CEUR-WS.org/Vol-2380/paper_270.pdf

Overview of the CLEF-2019 CheckThat! Lab: Automatic Identification and Verification of Claims. Task 2: Evidence and Factuality

Maram Hasanain[1], Reem Suwaileh[1], Tamer Elsayed[1],
Alberto Barrón-Cedeño[2], and Preslav Nakov[3]

[1] Computer Science and Engineering Department, Qatar University, Doha, Qatar
{maram.hasanain,rs081123,telsayed}@qu.edu.qa
[2] DIT, Università di Bologna, Forlì, Italy
a.barron@unibo.it
[3] Qatar Computing Research Institute, HBKU, Doha, Qatar
pnakov@qf.org.qa

Abstract. We present an overview of Task 2 of the second edition of the CheckThat! Lab at CLEF 2019. Task 2 asked (A) to rank a given set of Web pages with respect to a check-worthy claim based on their usefulness for fact-checking that claim, (B) to classify these same Web pages according to their degree of usefulness for fact-checking the target claim, (C) to identify useful passages from these pages, and (D) to use the useful pages to predict the claim's factuality. Task 2 at CheckThat! provided a full evaluation framework, consisting of data in Arabic (gathered and annotated from scratch) and evaluation based on normalized discounted cumulative gain (nDCG) for ranking, and F_1 for classification. Four teams submitted runs. The most successful approach to subtask A used learning-to-rank, while different classifiers were used in the other subtasks. We release to the research community all datasets from the lab as well as the evaluation scripts, which should enable further research in the important task of evidence-based automatic claim verification.

Keywords: Fact-Checking · Veracity · Evidence-based Verification · Fake News Detection · Computational Journalism

1 Introduction

The spread of "fake news" in all types of online media created a pressing need for automatic fake news detection systems [23]. The problem has various aspects [24], but here we are interested in identifying the information that is useful for fact-checking a given claim, and then also in predicting its factuality [5,13,20,22,25,28].

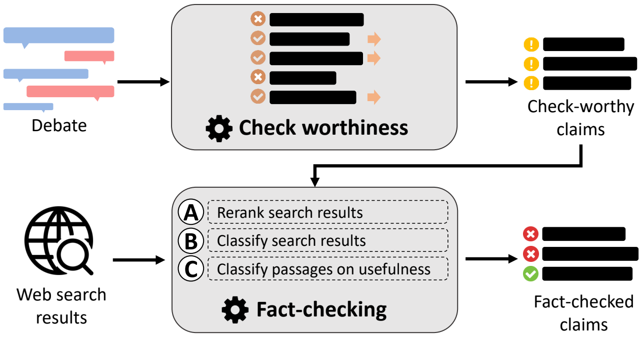


Fig. 1: Information verification pipeline with the two tasks in the CheckThat! lab: check-worthiness estimation and factuality verification.

Evidence-based fake news detection systems can serve fact-checking in two ways: (*i*) by facilitating the job of a human fact-checker, but not replacing her, and (*ii*) by increasing her trust in a system's decision [19,22,25]. We focus on the problem of checking the factuality of a claim, which has been studied before but rarely in the context of evidence-based fake news detection systems [3,4,7,15,17,21,27,29].

There are several challenges that make the development of automatic fake news detection systems difficult:

1. A fact-checking system is effective if it is able to identify a false claim before it reaches a large audience. Thus, the current speed at which claims spread on the Internet and social media imposes strict efficiency constraints on fact-checking systems.
2. The problem is difficult to the extent that, in some cases, even humans can hardly distinguish between fake and true news [24].
3. There are very few *large-scale* benchmark datasets that could be used to test and improve fake news detection systems [24,25].

Thus, in 2018 we started the CheckThat! lab on Automatic Identification and Verification of Political Claims [1,6,18]. We organized a second edition of the lab in 2019 [2,8,9], which aims at providing a full evaluation framework along with large-scale evaluation datasets. The lab this year is organized around two different tasks, which correspond to the main blocks in the verification pipeline, as depicted in Figure 1. This paper describes **Task 2: Evidence and Factuality**. This task focuses on extracting evidence from the Web to support the making of a veracity judgment for a given target claim. We divide Task 2 into the following four subtasks: (A) ranking Web pages with respect to a check-worthy claim based on their potential usefulness for fact-checking that claim; (B) classifying Web pages according to their degree of usefulness for fact-checking the target claim; (C) extracting passages from these Web pages that would be useful for fact-checking the target claim; and (D) using these useful pages to verify whether the target claim is factually true or not.

Since Task 2 in this edition of the lab had a different goal from last year's [18], we built a new dataset from scratch by manually curating claims, retrieving Web pages through a commercial search engine, and then hiring both in-house and crowd annotators to collect judgments for the four subtasks. As a result of our efforts, we release the CT19-T2 dataset, which contains Arabic claims as well as retrieved Web pages, along with three sets of annotations for the four subtasks.

Four teams participated in this year's Task 2, and they submitted 55% more runs compared to the 2018 edition [18]. The most successful systems relied on supervised machine learning models for both ranking and classification. We believe that there is still large room for improvement, and thus we release the annotated corpora and the evaluation scripts, which should enable further research on evidence-supported automatic claim verification.[1]

The remainder of this paper is organized as follows. Section 2 discusses the task in detail. Section 3 describes the dataset. Section 4 describes the participans' approaches and their performance on the four subtasks. Finally, Section 5 draws some conclusions and points to possible directions for future work.

2 Task Definition

Task 2 focuses on building tools to verify the factuality of a given claim. This is the first-ever version of this task, and we run it in Arabic.[2] The task is formally defined as follows:

> Given a check-worthy claim c and a set of Web pages P (the retrieved results of Web search in response to a search query representing c), identify which of the Web pages (and passages A of those Web pages) can be useful for assisting a human in fact-checking the claim. Finally, determine the factuality of the claim according to the supporting information in the useful pages and passages.

As Figure 2 shows, the task is divided into four subtasks that target different aspects of the problem.

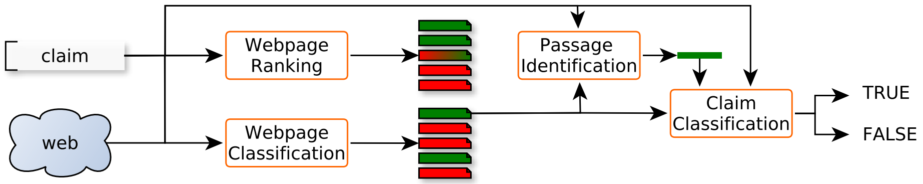


Fig. 2: A zoom into the four subtasks in Task 2.

[1] http://sites.google.com/view/clef2019-checkthat/datasets-tools

[2] In 2018, we had a different fact-checking task, where no retrieved Web pages were provided [6].

Subtask A, Webpage ranking: *Rank the Web pages P based on how useful they are for verifying the target claim.* The systems are asked to produce a score for each page, based on which the pages would be ranked. See the definition of "useful" below.

Subtask B, Webpage classification: *Classify each Web page $p \in P$ as "very useful for verification", "useful", "not useful", or "not relevant."* A page p is considered *very useful* for verification if it is *relevant* with respect to c (i.e., on-topic and discussing the claim) and it *provides sufficient evidence* to verify the veracity of c, such that there is no need for another document to be considered for verifying this claim. A page is *useful* for verification if it is relevant to the claim and provides some valid evidence, but it is *not solely sufficient* to determine the c's veracity on its own. The evidence can be a source, some statistics, a quote, etc.

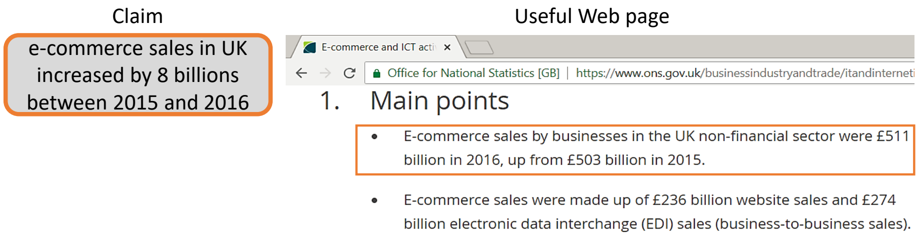
A particular piece of evidence is considered not valid if the source cannot be verified or is ambiguous (e.g., expressing that "experts say that..." without mentioning who those experts are), or it is just an opinion of a person/expert instead of an objective analysis.

Notice that this is different from *stance detection* as a page might agree with a claim, but it might still lack evidence to verify it.

Subtask C, Passage identification: *Find passages within the Web pages P that are* ***useful*** *for claim verification.* Again, notice that this is different from stance detection.

Subtask D, Claim classification: *Classify the claim's factuality as "true" or "false."* The claim is considered true if it is accurate as stated (or there is sufficient reliable evidence supporting it), otherwise it is considered false.

Figure 3 shows an example: a Web page considered as useful for verifying the given claim, since it has evidence showing the claim to be true and it is an official United Kingdom page on national statistics. The useful passage in the page is the one reporting the supporting statistics. For the sake of readability, the example is given in English, but this year the task was offered only in Arabic.


Claim
Useful Web page
E-commerce and ICT acti
Office for National Statistics [GB]
https://www.ons.gov.uk/businessindustryandtrade/itandinterneti

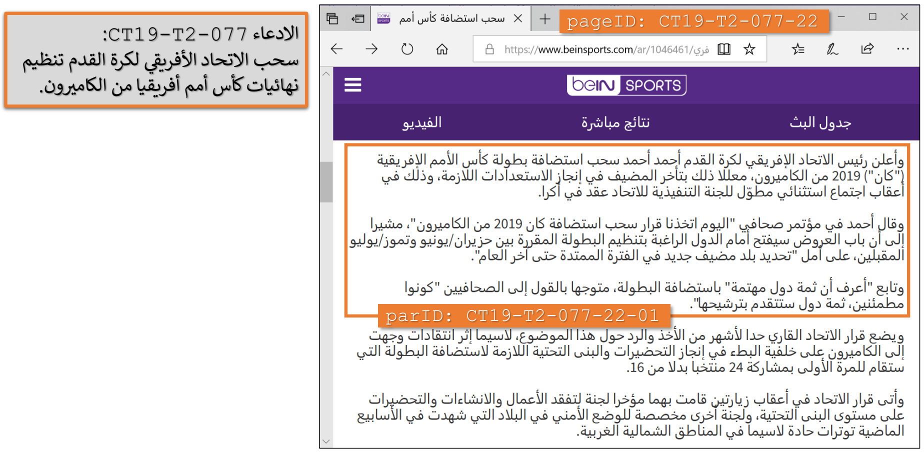
e-commerce sales in UK increased by 8 billions between 2015 and 2016

1. Main points

- E-commerce sales by businesses in the UK non-financial sector were £511 billion in 2016, up from £503 billion in 2015.
- E-commerce sales were made up of £236 billion website sales and £274 billion electronic data interchange (EDI) sales (business-to-business sales).

Fig. 3: English claim, a useful Web page, and a useful passage (in the orange rectangle on the right).

Figure 4 shows an Arabic example of an actual claim, a useful Web page, and a paragraph from our training dataset. The claim translates to English as follows: “The Confederation of African Football has withdrawn the organization of the Africa Cup of Nations from Cameroon.” The page shows a news article reporting the news; it is useful for facrt-checking since it contains a quotation of an official statement confirming the claim.



الادعاء CT19-T2-077:
سحب الاتحاد الأفريقي لكرة القدم تنظيم
نهائيات كأس أمم أفريقيا من الكاميرون.

وأعلن رئيس الاتحاد الإفريقي لكرة القدم أحمد أحمد سحب استضافة بطولة كأس الأمم الإفريقية ("كان") 2019 من الكاميرون، معللا ذلك بتأخر المضيف في إنجاز الاستعدادات اللازمة، وذلك في أعقاب اجتماع استثنائي مطوّل للجنة التنفيذية للاتحاد عقد في أكرا.

وقال أحمد في مؤتمر صحافي "اليوم اتخذنا قرار سحب استضافة كان 2019 من الكاميرون"، مشيرا إلى أن باب العروض سيفتح أمام الدول الراغبة بتنظيم البطولة المقررة بين حزيران/يونيو وتموز/يوليو المقبلين، على أمل "تحديد بلد مضيف جديد في الفترة الممتدة حتى آخر العام".

وتابع "أعرف أن ثمة دول مهتمة" باستضافة البطولة، متوجها بالقول إلى الصحافيين "كونوا مطمئنين، ثمة دول ستتقدم بترشيحها".

ويضع قرار الاتحاد القاري حدا لأشهر من الأخذ والرد حول هذا الموضوع، لاسيما إثر انتقادات وجهت إلى الكاميرون على خلفية البطء في إنجاز التحضيرات والبنى التحتية اللازمة لاستضافة البطولة التي ستقام للمرة الأولى بمشاركة 24 منتخبا بدلا من 16.

وأتى قرار الاتحاد في أعقاب زيارتين قامت بهما مؤخرا لجنة لتفقد الأعمال والانشاءات والتحضيرات على مستوى البنى التحتية، ولجنة أخرى مخصصة للوضع الأمني في البلاد التي شهدت في الأسابيع الماضية توترات حادة لاسيما في المناطق الشمالية الغربية.

Fig. 4: Arabic claim, a useful Web page, and a useful passage (in the orange rectangle on the right) from the training data.

3 Dataset

Collecting claims. Subtasks A, B, and C are all new to the lab this year. As a result, we built a new evaluation dataset to support all subtasks —the CT19-T2 corpus. We selected 69 claims from multiple sources including a pre-existing set of Arabic claims [5], a survey in which we asked the public to provide examples of claims they have heard of, and some headlines from six Arabic news agencies that we rewrote into claims. The news agencies selected are well-known in the Arab world: Al Jazeera, BBC Arabic, CNN Arabic, Al Youm Al Sabea, Al Arabiya, and RT Arabic. We made sure the claims span different topical domains, e.g., health or sports, besides politics. Ten claims were released for training and the rest were used for testing.

Labeling claims. We acquired the veracity labels for the claims in two steps. First, two of the lab organizers labelled each of the 69 claims independently. Then, they met to resolve any disagreements, and thus reach consensus on the veracity labels for all claims.

Labeling pages and passages. We formulated a query representing each claim, and we issued it against the Google search engine in order to retrieve the top 100 Web pages. We used a language detection tool to filter out non-Arabic pages, and we eventually used the top-50 of the remaining pages. The labeling pipeline was carried out as follows:

1. **Relevance**. We first identified relevant pages, since we assume that non-relevant pages cannot be useful for claim verification, and thus should be filtered out from any further labeling. In order to speedup the relevance labeling process, we hired two types of annotators: Amazon Mechanical Turk crowd-workers and in-house annotators. Each page was labeled by *three* annotators, and the majority label was used as the final page label.
2. **Usefulness as a whole**. Relevant pages were then given to in-house annotators to be labeled for usefulness using a two-way classification scheme: *useful* (including *very useful*, but not distinguishing between the two) and *not useful*. Similarly to relevance labeling, each page was labeled by three annotators, and the final page label was the majority label.
3. **Useful vs. very useful**. One of the lab organizers went over the useful pages from step 2 and further classified them into *useful* and *very useful*. We opted for this design since we found through pilot studies that the annotators found it difficult to differentiate between *useful* and *very useful* pages.
4. **Splitting into passages**. We manually split the *useful* and the *very useful* pages into passages, as we found that the automatic techniques for splitting pages into passages were not accurate enough.
5. **Useful passages**. Finally, one of the lab organizers labelled each passage for usefulness. Due to time constraints, we could not split the pages and label the resulting passages for all the claims in the *testing set*. Thus, we only release labels for passages of pages corresponding to 33 out of the 59 testing claims. Note that this only affects subtask C.

Table 1 summarizes the statistics about the training and the test data. Note that the passages in the test set are for 33 claims only (see above).

Table 1: Statistics about the CT19-T2 corpus for Task 2.

	Claims		**Pages**		**Passages**	
Set	**Total**	**True**	**Total**	**Useful**	**Total**	**Useful**
Training	10	5	395	32	167	54
Test	59	30	2,641	575	1,722	578

Table 2: Summary of participants' approaches.

Subtask	**A**			**B**				**C**		**D2**		
Team	[10]	[12]	[26]	[10]	[11]	[12]	[26]	[10]	[12]	[11]	[12]	[26]
Representation												
BERT embeddings	☑			☑				☑				
Word embeddings		☑			☑	☑				☑	☑	
Bag of words			☑				☑	☑	☑		☑	☑
Models												
Feed-Forward DNN	☑			☑				☑				
Naïve Bayes		☑				☑			☑			
Random Forest		☑				☑			☑		☑	
Gradient Boosting						☑					☑	
Support vector machine		☑				☑			☑			
Enhanced Sequential Inference					☑					☑		
Rule-based					☑					☑		
Features												
Content	☑		☑	☑	☑	☑	☑	☑	☑	☑		☑
Credibility		☑				☑					☑	
Similarity		☑			☑			☑	☑	☑	☑	
Statistical		☑	☑		☑			☑	☑	☑	☑	
External data	☑			☑				☑				

Teams
[10] TheEarthIsFlat
[11] UPV-UMA
[12] bigIR
[26] EvolutionTeam

4 Evaluation

In this section we describe the participants' approaches to the different subtasks. Table 2 summarizes the approaches. We also present the evaluation set-up used to evaluate each subtask, and then we present and discuss the results.

4.1 Subtask A

Runs. Three teams participated in this subtask submitting a total of seven runs [10,12,26]. There were two kinds of approaches. In the first kind, token-level BERT embeddings were used with text classification to rank pages [10]. In the second kind, the runs used a learning-to-rank model based on different classifiers, including Naïve Bayes and Random Forest, with a variety of features [12]. In one run, external data was used to train the text classifier [10], while all other runs represent systems trained on the provided labelled data only.

Table 3: Results for Subtask 2.A, ordered by nDCG@10 score. The runs that used external data are marked with *.

Team	**Run**	**nDCG@5**	**nDCG@10**	**nDCG@15**	**nDCG@20**
Baseline	–	**0.52**	**0.55**	**0.58**	**0.61**
bigIR	1	0.47	0.50	0.54	0.55
bigIR	3	0.41	0.47	0.50	0.52
EvolutionTeam	1	0.40	0.45	0.48	0.51
bigIR	4	0.39	0.45	0.48	0.51
bigIR	2	0.38	0.41	0.45	0.47
TheEarthIsFlat2A	1	0.08	0.10	0.12	0.14
TheEarthIsFlat2A*	2	0.05	0.07	0.10	0.12

Evaluation measures. Subtask A was modeled as a ranking problem, in which *very useful* and *useful* pages should be ranked on top. Since this is a graded usefulness problem, we evaluate it using the mean of Normalized Discounted Cumulative Gain (nDCG) [14,16]. In particular, we consider nDCG@10 (i.e., nDCG computed at cutoff 10) as the official evaluation measure for this subtask, but we report nDCG at cutoffs 5, 15, and 20 as well. For all measures, we used macro-averaging over the testing claims.

Results. Table 3 shows the results for all seven runs. It also includes the results for a simple baseline: the original ranking in the search result list. We can see that the baseline surprisingly performs the best. This is due to the fact that in our definition of usefulness, useful pages must be relevant, and Google, as an effective search engine, has managed to rank relevant pages (and consequently, many of the *useful* pages) first. This result indicates that the task of ranking pages by usefulness is not easy and systems need to be further developed in order to differentiate between relevance and usefulness, while also benefiting from the relevance-based rank of a page.

4.2 Subtask B

Runs. Four teams participated in this subtask, submitting a total of eight runs [10,11,12,26]. All runs used supervised text classification models, such as Random Forest and Gradient Boosting [12]. Two teams opted for using embedding-based language representations: one considered word embeddings [11] and another BERT-based token-level embeddings [10]. In one run, external data was used to train the model [10], while all the remaining runs were trained on the provided training data only.

Evaluation measures. Similarly to Subtask A, Subtask B also aims at identifying useful pages for claim verification, but it is modeled as a *classification*, rather than a ranking problem. Thus, here we use standard evaluation measures for text classification: Precision, Recall, F_1, and Accuracy, with F_1 being the official score for the task.

Table 4: Results for Subtask 2.B for 2-way and 4-way classification. The runs are ranked by F_1 score. Runs tagged with * used external data.

(a) 2-way classification

Team	**Run**	$\mathbf{F_1}$	**P**	**R**	**Acc**
Baseline	–	**0.42**	**0.30**	**0.72**	**0.57**
UPV-UMA	1	0.38	0.26	0.73	0.49
bigIR	1	0.08	0.40	0.04	0.78
bigIR	3	0.07	0.39	0.04	0.78
bigIR	4	0.07	0.57	0.04	0.78
bigIR	2	0.04	0.22	0.02	0.77
TheEarthIsFlat	1	0.00	0.00	0.00	0.78
TheEarthIsFlat*	2	0.00	0.00	0.00	0.78
EvolutionTeam	1	0.00	0.00	0.00	0.78

(b) 4-way classification

Team	**Run**	$\mathbf{F_1}$	**P**	**R**	**Acc**
TheEarthIsFlat	1	0.31	0.28	0.36	0.59
bigIR	3	0.31	0.37	0.33	0.58
TheEarthIsFlat*	2	0.30	0.27	0.35	0.60
bigIR	4	0.30	0.41	0.32	0.57
EvolutionTeam	1	0.29	0.26	0.33	0.58
Baseline	–	**0.28**	**0.32**	**0.32**	**0.30**
UPV-UMA	1	0.23	0.30	0.29	0.24
bigIR	1	0.16	0.25	0.23	0.26
bigIR	2	0.16	0.25	0.22	0.25

Results. Table 4a reports the results for 2-way classification —*useful/very useful* vs. *not useful/not relevant*—, reporting results for predicting the *useful* class. Table 4b shows the results for 4-way classification —*very useful* vs. *useful* vs. *not useful* vs. *not relevant*—, reporting macro-averaged scores over the four classes, for each of the evaluation measures.

We include a baseline: the original ranking from the search results list. The baseline assumes the top-50% of the results to be *useful* and the rest *not useful* for the 2-way classification. For the 4-way classification, the baseline assumes the top-25% to be *very useful*, the next 25% to be *useful*, the third 25% to be *not useful*, and the rest to be *not relevant*.

Table 4a shows that almost all systems struggled to retrieve any *useful* pages at all. Team UPV-UMA is the only one that managed to achieve high recall. This is probably due to the *useful* class being under-represented in the training dataset, while being much more frequent in the test dataset: we can see in Table 1 that it covers just 8% of the training examples, but 22% of the testing ones. Training the models with a limited number of *useful* pages might have caused them to learn to underpredict this class. Similarly to Subtask A, the simple baseline that assumes the top-ranked pages to be more useful is most effective. This again can be due to the correlation between usefulness and relevance.

Comparing the results in Table 4a to those in Table 4b, we notice a very different performance ranking; runs that had the worst performance at finding *useful* pages, are actually among the best runs in the 4-way classification. These runs were able to effectively detect the *not relevant* and *not useful* pages as compared to *useful* ones. The baseline, which was effective at identifying *useful* pages, is not as effective at identifying pages in the other classes. This might indicate that *not useful* and *not relevant* pages are not always at the bottom of the ranked list as this baseline assumes, which sheds some light on the importance of usefulness estimation to aid fact-checking.

Table 5: Performance of the models when predicting useful passages for Subtask 2.C. Precision, recall and F_1 are calculated with respect to the positive class, i.e., *useful.* The runs are ranked by F_1.

Team	**Run**	$\mathbf{F_1}$	**P**	**R**	**Acc**
TheEarthIsFlat2Cnoext	1	0.56	0.40	0.94	0.51
TheEarthIsFlat2Cnoext	2	0.55	0.41	0.87	0.53
bigIR	2	0.40	0.39	0.42	0.58
bigIR	1	0.39	0.38	0.41	0.58
bigIR	4	0.37	0.37	0.38	0.57
Baseline		**0.37**	**0.42**	**0.39**	**0.57**
bigIR	3	0.19	0.33	0.14	0.61

One additional factor that might have caused such a varied ranking of runs is our own observation on the difficulty and subjectivity of differentiating between *useful* and *very useful* pages. At annotation time, we observed that annotators and even lab organizers were not able to easily distinguish between these two types of pages.

4.3 Subtask C

Runs. Two teams participated in this subtask [10,12], submitting a total of seven runs. One of the teams used text classifiers including Naïve Bayes and SVM with a variety of features such as bag-of-words and named entities [12]. All runs also considered using the similarity between the claim and the passages as a feature in their models.

Evaluation measures. Subtask C aims at identifying useful passages for claim verification and we modeled it as a classification problem. As in typical classification problems, we evaluated it using Precision, Recall, F_1, and Accuracy, with F_1 being the official evaluation measure.

Results. Table 5 shows the evaluation results, including a simple baseline that assumes the first passage in a page to be *not useful*, the next two passages to be *useful*, and the remaining passages to be *not useful.* This baseline is motivated by our observation that *useful* passages are typically located at the heart of the document following some introductory passage(s).

Team TheEarthIsFlat managed to identify most *useful* passages, thus achieving a very high recall (0.94 for its run 1), with a relatively similar precision to the other runs, and the baseline. Note that in all the runs by the bigIR system, as well as in the baseline system, the precision and the recall are fairly balanced. The baseline performs almost as well as the four runs by bigIR. This indicates that considering the position of the passage in a page might be a useful feature when predicting the passage usefulness, and thus it should be considered when addressing the problem.

Table 6: Results for Subtask 2.D for both cycles 1 and 2. The runs are ranked by F_1 score. The runs tagged with a * used external data.

(a) Cycle 1, where the usefulness of the Web pages was unknown.

Team	**F$_1$**	**P**	**R**	**Acc**
EvolutionTeam	0.48	0.55	0.53	0.53
Baseline	**0.34**	**0.25**	**0.50**	**0.51**

(b) Cycle 2, where the the usefulness of the Web pages was known.

Team	**Run**	**F$_1$**	**P**	**R**	**Acc**
UPV-UMA*	21	0.62	0.63	0.63	0.63
UPV-UMA*	11	0.55	0.56	0.56	0.56
UPV-UMA*	22	0.54	0.60	0.57	0.58
bigIR	1	0.53	0.55	0.55	0.54
bigIR	3	0.53	0.55	0.54	0.54
bigIR	2	0.51	0.53	0.53	0.53
bigIR	4	0.51	0.53	0.53	0.53
UPV-UMA*	12	0.51	0.65	0.57	0.58
EvolutionTeam	1	0.43	0.45	0.46	0.46
Baseline		**0.34**	**0.25**	**0.50**	**0.51**

4.4 Subtask D

The main aim of Task 2 was to study the effect of using identified *useful* and *very useful* pages for claim verification. Thus, we had two evaluation cycles for Subtask D. In the first cycle, the teams were asked to fact-check claims using all the Web pages, without knowing which were *useful*/*very useful*. In the second cycle, the usefulness labels were released in order to allow the systems to fact-check the claims using only *useful*/*very useful* Web pages.

Runs. Two teams participated in cycle 1, submitting one run each [12,26], but one of the runs was invalid, and thus there is only one official run. Cycle 2 attracted more participation: three teams with nine runs [11,12,26]. Thus, we will focus our discussion on cycle 2. One team opted for using textual entailment with embedding-based representations for classification [11]. Another team used text classifiers such as Gradient Boosting and Random Forests [12]. External data was used to train the textual entailment component of the system in four runs, whereas the remaining runs were trained on the provided data only.

Evaluation measures. Subtask D aims at predicting a claim's veracity. It is a classification task, and thus we evaluate it using Precision, Recall, F_1, and Accuracy, with F_1 being the official measure.

Results. Table 6 shows the results for cycles 1 and 2, where we macro-average precision, recall, and F_1 over the two classes. We show the results for a simple majority-class baseline, which all runs manage to beat for both cycles.

Due to the low participation in cycle 1, it is difficult to draw conclusions about whether providing systems with useful pages helps to improve their performance.

5 Conclusion and Future Work

We have presented an overview of Task 2 of the CLEF–2019 `CheckThat!` Lab on Automatic Identification and Verification of Claims, which is the second edition of the lab. Task 2 was designed to aid a human who is fact-checking a claim. It asked systems (A) to rank Web pages with respect to a check-worthy claim based on their usefulness for fact-checking that claim, (B) to classify the Web pages according to their degree of usefulness, (C) to identify useful passages from these pages, and (D) to use the useful pages to predict the claim's factuality. As part of the lab, we release a dataset in Arabic in order to enable further research in automatic claim verification.

A total of four teams participated in the task (compared to two in 2018) submitting a total of 31 runs. The evaluation results show that the most successful approaches to Task 2 used learning-to-rank for subtask A, while different classifiers were used in the other subtasks.

Although one of the aims of the lab was to study the effect of using *useful* pages for claim verification, the low participation in the first cycle of subtask D has hindered carrying such a study. In the future, we plan to setup this subtask, so that the teams would need to participate in both cycles in order for their runs to be considered valid. We also plan to extend the dataset for Task 2 to include claims in at least one language other than Arabic.

Acknowledgments

This work was made possible in part by grant# NPRP 7-1330-2-483 from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors.

This research is also part of the Tanbih project,[3] which aims to limit the effect of "fake news", propaganda and media bias by making users aware of what they are reading. The project is developed in collaboration between the Qatar Computing Research Institute (QCRI), HBKU and the MIT Computer Science and Artificial Intelligence Laboratory (CSAIL).

References

1. Atanasova, P., Màrquez, L., Barrón-Cedeño, A., Elsayed, T., Suwaileh, R., Zaghouani, W., Kyuchukov, S., Da San Martino, G., Nakov, P.: Overview of the CLEF-2018 CheckThat! Lab on automatic identification and verification of political claims, Task 1: Check-worthiness. In: CLEF 2018 Working Notes. Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum. CEUR Workshop Proceedings, CEUR-WS.org (2018)
2. Atanasova, P., Nakov, P., Karadzhov, G., Mohtarami, M., Martino, G.D.S.: Overview of the CLEF-2019 CheckThat! Lab on Automatic Identification and Verification of Claims. Task 1: Check-Worthiness. In: Cappellato, L., Ferro, N., Losada,

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

D., Müller, H. (eds.) CLEF 2019 Working Notes. Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum. CEUR Workshop Proceedings, CEUR-WS.org, Lugano, Switzerland (2019)

3. Ba, M.L., Berti-Equille, L., Shah, K., Hammady, H.M.: VERA: A platform for veracity estimation over web data. In: Proceedings of the 25th International Conference Companion on World Wide Web. pp. 159–162. WWW '16 (2016)
4. Baly, R., Karadzhov, G., Saleh, A., Glass, J., Nakov, P.: Multi-task ordinal regression for jointly predicting the trustworthiness and the leading political ideology of news media. In: Proceedings of the 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 2109–2116. NAACL-HLT '19, Minneapolis, MN, USA (2019)
5. Baly, R., Mohtarami, M., Glass, J., Màrquez, L., Moschitti, A., Nakov, P.: Integrating stance detection and fact checking in a unified corpus. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 21–27. NAACL-HLT '18, New Orleans, Louisiana, USA (2018)
6. Barrón-Cedeño, A., Elsayed, T., Suwaileh, R., Màrquez, L., Atanasova, P., Zaghouani, W., Kyuchukov, S., Da San Martino, G., Nakov, P.: Overview of the CLEF-2018 CheckThat! Lab on automatic identification and verification of political claims, Task 2: Factuality. In: CLEF 2018 Working Notes. Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum. CEUR Workshop Proceedings, CEUR-WS.org (2018)
7. 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)
8. Elsayed, T., Nakov, P., Barrón-Cedeño, A., Hasanain, M., Suwaileh, R., Da San Martino, G., Atanasova, P.: CheckThat! at CLEF 2019: Automatic identification and verification of claims. In: Azzopardi, L., Stein, B., Fuhr, N., Mayr, P., Hauff, C., Hiemstra, D. (eds.) Advances in Information Retrieval. pp. 309–315. Springer International Publishing (2019)
9. Elsayed, T., Nakov, P., Barrón-Cedeño, A., Hasanain, M., Suwaileh, R., Da San Martino, G., Atanasova, P.: Overview of the CLEF-2019 CheckThat!: Automatic identification and verification of claims. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. LNCS, Lugano, Switzerland (2019)
10. Favano, L., Carman, M., Lanzi, P.: TheEarthIsFlat's submission to CLEF'19 CheckThat! challenge. In: CLEF 2019 Working Notes. Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum. CEUR Workshop Proceedings, CEUR-WS.org, Lugano, Switzerland (2019)
11. Ghanem, B., Glavaš, G., Giachanou, A., Ponzetto, S., Rosso, P., Rangel, F.: UPV-UMA at CheckThat! Lab: Verifying Arabic claims using cross lingual approach. In: CLEF 2019 Working Notes. Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum. CEUR Workshop Proceedings, CEUR-WS.org, Lugano, Switzerland (2019)
12. Haouari, F., Ali, Z., Elsayed, T.: bigIR at CLEF 2019: Automatic verification of Arabic claims over the web. In: CLEF 2019 Working Notes. Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum. CEUR Workshop Proceedings, CEUR-WS.org, Lugano, Switzerland (2019)
13. 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 Compu-

tational Linguistics. pp. 26–30. NAACL-HLT '18, New Orleans, Louisiana, USA (2018)

14. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information Systems (TOIS) **20**(4), 422–446 (2002)
15. Ma, J., Gao, W., Mitra, P., Kwon, S., Jansen, B.J., Wong, K.F., Cha, M.: Detecting rumors from microblogs with recurrent neural networks. In: Proceedings of the 25th International Joint Conference on Artificial Intelligence. pp. 3818–3824. IJCAI '16, New York, New York, USA (2016)
16. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press, New York, NY, USA (2008)
17. Mukherjee, S., Weikum, G.: Leveraging joint interactions for credibility analysis in news communities. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. pp. 353–362. CIKM '15, Melbourne, Australia (2015)
18. Nakov, P., Barrón-Cedeño, A., Elsayed, T., Suwaileh, R., Màrquez, L., Zaghouani, W., Atanasova, P., Kyuchukov, S., Da San Martino, G.: Overview of the CLEF-2018 CheckThat! Lab on automatic identification and verification of political claims. In: Proceedings of the Ninth International Conference of the CLEF Association: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Lecture Notes in Computer Science, Springer (2018)
19. Nguyen, A.T., Kharosekar, A., Lease, M., Wallace, B.: An interpretable joint graphical model for fact-checking from crowds. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 1511–1518. AAAI '18, New Orleans, LA, USA (2018)
20. Nie, Y., Chen, H., Bansal, M.: Combining fact extraction and verification with neural semantic matching networks. In: Proceedings of the 33rd AAAI Conference on Artificial Intelligence. AAAI '19, Honolulu, Hawaii, USA (2019)
21. Popat, K., Mukherjee, S., Strötgen, J., Weikum, G.: Credibility assessment of textual claims on the web. In: Proceedings of the 25th ACM International Conference on Information and Knowledge Management. pp. 2173–2178. CIKM '16, Indianapolis, Indiana, USA (2016)
22. Popat, K., Mukherjee, S., Yates, A., Weikum, G.: DeClarE: Debunking fake news and false claims using evidence-aware deep learning. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 22–32. EMNLP '18, Brussels, Belgium (2018)
23. Rubin, V.L., Chen, Y., Conroy, N.J.: Deception detection for news: three types of fakes. In: Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community. p. 83. American Society for Information Science (2015)
24. Shu, K., Sliva, A., Wang, S., Tang, J., Liu, H.: Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations Newsletter **19**(1), 22–36 (2017)
25. Thorne, J., Vlachos, A., Christodoulopoulos, C., Mittal, A.: FEVER: a large-scale dataset for Fact Extraction and VERification. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 809–819. NAACL-HLT '18, New Orleans, LA, USA (2018)
26. Touahri, I., Mazroui, A.: Automatic identification and verification of political claims. In: CLEF 2019 Working Notes. Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum. CEUR Workshop Proceedings, CEUR-WS.org, Lugano, Switzerland (2019)

27. Yasser, K., Kutlu, M., Elsayed, T.: Re-ranking web search results for better fact-checking: A preliminary study. In: Proceedings of 27th ACM International Conference on Information and Knowledge Management. pp. 1783–1786. CIKM '19, Turin, Italy (2018)
28. Yoneda, T., Mitchell, J., Welbl, J., Stenetorp, P., Riedel, S.: UCL machine reading group: Four factor framework for fact finding (HexaF). In: Proceedings of the First Workshop on Fact Extraction and VERification. pp. 97–102. FEVER '18, Brussels, Belgium (2018)
29. Zubiaga, A., Liakata, M., Procter, R., Hoi, G.W.S., Tolmie, P.: Analysing how people orient to and spread rumours in social media by looking at conversational threads. PloS one **11**(3), e0150989 (2016)