Overview of CheckThat! 2020 Arabic: Automatic Identification and Verification of Claims in Social Media

Maram Hasanain[1], Fatima Haouari[1], Reem Suwaileh[1], Zien Sheikh Ali[1], Bayan Hamdan[2], Tamer Elsayed[1], Alberto Barrón-Cedeño[3], Giovanni Da San Martino[4], and Preslav Nakov[4]

[1] Computer Science and Engineering Department, Qatar University, Doha, Qatar
{maram.hasanain, 200159617, rs081123, zs1407404, telsayed}@qu.edu.qa
[2] Research Consultant
bayan.hamdan995@gmail.com
[3] DIT, Università di Bologna, Forlì, Italy
a.barron@unibo.it
[4] Qatar Computing Research Institute, HBKU, Doha, Qatar
{pnakov, gmartino}@hbku.edu.qa

Abstract. In this paper, we present an overview of the Arabic tasks of the third edition of the CheckThat! Lab at CLEF 2020. The lab featured three Arabic tasks over social media (and the Web): Task 1 on check-worthiness estimation, Task 3 on evidence retrieval, and Task 4 on claim verification. For evaluation, we collected a dataset of Arabic tweets and Web pages consisting of 7.5K tweets and 14,742 Web pages. The systems in the ranking tasks (Task 1 and Task 3) were evaluated using precision at 30 (P@30) and precision at 10 (P@10), respectively. F_1 was the official evaluation measure for Task 4. Eight teams submitted runs to the Arabic tasks, which is double the number of teams participating in the Arabic tasks of the CheckThat! lab at CLEF 2019. The most successful approach to Task 1 used an Arabic pre-trained language model, while text similarity measures and linguistic features were used in the other tasks. We release to the research community all datasets from the lab, which should enable further research on automatic claim verification in Arabic social media.

1 Introduction

With the rapid growth of social media such as Twitter, large amounts of fake and unverified claims have emerged and have been propagated to affect online social media users as well as the offline society. Thus, the automatic detection and verification of fake claims could help mitigate this negative development and benefit not only normal users, but also journalists and news agencies.

A plethora of studies addressed the problem of claim identification [14, 15, 18, 28, 30, 34] and verification [9, 26, 27, 29, 38] in social media, but addressing these tasks in Arabic is severely under-explored [2, 3]. Similarly, check-worthiness estimation is under-explored in social media [1]. A considerable body of literature on check-worthiness estimation exists, but the focus has been mainly on political debates and speeches [21, 22, 24, 33].

In its third edition, the CheckThat! lab [7][5] focused on social media, specifically Twitter, with the aim of enabling the automatic verification of claims. This paper focuses on the three Arabic tasks offered by the CheckThat! lab in 2020:[6]

Task 1 Check-worthiness estimation for tweets: predict which tweet from a stream of tweets on a topic should be prioritized for fact-checking.

Task 3 Evidence retrieval: given a check-worthy claim in a tweet on a specific topic and a set of text snippets extracted from potentially relevant Web pages, return a ranked list of evidence snippets for the claim.

Task 4 Claim verification: given a check-worthy claim in a tweet and a set of potentially-relevant Web pages, estimate the veracity of the claim.

The Arabic tasks attracted 8 teams, which submitted a total of 30 runs, and the most successful approaches adopted fine-tuning existing pretrained models, namely AraBERT [4] and multilingual BERT [16]. The datasets for the three tasks were created from scratch as the goal for this year was to focus on social media for the first time, as opposed to previous editions of the lab, which featured automatic identification and verification of political claims [32, 5, 8], and evidence-based claim verification [17, 6, 20]. We make the datasets available to the research community to support further research on the three tasks.[7]

For each of the Arabic tasks, we describe below the evaluation dataset created to support that task, we present a summary of the approaches used by the participating systems, and we discuss the evaluation results.

2 Task 1_{ar}: Check-Worthiness on Tweets

Since check-worthiness estimation for tweets in general, and for *Arabic* tweets in particular, is a relatively new task, we constructed a new dataset specifically designed for training and testing systems for this task. We identified the need for a "context" that affects the check-worthiness of tweets, and we used "topics" to represent that context. Given a topic, we define a check-worthy tweet as a tweet that is relevant to the topic, contains one main claim that can be fact-checked by consulting reliable sources, and is important enough to be worthy of verification. More on the annotation criteria is presented later in this section.

[5] `https://sites.google.com/view/clef2020-checkthat/`
[6] Refer to [35] to read about the English tasks of the CheckThat! 2020 lab.
[7] `https://sites.google.com/view/clef2020-checkthat/`

عنوان الموضوع: تدخل تركيا في سوريا

شرح الموضوع: بعد استمرار الحرب في سوريا ما يُقارب التسع سنين بعد اشتعال الثورة عام 2011، قررت تركيا الدخول بهدف معلن وهو حماية المدنيين وردع القوات السورية والأجنبية عن توتير الأوضاع وقتل وتشريد المدنيين، وهو ما أثار الرأي العام في العالم. يتحدث هذا الموضوع عن تطورات التدخل التركي العسكري في سوريا على جميع الأصعدة.

Topic title: Intervention of Turkey in Syria

Topic description: After 9 years of war in Syria since the eruption of the revolution in 2011, Turkey decided to intervene in Syria with the declared aim of protecting civilians and deterring Syrian and foreign forces from aggravating the situation, and killing and displacing civilians, which ignited public opinion in the world. This topic talks about developments related to the Turkish military intervention in Syria on all aspects.

Fig. 1: **Task 1:** topic CT20-AR-19 from the training dataset.

2.1 Dataset

In order to construct the dataset for this task, we first manually created fifteen topics over the period of several months. We then selected trending topics at the time among Arab social media users. Each topic was represented using a short title and a much longer text description. Figure 1 shows an example topic from the training dataset.

Examples of other topic titles include "Coronavirus in the Arab World", "Sudan and normalization", and "Deal of the century". We augmented each topic with a set of keywords, hashtags, and usernames to track in Twitter. Once we had created a topic, we immediately crawled a one-week stream of tweets using the constructed search terms, where we searched Twitter (via the Twitter search API) using each term by the end of each day. We limited the search to original Arabic tweets (i.e., we excluded retweets). We then de-duplicated the tweets and we dropped those matching our qualification filter that excludes tweets containing terms from a blacklist of explicit terms and tweets that contain more than four hashtags or more than two URLs. Afterwards, we ranked the tweets by popularity (defined by the sum of their retweets and likes), and we selected the top-500 to be annotated.

The annotation process was performed in two steps; we first identified the tweets that are relevant to the topic and contain factual claims, then we identified the check-worthy tweets among those relevant tweets.

We first recruited one annotator to annotate each tweet for its relevance with respect to the target topic. In this step, we labeled each tweet as one of three categories:

- Non-relevant tweet for the target topic.
- Relevant tweet but *with no factual claims*, such as a tweet expressing an opinion about the topic, reference, or speculation about the future.
- Relevant tweet that contains a factual claim and that can be fact-checked by consulting reliable sources.

Translation. Defenses of the criminal Russian-Iranian occupation militias on the Nairab axis have collapsed and there is a state of chaos with some occupation mercenaries fleeing after the destruction of their tanks due to rebels' strikes and the support of the Turkish army. The Russian occupation brought air forces into the battle, trying to change the course of the battles.

Fig. 2: **Task 1:** tweet with a check-worthy claim related to topic CT20-AR-19.

Only relevant tweets with factual claims were labelled for check-worthiness. Two annotators annotated those tweets first. A third *expert* annotator performed disagreement resolution whenever needed. Due to the subjective nature of check-worthiness, we chose to represent the check-worthiness criteria by several questions, to help the annotators think about different aspects of check-worthiness. The annotators were asked to answer the following three questions for each tweet, using a Likert scale between 1 and 5:

1. Do you think the claim in the tweet is of interest to the public?
2. To what extent do you think the claim can negatively affect the reputation of an entity, country, etc.?
3. Do you think journalists will be interested in covering the spread of the claim or the information discussed by the claim?

Once the annotator has answered the above questions, s/he is required to answer the following fourth question considering all the ratings given previously:

Do you think the claim in the tweet is check-worthy?

This is a yes/no question, and the resulting answer is the label we use to represent check-worthiness in this dataset. Figure 2 shows an example of a tweet making a check-worthy claim.

For the final set, all tweets but those labelled as check-worthy were considered not check-worthy. Given 500 tweets annotated for each of the fifteen topics, the annotated set contained 2,062 check-worthy claims (27.5%). Three topics constituted the training set, and the remaining twelve topics were used to evaluate the participating systems.

Table 1: **Task 1:** summary of the approaches. We show information about the learning models (including Transformers), about the main representations, whether the participants used external data, and whether they used machine translation (MT) to be able to use additional data in English.

Team		**Models**					**Distrib.**					**Represent.**							**Other**	
		BERT	**Bi–LSTM**	**NN**	**SVM**	**SGD**	**LASER**	**FastText**	**GloVe**	**word2vec**	**PCA**	**One-hot**	**Morphology**	**Syntax**	**Sentiment**	**Dependencies**	**NER**		**External data**	**MT**
Accenture	[37]	●																	●	●
bigIR	[19]	●																	●	
check_square	[13]				●					●	●		●			●				
DamascusTeam	[23]		●									●							●	
EvolutionTeam	[36]														●		●		●	
NLP&IR@UNED	[31]			●					●											
TOBB ETU	[25]	●						●												
WSSC_UPF	–					●	●						●	●					●	

2.2 Overview of the Approaches

Eight teams participated in this task submitting a total of 26 runs. Table 1 shows an overview of the approaches. The most successful runs fine-tuned existing pre-trained models, namely AraBERT and multilingual BERT. Other approaches relied on pre-trained models such as Glove, Word2vec, and Language-Agnostic SEntence Representations (LASER) to obtain embeddings for the tweets, which were fed either into a neural network or other machine learning models, such as SVM. In addition to text representations, some teams used other features, namely morphological and syntactic features, part-of-speech (POS) tags, named entities, and sentiment features.

2.3 Evaluation

We treated Task 1 as a ranking problem, where we expect check-worthy tweets to be ranked at the top. We evaluated the runs using precision at k ($P@k$) and Mean Average Precision (MAP). We considered $P@30$ as the official evaluation measure, as we anticipated that the user would check a maximum of 30 claims per week. We also developed two simple baselines: *baseline 1* which ranks tweets in descending order based on their popularity score (sum of likes and retweets a tweet has received), and *baseline 2* which ranks tweets in reverse chronological order, i.e., the most recent ones first. Table 2 shows the performance of the best run per team in addition to the two baselines, ranked by the official measure. We can see that most teams managed to improve over the two baselines by a large margin.

Table 2: **Task 1:** performance of the best run from each team.

RunID	*P*@10	*P*@20	*P*@30	MAP
Accenture-AraBERT	0.7167	0.6875	0.7000	0.6232
TobbEtu-AF	0.7000	0.6625	0.6444	0.5816
bigIR-bert	0.6417	0.6333	0.6417	0.5511
check_square-w2vposRun2	0.6083	0.6000	0.5778	0.4949
DamascusTeam-Run03	0.5833	0.5750	0.5472	0.4539
nlpir01-run4	0.6083	0.5625	0.5333	0.4614
baseline2	0.3500	0.3625	0.3472	0.3149
baseline1	0.3250	0.3333	0.3417	0.3244
EvolutionTeam-Run1	0.2500	0.2667	0.2833	0.2675
WSSC_UPF-RF01	0.1917	0.1667	0.2028	0.2542

3 Task 3_{ar}: Evidence Retrieval

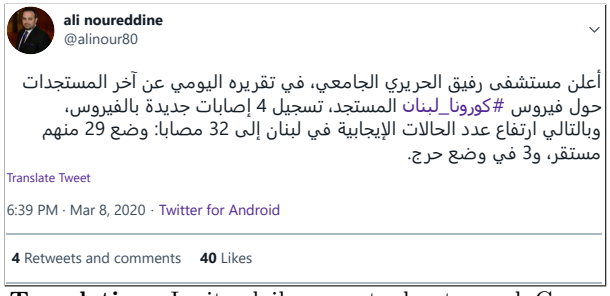
Evidence retrieval represents the second major step in an automatic fact-checking system where evidence is collected to be used for claim verification. Potentially, systems can extract evidence from any source. However, in order to unify the evaluation setup and to ensure that all systems have access to the same source of evidence, this was defined as a ranking task over a set of text snippets provided along with check-worthy claims. We define an evidence snippet as a text snippet from a Web page that constitutes evidence supporting or refuting the claim.

3.1 Dataset

For this task, we needed a set of claims and a set of potentially relevant Web pages, from which evidence snippets would be extracted. We first collected the set of Web pages using the topics for Task 1. While developing the topics, we represented each one by a set of search phrases. We used these phrases as queries to Google Web search daily, and in a week we collected a set of Web pages, which we then used to construct a dataset for the task.

As for the set of claims, we draw a random sample from the check-worthy tweets identified for each topic from Task 1. Since the data from Task 2, Subtask C in last year's edition of the lab could be used for training [20], we only released test claims and Web pages for the twelve test topics used in Task 1. The dataset for this task contains a total of 200 claims and 14,742 corresponding Web pages.

Since we seek a controlled method to allow systems to return snippets, which in turn would allow us to label a consistent set of potential evidence snippets, we automatically pre-split these pages into snippets, which we eventually released for each page. To extract snippets, we first de-duplicated the crawled Web pages using the URL. Then, we extracted the textual content from the HTML document after removing any markup and scripts. Finally, we detected the Arabic text and we split it into snippets, using full-stops, question marks, or exclamation marks as delimiters. Overall, we extracted 169,902 snippets.



Translation. In its daily report about novel Coronavirus development, Rafiq AlHariri University Hospital announced that 4 new cases were reported, and thus the number of positive cases in Lebanon increased to 32: 29 of them are stable, and 3 are in a critical situation.

"واشار البيان الى تسجيل 4 إصابات جديدة بالفايروس، وبالتالي ارتفاع عدد الحالات الإيجابية في لبنان إلى 32 مصابا، مؤكدا ان وضع 29 منهم مستقر، و3 في وضع حرج."

"The statement indicated that 4 new infections were recorded, and thus the number of positive cases in Lebanon increased to 32, stressing that 29 of them are stable, and 3 are in a critical situation"

(a) Evidence Snippet (CT20-AR-29-2043-03)

"وفي لبنان، أعلن مستشفى رفيق الحريري (حكومي) بالعاصمة بيروت، ارتفاع إصابات "كورونا" إلى 22 بعد تسجيل 6 إصابات جديدة."

"In Lebanon, the Rafic Hariri Hospital (governmental) in the capital, Beirut, announced that Corona infections had risen to 22, after 6 new infections were recorded."

(b) Non-evidence snippet (CT20-AR-29-0619-03)

Fig. 3: **Task 3:** Example of a check-worthy tweet and two annotated snippets for the topic "Coronavirus in the Arab world."

Due to the large number of snippets collected for the claims, annotating all pairs of claims and snippets was infeasible given the limited amount of time we had. Therefore, we followed a *pooling* method: we annotated pooled evidence snippets returned from the submitted runs by the participating systems. Since the official evaluation measure for the task was $P@10$, we first extracted the top 10 evidence snippets returned by each run for each claim. We then created a pool of unique snippets per claim (considering both snippet IDs and content for de-duplication). Finally, a single person annotated each snippet for a claim. The annotators were asked to decide whether a snippet contained evidence that would be useful for verifying the input claim. This evidence can be statistics, quotes, facts extracted from verified sources, etc.

Figure 3 shows an example of a check-worthy tweet. We observe that the example evidence snippet (Fig. 3a) repeats the same information from the tweet referring to a report as the source of the information. While the non-evidence snippet (Fig. 3b) is also very related to the tweet, it states a smaller number of infections since the snippet was extracted from a Web page posted a day before the tweet posting time.

Overall, we annotated 3,380 snippets. After label propagation, we had 3,720 annotated snippets of which only 95 were evidence snippets. Our annotation volume was limited due to the very small number of runs participating in the task (two runs submitted by one team).

3.2 Overview of the Approaches

Only one team, EvolutionTeam [36], participated in the task and they submitted two runs. They used the cosine similarity between the claim and the snippet as their ranking score to rank the candidate evidence snippets. In a second run, the similarity was weighted by the intersection between the snippet and a lexicon of sentiment words.

3.3 Evaluation

This task was modeled as a ranking problem, where the system is expected to rank the evidence snippets at the top of the list. In order to evaluate the submitted runs, we computed $P@k$ at different cutoffs (k = 1, 5, and 10). The official measure was $P@10$.

The participating team's best-performing run achieved an average $P@10$ of 0.0456 over the claims.

4 Task 4_{ar}: Claim Verification

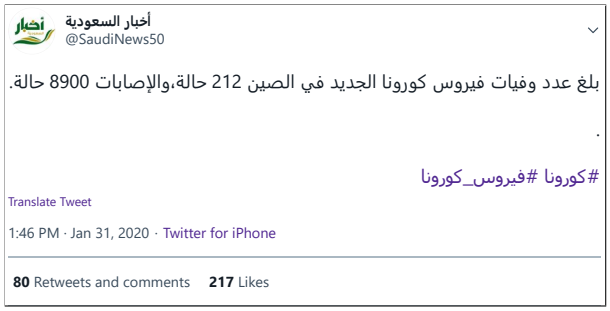
Starting with the same 200 claims used in Task 3, one expert fact-checker verified each claim's veracity. We limited the annotation categories to two, True and False, excluding partially-true claims. A True claim is a claim that is supported by a reliable source that confirms the authenticity of the information published in the tweet. A False claim can be a claim that mentions information contradicting that in a reliable source or that has been explicitly refuted by a reliable source.

4.1 Dataset

The claims in the tweets were annotated considering two main factors; the content of the tweet (claim) and the date of the tweet publication. For the annotation, we considered supporting or refuting information that was reported before, on, or a few days after the time of the claim. We consulted several reliable sources to verify the claims. These sources differed depending on the topic of the claim. For example, for health-related claims, we consulted refereed studies or articles published in reliable medical journals or websites, such as APA.

Out of the initial 200 claims, we ended up with 165 claims for which we managed to find a definite label. Only six claims among these 165 were found to be False. Since data from Task 2-Subtask D in the last year's edition of the lab can be used for training [20], the final set of 165 annotated claims was used to evaluate the submitted runs.

(a) Tweet with a True claim



(b) Tweet with a False claim

Translation (a). Ministry of Health announced the return of 10 Saudi students from China. The students were placed in precautionary isolation for a period of two weeks in appropriate housing, accompanied by specialized medical teams. This comes as part of the precautionary measures and continuous monitoring of the Novel Coronavirus.

Translation (b). The number of new deaths due to Coronavirus in China has reached 212, and 8,900 are infected.

Fig. 4: **Task 4:** example True and False claims on the topic "Novel Coronavirus."

Figure 4 shows an example of two claims for topic CT20-AR-03 "Novel Coronavirus". The second claim is False since it reports a wrong number of cases at the time of tweet posting, as compared to official Chinese sources.

4.2 Overview of the Approaches

There were two runs submitted by EvolutionTeam [36]. They used a scoring function that computes the degree of concordance and negation (using a manual list) between a claim and all input text snippets for that claim.

4.3 Evaluation

We treated the task as a classification problem and we used typical evaluation measures for such tasks in the case of class imbalance: F_1 measure (official), Precision, and Recall. The best run achieved a macro-averaged F_1 score of 0.5524.

5 Conclusion and Future Work

In this overview paper, we presented a description of the three Arabic tasks that were opffered as part of the third edition of the CheckThat! lab at CLEF 2020. Unliek previous editions of the lab, this time we focused on false information propagated on Arabic social media (specifically, on Twitter). Task 1 on check-worthiness ranking of tweets attracted the highest number of participating teams. Generally, the best approaches for that task relied on pre-trained language models such as multi-lingual BERT and AraBERT. Moreover, one team participated in Tasks 3 and 4. We suspect that the low number of participants in these two tasks was due to the lack of new training data provided for this edition of the lab.

For future editions of the lab, we plan to focus on Task 1, since it is a very critical step in the process of automatic verification over social media, where a huge stream of tweets needs to be processed in order to identify claims that are worth fact-checking.

Acknowledgments

This work was made possible in part by NPRP grant# NPRP11S-1204-170060 from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors. 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.

This research is also part of the Tanbih project, which aims to limit the effect of disinformation, "fake news", propaganda, and media bias.

References

1. Alam, F., Shaar, S., Nikolov, A., Mubarak, H., Martino, G.D.S., Abdelali, A., Dalvi, F., Durrani, N., Sajjad, H., Darwish, K., et al.: Fighting the COVID-19 infodemic: Modeling the perspective of journalists, fact-checkers, social media platforms, policy makers, and the society. arXiv preprint arXiv:2005.00033 (2020)
2. Alkhair, M., Meftouh, K., Smaïli, K., Othman, N.: An Arabic corpus of fake news: Collection, analysis and classification. In: Proceedings of the International Conference on Arabic Language Processing. pp. 292–302. ICALP '19, Springer, Nancy, France (2019)
3. Alzanin, S.M., Azmi, A.M.: Rumor detection in Arabic tweets using semi-supervised and unsupervised expectation–maximization. Knowledge-Based Systems **185**, 104945 (2019)
4. Antoun, W., Baly, F., Hajj, H.: AraBERT: Transformer-based model for Arabic language understanding. In: Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools. pp. 9–15. OSAC '20, Marseille, France (2020)

5. 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. [12]
6. 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. [11]
7. Barrón-Cedeño, A., Elsayed, T., Nakov, P., Da San Martino, G., Hasanain, M., Suwaileh, R., Haouari, F., Babulkov, N., Hamdan, B., Nikolov, A., Shaar, S., Sheikh Ali, Z.: Experimental ir meets multilinguality, multimodality, and interaction proceedings of the eleventh international conference of the clef association (clef 2020). In: Arampatzis, A., Kanoulas, E., Tsikrika, T., Vrochidis, S., Joho, H., Lioma, C., Eickhoff, C., Névéol, A., Cappellato, L., Ferro, N. (eds.) Overview of CheckThat! 2020: Automatic Identification and Verification of Claims in Social Media. LNCS (12260), Springer (2020)
8. 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: Cappellato et al. [12]
9. Bian, T., Xiao, X., Xu, T., Zhao, P., Huang, W., Rong, Y., Huang, J.: Rumor detection on social media with bi-directional graph convolutional networks. In: Proceedings of the AAAI Conference on Artificial Intelligence. AAAI '20, vol. 34, pp. 549–556. New York, NY, USA (2020)
10. Cappellato, L., Eickhoff, C., Ferro, N., Névéol, A. (eds.): CLEF 2020 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org (2020)
11. 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)
12. 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)
13. Cheema, G.S., Hakimov, S., Ewerth, R.: Check_square at CheckThat! 2020: Claim detection in social media via fusion of transformer and syntactic features. In: Cappellato et al. [10]
14. Chen, T., Li, X., Yin, H., Zhang, J.: Call attention to rumors: Deep attention based recurrent neural networks for early rumor detection. In: Pacific-Asia conference on knowledge discovery and data mining. pp. 40–52. PAKDD '18, Springer, Melbourne, Australia (2018)
15. Chen, Y., Sui, J., Hu, L., Gong, W.: Attention-residual network with CNN for rumor detection. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management. pp. 1121–1130. CIKM '19, Beijing, China (2019)
16. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 4171–4186. NAACL-HLT '19, Minneapolis, Minnesota, USA (2019)
17. 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)

18. Gao, J., Han, S., Song, X., Ciravegna, F.: RP-DNN: A tweet level propagation context based deep neural networks for early rumor detection in social media. In: Proceedings of the 12th Language Resources and Evaluation Conference. pp. 6094–6105. LREC '20, Marseille, France (2020)
19. Hasanain, M., Elsayed, T.: bigIR at CheckThat! 2020: Multilingual BERT for ranking Arabic tweets by check-worthiness. In: Cappellato et al. [10]
20. 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. [11]
21. Hassan, N., Arslan, F., Li, C., Tremayne, M.: Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1803–1812. Halifax, NS (2017)
22. 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)
23. Hussein, A., Hussein, A., Ghneim, N., Joukhadar, A.: DamascusTeam at CheckThat! 2020: Check worthiness on Twitter with hybrid CNN and RNN models. In: Cappellato et al. [10]
24. 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. pp. 26–30. NAACL-HLT '18, New Orleans, Louisiana, USA (2018)
25. Kartal, Y.S., Kutlu, M.: TOBB ETU at CheckThat! 2020: Prioritizing English and Arabic claims based on check-worthiness. In: Cappellato et al. [10]
26. Khoo, L.M.S., Chieu, H.L., Qian, Z., Jiang, J.: Interpretable rumor detection in microblogs by attending to user interactions. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 8783–8790. AAAI' 20, New York, NY, USA (2020)
27. Liu, Y., Wu, Y.F.B.: Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence. pp. 354–361. AAAI '18, New Orleans, Louisiana, USA (2018)
28. 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, NY, USA (2016)
29. Ma, J., Gao, W., Wong, K.F.: Detect rumors in microblog posts using propagation structure via kernel learning. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. pp. 708–717. ACL '17, Vancouver, Canada (2017)
30. Ma, J., Gao, W., Wong, K.F.: Detect rumors on Twitter by promoting information campaigns with generative adversarial learning. In: Proceedings of the World Wide Web Conference. pp. 3049–3055. WWW '19, San Francisco, CA, USA (2019)
31. Martinez-Rico, J., Araujo, L., Martinez-Romo, J.: NLP&IR@UNED at CheckThat! 2020: A preliminary approach for check-worthiness and claim retrieval tasks using neural networks and graphs. In: Cappellato et al. [10]

32. 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)
33. 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)
34. Santhoshkumar, S., Babu, L.D.: Earlier detection of rumors in online social networks using certainty-factor-based convolutional neural networks. Social Network Analysis and Mining **10**(1), 1–17 (2020)
35. Shaar, S., Nikolov, A., Babulkov, N., Alam, F., Barrón-Cedeño, A., Elsayed, T., Hasanain, M., Suwaileh, R., Haouari, F., Da San Martino, G., Nakov, P.: Overview of CheckThat! 2020 English: Automatic identification and verification of claims in social media. In: Cappellato et al. [10]
36. Touahri, I., Mazroui, A.: EvolutionTeam at CheckThat! 2020: Integration of linguistic and sentimental features in a fake news detection approach. In: Cappellato et al. [10]
37. Williams, E., Rodrigues, P., Novak, V.: Accenture at CheckThat! 2020: If you say so: Post-hoc fact-checking of claims using transformer-based models. In: Cappellato et al. [10]
38. Zhang, Q., Lipani, A., Liang, S., Yilmaz, E.: Reply-aided detection of misinformation via bayesian deep learning. In: Proceedings of the World Wide Web Conference. p. 2333–2343. WWW '19, San Francisco, CA, USA (2019)