QMUL-SDS at CheckThat! 2021: Enriching Pre-Trained Language Models for the Estimation of Check-Worthiness of Arabic Tweets

Amani S. Abumansour[1,2], Arkaitz Zubiaga[1]

[1]*Queen Mary University of London, United Kingdom*
[2]*Taif University, Saudi Arabia*

Abstract
This paper describes our submission to the CheckThat! Lab at CLEF 2021, where we participated in Subtask 1A (check-worthy claim detection) in Arabic. We introduce our approach to estimate the check-worthiness of tweets as a ranking task. In our approach, we propose to fine-tune state-of-art transformer based models for Arabic such as AraBERTv0.2-base as well as to leverage additional training data from last year's shared task (CheckThat! Lab 2020) along with the dataset provided this year. According to the official evaluation, our submission obtained a joint 4th position in the competition where seven other groups participated.

Keywords
checkworthiness, checkworthy claim detection, fact-checking, Arabic NLP.

1. Introduction

Sifting through large volumes of social media content can become a burdensome task for journalists performing fact-checking, where computational journalism approaches to automated fact-checking can help alleviate the task [1]. The automated fact-checking pipeline encompasses an important sub-task consisting of claim check-worthiness detection, i.e. given a collection of sentences as input, identify the most prominent claims ranked by check-worthiness [2]. This sub-task can be treated as a text classification task, consisting of determining check-worthy statements, followed by a ranking step based on the check-worthiness score.

There has been a body of work in claim (check-worthiness) detection in recent years. Both ClaimBuster [3] and CNC [4] used traditional classifiers in combination, such as SVM and Logistic regression. In addition, others have used neural networks as is the case of Atanasova et al. [5] who considered context and features along with Feed-Forward Neural Network (FNN). Neural network models were also used by ClaimRank [6] as well as participants of recent shared tasks at CheckThat! [7, 8].

Subsequently, the emergence of Bidirectional Encoder Representations (BERT) has meant a significant milestone in the history of NLP since it attained state-of-art results when applied on several tasks including text classification [9]. The effectiveness of pre-trained language models

CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania
a.s.a.abumansour@qmul.ac.uk (A. S. Abumansour); a.zubiaga@qmul.ac.uk (A. Zubiaga)

stems from the ability of transformers to create embeddings as an output of the pre-training process. Since then, many efforts have focused on fine-tuning pre-trained language models. For example, Hasanain and Elsayed [10] fine-tuned multilingual BERT (mBERT), and others fine-tuned AraBERT [11, 12] in the CheckThat! Lab 2020 [8].

In our contribution to Substask 1A [13] in CheckThat! Lab at CLEF 2021 [14], we started by fine-tuning the latest version of AraBERT. In addition, we investigated the benefits of incorporating the CT20-AR dataset from last year's edition of the shared task (CheckThat! 2020), besides preprocessing functions in our test. In what follows we describe our approach and discuss the results we achieved.

2. Approach

This section describes in greater detail the process we follow to handle the task, and is divided into three parts: datasets, data preprocessing, and ranking methodology. The first part illustrates the datasets we used in our experiments. The second part describes the pre-processing techniques performed prior to training and testing. The third part describes our method to rank Arabic tweets using AraBERTv0.2-base.

2.1. Datasets

The organisers provided the CT21-AR training dataset which contains Arabic tweets [13]. The CT21-AR dataset provides labels for each entry (sentence) as to whether it is a claim or not, as well as whether it is check-worthy or not. For the purposes of this task, we solely considered the check-worthiness label where a value of 1 indicates a "check-worthy" tweet, and a 0 indicates a "not check worthy" tweet.

We also looked into expanding the training data by leveraging additional datasets from previous editions of the CheckThat! shared task. In order to increase the training data, we incorporated the CT20-AR dataset from CLEF 2020 along with the CT21-AR dataset for training [15]. In the case of CT20-AR, it only contains one label pertaining to "check-worthy" (1), and "not check worthy" (0); labels for "claim" or "not claim" were not provided.

Both datasets, CT20-AR and CT21-AR, are imbalanced. In particular, 25% of the CT21-AR training data are check-worthy claims, with a slightly higher ratio (27.5%) for CT20-AR.

2.2. Pre-processing

The provided dataset contains tweets written by different users and therefore with variations in style and writing. This can be seen for example in the presence of emojis, hyperlinks, and other symbols in some of the tweets. In addition, users might mention other users or type hashtags. Last year's CheckThat! participants, such as Novak [11], used AraBERT but did not apply any preprocessing step in their attempts. Use of preprocessing has however been recommended by Antoun et al. [16], given that it converts the data into a more standard format prior to fine-tune AraBERT. Therefore, we considered it would be useful to leverage the preprocessor in the setup of our experiments.

Hence, we used AraBERT's preprocess function[1] to perform the following:

- Substitute all URLs, email addresses, and user mentions with [رابط], [بريد], and [مستخدم] respectively.
- Eliminate line breaks and markup written in HTML, repeated characters, extra spaces, and unwanted characters including emotion icons.
- Handling white spaces between words and digits (non-Arabic, or English), and/or a combination of both, and before and after two brackets.

Additionally, we applied extra functions to replace digits with [رقم], and to remove punctuation marks that were not treated in the previous function such as # ,and _. Afterwards, we tokenised all sentences using the BERT Fast tokeniser[2].

2.3. Ranking Methodology

The subtask 1A in CheckThat 2021 is evaluated as a ranking task. Therefore, we used huggingface transformers [17] for fine tuning a newly released AraBERTv0.2-base with Sequence Classification. Then, the results of the neural network output layer are passed into a softmax function in order to acquire the probability distribution for each predicted output class; we use the value output by the softmax function to rank the sentences by check-worthiness. Thus, we estimated the level of check-worthiness for each tweet in the test set.

3. Results and Discussion

For the experiments, we developed four models as shown in Table 1. Models vary in two aspects: (1) whether or not the pre-processing component is used, and (2) whether or not the CT20-AR dataset is used to expand the training data. These variations allowed us to assess the extent to which these two variations could lead to improved performance and subsequently for us to choose the model to submit to the competition. In models 1 and 3, both datasets (CT20-AR and CT21-AR) are utilised for the training phase. The other models (2 and 4) are only trained on the current CT21-AR training set. When it comes to the pre-processing, we adopted it in two models, 1 and 2; hence testing all combinations of pre-processing (yes/no) and use of extra CT20-AR dataset (yes/no). We consistently use the ranking methodology described in §2.3 throughout the experiments.

Experiments on the development set enabled us to choose the optimal model to submit to the competition. We found that both model 2 and model 4 were overfitting, with better results for models 1 and 3 incorporating the CT20-AR data. Thus, models 1 and 3 seemed to be better options, so we continued further exploring the pre-processing step. Through further exploration, we found that the pre-processing was leading to noticeable improvements in the performance. This ultimately led to our decision of submitting model 1 to the shared task.

Further, Table 2 presents the performance of our four models based on test set. All the results outperformed the n-gram baseline in all metrics. In terms of mean average precision (MAP)

[1] https://github.com/aub-mind/arabert

[2] https://huggingface.co/transformers/model_doc/bert.html#berttokenizerfast

Table 1
Description of our models using different variants of training data and pre-processing.

Model	Datasets	Preprocessing
Model_1	CT20-AR + CT21-AR	Yes
Model_2	CT21-AR	Yes
Model_3	CT20-AR + CT21-AR	No
Model_4	CT21-AR	No

Table 2
Performance of our models on test set. The primary model is boldfaced

Model	MAP	MRR	RP	P@1	P@3	P@5	P@10	P@20	P@30
Model_1	0.597	0.5	0.603	0	0.667	0.6	0.7	0.65	0.72
Model_2	0.5997	1	0.5868	1	0.6667	0.8	0.9	0.8	0.7
Model_3	0.5815	0.3333	0.5868	0	0.3333	0.6	0.8	0.7	0.6667
Model_4	0.5924	1	0.5868	1	0.6667	0.8	0.9	0.8	0.8333
ngram-baseline	0.428	0.5	0.409	0	0.667	0.6	0.5	0.45	0.44

Table 3
Official results of subtask 1A for Arabic

Rank	Model	MAP	MRR	RP	P@1	P@3	P@5	P@10	P@20	P@30
1	Accenture	0.658	1	0.599	1	1	1	1	0.95	0.84
2	bigIR	0.615	0.5	0.579	0	0.667	0.6	0.6	0.8	0.74
3	SCUoL	0.612	1	0.599	1	1	1	1	0.95	0.78
4	ICompass	0.597	0.333	0.624	0	0.333	0.4	0.4	0.5	0.64
4	QMUL-SDS	0.597	0.5	0.603	0	0.667	0.6	0.7	0.65	0.72
5	TOBB ETU	0.575	0.333	0.574	0	0.333	0.4	0.4	0.5	0.68
6	DamascusTeam	0.571	0.5	0.558	0.667	0.6	0.8	0.7	0.64	
7	ibaris	0.548	1	0.55	1	0.667	0.6	0.5	0.4	0.58
	ngram-baseline	0.428	0.5	0.409	0	0.667	0.6	0.5	0.45	0.44

specifically, the estimated result is very similar in comparison with other results. Also, it got the higher scores for R-Precision (RP) and p@3. However, we observe that both model 1 and model 3 get the first position in the ranking wrong (P@1=0), which requires further analysis.

Overall, the mean average precision (MAP) is the official metric used for the competition. Table 3 shows the final results compared to other participants, where our team ranked as joint 4^{th} position.

4. Conclusion

We have described our submission to CLEF CheckThat! Lab 2021 subtask 1A (claim check-worthiness detection in Arabic). We propose two variations to further improve a fine-tuned

AraBERT model. More specifically, we propose to test variations performing text pre-processing (yes/no) as well as incorporating additional training data from the CT20-AR dataset (yes/no). We then rank our predictions by using a softmax function, which leads to the final ranking. Through development, we observed that the model making use of both variants (using pre-processing and incorporating additional data) led to best performance, and hence chose to submit this model to the competition. The improved performance with the use of a pre-processing step reinforces our findings from the participation in CheckThat! 2020 [18] showing that processing special tokens, such as numeric expressions, can be beneficial for the task.

Our primary model achieved the joint 4th place according to the official evaluation, with a MAP score of 0.597. We make some observations that are left for further exploration in future work. First, we plan to dig into the predictions of our models to investigate extreme cases where p@k equals to 0 or 1. Second, we aim to tackle the imbalance of the datasets with the aim of improving performance. Lastly, we plan to experiment with other Arabic pre-trained language models in the future.

Acknowledgments

This work was supported by the Engineering and Physical Sciences Research Council (grant EP/V048597/1). Amani S. Abumansour holds a scholarship from Taif University, Saudi Arabia.

References

[1] A. Zubiaga, Mining social media for newsgathering: A review, Online Social Networks and Media 13 (2019) 100049.

[2] N. Hassan, F. Arslan, C. Li, M. Tremayne, 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, 2017, pp. 1803–1812.

[3] N. Hassan, G. Zhang, F. Arslan, J. Caraballo, D. Jimenez, S. Gawsane, S. Hasan, M. Joseph, A. Kulkarni, A. K. Nayak, et al., Claimbuster: The first-ever end-to-end fact-checking system, Proceedings of the VLDB Endowment 10 (2017) 1945–1948.

[4] L. Konstantinovskiy, O. Price, M. Babakar, A. Zubiaga, Toward automated factchecking: Developing an annotation schema and benchmark for consistent automated claim detection, Digital Threats: Research and Practice 2 (2021). URL: https://doi.org/10.1145/3412869. doi:10.1145/3412869.

[5] P. Atanasova, P. Nakov, L. Màrquez, A. Barrón-Cedeño, G. Karadzhov, T. Mihaylova, M. Mohtarami, J. Glass, Automatic fact-checking using context and discourse information, J. Data and Information Quality 11 (2019). URL: https://doi.org/10.1145/3297722. doi:10.1145/3297722.

[6] 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, Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 26–30. URL: https://www.aclweb.org/anthology/N18-5006. doi:10.18653/v1/N18-5006.

[7] T. Elsayed, P. Nakov, A. Barrón-Cedeño, M. Hasanain, R. Suwaileh, G. Martino, P. Atanasova, Overview of the CLEF-2019 CheckThat! Lab: Automatic Identification and Verification of Claims, 2019, pp. 301–321. doi:10.1007/978-3-030-28577-7_25.

[8] A. Barrón-Cedeno, T. Elsayed, P. Nakov, G. Da San Martino, M. Hasanain, R. Suwaileh, F. Haouari, N. Babulkov, B. Hamdan, A. Nikolov, et al., Overview of checkthat! 2020: Automatic identification and verification of claims in social media, in: Proceedings of CLEF, Springer, 2020, pp. 215–236.

[9] 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, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://www.aclweb.org/anthology/N19-1423. doi:10.18653/v1/N19-1423.

[10] M. Hasanain, T. Elsayed, bigir at checkthat! 2020: Multilingual bert for ranking arabic tweets by check-worthiness., in: Proceedings of CLEF (Working Notes), 2020.

[11] V. Novak, Accenture at checkthat! 2020: If you say so: Post-hoc fact-checking of claims using transformer-based models, in: Proceedings of CLEF (Working Notes), 2020.

[12] Y. S. Kartal, M. Kutlu, Tobb etu at checkthat! 2020: Prioritizing english and arabic claims based on check-worthiness, in: Proceedings of CLEF (Working Notes), 2020.

[13] S. Shaar, M. Hasanain, B. Hamdan, Z. S. Ali, F. Haouari, M. K. Alex Nikolov, F. A. Yavuz Selim Kartal, G. Da San Martino, A. Barrón-Cedeño, R. Míguez, T. Elsayed, P. Nakov, Overview of the CLEF-2021 CheckThat! lab task 1 on check-worthiness estimation in tweets and political debates, in: Working Notes of CLEF 2021—Conference and Labs of the Evaluation Forum, CLEF ’2021, Bucharest, Romania (online), 2021.

[14] 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: Proceedings of the 12th International Conference of the CLEF Association: Information Access Evaluation Meets Multiliguality, Multimodality, and Visualization, CLEF ’2021, Bucharest, Romania (online), 2021.

[15] M. Hasanain, F. Haouari, R. Suwaileh, Z. S. Ali, B. Hamdan, T. Elsayed, A. Barrón-Cedeño, G. D. S. Martino, P. Nakov, Overview of checkthat! 2020i arabic: Automatic identification and verification of claims in social media, in: Proceedings of CLEF, 2020.

[16] W. Antoun, F. Baly, H. Hajj, Arabert: Transformer-based model for arabic language understanding, in: Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection, 2020, pp. 9–15.

[17] T. Wolf, J. Chaumond, L. Debut, V. Sanh, C. Delangue, A. Moi, P. Cistac, M. Funtowicz, J. Davison, S. Shleifer, et al., Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2020, pp. 38–45.

[18] R. Alkhalifa, T. Yoong, E. Kochkina, A. Zubiaga, M. Liakata, QMUL-SDS at CheckThat! 2020: determining COVID-19 tweet check-worthiness using an enhanced CT-BERT with numeric expressions, in: Proceedings of CLEF (Working Notes), 2020.