

Z-Index at CheckThat! Lab 2022: Check-Worthiness Identification on Tweet Text

Prerona Tarannum¹, Md. Arid Hasan¹, Firoj Alam² and Sheak Rashed Haider Noori¹

¹*Daffodil International University*

²*Qatar Computing Research Institute*

Abstract

The wide use of social media and digital technologies facilitates sharing various news and information about events and activities. Despite sharing positive information misleading and false information is also spreading on social media. There have been efforts in identifying such misleading information both manually by human experts and automatic tools. Manual effort does not scale well due to the high volume of information, containing factual claims, are appearing online. Therefore, automatically identifying check-worthy claims can be very useful for human experts. In this study, we describe our participation in Subtask-1A: Check-worthiness of tweets (English, Dutch and Spanish) of CheckThat! lab at CLEF 2022. We performed standard preprocessing steps and applied different models to identify whether a given text is worthy of fact-checking or not. We use the oversampling technique to balance the dataset and applied SVM and Random Forest (RF) with TF-IDF representations. We also used BERT multilingual (BERT-m) and XLM-RoBERTa-base pre-trained models for the experiments. We used BERT-m for the official submissions and our systems ranked as 3rd, 5th, and 12th in Spanish, Dutch, and English, respectively. In further experiments, our evaluation shows that transformer models (BERT-m and XLM-RoBERTa-base) outperform the SVM and RF in Dutch and English languages where a different scenario is observed for Spanish.

Keywords

Check-worthiness, Check-worthy claim detection, Fact-checking, Disinformation, Misinformation, Social Media Text, Transformer Models,

1. Introduction

Recently, social media became the main communication channel to exchanging information among people. As a result, it becomes the primary source of news [1]. In our daily activities, such information is helpful, however, a major part of them contains misleading content that is harmful to individuals, society, or organizations [2, 3]. The harmful or misleading content includes hate speech [4], hostility [5, 6], propagandistic news and memes [7, 8, 9, 10], harmful memes [11], abusive language [12], cyberbullying and cyber-aggression [13, 14] and rumours [15]. The misleading or harmful aspects of such information raised the interest to identify and flag them to reduce their spread, further. There have been significant research efforts to automatically identify


CLEF 2022: Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy

✉ prerona15-14134@diu.edu.bd (P. Tarannum); arid.cse0325.c@diu.edu.bd (Md. A. Hasan); fialam@hbku.edu.qa (F. Alam); drnoori@daffodilvarsity.edu.bd (S. R. H. Noori)

🆔 0000-0002-3292-1870 (P. Tarannum); 0000-0001-7916-614X (Md. A. Hasan); 0000-0001-7172-1997 (F. Alam); 0000-0001-6937-6039 (S. R. H. Noori)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

such content. Recent surveys on fake news [16] disinformation [17], rumours [15], propaganda [7], multimodal memes [18], hate speech [4], cyberbullying [19], and offensive content [20] highlight the importance of the problem and relevant approaches to address them.

Most often information is disseminated with facts to make people believe it is true, which are typically found in political debates, and social and global agendas. Identifying whether such facts are true or false is an important step in fighting misleading information. There have been manual efforts by fact-checking organizations to identify the truthfulness of such factual statements. As such manual efforts do not scale well, therefore, it is important to automatically identify them. However, there is a reliability issue with the automated approach [21]. A trade-off is to support human fact-checkers using an automated approach, which includes different steps in the fact-checking pipeline [22]. The first step of the fact checking pipeline is to find content that is check-worthy. The CheckThat! Lab (CTL) shared tasks is addressing this problem for the past several years. As an ongoing effort, this year CheckThat! Lab offered check worthiness subtask in six different languages such as Arabic, Bulgarian, Dutch, English, Spanish, and Turkish where data was collected from Twitter [23, 24, 25]. We participated in check worthiness subtask and focused on Dutch, English, and Spanish. For the experiments, we used different pretrained transformer based models, which have been widely used in several NLP tasks [2, 26]. The difficulties arise when multilingual pretrained model is used in such tasks where facts and claims vary by country [27] and knowledge transferring across the language could spread the disinformation. We used multilingual transformer models (m-BERT and XLM-RoBERTa) for our experiments. In addition to the transformer models, we also used SVM and RF with TF-IDF representations.

The rest of this paper is organized as follows. In section 2, we provided related works that are relevant for this study. We then discuss the methodology in Section 3. Results of the experiments and detailed discussions are provided in Section 4. Finally, we conclude our study in Section 5.

2. Related Work

To deal with the factuality of statements there have been initiatives to manually check them and as result many fact-checking organizations have emerged, such as FactCheck.org¹, Snopes², PolitiFact³, and FullFact⁴. In addition, there have also been some international initiatives such as the *Credibility Coalition*⁵ and *Eufactcheck*⁶ [28].

One of the earlier efforts in this direction is the ClaimBuster system [29], which has been developed using the transcripts of 30 historical US election debates with a total of 28,029 transcribed sentences. The annotation includes *non-factual*, *unimportant factual*, and *check-worthy factual* class labels and has been carried out by students, professors, and journalists. Gencheva et al. [30] also focused on the 2016 US Presidential debates for which they obtained annotations from different fact-checking organizations. An extension of this work resulted in

¹<http://www.factcheck.org/>

²<http://www.snopes.com/fact-check/>

³<http://www.politifact.com/>

⁴<http://fullfact.org/>

⁵<https://credibilitycoalition.org/>

⁶<https://eufactcheck.eu/>

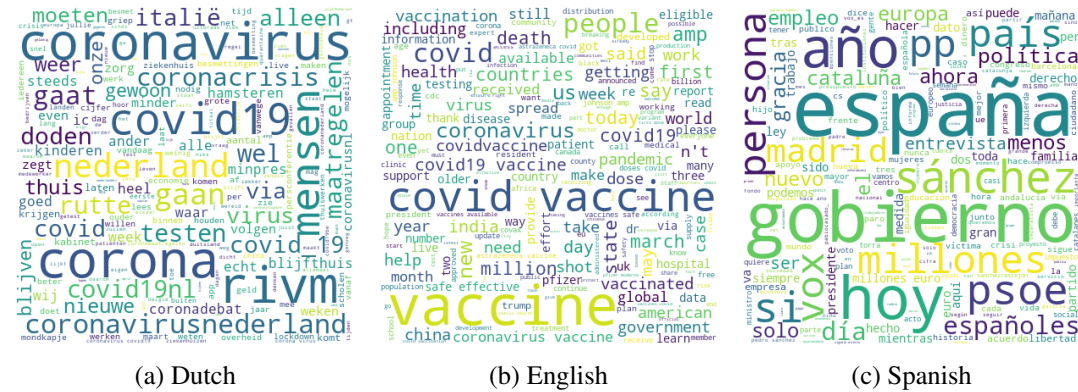


Figure 1: Word-cloud representing top frequent words in different languages.

the development of ClaimRank, where the authors used more data and also included Arabic content Jaradat et al. [31]. Alam et al. [3] focused on COVID 19 topics in languages which are Arabic, Bulgarian, Dutch, and English, and achieved strong performances using pre-trained language models. The study also discussed the utility of single-task and multitask settings. The positive unlabelled learning technique for check-worthiness tasks has been introduced by Wright and Augenstein [32] where authors experimented with this technique with the BERT model on different datasets and achieved the best results on two datasets out of three. The study of Alhindi et al. [33] introduced a multi-layer annotated news corpus and augmented discourse structure to understand the relation between fact-checking and argumentation. The first Turkish dataset for check-worthiness has been studied by Kartal and Kutlu [34], where BERT multilingual outperforms other models.

Some notable research outcomes came from shared tasks. For example, the CLEF Check-That! labs’ shared tasks [35, 36, 37, 38] in the past few years featured challenges on automatic identification [39, 40] and verification [41, 42] of claims in political debates, and tweets [43].

3. Methodology

3.1. Data

The dataset we used in our study is obtained from CLEF CheckThat!2022 lab task1: *Identifying Relevant Claims in Tweets* [25]. The data is based on the COVID-19 topic for Dutch and English where Spanish is mixed of politics and COVID-19 topics, which is collected from Twitter. In table 1, we present the distribution of the datasets that we used in this shared task to run our experiments. In Figure 1, we present the word cloud for all three languages to understand the most common words present in the datasets. We first removed the stopwords from the data and then used the rest of the words to generate the most frequent words.

Table 1
Data splits and distributions of Subtask 1A: Check-worthiness of tweets

Class label	Train	Dev	Test	Total
Dutch				
No	546	44	350	940
Yes	377	28	316	721
Total	923	72	666	1661
English				
No	1675	151	110	1936
Yes	447	44	39	530
Total	2122	195	149	2466
Spanish				
No	3087	2195	4296	9578
Yes	1903	305	704	2912
Total	4990	2500	5000	12490

3.2. Preprocessing

The CTL subtask-1A datasets are collected from Twitter. As a result, the data contains many symbols, URLs, and invisible characters. We performed several preprocessing steps to clean the noisy data. First, we perform URLs and unnecessary character removal steps by following the approach discussed in [44]. Then, we removed the stopwords from the data. Finally, we removed hashtag signs and usernames.

3.3. Models

We used both deep learning and traditional models to run classification experiments. As deep learning algorithms, we used two transformer based models, BERT [45] and XLM-RoBERTa [46]. Several factors were considered while choosing the algorithms. Among the transformer based models, BERT and XLM-RoBERTa are larger in parameter size.⁷ The number of parameters and network size is responsible for computation time and performance of the learning. For these two models, we used the multilingual version of the models. For the later case, we used the two most popular algorithms such as (i) Random Forest (RF) [47], and (ii) Support Vector Machines (SVM) [48].

3.4. Experiments

Transformers models We use the Transformer Toolkit [49] for transformer-based models. We used learning rate of $1e - 5$ to fine-tune each model [45]. Model specific tokenizer is available with Transformer Toolkit that we used in our study. For transformer based model, we run 4, 2,

⁷110 million parameters in *BERT multilingual* and 125 million in *XLM-RoBERTa base*

Table 2

Hyper-parameters for traditional models to reproduce the results.

Parameters	Dutch		English		Spanish	
	SVM	RF	SVM	RF	SVM	RF
Number of Feature	1850	1500	1750	2800	3200	1700
N-gram	3	3	4	3	4	3
Random Seed	2814	2814	2814	2814	2814	2814

Table 3

Official results on the test set and overall ranking of Subtask 1A: Check-worthiness of tweets

Language	Model	F1 (positive class)	Rank
Dutch	BERT-m	0.497	5 th
English	BERT-m	0.478	12 th
Spanish	BERT-m	0.303	3 rd

and 8 epochs for BERT-m model for Dutch, English, and Spanish languages, and 4, 4, and 8 epochs for Dutch, English, and Spanish languages for XLM-RoBERTa-base model.

Traditional Algorithms To train the classifiers using the above-mentioned traditional models, we first transformed the preprocessed data into tf-idf vectors with weighted n -gram (unigram, bigram and trigram) to use contextual information. The class distribution of provided dataset for English and Spanish is not well balanced. Therefore, to balance the class distribution, we applied oversampling techniques [50] for all three languages. We merged the train and dev-test set to train the model. We applied the upsampling technique to the combined dataset with a ratio of 1.0 with respect to the negative class. In Table 2, we report the hyper-parameters with the values to reproduce our results.

4. Results and Discussion

In Table 3, we report the official results and ranking evaluated by the lab organizers. The official evaluation metric for subtask 1A is F1 measure with respect to the positive class.

In Table 4, we report the detailed classification results for each language. After releasing the gold set once the submission period ends, we re-run all the experiments and reported the detailed results. From the table, we can conclude that among the traditional models the performance of SVM is much better than RF except for Spanish data where RF is 0.25% higher. The upsampling technique for traditional models improves from 0.10% to 1.10% on different languages with respect to the positive class. We know from the literature, transformer based models are well-known for their performances and capabilities. Although XLM-Roberta base and BERT-m models provide the best results for Dutch and English languages with respect to positive class, where the traditional model outperforms the transformer models on Spanish language by a large margin.

Table 4

Detail results on the test set of Subtask 1A: Check-worthiness of tweets. **Bold** indicates positive class F1 score. Underline indicates best F1 score for each language.

Class label	Model	Accuracy	Precision	Recall	F1 Score
Dutch					
No	SVM	59.01	60.85	61.71	61.28
Yes			56.91	56.01	56.46
No	RF	57.96	57.85	73.71	64.82
Yes			58.18	40.51	47.76
No	BERT-m	60.06	60.82	67.43	63.96
Yes			58.99	51.90	55.22
No	XLM-RoBERTa base	56.76	60.00	53.14	56.36
Yes			53.93	60.76	<u>57.14</u>
English					
No	SVM	69.80	85.71	70.91	77.61
Yes			44.83	66.67	53.61
No	RF	75.17	76.64	95.45	85.02
Yes			58.33	17.95	27.45
No	BERT-m	63.09	89.86	56.36	69.27
Yes			40.00	82.05	<u>53.78</u>
No	XLM-RoBERTa base	51.01	91.11	37.27	52.90
Yes			33.65	89.74	48.95
Spanish					
No	SVM	84.76	92.89	89.08	90.95
Yes			46.70	58.38	51.89
No	RF	88.62	91.27	95.93	93.54
Yes			63.92	44.03	<u>52.14</u>
No	BERT-m	68.30	91.75	69.34	78.99
Yes			24.87	61.93	35.49
No	XLM-RoBERTa base	70.64	90.33	73.72	81.18
Yes			24.43	51.85	33.21

5. Conclusion

In this study, we have run comparative experiments using different check-worthiness claim datasets consisting of Dutch, English, and Spanish languages, which are provided by CLEF CheckThat! lab 2022 organizers as a part of shared tasks. We cleaned the data to run the classification experiments. We investigated different machine learning algorithms including traditional (i.e., SVM) and deep learning models (i.e., BERT multilingual). Despite the cost

of increased resource and time complexity, transformer based models did not perform well for Spanish language, however, outperformed the Dutch and English languages. Our study reveals that the transformer based models outperforms the traditional machine learning approach for Dutch and English language tasks.

6. Acknowledgments

We would like to thank the organizers and other participants in the challenge. We are thankful to DIU NLP and ML Research Lab for the workplace support. Finally, thanks to all the anonymous reviewers for their suggestions.

Part of this work is made within the Tanbih mega-project,⁸ developed at the Qatar Computing Research Institute, HBKU, which aims to limit the impact of “fake news”, propaganda, and media bias by making users aware of what they are reading, thus promoting media literacy and critical thinking.

References

- [1] A. Perrin, Social media usage. pew research center 2015: 52-68, 2020.
- [2] F. Alam, F. Dalvi, S. Shaar, N. Durrani, H. Mubarak, A. Nikolov, G. Da San Martino, A. Abdelali, H. Sajjad, K. Darwish, P. Nakov, Fighting the COVID-19 infodemic in social media: A holistic perspective and a call to arms, in: Proceedings of the International AAAI Conference on Web and Social Media, ICWSM '21, 2021, pp. 913–922. URL: <https://ojs.aaai.org/index.php/ICWSM/article/view/18114>.
- [3] F. Alam, S. Shaar, F. Dalvi, H. Sajjad, A. Nikolov, H. Mubarak, G. Da San Martino, A. Abdelali, N. Durrani, K. Darwish, A. Al-Homaid, W. Zaghouni, T. Caselli, G. Danoe, F. Stolk, B. Bruntink, P. Nakov, Fighting the COVID-19 infodemic: Modeling the perspective of journalists, fact-checkers, social media platforms, policy makers, and the society, in: Findings of the Association for Computational Linguistics: EMNLP 2021, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 611–649. URL: <https://aclanthology.org/2021.findings-emnlp.56>. doi:10.18653/v1/2021.findings-emnlp.56.
- [4] P. Fortuna, S. Nunes, A survey on automatic detection of hate speech in text, ACM Computing Surveys (CSUR) 51 (2018) 1–30.
- [5] S. Brooke, “Condescending, Rude, Assholes”: Framing gender and hostility on Stack Overflow, in: Proceedings of the Third Workshop on Abusive Language Online, Association for Computational Linguistics, Florence, Italy, 2019, pp. 172–180. URL: <https://aclanthology.org/W19-3519>. doi:10.18653/v1/W19-3519.
- [6] S. Joksimovic, R. S. Baker, J. Ocumpaugh, J. M. L. Andres, I. Tot, E. Y. Wang, S. Dawson, Automated identification of verbally abusive behaviors in online discussions, in: Proceedings of the Third Workshop on Abusive Language Online, Association for Computational

⁸<http://tanbih.qcri.org>

- Linguistics, Florence, Italy, 2019, pp. 36–45. URL: <https://aclanthology.org/W19-3505>. doi:10.18653/v1/W19-3505.
- [7] G. Da San Martino, S. Cresci, A. Barrón-Cedeño, S. Yu, R. D. Pietro, P. Nakov, A survey on computational propaganda detection, in: C. Bessiere (Ed.), Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020, ijcai.org, 2020, pp. 4826–4832. URL: <https://doi.org/10.24963/ijcai.2020/672>. doi:10.24963/ijcai.2020/672.
- [8] G. Da San Martino, S. Yu, A. Barrón-Cedeño, R. Petrov, P. Nakov, Fine-grained analysis of propaganda in news article, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 5636–5646. URL: <https://www.aclweb.org/anthology/D19-1565>. doi:10.18653/v1/D19-1565.
- [9] D. Dimitrov, B. Bin Ali, S. Shaar, F. Alam, F. Silvestri, H. Firooz, P. Nakov, G. Da San Martino, Detecting propaganda techniques in memes, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL-IJCNLP '21, Association for Computational Linguistics, Online, 2021, pp. 6603–6617. URL: <https://aclanthology.org/2021.acl-long.516>. doi:10.18653/v1/2021.acl-long.516.
- [10] D. Dimitrov, B. Bin Ali, S. Shaar, F. Alam, F. Silvestri, H. Firooz, P. Nakov, G. Da San Martino, SemEval-2021 task 6: Detection of persuasion techniques in texts and images, in: Proceedings of the 15th International Workshop on Semantic Evaluation, SemEval '21, Association for Computational Linguistics, Online, 2021, pp. 70–98. URL: <https://aclanthology.org/2021.semeval-1.7>. doi:10.18653/v1/2021.semeval-1.7.
- [11] S. Pramanick, S. Sharma, D. Dimitrov, M. S. Akhtar, P. Nakov, T. Chakraborty, MOMENTA: A multimodal framework for detecting harmful memes and their targets, in: Findings of the Association for Computational Linguistics: EMNLP 2021, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 4439–4455. URL: <https://aclanthology.org/2021.findings-emnlp.379>. doi:10.18653/v1/2021.findings-emnlp.379.
- [12] H. Mubarak, K. Darwish, W. Magdy, Abusive language detection on arabic social media, in: Proceedings of the first workshop on abusive language online, 2017, pp. 52–56.
- [13] C. Van Hee, E. Lefever, B. Verhoeven, J. Mennes, B. Desmet, G. De Pauw, W. Daelemans, V. Hoste, Detection and fine-grained classification of cyberbullying events, in: Proceedings of the International Conference Recent Advances in Natural Language Processing, INCOMA Ltd. Shoumen, BULGARIA, Hissar, Bulgaria, 2015, pp. 672–680. URL: <https://aclanthology.org/R15-1086>.
- [14] R. Kumar, A. K. Ojha, S. Malmasi, M. Zampieri, Benchmarking aggression identification in social media, in: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying, 2018, pp. 1–11.
- [15] A. Bondielli, F. Marcelloni, A survey on fake news and rumour detection techniques, *Information Sciences* 497 (2019) 38–55.
- [16] X. Zhou, R. Zafarani, A survey of fake news: Fundamental theories, detection methods, and opportunities, *CSUR* 53 (2020) 1–40.

- [17] F. Alam, S. Cresci, T. Chakraborty, F. Silvestri, D. Dimitrov, G. D. S. Martino, S. Shaar, H. Firooz, P. Nakov, A survey on multimodal disinformation detection, arXiv:2103.12541 (2021).
- [18] T. H. Afridi, A. Alam, M. N. Khan, J. Khan, Y. K. Lee, A multimodal memes classification: A survey and open research issues, in: 5th International Conference on Smart City Applications, SCA 2020, Springer Science and Business Media Deutschland GmbH, 2021, pp. 1451–1466.
- [19] B. Haidar, M. Chamoun, F. Yamout, Cyberbullying detection: A survey on multilingual techniques, in: 2016 European Modelling Symposium (EMS), 2016, pp. 165–171. doi:10.1109/EMS.2016.037.
- [20] F. Husain, O. Uzuner, A survey of offensive language detection for the arabic language, ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP) 20 (2021) 1–44.
- [21] S. Shaar, F. Alam, G. D. S. Martino, P. Nakov, Assisting the human fact-checkers: Detecting all previously fact-checked claims in a document, arXiv:2109.07410 (2021). arXiv:2109.07410.
- [22] P. Nakov, D. Corney, M. Hasanain, F. Alam, T. Elsayed, A. Barrón-Cedeño, P. Papotti, S. Shaar, G. D. S. Martino, Automated fact-checking for assisting human fact-checkers, in: Proceedings of the 30th International Joint Conference on Artificial Intelligence, IJCAI '21, 2021, pp. 4551–4558.
- [23] P. Nakov, A. Barrón-Cedeño, G. Da San Martino, F. Alam, J. M. Struß, T. Mandl, R. Míguez, T. Caselli, M. Kutlu, W. Zaghouni, C. Li, S. Shaar, G. K. Shahi, H. Mubarak, A. Nikolov, N. Babulkov, Y. S. Kartal, J. Beltrán, The clef-2022 checkthat! lab on fighting the covid-19 infodemic and fake news detection, in: M. Hagen, S. Verberne, C. Macdonald, C. Seifert, K. Balog, K. Nørvåg, V. Setty (Eds.), Advances in Information Retrieval, Springer International Publishing, Cham, 2022, pp. 416–428.
- [24] P. Nakov, A. Barrón-Cedeño, G. Da San Martino, F. Alam, J. M. Struß, T. Mandl, R. Míguez, T. Caselli, M. Kutlu, W. Zaghouni, C. Li, S. Shaar, G. K. Shahi, H. Mubarak, A. Nikolov, N. Babulkov, Y. S. Kartal, J. Beltrán, M. Wiegand, M. Siegel, J. Köhler, Overview of the CLEF-2022 CheckThat! lab on fighting the COVID-19 infodemic and fake news detection, in: Proceedings of the 13th International Conference of the CLEF Association: Information Access Evaluation meets Multilinguality, Multimodality, and Visualization, CLEF '2022, Bologna, Italy, 2022.
- [25] P. Nakov, A. Barrón-Cedeño, G. Da San Martino, F. Alam, R. Míguez, T. Caselli, M. Kutlu, W. Zaghouni, C. Li, S. Shaar, H. Mubarak, A. Nikolov, Y. S. Kartal, J. Beltrán, Overview of the CLEF-2022 CheckThat! lab task 1 on identifying relevant claims in tweets, in: Working Notes of CLEF 2022—Conference and Labs of the Evaluation Forum, CLEF '2022, Bologna, Italy, 2022.
- [26] F. Alam, A. Hasan, T. Alam, A. Khan, J. Tajrin, N. Khan, S. A. Chowdhury, A review of bangla natural language processing tasks and the utility of transformer models, arXiv preprint arXiv:2107.03844 (2021).
- [27] K. Singh, G. Lima, M. Cha, C. Cha, J. Kulshrestha, Y.-Y. Ahn, O. Varol, Misinformation, believability, and vaccine acceptance over 40 countries: Takeaways from the initial phase of the covid-19 infodemic, Plos one 17 (2022) e0263381.

- [28] M. Stencel, Number of fact-checking outlets surges to 188 in more than 60 countries, *Duke Reporters' LAB* (2019) 12–17.
- [29] N. Hassan, C. Li, M. Tremayne, Detecting check-worthy factual claims in presidential debates, in: *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, CIKM '15*, Association for Computing Machinery, Melbourne, Australia, 2015, pp. 1835–1838. URL: <https://doi.org/10.1145/2806416.2806652>. doi:10.1145/2806416.2806652.
- [30] P. Gencheva, P. Nakov, L. Màrquez, A. Barrón-Cedeño, I. Koychev, A context-aware approach for detecting worth-checking claims in political debates, in: *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP '17*, INCOMA Ltd., Varna, Bulgaria, 2017, pp. 267–276. URL: https://doi.org/10.26615/978-954-452-049-6_037.
- [31] I. Jaradat, P. Gencheva, A. Barrón-Cedeño, L. Màrquez, P. Nakov, ClaimRank: Detecting check-worthy claims in Arabic and English, in: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, NAACL-HLT '18*, Association for Computational Linguistics, New Orleans, Louisiana, USA, 2018, pp. 26–30. URL: <https://aclanthology.org/N18-5006>. doi:10.18653/v1/N18-5006.
- [32] D. Wright, I. Augenstein, Claim check-worthiness detection as positive unlabelled learning, *arXiv preprint arXiv:2003.02736* (2020).
- [33] T. Alhindi, B. McManus, S. Muresan, What to fact-check: Guiding check-worthy information detection in news articles through argumentative discourse structure, in: *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2021, pp. 380–391.
- [34] Y. S. Kartal, M. Kutlu, Trclaim-19: The first collection for turkish check-worthy claim detection with annotator rationales, in: *Proceedings of the 24th Conference on Computational Natural Language Learning*, 2020, pp. 386–395.
- [35] P. Nakov, A. Barrón-Cedeño, T. Elsayed, R. Suwaileh, L. Màrquez, W. Zaghouni, P. Atanasova, S. Kyuchukov, G. Da San Martino, Overview of the CLEF-2018 Check-That! lab on automatic identification and verification of political claims, in: *CLEF, Lecture Notes in Computer Science*, Springer, Avignon, France, 2018, pp. 372–387. URL: https://link.springer.com/chapter/10.1007/978-3-319-98932-7_32#citeas.
- [36] T. Elsayed, P. Nakov, A. Barrón-Cedeño, M. Hasanain, R. Suwaileh, G. Da San Martino, P. Atanasova, Overview of the CLEF-2019 CheckThat!: Automatic identification and verification of claims, in: *Experimental IR Meets Multilinguality, Multimodality, and Interaction, LNCS*, Lugano, Switzerland, 2019.
- [37] T. Elsayed, P. Nakov, A. Barrón-Cedeño, M. Hasanain, R. Suwaileh, G. Da San Martino, P. Atanasova, CheckThat! at CLEF 2019: Automatic identification and verification of claims, in: *Advances in Information Retrieval, ECIR '19*, Springer International Publishing, Cologne, Germany, 2019, pp. 309–315. URL: https://link.springer.com/chapter/10.1007/978-3-030-15719-7_41.
- [38] S. Shaar, F. Alam, G. Da San Martino, A. Nikolov, W. Zaghouni, P. Nakov, A. Feldman, Findings of the NLP4IF-2021 shared tasks on fighting the COVID-19 infodemic and censorship detection, in: *Proceedings of the Fourth Workshop on NLP for Internet Freedom*:

Censorship, Disinformation, and Propaganda, NLP4IF '21', Association for Computational Linguistics, Online, 2021, pp. 82–92. URL: <https://aclanthology.org/2021.nlp4if-1.12>. doi:10.18653/v1/2021.nlp4if-1.12.

- [39] P. Atanasova, L. Màrquez, A. Barrón-Cedeño, T. Elsayed, R. Suwaileh, W. Zaghouani, S. Kyuchukov, G. Da San Martino, P. Nakov, 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.
- [40] P. Atanasova, P. Nakov, G. Karadzhov, M. Mohtarami, G. D. S. Martino, Overview of the CLEF-2019 CheckThat! Lab on Automatic Identification and Verification of Claims. Task 1: Check-Worthiness, in: CLEF 2019 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Lugano, Switzerland, 2019.
- [41] A. Barrón-Cedeño, T. Elsayed, R. Suwaileh, L. Màrquez, P. Atanasova, W. Zaghouani, S. Kyuchukov, G. Da San Martino, P. Nakov, Overview of the CLEF-2018 CheckThat! lab on automatic identification and verification of political claims, task 2: Factuality, in: CLEF 2018 Working Notes. Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS.org, Avignon, France, 2018.
- [42] M. Hasanain, R. Suwaileh, T. Elsayed, A. Barrón-Cedeño, P. Nakov, Overview of the CLEF-2019 CheckThat! Lab on Automatic Identification and Verification of Claims. Task 2: Evidence and Factuality, in: CLEF 2019 Working Notes. Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS.org, Lugano, Switzerland, 2019.
- [43] P. Nakov, G. Da San Martino, T. Elsayed, A. Barrón-Cedeño, R. Míguez, S. Shaar, F. Alam, F. Haouari, M. Hasanain, W. Mansour, B. Hamdan, Z. S. Ali, N. Babulkov, A. Nikolov, G. K. Shahi, J. M. Struß, T. Mandl, M. Kutlu, Y. S. Kartal, Overview of the CLEF-2021 CheckThat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news, in: K. Candan, B. Ionescu, L. Goeuriot, B. Larsen, H. Müller, A. Joly, M. Maistro, F. Piroi, G. Faggioli, N. Ferro (Eds.), *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Twelfth International Conference of the CLEF Association, LNCS (12880)*, Springer, 2021. URL: https://link.springer.com/chapter/10.1007/978-3-030-72240-1_75.
- [44] F. Alam, H. Sajjad, M. Imran, F. Ofli, CrisisBench: Benchmarking crisis-related social media datasets for humanitarian information processing, in: *Proceedings of the International AAAI Conference on Web and Social Media, ICWSM '21*, 2021, pp. 923–932. URL: <https://ojs.aaai.org/index.php/ICWSM/article/view/18115>.
- [45] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, 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, NAACL-HLT '19*, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 4171–4186. URL: <https://www.aclweb.org/anthology/N19-1423>. doi:10.18653/v1/N19-1423.
- [46] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: *Proceedings of the 58th Annual Meeting of the Association for Computa-*

tional Linguistics, ACL '20, Association for Computational Linguistics, Online, 2020, pp. 8440–8451. URL: <https://aclanthology.org/2020.acl-main.747>. doi:10.18653/v1/2020.acl-main.747.

- [47] A. Liaw, M. Wiener, et al., Classification and regression by random forest, *R news* 2 (2002) 18–22.
- [48] J. Platt, Sequential minimal optimization: A fast algorithm for training support vector machines, Technical Report, Microsoft, Redmond, USA., 1998.
- [49] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, A. Rush, Transformers: State-of-the-art natural language processing, in: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP '20*, Association for Computational Linguistics, Online, 2020, pp. 38–45. URL: <https://aclanthology.org/2020.emnlp-demos.6>. doi:10.18653/v1/2020.emnlp-demos.6.
- [50] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, Smote: synthetic minority over-sampling technique, *Journal of artificial intelligence research* 16 (2002) 321–357.