. 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, September 2016
16. 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 2018*, New Orleans, Louisiana, USA, pp. 26–30 (2018)
17. Karadzhov, G., Nakov, P., Màrquez, L., Barrón-Cedeño, A., Koychev, I.: Fully automated fact checking using external sources. In: *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pp. 344–353. INCOMA Ltd., Varna (2017)
18. 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, RANLP 2017*, Varna, Bulgaria, pp. 344–353 (2017)
19. Lal, Y.K., Khatrar, D., Kumar, V., Mishra, A., Varma, V.: Check it out : politics and neural networks. In: Cappellato et al. [5]
20. Ma, J., et al.: Detecting rumors from microblogs with recurrent neural networks. In: *Proceedings of the 25th International Joint Conference on Artificial Intelligence, IJCAI 2016*, New York, New York, USA, pp. 3818–3824 (2016)
21. Mihaylova, T., et al.: Fact checking in community forums. In: *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, AAAI 2018*, New Orleans, Louisiana, USA, pp. 879–886 (2018)
22. Mitra, T., Gilbert, E.: CREDBANK: a large-scale social media corpus with associated credibility annotations. In: Cha, M., Mascolo, C., Sandvig, C. (eds.) *Proceedings of the Ninth International Conference on Web and Social Media, ICWSM 2015*, Oxford, UK, pp. 258–267 (2015)
23. Nakov, P., Ritter, A., Rosenthal, S., Sebastiani, F., Stoyanov, V.: SemEval-2016 task 4: Sentiment analysis in Twitter. In: *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval 2016*, San Diego, California, USA, pp. 1–18 (2016)
24. 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 (2016)
25. 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, CIKM 2017*, Singapore, pp. 2259–2262 (2017)
26. Popat, K., Mukherjee, S., Strötgen, J., Weikum, G.: Credibility assessment of textual claims on the web. In: *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, CIKM 2016*, pp. 2173–2178. ACM, Indianapolis (2016)
27. 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, WWW 2017*, Perth, Australia, pp. 1003–1012 (2017)
28. 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, EMNLP 2017*, pp. 2931–2937 (2017)

29. 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 2017, New Orleans, Louisiana, USA, pp. 859–864 (2017)
30. 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. NAACL-HLT 2018, New Orleans, Louisiana, USA, pp. 809–819 (2018)
31. 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, Baltimore, Maryland, USA, pp. 18–22 (2014)
32. Wang, D., Simonsen, J., Larseny, B., Lioma, C.: The Copenhagen team participation in the factuality 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. [5]
33. 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, ACL 2017, Vancouver, Canada, pp. 422–426 (2017)
34. Yasser, K., Kutlu, M., Elsayed, T.: bigIR at CLEF 2018: detection and verification of check-worthy political claims. In: Cappellato et al. [5]
35. Zuo, C., Karakas, A., Banerjee, R.: A hybrid recognition system for check-worthy claims using heuristics and supervised learning. In: Cappellato et al. [5]