CheckThat! at CLEF 2019: Automatic identification and verification of claims

CheckThat! at CLEF 2019: Automatic Identification and Verification of Claims

Tamer Elsayed[1], Preslav Nakov[2], Alberto Barrón-Cedeño[2], Maram Hasanain[1], Reem Suwaileh[1], Giovanni Da San Martino[2], and Pepa Atanasova[3]

[1] Qatar University, Doha, Qatar
[2] Qatar Computing Research Institute, HBKU, Doha, Qatar
[3] Sofia University, Sofia, Bulgaria
{telsayed,maram.hasanain,reem.suwaileh}@qu.edu.qa
{pnakov,albarron,gmartino}@hbku.edu.qa
pepa.k.gencheva@gmail.com

Abstract. We introduce the second edition of the CheckThat! Lab, part of the 2019 Cross-Language Evaluation Forum (CLEF). CheckThat! proposes two complementary tasks. Task 1: predict which claims in a political debate should be prioritized for fact-checking. Task 2: rank Web-retrieved pages against a check-worthy claim based on their usefulness for fact-checking, extract useful passages from those pages, and then use them all to decide whether the claim is factually true or false. Checkthat! provides a full evaluation framework, consisting of data in English (derived from fact-checking sources) and Arabic (gathered and annotated from scratch) and evaluation based on mean average precision (MAP) for ranking and F_1 for classification tasks.

1 Overview

The current coverage of news 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, or a blog note can spread almost instantaneously. This proliferation speed leaves little time to double-check claims against the facts, which has proven critical in electoral campaigns, e.g., during the 2016 US presidential campaign in the USA and during Brexit. Indeed, some politicians were fast to notice that when it comes to shaping public opinion, facts were secondary and that appealing to emotions and beliefs worked better, especially in social media. It has been even proposed that this was marking the dawn of a *post-truth age*.

Investigative journalists and volunteers have been working hard trying to get to the root of a claim and to present solid evidence in favor or against it. However, manual fact-checking is very time-consuming, and thus automatic methods have been proposed as a way to speed-up the process. For instance, there has been work on checking the factuality/credibility of a claim, of a news article, or of an entire news outlet [2,4,6,8,9,10,11,13,15]. However, less attention has been paid to other steps of the fact-checking pipeline shown in Figure 1, e.g., check-worthiness estimation has been severely understudied as a problem [1,5,7].

A typical fact-checking pipeline includes the following steps. First, check-worthy text fragments are identified. Then, documents that might be useful for fact-checking the claim [14] are retrieved from various sources, and supporting evidence is extracted. By comparing a claim against the retrieved evidence, a system can determine whether the claim is likely true or likely false (or unsure, if no supporting evidence either way could be found). This **CheckThat!** CLEF 2019 lab addresses these understudied aspects through two tasks:

Task 1: Check-Worthiness. The task aims at predicting which claim in a political debate should be prioritized for fact-checking.

Task 2: Evidence and Factuality. The task focuses on extracting evidence to support fact-checking a claim.

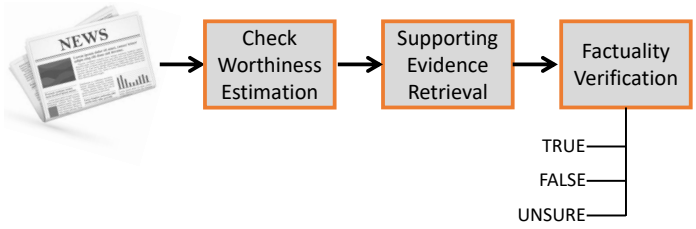

Fig. 1. Information verification pipeline: our two tasks cover all three steps.

2 Usage Scenarios

Automated systems for claim identification and verification could be very useful as supportive technology for investigative journalism. They provide assistance and guidance and save time. A system could automatically identify check-worthy claims and present them to the journalist as a ranking from more to less relevant. Additionally, for a claim, the system could identify documents that are *useful* for humans to manually fact-check and it could also estimate a *veracity score* supported by evidence extracted from such documents, which would help the journalist to focus on the most outstanding cases.

Another useful scenario, with the potential of impacting larger communities, would be helping the social media users who get a large flow of claims daily and want help in verifying them.

3 Target Audience

The main targets for nourishing the list of participants are the information retrieval, computational linguistics, and machine learning communities. We also hope that the lab attracts neighboring communities that would be interested in the problem, maybe from slightly different angles, e.g., social computing, social sciences, and investigative journalism.

4 Description of the Tasks

CheckThat! 2019[4] is a continuation of the evaluation lab at CLEF-2018 [12].[5] It is organized around two different tasks, which correspond to the three main blocks in the verification pipeline (Figure 1): check-worthiness estimation (**Task 1**), and extracting supporting evidence and factuality verification (**Task 2**). We address these two tasks separately in order to ease the participation and to have independent evaluations and more meaningful comparisons of systems.

4.1 Task 1: Check-Worthiness

Task 1 is defined as follows: *Given a political debate or a transcribed speech, segmented into sentences, with speakers annotated, identify which sentence should be prioritized for fact-checking.* This is a ranking task and systems are required to produce a score per sentence, according to which the ranking will be performed. This task will be run in English.

Dataset The training data for *Task 1* is ready. We selected four transcripts of the 2016 US election: one vice-presidential and three presidential debates. For each debate, we used the publicly-available manual analysis about it from nine reputable fact-checking sources (ABC News, Chicago Tribune, CNN, FactCheck.org, NPR, PolitiFact, The Guardian, The New York Times, and The Washington Post). This could include not just a statement about factuality, but any free text that journalists decided to add, e.g., links to biographies or behavioral analysis of the opponents and moderators. We converted this to binary annotation about whether a particular sentence was annotated for factuality by a given source.

The training dataset of four debates, contains a total of 5,415 annotated sentences in context, with 880 of them being identified as check-worthy by at least one of the sources. The agreement between the sources is not high. The reason for this is that different media aim at annotating sentences according to their own editorial line, rather than trying to be exhaustive in any way. This suggests that the task of predicting which sentence would contain check-worthy claims will be challenging. Thus, we focus on a ranking task rather than on absolute predictions.

The test set for *Task 1* will be created following the same approach as the training data: more debates will be collected with the same nine fact-checking sources to annotate the check-worthy claims. The volume of this data will be around 25% of the training set size.

Table 1 shows an excerpt from the first presidential debate in the American elections in 2016 together with the annotation flag (0 or 1) indicating whether each of the media has fact-checked the claim. The positive examples for *Task 1* will be those fact-checked by at least one source (those highlighted in blue in Table 1).

[4] http://alt.qcri.org/clef2019-checkthat/
[5] http://alt.qcri.org/clef2018-factcheck/

Speaker	Total	CT	ABC	CNN	WP	NPR	PF	TG	NYT	FC	Text
Clinton	0	0	0	0	0	0	0	0	0	0	So we're now on the precipice of having a potentially much better economy, but the last thing we need to do is to go back to the policies that failed us in the first place.
Clinton	6	1	1	0	0	1	1	0	1	1	Independent experts have looked at what I've proposed and looked at what Donald's proposed, and basically they've said this, that if his tax plan, which would blow up the debt by over $5 trillion and would in some instances disadvantage middle-class families compared to the wealthy, were to go into effect, we would lose 3.5 million jobs and maybe have another recession.
Clinton	1	1	0	0	0	0	0	0	0	0	They've looked at my plans and they've said, OK, if we can do this, and I intend to get it done, we will have 10 million more new jobs, because we will be making investments where we can grow the economy.
Clinton	0	0	0	0	0	0	0	0	0	0	Take clean energy.
Clinton	0	0	0	0	0	0	0	0	0	0	Some country is going to be the clean- energy superpower of the 21st century.
Clinton	6	1	1	1	1	0	0	1	0	1	Donald thinks that climate change is a hoax perpetrated by the Chinese.
Clinton	0	0	0	0	0	0	0	0	0	0	I think it's real.
Trump	5	1	1	0	1	1	1	0	0	0	I did not.

Table 1. Excerpt from the transcript of the first US Presidential Debate in 2016, annotated by nine sources: Chicago Tribune, ABC News, CNN, Washington Post, NPR, PolitiFact, The Guardian, The New York Times and Factcheck.org. Whether the media fact-checked the claim or not is indicated by a 1 or 0, respectively. The blue examples are the positive examples for *Task 1* (i.e., those with a positive number of sources that commented on the claim).

Evaluation We approach *Task 1* as a ranking task. As in the first edition, we plan to use *Mean Average Precision* (MAP) as the official evaluation measure. Most media rarely check more than 50 claims per debate. Thus, we plan to add $P@k$ for $k \in \{5, 10, 20, 50\}$ as well.

4.2 Task 2: Evidence and Factuality

Task 2 is defined as follows: *Given a claim associated with a set of Web pages P (that constitute the results of Web search in response to using the claim as a search query), identify which of the Web pages (and passages of those Web pages) can be useful in assisting a human who is fact-checking the claim. Finally, judge the claim factuality according to the supporting information in the passages of the Web documents.* This task will be run in Arabic.

The task is divided into several subtasks that target different aspects of the problem:

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

Subtask B: ***Classify each of the Web pages*** *as "very useful for verification", "useful", "not useful", or "unsure."* A page is considered *very useful* for verification if it is *relevant* with respect to the claim (i.e., on-topic and discussing the claim) and it *provides sufficient evidence* to verify the veracity of the claim such that there is no need for another document to be checked for this claim. A page is useful for verification if it is relevant to the claim and provides some valid evidence, but it is not sufficient to determine the claim's veracity on its own. The evidence can be a source, some statistics, a quote, etc. However, a particular piece of evidence is considered not valid if the source cannot be verified (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 objective analysis. Notice that this is different from *stance detection*, as a Web page might agree with a claim, but it might still lack evidence to verify it.

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

Subtask D: ***Find the claim's factuality*** *as "true" or "false."* The claim is considered true if it is accurate and there is nothing significant missing. A claim is false if it is not accurate.

Dataset *Task 2* is completely new to the lab this year. For the dataset, we will select a set of about 75 claims from multiple sources including an pre-existing set of Arabic claims [3], 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.

For each claim, we will search (using the claim as a query) a commercial search engine (e.g., Google or Bing) and we will extract the top 50 resulting Web pages. Crowd workers will then be hired to annotate the Web pages for relevance. In-house annotators will then be hired to annotate the relevant pages for the first two subtasks (i.e., those based on usefulness of pages). As for Subtask C, only pages that are labeled as useful for claim verification will be used, and we will split each page into paragraphs (assuming each paragraph is a passage). In-house annotators will then label a paragraph as whether it is useful for verifying the claim or not. Majority voting will be used to determine the final label of a page/passage at the different labelling tasks. The final dataset will consist of 25 training claims and 50 testing claim. As this is a new task, we will work on releasing the training claims with annotations early in the lab schedule in order to support experiments by the participating teams.

Figure 2 is an example for *Task 2*. For the sake of readability, the example is given in English, but this year the task will be offered in Arabic only.

Evaluation *Task 2* includes both ranking and classification subtasks. For Subtask A, we plan to use ranking measures such as *Mean Average Precision* (MAP) as the official evaluation measure and *Precision at k* ($P@k$). For the classification subtasks B, C, and D, we will use Accuracy, Precision, Recall, and F_1. F_1 will be the official evaluation measure.

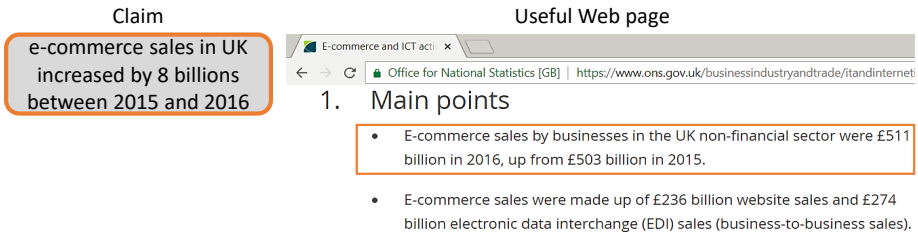

Fig. 2. A claim associated with a useful Web page, and a useful passage (in a box).

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, Avignon, France (2018)
2. 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. International World Wide Web Conferences Steering Committee (2016)
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, Volume 2 (Short Papers). vol. 2, pp. 21–27 (2018)
4. Castillo, C., Mendoza, M., Poblete, B.: Information credibility on Twitter. In: Proceedings of the 20th international conference on World wide web. pp. 675–684. ACM (2011)
5. Gencheva, P., Nakov, P., Màrquez, L., no, A.B.C., Koychev, I.: A context-aware approach for detecting worth-checking claims in political debates. In: Proceedings of the 2017 International Conference on Recent Advances in Natural Language Processing. RANLP '17, Varna, Bulgaria (2017)
6. 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. pp. 172–180. AIMSA '16, Varna, Bulgaria (2016)
7. Hassan, N., Li, C., Tremayne, M.: Detecting check-worthy factual claims in presidential debates. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. pp. 1835–1838. ACM (2015)
8. Karadzhov, G., Gencheva, P., Nakov, P., Koychev, I.: We built a fake news & clickbait filter: What happened next will blow your mind! In: Proceedings of the 2017 International Conference on Recent Advances in Natural Language Processing. RANLP '17, Varna, Bulgaria (2017)
9. Karadzhov, G., Nakov, P., Màrquez, L., no, A.B.C., Koychev, I.: Fully automated fact checking using external sources. In: Proceedings of the 2017 International Conference on Recent Advances in Natural Language Processing. RANLP '17, Varna, Bulgaria (2017)

10. 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 IJCAI (2016)
11. 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. ACM (2015)
12. Nakov, P., Barrón-Cedeño, A., Elsayed, T., Suwaileh, R., Màrquez, L., Zaghouani, W., Atanasova, P., Kyuchukov, S., Da San Martino, G.: Overview of the CLEF-2018 CheckThat! lab on automatic identification and verification of political claims. In: Proceedings of the Ninth International Conference of the CLEF Association: Experimental IR Meets Multilinguality, Multimodality, and Interaction. pp. 372–387. Lecture Notes in Computer Science, Springer, Avignon, France (2018)
13. 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. pp. 2173–2178. ACM (2016)
14. 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 (CIKM). pp. 1783–1786. ACM, Turin, Italy (2018)
15. 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)