Overview of the CLEF-2018 CheckThat! Lab on Automatic Identification and Verification of Political Claims. Task 2: Factuality*

Alberto Barrón-Cedeño[1], Tamer Elsayed[2], Reem Suwaileh[2], Lluís Màrquez[3], Pepa Atanasova[4], Wajdi Zaghouani[5], Spas Kyuchukov[6], Giovanni Da San Martino[1], and Preslav Nakov[1]

[1] Qatar Computing Research Institute, HBKU, Doha, Qatar
[2] Computer Science and Engineering Department, Qatar University, Doha, Qatar
[3] Amazon, Barcelona, Spain
[4] SiteGround, Sofia, Bulgaria
[5] College of Humanities and Social Sciences, HBKU, Doha, Qatar
[6] Sofia University "St Kliment Ohridski", Sofia, Bulgaria
{albarron, gmartino, pnakov}@qf.org.qa
{telsayed, reem.suwaileh}@qu.edu.qa
lluismv@amazon.com pepa.gencheva@siteground.com
wzaghouani@hbku.edu.qa spas.kyuchukov@gmail.com

Abstract. We present an overview of the CLEF-2018 CheckThat! Lab on Automatic Identification and Verification of Political Claims, with focus on Task 2: Factuality. The task asked to assess whether a given check-worthy claim made by a politician in the context of a debate/speech is factually true, half-true, or false. In terms of data, we focused on debates from the 2016 US Presidential Campaign, as well as on some speeches during and after the campaign (we also provided translations in Arabic), and we relied on comments and factuality judgments from factcheck.org and snopes.com, which we further refined manually. A total of 30 teams registered to participate in the lab, and five of them actually submitted runs. The most successful approaches used by the participants relied on the automatic retrieval of evidence from the Web. Similarities and other relationships between the claim and the retrieved documents were used as input to classifiers in order to make a decision. The best-performing official submissions achieved mean absolute error of .705 and .658 for the English and for the Arabic test sets, respectively. This leaves plenty of room for further improvement, and thus we release all datasets and the scoring scripts, which should enable further research in fact-checking.

Keywords: computational journalism · factuality · fact-checking · veracity

* This paper focuses on Task 2 (Factuality). For Task 1 (Check-Worthiness), see [1].

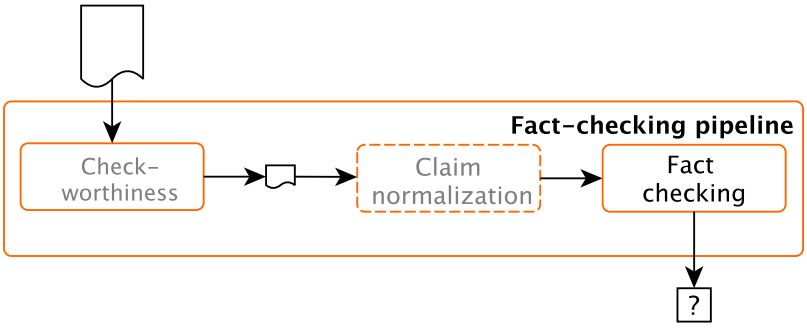

Fig. 1: The fact-checking pipeline. First, the input document is analyzed to identify sentences containing check-worthy claims [1]. Second, a check-worthy claim is extracted and normalized. Finally the claim is fact-checked (this task).

1 Introduction

The CheckThat! lab at CLEF-2018 [24] promotes the development of tools for computational journalism. It is divided into two tasks. This paper offers an overview of the CLEF 2018 CheckThat lab "Task 2: Factuality", which focuses on tools to verify (and possibly to provide evidence to an expert about) the factuality of a claim in a political debate or a speech. The reader interested in "Task 1: Check-worthiness" can refer to [1].

Task 2 represents the final step in the pipeline of the full fact-checking process, displayed in Figure 1. It is defined as follows:

> *Given a check-worthy claim in the form of a (transcribed) sentence, determine whether the claim is likely to be true, half-true, or false.*

We offered the task in two languages: English and Arabic, using translation for the latter. Table 1 shows two examples of debate fragments. In Table 1a, candidate Donald Trump claims that President Bill Clinton approved NAFTA. This is only half-true, as it was President George W. Bush who signed the approval for NAFTA, but Bill Clinton signed it into law. In Table 1b, Hillary Clinton claims Donald Trump has faced bankruptcy six times, which is true.[1]

The most successful approaches used by the participants based their veracity predictions on evidence retrieved from the Web, which they compared to the target claim. Then, they used a supervised model to predict whether the claim should be considered as true, half-true, or false.

The rest of the paper is organized as follows. Section 2 discusses related work. Section 3 describes the evaluation framework and the task setup. Section 4 gives an overview of the participating systems, followed by the official results in Section 5, and the discussion in Section 6. Finally, Section 7 draws conclusion.

[1] http://www.nbcnews.com/news/us-news/trump-bankruptcy-math-doesn-t-add-n598376

Hillary Clinton:	I think my husband did a pretty good job in the 1990s.	
Hillary Clinton:	I think a lot about what worked and how we can make it work again…	
Donald Trump:	Well, he approved NAFTA…	**half-true**

(a) On Bill Clinton's involvement in NAFTA.

Hillary Clinton:	He provided a good middle-class life for us, but the people he worked for, he expected the bargain to be kept on both sides.	
Hillary Clinton:	And when we talk about your business, you've taken business bankruptcy six times.	**true**

(b) On Donald Trump's bankruptcy-related history.

Table 1: Fragments from the 1st 2016 US presidential debate. The veracity of the check-worthy claims is shown on the right.

2 Related Work

The credibility of content on the Web has been questioned by researchers for a long time. While in the early days news portals were the main target [4, 11, 14], the interest has eventually shifted towards social media [6, 15, 27, 32], which are abundant in sophisticated malicious users, e.g., opinion manipulation *trolls*, *sockpuppets* [19], *Internet water army* [7], and *seminar users* [8].

There have been several datasets that focus on rumor detection. The gold labels are typically extracted from fact-checking websites such as Politifact with datasets ranging in size from 300 for the Emergent dataset [10] to 12.8K claims for the Liar dataset [30]. Another fact-checking source that has been used is snopes.com, with datasets ranging in size from 1k claims [18] to 5k claims [25]. Less popular as a source has been Wikipedia with datasets ranging in size from 100 [25] to 185k claims for the FEVER dataset [28]. These datasets rely on crowdsourced annotations, which allows them to get large-scale, but risks having lower quality standards compared to the rigorous annotations by fact-checking organizations. Other crowdsourced efforts include the SemEval-2017's shared task on Rumor Detection [9] with 5.5k annotated rumorous tweets, and CREDBANK with 60M annotated tweets [21]. Finally, there have been manual annotation efforts, e.g., for fact-checking the answers in a community question answering forum with size of 250 [20]. There have been also efforts in languages other than English, including Arabic [3], Bulgarian [14], and Chinese [18].

Several truth discovery algorithms are studied and combined in an ensemble classifier for veracity estimation in the VERA system [2]. However, the input to their model is structured data, while here we are interested in unstructured text as input.

Original	Normalized
Donald Trump: Well, he approved NAFTA	President Bill Clinton approved NAFTA
Donald Trump: Last year, we had almost $800 billion trade deficit.	In 2015, the USA had a trade deficit of almost $800 billion a year.

Table 2: Examples of claims as originally expressed and their normalized versions, as included in the CT-FCC-18 corpus for fact checking.

Similarly, the task defined in [23] combines three objectives: assessing the credibility of a set of posted articles, estimating the trustworthiness of sources, and predicting user's expertise. They considered a manifold of features characterizing language, topics and Web-specific statistics (e.g., review ratings) on top of a continuous conditional random fields model. In follow-up work, [26] proposed a model to support or refute claims from snopes.com and Wikipedia by considering supporting information gathered from the Web. In yet another follow-up work, [27] proposed a complex model that considers stance, source reliability, language style, and temporal information. Finally, [22] surveyed different methodologies to assess user-generated Web contents on the basis of various aspects, including credibility.

Many participants based their models upon [16] (cf. Section 4). In this case, keywords are selected from the claim and submitted as queries against search engines. Then, the returned results are fed into a neural network, which incorporates LSTM-derived representations of the claim and of the retrieved documents, together with similarities between the claim and the Web text.

3 Evaluation Framework

3.1 Corpus

We produced the corpus CT-FCC-18 , which stands for CheckThat! Fact-Checking Corpus 2018.[2] CT-FCC-18 includes claims from the 2016 US Presidential campaign, political speeches and a number of isolated claims. In order to derive the annotation, we used the publicly-available analysis carried out by FactCheck.org.[3] This analysis includes labeling a claim as true, half-true, or false and we adopt these same labels as gold standard. Understanding some of the claims depended heavily on their context. Hence, we manually produced normalized versions in order to make them self-contained. We also got rid of non-informative text fragments. In a real scenario, a model would be necessary to carry out this process, as illustrated in Figure 1. Table 2 shows two examples of normalized claims.

[2] CT-FCC-18 is available at https://github.com/clef2018-factchecking/clef2018-factchecking/.

[3] For instance, http://transcripts.factcheck.org/presidential-debate-hofstra-university-hempstead-new-york/

English	Arabic	Label
A U.S. postage stamp commemorates the Islamic holidays of Eid al-Fitr and Eid al-Adha.	تحتفل الطوابع البريدية الأمريكية بعطلة عيد الفطر وعيد الأضحى.	**true**
The first three digits of a bar code indicate a product's country of origin.	تشير الأرقام الثلاثة الأولى من الرمز الشريطي على المنتجات إلى بلد المنشأ.	**half-true**
Facebook CEO Mark Zuckerberg has converted to Islam.	دخل الرئيس التنفيذي لفيسبوك، مارك زوكربيرج، في الإسلام.	**false**

Table 3: Examples of claims from snopes.com, which we translated to Arabic.

For Arabic, we compiled additional claims without context. Different from the rest of the documents, we focused on Arab- and Islam-related claims from *Snopes.com*. We searched for relevant claims by querying the website with terms such as "Arab", "Islam", and "Palestine", and initially retrieved 400 claims. We then extracted both the text and the labels for those claims. We manually excluded: (*a*) claims of low interest to the Arab World; (*b*) near-duplicates; and (*c*) claims with ambiguous or unconfirmed labels. We gathered a total of 150 claims after filtering and normalization: 30 true, 10 half-true, and 110 false. We translated the claims into Arabic with Google Translate and manually post-edited the result. Table 3 shows examples of the original and translated claims.

Table 4 shows statistics about the full CT-FCC-18 corpus. The English partition includes claims from five debates and five speeches. The Arabic partition includes the claims from the same five debates and one of the speeches. These translations were produced by professional translators. Additionally, the Arabic partition includes 150 isolated claims. For both languages, the first three debates were released as training data, and the rest of the claims were used for testing.

3.2 Evaluation Measures

We have an ordering between the classes (*true*, *half-true*, and *false*), where confusing one extreme with the other one is more harmful than confusing it with a neighboring class. This is known as an *ordinal classification* problem (aka *ordinal regression*), and requires an evaluation measure that takes this ordering into account. We chose mean absolute error (MAE) as the official measure:

$$MAE = \frac{\sum_{c=1}^{C} d(y_c, x_c)}{C} \quad (1)$$

where y_c and x_c are the gold and the predicted labels for claim c, respectively, and $d \in \{0, 1, 2\}$ is the difference between them (*false*:0, *half-true*:1, *true*:2).

We also compute macro-average MAE, accuracy, macro-averaged F_1, and macro-averaged recall.[4]

[4] The implementation of the evaluation measures is available at http://github.com/clef2018-factchecking/clef2018-factchecking/

Training	True	Half True	False
Debates			
1st Presidential	**8**	**9**	**13**
H. Clinton	3	4	2
D. Trump	4	5	11
L. Holt	1	0	0
2nd Presidential	**4**	**7**	**14**
H. Clinton	2	2	2
D. Trump	1	5	12
A. Cooper	1	0	0
Vice-Presidential	**7**	**6**	**14**
T. Kaine	5	4	9
M. Pence	2	2	5

Test	True	Half True	False
Debates			
3rd Presidential	**19**	**8**	**21**
H. Clinton	13	4	4
D. Trump	5	3	17
Ch. Wallace	1	1	0
9th Democratic	**3**	**3**	**4**
H. Clinton	3	0	2
B. Sanders	0	3	2
D. Trump Speeches			
Acceptance	8	5	7
At WEF	6	2	3
At Tax Reform Event	4	4	4
Address to Congress	6	3	4
Miami Speech	4	9	12
Isolated claims			
● Snopes.com	30	10	110

Table 4: Overview of the claims in the CT-FCC-18 corpus. The instances that were translated into Arabic are marked with . Isolated claims from Snopes.com —released in Arabic only— are marked with ●. In speeches and debates, the number of true, half-true, and false claims is broken down to the speaker level.

4 Overview of Participants' Approaches

Table 5 offers a summary of the used approaches and representations; see the system description papers for more detail. Overall, participants chose to ignore the context of the claim, i.e., they did not use the rest of the debate/speech. They further only used the normalized version of the claim, ignoring the original sentence it originated in.

Copenhagen [29] used convolutional neural networks and support vector machines. In order to get information to support or to refute a claim, they retrieved a number of snippets by querying Google. Different from [16] —the model they were inspired by—, they did not select keywords, but queried the search engine with full texts (of decreasing size, in case no enough documents were retrieved). The text of the claim of the most similar retrieved supporting texts were then fed into their model.

UPV-INAOE-Autoritas [13] used a random forest. Similarly to the *Copenhagen* team [29], they retrieved evidence from the Web. In this case, both the Google and the Bing search engines were used to retrieve five snippets for a query consisting of the full claim. For each of the ten retrieved snippets, three features were computed: (*i*) the similarity between the claim and the snippet, calculated using word2vec embeddings, (*ii*) the similarity between the claim and the snippet, calculated over the tokens, and (*iii*) the Alexa rank of the website. These features were also combined, considering their mean and standard deviation.

	[13]	[17]	[29]	[31]
Learning Models				
Logistic regression				☑
Long short-term memory		☑		
Conv. neural network			☑	
Support vector machine			☑	
Random forest	☑			☑
Search Engines				
Google	☑		☑	☑
Bing	☑			
Representations				
Bag of words	☑		☑	☑
Word embeddings	☑	☑	☑	☑

	[13]	[17]	[29]	[31]
f(**claim, doc**)				
Similarity	☑		☑	
Alexa rank	☑			
Stance				☑
Contradiction				☑
NN concatenation			☑	

Teams
[13] UPV-INAOE-Autoritas
[29] Copenhagen
[17] Check it out
[31] bigIR
[–] FACTR

Table 5: Summary of the models and representations used by the participants.

BigIR [31] also retrieved supporting documents from the Web; in this case, following the same strategy as [16]. Still, they go further in trying to find the relevant fragments within the retrieved documents. Rather than using all the contents, they first compute the similarity between the claim and each sentence in the document and then they select those that pass a given threshold. The features for the supervised model are aggregations of the ones computed for each claim–sentence pair and include the stance of the sentence with respect to the claim and the degree of contradiction between the claim and the sentence, calculated at the term level.

Check it out [17] opted for a bidirectional long short-term memory network with attention. Different from the previous approaches, in this case no external information (e.g., no supporting documents) was used at all. Only the embedding representations of the claim itself were considered.

Note that the *bigIR* team [31] tried to identify the relevant fragments in the retrieved Web documents by considering only those with high similarity with respect to the claim. Most other approaches [29, 31] were based to some extent on [16], and only the *Check it out* team [17] approached the task without using any external supporting documents.

5 Results

The lab participants were allowed to submit one primary and no more than two contrastive runs. The latter were aimed at trying variations of their main approach or alternative models. However, for ranking purposes, only the primary submissions were considered. Five teams submitted runs for the English task; two of them did so for Arabic as well.

	MAE	Macro MAE	Acc	Macro F_1	Macro AvgR
[29] **Copenhagen**					
primary	**.7050**$_{(1)}$	**.6746**$_{(1)}$	**.4317**$_{(1)}$	**.4008**$_{(1)}$	**.4502**$_{(1)}$
cont. 1	.7698	.7339	.4676	.4681	.4721
[–] **FACTR**					
primary	.9137$_{(2)}$	.9280$_{(2)}$	.4101$_{(2)}$	.3236$_{(2)}$	.3684$_{(2)}$
cont. 1	.9209	.9358	.4029	.3063	.3611
cont. 2	.9281	.9314	.4101	.3420	.3759
[13] **UPV–INAOE–Autoritas**					
primary	.9496$_{(3)}$	.9706$_{(3)}$	.3885$_{(4)}$	.2613$_{(3)}$	.3403$_{(3)}$
[31] **bigIR**					
primary	.9640$_{(4)}$	1.0000$_{(4)}$	.3957$_{(3)}$	.1890$_{(4)}$	.3333$_{(4)}$
cont. 1	.9640	1.0000	.3957	.1890	.3333
cont. 2	.9424	.9256	.3525	.3297	.3405
[17] **Check It Out**					
primary	.9640$_{(4)}$	1.0000$_{(4)}$	.3957$_{(3)}$	.1890$_{(4)}$	.3333$_{(4)}$
Baselines					
n-gram	.9137	.9236	.3957	.3095	.3588
random	.8345	.8139	.3597	.3569	.3589

Table 6: English results, ranked based on MAE, the official evaluation measure. The best score for each evaluation measure is shown in bold.

English. Table 6 shows the results on the English dataset. Overall, the top-performing system is the one by the Copenhagen team [29]. One aspect that might explain the relatively large difference in performance compared to the other teams is the use of additional training material. The Copenhagen team incorporated hundreds of labeled claims from Politifact[5] to their training set. As described in Section 4, this model combines the claim and supporting texts to build representations. Their primary submission is an SVM, whereas their contrastive one uses a CNN.

The FACTR team, ranked second, used an approach similar to most other teams: converting the claim into a query for a search engine, computing stance, sentiment, and other features over the supporting documents, and using them in a supervised model. Unfortunately, no further information is available about it, as no paper was submitted to describe their system.

To put these results in perspective, the bottom of Table 6 shows the results for two baselines: (*i*) random label, and (*ii*) an n-gram based classifier. We can see that both baselines outperform many of the teams. In particular, in terms of MAE, only the Copenhagen team could improve over the random baseline, while the second best team FACTR is tied with the n-gram baseline. However, the baselines are weak on other evaluation measures, e.g., on Accuracy.

[5] http://www.politifact.com

Arabic. Table 7 shows the results for the two teams that participated in the Arabic task. The FACTR team translated all the claims into English and performed the rest of the experiments in that language. In contrast, UPV–INAOE–Autoritas [13] translated the claims into English, but only in order to query the search engines,[6] and then translated the retrieved evidence into Arabic in order to keep working in that language. Perhaps, the noise generated by using two imperfect translations caused their performance to decrease; the performance of the two teams in the English task was much closer.

Looking at the bottom of Table 7, we can see that once again the winning team FACTR managed to outperform both baselines. However, this time the random baseline was not as strong, and was clearly worse than the n-gram one.

6 Discussion

While the training set only included debates, the test set included speeches and single claims (cf. Section 3.1). Table 8 shows the performance of the models when dealing with each type of input text. For English, the top-performing models dealt better with speeches than with debates, and the lower the ranking, the smaller the differences. Perhaps having relatively more focused texts (as in a speech, the speaker usually follows a predefined script and does not need to adapt to other speakers) causes the factuality estimation to be simpler. Another reason could be that there is more online evidence to judge claims from speeches. We observe the same trend for Arabic: claims from speeches or isolated claims could be verified better than those coming from debates.

Overall, the performance of the models for Arabic was better than for English. The reason is that the isolated claims from Snopes.com —which were released only in Arabic (cf. Table 4)— were easier to verify.

7 Conclusion

We provided an overview of the CLEF-2018 CheckThat! Lab on Automatic Identification and Verification of Political Claims, with focus on Task 2 on assessing the veracity of claims. The task consisted of labeling isolated claims or from a political debate or a speech according to their factuality: true, half-true, or false. We offered the task in both English and Arabic.

Our evaluation framework consisted of a dataset of five debates and five speeches divided into training and test partitions both in English and Arabic. The evaluation was carried out using mean absolute error. Such a framework allowed five research teams to experiment with the use of different classifiers —convolutional neural networks, long short-term memory networks, support vector machines, random forests, and logistic regression— and multiple representations that aimed at assessing the factuality of a claim by considering evidence downloaded from the Web.

[6] The Arabic dataset was produced by translating the instances from English (cf. Section 3). Hence it was difficult to find evidence in Arabic.

	MAE	Macro MAE	Acc	Macro F1	Macro AvgR
FACTR					
primary	**.6579**$_{(1)}$	**.8914**$_{(1)}$	**.5921**$_{(1)}$	**.3730**$_{(1)}$	**.3804**$_{(1)}$
cont. 1	.7018	.9461	.5833	.3691	.3766
cont. 2	.6623	.9153	.5965	.3657	.3804
[13] **UPV–INAOE–Autoritas**					
primary	.8202$_{(2)}$	1.0417$_{(2)}$	.5175$_{(2)}$	.2796$_{(2)}$	.3027$_{(2)}$
Baselines					
n-gram	.6798	.9850	.5789	.2827	.3267
random	.9868	.9141	.3070	.2733	.2945

Table 7: Arabic results, ranked based on MAE, the official evaluation measure. The best score for each evaluation measure is shown in bold.

	English		**Arabic**		
	Debate	**Speech**	**Debate**	**Speech**	**Single**
[29] Copenhagen	.8103$_{(1)}$	.6420$_{(1)}$			
FACTR	.9310$_{(3)}$	.9012$_{(2)}$	1.0345$_{(2)}$	.8500$_{(1)}$	.4867$_{(1)}$
[13] UPV–INAOE–Aut.	.8966$_{(2)}$	.9877$_{(4)}$	.9483$_{(1)}$	.9500$_{(2)}$	.7533$_{(2)}$
[31] bigIR	.9483$_{(4)}$	.9753$_{(3)}$			
[17] Check It Out	.9483$_{(4)}$	.9753$_{(3)}$			

Table 8: Mean absolute error for the primary submissions when dealing with claims from different sources: debates, speeches, or isolated claims.

The best-performing models relied on convolutional neural networks and a manifold of similarities. Yet the performance on the test dataset remains ceiled at mean absolute error of 0.705. This leaves large room for further improvement, and thus we release[7] all datasets and the scoring scripts, which should enable further research in check-worthiness estimation.

In future iterations of the lab, we plan to add more debates and speeches, both annotated and unannotated, which would enable semi-supervised learning. We further want to add annotations for the same debates/speeches from different fact-checking organizations, which would allow using multi-task learning [12].

Acknowledgments

This work was made possible in part by NPRP grant# NPRP 7-1313-1-245 from the Qatar National Research Fund (a member of Qatar Foundation). Statements made herein are solely the responsibility of the authors.

[7] http://alt.qcri.org/clef2018-factcheck/

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: Cappellato et al. [5]
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. WWW '16, Montréal, Québec, Canada (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. pp. 21–27. NAACL-HLT '18, New Orleans, Louisiana, USA (2018)
4. Brill, A.M.: Online journalists embrace new marketing function. Newspaper Research Journal **22**(2), 28 (2001)
5. 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, Avignon, France (2018)
6. 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)
7. Chen, C., Wu, K., Srinivasan, V., Zhang, X.: Battling the Internet Water Army: detection of hidden paid posters. In: Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. pp. 116–120. ASONAM '13, Niagara, Ontario, Canada (2013)
8. Darwish, K., Alexandrov, D., Nakov, P., Mejova, Y.: Seminar users in the Arabic Twitter sphere. In: Proceedings of the 9th International Conference on Social Informatics. pp. 91–108. SocInfo '17, Oxford, UK (2017)
9. Derczynski, L., Bontcheva, K., Liakata, M., Procter, R., Wong Sak Hoi, G., Zubiaga, A.: SemEval-2017 Task 8: RumourEval: Determining rumour veracity and support for rumours. In: Proceedings of the 11th International Workshop on Semantic Evaluation. pp. 60–67. SemEval '17, Vancouver, Canada (2017)
10. Ferreira, W., Vlachos, A.: Emergent: a novel data-set for stance classification. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 1163–1168. NAACL-HLT '16, San Diego, California, USA (2016)
11. Finberg, H., Stone, M.L., Lynch, D.: Digital journalism credibility study. Online News Association. Retrieved November **3**, 2003 (2002)
12. Gencheva, P., Nakov, P., Màrquez, L., Barrón-Cedeño, A., Koychev, I.: A context-aware approach for detecting worth-checking claims in political debates. In: Proceedings of the International Conference Recent Advances in Natural Language Processing. pp. 267–276. RANLP '17, Varna, Bulgaria (2017)
13. Ghanem, B., Montes-y Gómez, M., Rangel, F., Rosso, P.: UPV-INAOE-Autoritas - Check That: An Approach based on External Sources to Detect Claims Credibility. In: Cappellato et al. [5]
14. 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)

15. Karadzhov, G., Gencheva, P., Nakov, P., Koychev, I.: We built a fake news & click-bait filter: What happened next will blow your mind! In: Proceedings of the International Conference on Recent Advances in Natural Language Processing. pp. 334–343. RANLP '17, Varna, Bulgaria (2017)
16. 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, Varna, Bulgaria (2017)
17. Lal, Y.K., Khattar, D., Kumar, V., Mishra, A., Varma, V.: Check It Out : Politics and Neural Networks. In: Cappellato et al. [5]
18. 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)
19. Mihaylov, T., Nakov, P.: Hunting for troll comments in news community forums. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. pp. 399–405. ACL '16, Berlin, Germany (2016)
20. Mihaylova, T., Nakov, P., Màrquez, L., Barrón-Cedeño, A., Mohtarami, M., Karadjov, G., Glass, J.: Fact checking in community forums. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence. pp. 879–886. AAAI '18, New Orleans, Louisiana, USA (2018)
21. Mitra, T., Gilbert, E.: CREDBANK: a large-scale social media corpus with associated credibility annotations. In: Cha, M., Mascolo, C., Sandvig, C. (eds.) Proceedings of the Ninth International Conference on Web and Social Media. pp. 258–267. ICWSM '15, Oxford, UK (2015)
22. Momeni, E., Cardie, C., Diakopoulos, N.: A survey on assessment and ranking methodologies for user-generated content on the web. ACM Comput. Surv. **48**(3), 41:1–41:49 (Dec 2015)
23. 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)
24. 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: Bellot, P., Trabelsi, C., Mothe, J., Murtagh, F., Nie, J., Soulier, L., Sanjuan, E., Cappellato, L., Ferro, N. (eds.) Proceedings of the Ninth International Conference of the CLEF Association: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Lecture Notes in Computer Science, Springer, Avignon, France (2018)
25. 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. CIKM '16, ACM, Indianapolis, Indiana, USA (2016)
26. 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. CIKM '16, Indianapolis, Indiana, USA (2016)
27. Popat, K., Mukherjee, S., Strötgen, J., Weikum, G.: Where the truth lies: Explaining the credibility of emerging claims on the web and social media. In: Proceed-

ings of the 26th International Conference on World Wide Web Companion. pp. 1003–1012. WWW '17, Perth, Australia (2017)
28. 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, Louisiana, USA (2018)
29. Wang, D., Simonsen, J., Larseny, B., Lioma, C.: The Copenhagen Team Participation in the Factuality Task of the Competition of Automatic Identification and Verification of Claims in Political Debates of the CLEF-2018 Fact Checking Lab. In: Cappellato et al. [5]
30. Wang, W.Y.: "Liar, liar pants on fire": A new benchmark dataset for fake news detection. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. pp. 422–426. ACL '17, Vancouver, Canada (2017)
31. Yasser, K., Kutlu, M., , Elsayed, T.: bigIR at CLEF 2018: Detection and Verifi cation of Check-Worthy Political Claims. In: Cappellato et al. [5]
32. Zubiaga, A., Liakata, M., Procter, R., Wong Sak Hoi, G., Tolmie, P.: Analysing how people orient to and spread rumours in social media by looking at conversational threads. PLoS ONE **11**(3), 1–29 (03 2016)