

Alma Mater Studiorum Università di Bologna
Archivio istituzionale della ricerca

Overview of the CLEF-2018 checkthat! lab on automatic identification and verification of political claims

This is the final peer-reviewed author's accepted manuscript (postprint) of the following publication:

Published Version:

Nakov P., Barron-Cedeno A., Elsayed T., Suwaileh R., Marquez L., Zaghouani W., et al. (2018).
Overview of the CLEF-2018 checkthat! lab on automatic identification and verification of political claims.
Springer Verlag [10.1007/978-3-319-98932-7_32].

Availability:

This version is available at: <https://hdl.handle.net/11585/709182> since: 2020-04-24

Published:

DOI: http://doi.org/10.1007/978-3-319-98932-7_32

Terms of use:

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (<https://cris.unibo.it/>).
When citing, please refer to the published version.

(Article begins on next page)

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

Preslav Nakov¹, Alberto Barrón-Cedeño¹, Tamer Elsayed²,
Reem Suwaileh², Lluís Màrquez³, Wajdi Zaghouani⁴,
Pepa Atanasova⁵, Spas Kyuchukov⁶, and Giovanni Da San Martino¹

¹ Qatar Computing Research Institute, HBKU, Doha, Qatar
{pnakov, albarron, gmartino}@qf.org.qa

² Computer Science and Engineering Department, Qatar University, Doha, Qatar
{telsayed, reem.suwaileh}@qu.edu.qa

³ Amazon, Barcelona, Spain
lluismv@amazon.com

⁴ College of Humanities and Social Sciences, HBKU, Doha, Qatar
wzaghouani@hbku.edu.qa

⁵ SiteGround, Sofia, Bulgaria
pepa.gencheva@siteground.com

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. 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 and 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 in automatic claim verification.

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

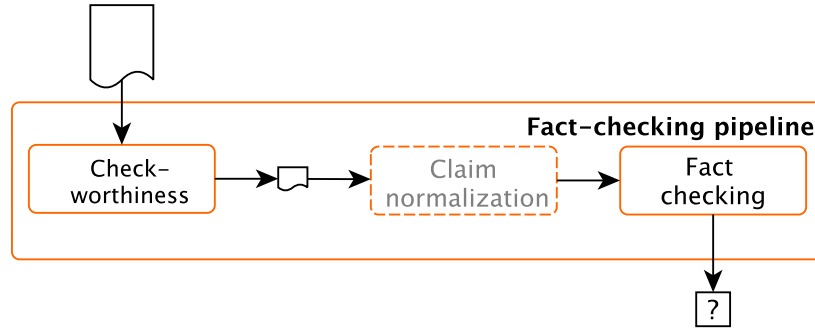
1 Introduction

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 FactCheck⁷ and Snopes⁸ 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.⁹ 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 former and on the latter 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, and finally they are fact-checked.



⁷ <http://www.factcheck.org>

⁸ <http://www.snopes.com>

⁹ 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.

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. The task is defined as follows:

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

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.

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 the normalized claim, but also the original sentence it originated in, 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 a challenging problem that requires dealing with anaphora resolution, paraphrasing, and dialogue analysis, and thus we decided to skip it and to provide participants readily-normalized claims.

We produced data starting from professional fact-checking annotations of debates and speeches from factcheck.org, thus creating CT-C-18, the CheckThat! 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.

Nine teams participated in CheckThat! 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 in check-worthiness estimation and automatic claim verification.¹⁰

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 detail for Task 2 (fact-checking). Finally, Section 6 discusses the lessons learned.

¹⁰ <https://github.com/clef2018-factchecking>

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 [20], and there is a Workshop on Fact Extraction and Verification at EMNLP’2018. Moreover, there is a SemEval-2017 shared task on Rumor Detection [6], 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 [25] as a multi-step process that includes (i) identifying check-worthy statements [8, 13, 16], (ii) generating questions to be asked about these statements [18], (iii) retrieving relevant information to create a knowledge base [24], and (iv) inferring the veracity of the statements, e.g., using text analysis [5, 23] or external sources [18, 22].

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 [7], 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 most 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.

We follow a similar setup for Task 1, 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 claims (as fact-checkers would only comment on one instance of a claim). We further have a larger dataset, and we have an Arabic version of the dataset; however, we are limited to a single fact-checking organization.

The work of [21] 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 asks to predict whether any of the fact-checking sources would select the target sentence. They use a boosting-like model that takes SVMs focusing on different clusters of the dataset and the final outcome is considered as that coming from the most confident classifier. The features considered go from LDA topic-modeling to POS tuples and bag-of-word representations.


The Fact Extraction and VERification corpus (FEVER) was released by the EMNLP’2018 Workshop on Fact Extraction and Verification to verify information against textual sources. FEVER consists of 185,445 claims created by modifying a selection of sentences from the Wikipedia and later on verified neglecting the knowledge of the sentence they were derived from. The claims are classified by the annotators as refuted, supported, or marked as lacking the necessary details to make a decision.







There have been several related shared tasks such as SemEval-2017’s shared task on Rumor Detection [6] with a total of 5599 annotated rumourous Tweets and the upcoming task at SemEval’2019 on Fact Checking in Community Question and Answering Forums.

3 Corpora

We produced the corpus CT-C-18, 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 second presidential debates and the vice-presidential debate in the 2016 US campaign. The labels for both tasks were derived from manual journalist judgments published at FactCheck.org. For Task 1, a claim was considered check-worthy if a journalist had fact-checked it. For Task 2 the judgment of the journalist was adopted: true, half-true, or false. We followed the same procedure for texts in the test set: two other debates and five speeches by D. Trump, which occurred after he took office as president. It is worth noting that there are cases in which the number of claims intended for the prediction of 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

Table 1. Overview of the debates and speeches 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. Documents translated into Arabic as well are marked with ; the Snopes.com claims were released in Arabic only (●).

| | Set | Claims | Check- worthy | Factuality true | half-true | false |
|---|----------|--------|------------------|--------------------|-----------|-------|
| Debates | | | | | | |
|  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 |
| D. Trump Speeches | | | | | | |
|  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 |
| ● Snopes.com claims | test | – | – | 30 | 10 | 110 |

do consider them as positive instances for Task 1, we consider them only once for Task 2.

The Arabic version of the corpus was produced manually by professional translators. We produced with this strategy all three training debates, both testing debates, and one out of the five testing speeches. In exchange, we included fresh Arabic-only instances by selecting 150 claims published at Snopes.com with the constraint of being either related with Islam or the Arab world. The language in Snopes.com is English as well. In this case we translated the claims with Google translate and manually post-edited them in order to come out with proper Arabic versions.

Further details about the construction of the two partitions of the CT-CWC-18 and CT-FCC-18 corpora can be found in [2, 3].

4 Task 1: Check-Worthiness

The participants approached this task as a classification problem; except for [27], who opted for a learning-to-rank approach. Multiple features were used, from embeddings to lexical representations, passing by estimators of sentiment and number of clauses, among others. In some cases, these models were combined with that of [7]. Table 2 gives a brief overview. Refer to [2] and the participants’ reports for further details.

Some additional strategies are worth mentioning. [28] generated sub-corpora from the training debate documents which included interventions from single characters. The purpose was to come out with a dataset closer to the speeches

included in the test set. [10] used a text distortion model [11] to try to remove irrelevant contents.

Table 2. Overview of learning models and representations used to approach Task 1: check-worthiness.

| Learning models | [1] | [10] | [12] | [27] | [28] | Representations | [1] | [10] | [12] | [27] | [28] | | | |
|-----------------------------|--------------------------|-------------------|------|------|------|------------------------|-----------------------|------|------|------|------|---|--|--|
| Recurrent neural network | | | ✓ | | | Bag of words | | | | | ✓ | | | |
| Multilayer perceptron | | | | | ✓ | <i>n</i> -grams | | ✓ | | | | | | |
| Support vector machine | ✓ | | | | ✓ | Part of speech tags | | | ✓ | ✓ | | | | |
| Random forests | ✓ | | | | | Verbal forms | | | | | ✓ | | | |
| <i>k</i> -nearest neighbors | | | ✓ | | | Negations | | | | | ✓ | | | |
| Gradient boosting | | | | ✓ | | Named entities | | | | ✓ | | | | |
| Teams | | | | | | Sentiment | | | | ✓ | ✓ | | | |
| | | | | | | Topics | | | | ✓ | ✓ | | | |
| | [1] RNCC | [27] bigIR | | | | | IR nutritional labels | ✓ | | | | | | |
| | [10] UPV-INAOE-Autoritas | [28] Prise de Fer | | | | | Clauses | | | | | ✓ | | |
| | [12] Copenhagen | | | | | Syntactic dependencies | | | ✓ | | | | | |
| | | | | | | Word embeddings | | ✓ | ✓ | ✓ | | | | |
| | | | | | | | | | | | | | | |

Table 3 shows the results on the English partition. As this is a retrieval problem, we use mean average precision as the primary evaluation measure. We also computed mean reciprocal rank (MRR), mean R-precision (MR-P), and macro-averaged precision at different values of k (MP@ k). The top-performing system was the contrasting 1 from the Copenhagen team [12]: a recurrent neural network on word2vec, part of speech, and syntactic dependencies. Interestingly, this participants submitted as primary a system that combined their neural network with the model of [7]; but it performed worst. The reason might be that [7] rely partially on structural information absent in the speeches of the test set (cf. Section 3).

Only two teams participated to Task 1 Arabic [10, 27]. Table 4 shows their results. Both approaches used identical models as the applied to the English partition. The model of [10] relies heavily on a number of lexicons and they opted for machine-translating them into Arabic and work in this language. The model of [27] relies on supervised models to determine the sentiment and the topic of the claim, which are trained on English corpora as well. Hence they opted for machine-translating the debates and speeches into English and working in this language.

5 Task 2: Factuality

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 of full documents, they performed different operations, including similarities or levels of contradiction and stance between

Table 3. Performance of the participants’ submissions to the CheckThat! Task 1: check-worthiness English. MAP is the metric used to rank the different teams. The runs include one primary and (at most) two contrastive submissions. The top-performing score per evaluation metric appears highlighted.

| | MAP | MRR | MR-P | MP@1 | MP@3 | MP@5 | MP@10 | MP@20 | MP@50 |
|---------------------------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|
| Prise de Fer [28] | | | | | | | | | |
| primary | 0.1332 | 0.4965 | 0.1352 | 0.4286 | 0.2857 | 0.2000 | 0.1429 | 0.1571 | 0.1200 |
| cont. 1 | 0.1366 | 0.5246 | 0.1475 | 0.4286 | 0.2857 | 0.2286 | 0.1571 | 0.1714 | 0.1229 |
| cont. 2 | 0.1317 | 0.4139 | 0.1523 | 0.2857 | 0.1905 | 0.1714 | 0.1571 | 0.1571 | 0.1429 |
| Copenhagen [12] | | | | | | | | | |
| primary | 0.1152 | 0.3159 | 0.1100 | 0.1429 | 0.1429 | 0.1143 | 0.1286 | 0.1286 | 0.1257 |
| cont. 1 | 0.1810 | 0.6224 | 0.1875 | 0.5714 | 0.4286 | 0.3143 | 0.2571 | 0.2357 | 0.1514 |
| UPV-INAOE-Autoritas [10] | | | | | | | | | |
| primary | 0.1130 | 0.4615 | 0.1315 | 0.2857 | 0.2381 | 0.3143 | 0.2286 | 0.1214 | 0.0886 |
| cont. 1 | 0.1232 | 0.3451 | 0.1022 | 0.1429 | 0.2857 | 0.2286 | 0.1429 | 0.1143 | 0.0771 |
| cont. 2 | 0.1253 | 0.5535 | 0.0849 | 0.4286 | 0.4286 | 0.2571 | 0.1429 | 0.1286 | 0.0771 |
| bigIR [27] | | | | | | | | | |
| primary | 0.1120 | 0.2621 | 0.1165 | 0.0000 | 0.1429 | 0.1143 | 0.1143 | 0.1000 | 0.1114 |
| cont. 1 | 0.1319 | 0.2675 | 0.1505 | 0.1429 | 0.0952 | 0.0857 | 0.1714 | 0.1786 | 0.1343 |
| cont. 2 | 0.1116 | 0.2195 | 0.1294 | 0.0000 | 0.1429 | 0.1429 | 0.1857 | 0.1429 | 0.0886 |
| fragarach | | | | | | | | | |
| primary | 0.0812 | 0.4477 | 0.1217 | 0.2857 | 0.1905 | 0.2000 | 0.1571 | 0.1071 | 0.0743 |
| blue | | | | | | | | | |
| primary | 0.0801 | 0.2459 | 0.0576 | 0.1429 | 0.0952 | 0.0571 | 0.0571 | 0.0857 | 0.0600 |
| RNCC [1] | | | | | | | | | |
| primary | 0.0632 | 0.3775 | 0.0639 | 0.2857 | 0.1429 | 0.1143 | 0.0571 | 0.0571 | 0.0486 |
| cont. 1 | 0.0886 | 0.4844 | 0.0945 | 0.4286 | 0.1429 | 0.1714 | 0.1286 | 0.1000 | 0.0714 |
| cont. 2 | 0.0747 | 0.2198 | 0.0984 | 0.0000 | 0.0952 | 0.1143 | 0.1000 | 0.1000 | 0.0829 |

the supporting document and the claim. [26] opted for concatenating the representations of claim and document in a neural network. Table 5 gives a brief overview. Refer to [3] and the corresponding participants’ reports for further details.

It is worth mentioning that [27] tried to identify the relevant fragments in the supporting documents by considering only those with a high similarity against the claim. Various of the approaches [27, 26] are based at some extent in [17]. Only [19] approached the task neglecting any supporting document.

Table 6 shows the evaluation results. As this is a classification task with ordering between the classes (aka *ordinal classification* or *ordinal regression*), we use mean absolute error (MAE) as the primary evaluation measure. We also computed accuracy (ACC) and macro-averaged F_1 and recall. Overall, the top-performing system is the primary one from the Copenhagen team [26]. One aspect that might explain the relatively-large difference between the performance of this approach and the others is the use of additional training material. Team

Table 4. Performance of the participants’ submissions to the CheckThat! Task 1: check-worthiness Arabic. MAP is the metric used to rank the different teams. The runs include one primary and (at most) two contrastive submissions. The top-performing score per evaluation metric appears highlighted.

| | MAP | MRR | MR-P | MP@1 | MP@3 | MP@5 | MP@10 | MP@20 | MP@50 |
|-----------------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|
| bigIR [27] | | | | | | | | | |
| primary | 0.0899 | 0.1180 | 0.1105 | 0.0000 | 0.0000 | 0.0000 | 0.1333 | 0.1000 | 0.1133 |
| cont. 1 | 0.1497 | 0.2805 | 0.1760 | 0.0000 | 0.3333 | 0.3333 | 0.2667 | 0.2333 | 0.1533 |
| cont. 2 | 0.0962 | 0.1660 | 0.0895 | 0.0000 | 0.1111 | 0.2000 | 0.1667 | 0.1000 | 0.0867 |
| UPV-INAEO [10] | | | | | | | | | |
| primary | 0.0585 | 0.3488 | 0.0087 | 0.3333 | 0.1111 | 0.0667 | 0.0333 | 0.0167 | 0.0400 |
| cont. 1 | 0.1168 | 0.6714 | 0.0649 | 0.6667 | 0.6667 | 0.4000 | 0.2000 | 0.1000 | 0.0733 |

Table 5. Overview of learning models, search engines used to retrieve supporting documents, representations, and operations f between claims and supporting documents used to approach Task 2: fact checking.

| | [9] | [19] | [26] | [27] | $f(\text{claim}, \text{doc})$ | [9] | [19] | [26] | [27] |
|------------------------|-----|------|------|------|-------------------------------|------|------------|------|------|
| Learning models | | | | | Similarity | ✓ | | | |
| Logistic regression | | | | ✓ | Alexa rank | ✓ | | | |
| Long short-term memory | | ✓ | | | Stance | | | | ✓ |
| Conv. neural network | | | ✓ | ✓ | Contradiction | | | | ✓ |
| Support vector machine | | | ✓ | | NN concatenation | | | ✓ | |
| Random forests | ✓ | | | ✓ | | | | | |
| Search engines | | | | | | | | | |
| Google | | ✓ | | ✓ | | | | | |
| Bing | | ✓ | | ✓ | | | | | |
| Representations | | | | | | | | | |
| Bag of words | | ✓ | | ✓ | | | | | |
| Word embeddings | ✓ | ✓ | ✓ | ✓ | | | | | |
| | | | | | Teams | | | | |
| | | | | | [9] UPV-INAEO-Autoritas | [26] | Copenhagen | | |
| | | | | | [19] Check it out | [27] | bigIR | | |

Copenhagen incorporated hundreds of labeled claims to their training set from Politifact.¹¹

Once again, only two teams participated to the Arabic task. Table 7 shows the results. In order to deal with it, team FACTR translated all the claims into English and performed the rest of the process in that language. Meanwhile, [9] translated the claims into English only to query the search engines¹² and then translated the retrieved evidence into Arabic in order to keep working in this language. Perhaps the noise generated by these iterations of imperfect translations caused their performance to decrease (note that the performance of the two teams in the English task is much closer).

¹¹ <http://www.politifact.com/>

¹² 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 6. Performance of the participants’ submissions to the CheckThat! Task 2: fact checking English. MAE is the metric used to rank the different teams. The runs include one primary and (at most) two contrasting submissions. The top-performing score per evaluation metric appears highlighted.

| | MAE | Macro MAE | ACC | Macro F1 | Macro Recall |
|--------------------------------|--------|---------------|---------------|---------------|---------------|
| Copenhagen [26] | | | | | |
| primary | 0.7050 | 0.6746 | 0.4317 | 0.4008 | 0.4502 |
| cont. 1 | 0.7698 | 0.7339 | 0.4676 | 0.4681 | 0.4721 |
| FACTR | | | | | |
| primary | 0.9137 | 0.9280 | 0.4101 | 0.3236 | 0.3684 |
| cont. 1 | 0.9209 | 0.9358 | 0.4029 | 0.3063 | 0.3611 |
| cont. 2 | 0.9281 | 0.9314 | 0.4101 | 0.3420 | 0.3759 |
| UPV-INAOE-Autoritas [9] | | | | | |
| primary | 0.9496 | 0.9706 | 0.3885 | 0.2613 | 0.3403 |
| bigIR [27] | | | | | |
| primary | 0.9640 | 1.0000 | 0.3957 | 0.1890 | 0.3333 |
| cont. 1 | 0.9640 | 1.0000 | 0.3957 | 0.1890 | 0.3333 |
| cont. 2 | 0.9424 | 0.9256 | 0.3525 | 0.3297 | 0.3405 |
| Check It Out [19] | | | | | |
| primary | 0.9640 | 1.0000 | 0.3957 | 0.1890 | 0.3333 |

Table 7. Performance of the participants’ submissions to the CheckThat! Task 2: fact checking Arabic. MAE is the metric used to rank the different teams. The runs include one primary and (at most) two contrasting submissions. The top-performing score per evaluation metric appears highlighted.

| | MAE | Macro MAE | ACC | Macro F1 | Macro Recall |
|--------------------------------|--------|-----------|--------|----------|--------------|
| FACTR | | | | | |
| primary | 0.6579 | 0.8914 | 0.5921 | 0.3730 | 0.3804 |
| cont. 1 | 0.7018 | 0.9461 | 0.5833 | 0.3691 | 0.3766 |
| cont. 2 | 0.6623 | 0.9153 | 0.5965 | 0.3657 | 0.3804 |
| UPV-INAOE-Autoritas [9] | | | | | |
| primary | 0.8202 | 1.0417 | 0.5175 | 0.2796 | 0.3027 |

6 Conclusions

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 in Arabic, relying on comments and factuality judgments from both factcheck.org and 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 evaluation metrics we have

released as a result of this lab should enable further research in check-worthiness estimation and in automatic claim verification.

References

1. Agez, R., Bosc, C., Lespagnol, C., Mothe, J., Petitcol, N.: IRIT at CheckThat! 2018. In: Cappellato et al. [4]
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: Cappellato et al. [4]
3. Barrón-Cedeño, A., Elsayed, T., Suwaileh, R., Màrquez, L., Gencheva, 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. [4]
4. Cappellato, L., Ferro, N., Nie, J.Y., Soulier, L. (eds.): Working Notes of CLEF 2018—Conference and Labs of the Evaluation Forum (2018)
5. 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)
6. 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)
7. 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)
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 on 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: An Approach based on External Sources to Detect Claims Credibility. In: Cappellato et al. [4]
10. 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. [4]
11. 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)
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. [4]
13. 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)

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. pp. 1835–1838. CIKM '15, Melbourne, Australia (10 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 (09 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 '18, New Orleans, LA, USA (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, Bulgaria (September 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. pp. 344–353. RANLP '17, Varna, Bulgaria (2017)
19. Lal, Y.K., Khattar, D., Kumar, V., Mishra, A., Varma, V.: Check It Out : Politics and Neural Networks. In: Cappellato et al. [4]
20. 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)
21. 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)
22. 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)
23. 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)
24. Shiralkar, P., Flammmini, 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)
25. 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)
26. 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. [4]
27. Yasser, K., Kutlu, M., , Elsayed, T.: bigIR at CLEF 2018: Detection and Verification of Check-Worthy Political Claims. In: Cappellato et al. [4]
28. Zuo, C., Karakas, A., Banerjee, R.: A Hybrid Recognition System for Check-worthy Claims Using Heuristics and Supervised Learning. In: Cappellato et al. [4]