CheckThat! at CLEF 2020: Enabling the Automatic Identification and Verification of Claims in Social Media

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

[1] DIT–Università di Bologna, Forlì, Italy
[2] Qatar University, Doha, Qatar
[3] Qatar Computing Research Institute, HBKU, Doha, Qatar
a.barron@unibo.it
{telsayed,maram.hasanain,rs081123,200159617}@qu.edu.qa
{pnakov,gmartino}@hbku.edu.qa

Abstract. We describe the third edition of the CheckThat! Lab, which is part of the 2020 Cross-Language Evaluation Forum (CLEF). CheckThat! proposes four complementary tasks and a related task from previous lab editions, offered in English, Arabic, and Spanish. Task 1 asks to predict which tweets in a Twitter stream are worth fact-checking. Task 2 asks to determine whether a claim posted in a tweet can be verified using a set of previously fact-checked claims. Task 3 asks to retrieve text snippets from a given set of Web pages that would be useful for verifying a target tweet's claim. Task 4 asks to predict the veracity of a target tweet's claim using a set of potentially-relevant Web pages. Finally, the lab offers a fifth task that asks to predict the check-worthiness of the claims made in English political debates and speeches. CheckThat! features a full evaluation framework. The evaluation is carried out using mean average precision or precision at rank k for ranking tasks, and F_1 for classification tasks.

1 Introduction

The mission of the CheckThat! lab is to foster the development of technology that would enable the automatic verification of claims. Automated systems for claim identification and verification can be very useful as supportive technology for investigative journalism, as they could provide help and guidance, thus saving time [14,22,24,33]. A system could automatically identify check-worthy claims, make sure they have not been fact-checked already by a reputable fact-checking organization, and then present them to a journalist for further analysis in a ranked list. Additionally, the system could identify documents that are potentially *useful* for humans to perform manual fact-checking of a claim, and it could also estimate a *veracity score* supported by evidence to increase the journalist's understanding and the trust in the system's decision.

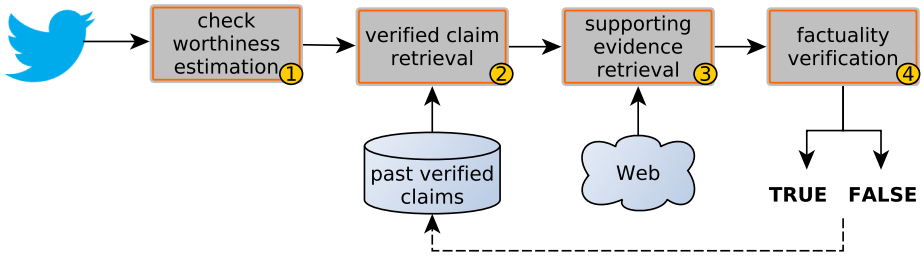

Fig. 1. Information verification pipeline. Our tasks cover all four steps. (Box 1 maps to task 1 whereas boxes 3–4 map to task 2 of the 2018 and 2019 editions [10,29].)

`CheckThat!` at CLEF 2020 is the third edition of the lab.[4] The 2018 edition [29] of `CheckThat!` focused on the identification and verification of claims in political debates.[5] Whereas the 2019 edition [9,10] also focused on political debates, isolated claims were considered as well, in conjunction with a closed set of Web documents to retrieve evidence from.[6]

In 2020, `CheckThat!` turns its attention to social media —in particular to *Twitter*— as information posted on that platform is not checked by an authoritative entity before publication and such information tends to disseminate very quickly. Moreover, social media posts lack context due to their short length and conversational nature; thus, identifying a claim's context is sometimes key for enabling effective fact-checking [7].

2 Description of the Tasks

The lab is mainly organized around four tasks, which correspond to the four main blocks in the verification pipeline, as illustrated in Figure 1. Tasks 1, 3, and 4 can be seen as reformulations of corresponding tasks in 2019, which enables re-use of training data and systems from previous editions of the lab (cf. Section 3). Task 2 runs for the first time. While Tasks 1–4 are focused on Twitter, Task 5 (not in Figure 1) focuses on political debates as in the previous two editions of the lab. All tasks are run in English. Additionally, Tasks 1, 3, and 4 are also offered in Arabic and/or Spanish.

2.1 Task 1: Check-Worthiness on Tweets

Task 1 is formulated as follows: *Given a topic and a stream of potentially-related tweets, rank the tweets according to their check-worthiness for the topic.*

Previous work on check-worthiness focused primarily on political debates and speeches, but here we focus on tweets instead.

[4] https://sites.google.com/view/clef2020-checkthat/
[5] http://alt.qcri.org/clef2018-factcheck/
[6] https://sites.google.com/view/clef2019-checkthat

Dataset We include "topics" this year, as we want to have a scenario that is close to that from 2019; a topic gives a context just like a debate did. We construct the dataset by tracking a set of manually-created topics in Twitter. A sample of tweets from the tracked stream (per topic) is shared with the participating systems as input for Task 1. The systems are asked to submit a ranked list of the tweets for each topic. Finally, using pooling, a set of tweets is selected and then judged by in-house annotators.

Evaluation We treat Task 1 as a ranking problem. Systems are evaluated using ranking evaluation measures, namely Mean Average Precision (MAP) and precision at rank k (P@k). The official measure is P@30.

2.2 Task 2: Verified Claim Retrieval

Task 2 is defined as follows: *Given a check-worthy claim and a dataset of verified claims, rank the verified claims, so that those that verify the input claim (or a sub-claim in it) are ranked on top.*

Given an input claim c and a set $V_c = \{v_i\}$ of verified claims, we consider each pair (c, v_i) as *Relevant* if v_i would save the process of verifying c from scratch, and as *Irrelevant* otherwise. Note that there might be more than one *relevant* verified claim per input claim, e.g., because the input claim might be composed of multiple claims. The task is similar to paraphrasing and textual similarity tasks, as well as to textual entailment [8,12,30].

Dataset Verified claims are retrieved from fact-checking websites such as *Snopes* and *PolitiFact*.

Evaluation Mean Average Precision on the first 5 retrieved claims (MAP@5) is used to assess the quality of the rankings submitted by the participants. A perfect ranking will have on top all v_i such that (c, v_i) is *Relevant*, in any order, followed by all *Irrelevant* claims. In addition to MAP@5, we also report MRR, MAP@k ($k = 3, 10, 20, all$) and Recall@k for $k = 3, 5, 10, 20$ in order to provide participants with more information about their systems.

2.3 Task 3: Evidence Retrieval

Task 3 is defined as follows: *Given a check-worthy claim on a specific topic and a set of text snippets extracted from potentially-relevant webpages, return a ranked list of all evidence snippets for the claim. Evidence snippets are those snippets that are useful in verifying the given claim.*

Dataset While tracking on-topic tweets, we search the Web to retrieve top-m Web pages using topic-related queries. This would ensure the freshness of the retrieved pages and enable reusability of the dataset for real-time verification tasks. Once we acquire annotations for Task 1, we share with participants the Web pages and text snippets from them solely for the check-worthy claims, which would enable the start of the evaluation cycle for Task 3. In-house annotators will label each snippet as evidence or not for a target claim.

Evaluation Tasks 3 is a ranking problem. We evaluate the ranked list per topic using MAP and P@k. The official measure is P@10.

2.4 Task 4: Claim Verification

Task 4 is defined as follows: *Given a check-worthy claim on a specific topic and a set of potentially-relevant Web pages, predict the veracity of the claim.* This task closes the verification pipeline.

Dataset The dataset for this task is the same as for Task 3. The only difference is that the in-house annotators judge each claim as true or false.

Evaluation Task 4 is a binary classification problem. Therefore, it is evaluated using standard classification evaluation measures: Precision, Recall, F_1, and Accuracy. The official measure is macro-averaged F_1.

2.5 Task 5: Check-Worthiness on Debates

Task 5 is defined as follows: *Given a debate segmented into sentences, together with speaker information, prioritize sentences for fact-checking.* This is a ranking task and each sentence should be associated with a score.

Dataset This is the third iteration of this task. We believe it is important to keep it alive as we have a large body of annotated data already and new material arrives with the coming 2020 US Presidential elections.

Evaluation Task 5 is yet another ranking problem. We use MAP as the official evaluation measure. We further report P@k for $k \in \{5, 10, 20, 50\}$.

3 Previously on `CheckThat!`

Two editions of `CheckThat!` have been held so far. While the datasets come from different genres, some of the tasks in the 2020 edition are reformulated. Hence, considering some of the most successful approaches applied in the past represents a good starting point to address the current challenges.

3.1 `CheckThat!` 2019

The 2019 edition featured two tasks [10]:

Task 1$_{2019}$*. Given a political debate, interview, or speech, transcribed and segmented into sentences, rank the sentences by the priority with which they should be fact-checked.*

The most successful approaches used neural networks for the individual classification of the instances. For example, Hansen et al. [19] learned domain-specific word embeddings and syntactic dependencies and applied an LSTM classifier.

Using some external knowledge paid off —they pre-trained the network with previous Trump and Clinton debates, supervised weakly with the ClaimBuster system. Some efforts were carried out in order to consider context. Favano et al. [11] trained a feed-forward neural network, including the two previous sentences as context. Whereas many approaches opted for embedding representations, feature engineering was also popular [13].

Task 2_{2019}. *Given a claim and a set of Web pages potentially relevant with respect to the claim, identify which of the pages (and passages thereof) are useful for assisting a human in fact-checking the claim. Finally, determine the factuality of the claim.*

The systems for evidence passage identification followed two approaches. BERT was trained and used to predict whether an input passage is useful to fact-check a claim [11]. Other participating systems used classifiers (e.g., SVM) with a variety of features including similarity between the claim and a passage, bag of words, and named entities [20]. As for predicting claim veracity, the most effective approach used a textual entailment model. The input was represented using word embeddings and external data was also used in training [15].

In the 2020 edition, Task 1_{2019} becomes Task 5, and Task 1 is a reformulation based on tweets (cf. Section 2.1). See [2] for further details. Task 2_{2019} becomes Tasks 3 and 4 (cf. Sections 2.3 and 2.4). See [21] for further details.

3.2 `CheckThat!` 2018

The 2018 edition featured two tasks [29]:

Task 1_{2018} was identical to Task 1_{2019}.

The most successful approaches used either a multilayer perceptron or an SVM. Zuo et al. [36] enriched the dataset by producing *pseudo-speeches* as a concatenation of all interventions by a debater. They used averaged word embeddings and bag-of-words as representations. Hansen et al. [18] represented the entries with embeddings, part of speech tags, and syntactic dependencies. They used a GRU neural network with attention. See [1] for further details.

Task 2_{2018}. *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.*

The best way to address this task was to retrieve relevant information from the Web, followed by a comparison to the claim in order to assess its factuality.[7] After retrieving such *evidence*, it is fed into the supervised model, together with the claim in order to assess its veracity. In the case of [18], they fed the claim and the most similar Web-retrieved text to convolutional neural networks and SVMs. Meanwhile, Ghanem et al. [16] computed features, such as the similarity between the claim and the Web text, and the Alexa rank for the website. See [4] for further details.

[7] While this year a similar procedure had to be carried out, we decompose it into three tasks (cf. Section 2).

4 Related Work

There has been work on checking the factuality/credibility of a claim, of a news article, or of an information source [3,25,26,28,31,35]. Claims can come from different sources, but special attention has been given to those from social media [17,27,32,34]. Check worthiness estimation is still a fairly-new problem especially in the context of social media [14,22,23,24].

`CheckThat!` further shares some aspects with other initiatives that have been run with high success in the past, e.g., stance detection (Fake News[8]), semantic textual similarity (STS at SemEval[9]), and community question answering (cQA at SemEval[10]).

5 Conclusion

We have presented the 2020 edition of the `CheckThat!` Lab, which features tasks that span the full verification pipeline: from spotting check-worthy claims to checking whether they have been fact-checked elsewhere already, to retrieving useful passages within relevant pages, to finally making a prediction about the factuality of a claim. To the best of our knowledge, this is the first shared task that addresses all steps of the fact-checking process. Moreover, unlike previous editions of the `CheckThat!` Lab, our main focus here is on social media, which are the center of "fake news" and disinformation. We further feature a more realistic information retrieval scenario with pooling for evaluation, as done at IR venues such as TREC. Last but not least, in-line with the general mission of CLEF, we promote multi-linguality by offering our tasks in different languages.

We hope that these tasks and the associated datasets will serve the mission of the `CheckThat!` initiative, which is to foster the development of datasets, tools and technology that would enable the automatic verification of claims and will support human fact-checkers in their fight against "fake news" and disinformation.

Acknowledgments

The work of Tamer Elsayed and Maram Hasanain was made possible by NPRP grant# NPRP 11S-1204-170060 from the Qatar National Research Fund (a member of Qatar Foundation). The work of Reem Suwaileh was supported by GSRA grant# GSRA5-1-0527-18082 from the Qatar National Research Fund and the work of Fatima Haouari was supported by GSRA grant# GSRA6-1-0611-19074 from the Qatar National Research Fund. The statements made herein are solely the responsibility of the authors. This research is also part of the Tanbih project, developed by the Qatar Computing Research Institute, HBKU and MIT-CSAIL, which aims to limit the effect of "fake news", propaganda, and media bias.

[8] Official Challenge website: http://www.fakenewschallenge.org/
[9] STS task at the SemEval 2017 edition: http://alt.qcri.org/semeval2017/task1/
[10] cQA task at the SemEval 2017 edition: http://alt.qcri.org/semeval2017/task3/

References

1. Atanasova, P., Marquez, 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. [6]
2. Atanasova, P., Nakov, P., Karadzhov, G., Mohtarami, M., Da San Martino, G.: Overview of the CLEF-2019 CheckThat! lab on automatic identification and verification of claims. Task 1: Check-worthiness. In: Cappellato et al. [5]
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 Companion (2016)
4. Barrón-Cedeño, A., Elsayed, T., Suwaileh, R., Marquez, L., Atanasova, P., Zaghouani, W., Kyuchukov, S., Da San Martino, G., Nakov, P.: Overview of the CLEF-2018 CheckThat! lab on automatic identification and verification of political claims. Task 2: Factuality. In: Cappellato et al. [6]
5. Cappellato, L., Ferro, N., Losada, D., Müller, H. (eds.): Working Notes of CLEF 2019 Conference and Labs of the Evaluation Forum. CEUR Workshop Proceedings, CEUR-WS.org (2019)
6. Cappellato, L., Ferro, N., Nie, J.Y., Soulier, L. (eds.): Working Notes of CLEF 2018–Conference and Labs of the Evaluation Forum. CEUR Workshop Proceedings, CEUR-WS.org (2018)
7. Cazalens, S., Lamarre, P., Leblay, J., Manolescu, I., Tannier, X.: A content management perspective on fact-checking. In: Proceedings of The Web Conference 2018. pp. 565–574. WWW '18 (2018)
8. Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., Specia, L.: SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In: Proceedings of the 11th International Workshop on Semantic Evaluation. pp. 1–14. SemEval '17 (2017)
9. 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: Advances in Information Retrieval. pp. 309–315. Springer International Publishing (2019)
10. 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. pp. 301–321. LNCS, Springer (2019)
11. Favano, L., Carman, M., Lanzi, P.: TheEarthIsFlat's submission to CLEF19 CheckThat! Challenge. In: Cappellato et al. [5]
12. Filice, S., Da San Martino, G., Moschitti, A.: Structural representations for learning relations between pairs of texts pp. 1003–1013 (2015)
13. Gasior, J., Przybyła, P.: The IPIPAN team participation in the check-worthiness task of the CLEF2019 CheckThat! Lab. In: Cappellato et al. [5]
14. 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 (2017)
15. Ghanem, B., Glava, G., Giachanou, A., Ponzetto, S., Rosso, P., Rangel, F.: UPV-UMA at CheckThat! lab: Verifying Arabic claims using cross lingual approach. In: Cappellato et al. [5]

16. 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. [6]
17. Gupta, A., Kumaraguru, P., Castillo, C., Meier, P.: TweetCred: Real-time credibility assessment of content on Twitter. In: Proceeding of the 6th International Social Informatics Conference. pp. 228–243. SocInfo'142 (2014)
18. 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. [6]
19. Hansen, C., Hansen, C., Simonsen, J., Lioma, C.: Neural weakly supervised fact check-worthiness detection with contrastive sampling-based ranking loss. In: Cappellato et al. [5]
20. Haouari, F., Ali, Z., Elsayed, T.: bigIR at CLEF 2019: Automatic verification of Arabic claims over the web. In: Cappellato et al. [5]
21. Hasanain, M., Suwaileh, R., Elsayed, T., Barrón-Cedeño, A., Nakov, P.: Overview of the CLEF-2019 CheckThat! lab on automatic identification and verification of claims. Task 2: Evidence and factuality. In: Cappellato et al. [5]
22. 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. CIKM 15 (2015)
23. Hassan, N., Tremayne, M., Arslan, F., Li, C.: Comparing automated factual claim detection against judgments of journalism organizations. In: Computation+Journalism Symposium (2016)
24. Hassan, N., Zhang, G., Arslan, F., Caraballo, J., Jimenez, D., Gawsane, S., Hasan, S., Joseph, M., Kulkarni, A., Nayak, A.K., et al.: Claimbuster: The first-ever end-to-end fact-checking system. Proceedings of the VLDB Endowment **10**(12), 1945–1948 (2017)
25. 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. pp. 344–353. RANLP' 17 (2017)
26. 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 International Joint Conference on Artificial Intelligence. pp. 3818–3824. IJCAI '16 (2016)
27. Mitra, T., Gilbert, E.: Credbank: A large-scale social media corpus with associated credibility annotations. In: Proceedings of the Ninth International AAAI Conference on Web and Social Media. ICWSM '15 (2015)
28. 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 (2015)
29. Nakov, P., Barrón-Cedeño, A., Elsayed, T., Suwaileh, R., Màrquez, L., Zaghouani, W., Gencheva, P., Kyuchukov, S., Da San Martino, G.: Overview of the CLEF-2018 lab on automatic identification and verification of claims in political debates. In: Working Notes of CLEF 2018 – Conference and Labs of the Evaluation Forum. CLEF '18, Avignon, France (2018)
30. Nakov, P., Màrquez, L., Moschitti, A., Magdy, W., Mubarak, H., Freihat, A.A., Glass, J., Randeree, B.: SemEval-2016 Task 3: Community question answering. In: Proceedings of the 10th International Workshop on Semantic Evaluation. pp. 525–545. SemEval '15 (2016)

31. 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 (2016)
32. Shu, K., Sliva, A., Wang, S., Tang, J., Liu, H.: Fake news detection on social media: A data mining perspective. SIGKDD Explor. Newsl. **19**(1), 2236 (2017)
33. Vasileva, S., Atanasova, P., Màrquez, L., Barrón-Cedeño, A., Nakov, P.: It takes nine to smell a rat: Neural multi-task learning for check-worthiness prediction. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing. RANLP '19 (2019)
34. Zhao, Z., Resnick, P., Mei, Q.: Enquiring minds: Early detection of rumors in social media from enquiry posts. In: Proceedings of the 24th International Conference on World Wide Web. pp. 1395–1405. WWW'15 (2015)
35. 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) (2016)
36. Zuo, C., Karakas, A., Banerjee, R.: A hybrid recognition system for check-worthy claims using heuristics and supervised learning. In: Cappellato et al. [6]