bigIR at CheckThat! 2020: Multilingual BERT for Ranking Arabic Tweets by Check-worthiness

Maram Hasanain and Tamer Elsayed

Computer Science and Engineering Department, Qatar University, Doha, Qatar
{maram.hasanain, telsayed}@qu.edu.qa

Abstract. This paper describes the third-year participation of our bigIR group at Qatar University in CheckThat! lab at CLEF. This year we participated only in Arabic Task 1 that focuses on detecting check-worthy tweets on a given topic. We submitted four runs using both traditional classification models and a pre-trained language model: multilingual BERT (mBERT). Official results showed that our run using mBERT was the best among all our submitted runs. Furthermore, bigIR team was ranked third among all eight teams participated in the lab, with our best run ranked 6th among 28 runs.

1 Introduction

With the huge flood of false information on the Web and social media, verification of all claims that a user face is becoming infeasible. The situation is even more challenging for professional fact-checkers and journalists who usually track multiple topics simultaneously with each having many claims. Twitter poses even more challenges with the tweets being limited in size and very quickly spreading. Moreover, there is a huge volume of tweets that might not even contain any factual claims to begin with. This situation motivated work on prioritization of tweets by their importance of verification for a given topic. Task 1 in the CheckThat! lab at CLEF 2020 was designed to support research solving that specific problem [3]. In the lab, the problem of tweets check-worthiness estimation targeted by Task 1 was defined as follows: "Predict which tweet from a stream of tweets on a topic should be prioritized for fact-checking."

Although the task was offered for both English and Arabic tweets, the bigIR group at Qatar university decided to participate specifically in the Arabic task, since Arabic is one of the most dominant languages in Twitter [2], yet still understudied in the fact-checking domain in general. This is our participation for the third year in a row in Arabic tasks of CheckThat! lab [7,13].

In Arabic Task 1, organizers provided participants with two datasets. The training dataset includes three topics with each having 500 Arabic tweets annotated by check-worthiness. The test dataset includes twelve topics, each with 500

Arabic tweets [8]. For each test topic, we were asked to return a list of the 500 tweets for the topic ranked by their check-worthiness. We tackled this problem in two ways. In the first, we use traditional learning-based classifiers with hand-crafted features. In the second, we fine-tune a multilingual BERT (mBERT) pre-trained model [6] with a classification layer. The run using mBERT was the best-performing among all of our submitted runs and was ranked 6th among all 28 runs in the lab for this task. These results demonstrate the effectiveness of pre-trained models (and BERT specifically) for the problem of check-worthiness estimation which is consistent with very recent studies on the problem including other submissions to the same task [9,10,12].

We discuss the approach we followed in details in Section 2 and briefly present our results in comparison to top teams in the lab in Section 3. We finally provide some concluding remarks and directions for future work in Section 4.

2 Approach

We approach check-worthiness ranking by training different classification models. We choose two main approaches to the problem. We train several common text classification models with hand-crafted features hypothesizing they are good discriminators of claim check-worthiness. In the other approach, we fine-tune a multilingual BERT model [6]. BERT has shown strong performance in multiple text classification tasks, and very recent applications of BERT in the specific problem at hand showed promising results [9,10]. Details on both approaches are presented in this section.

2.1 Traditional Classification

We start by developing 13 features hand-crafted for this task. These features were selected and inspired by many existing studies on fact-checking and check-worthiness ranking. The features are categorized as follows:

- **Social features**
  - hasURL: whether the tweet has a URL or not. We observe that many non-check-worthy claims have URLs citing official news agencies.
  - Number of hashtags
  - isVerified: whether the author of the tweet is a verified user or not. Less check-worthy claims were observed from verified accounts in the training set.
  - Tweet popularity score: The sum of the number of retweets and likes the tweet received.
  - User social connection score: The sum of the number of followers and friends the tweet author has.
  - User engagement score: The sum of the number of tweets the user posted and liked in Twitter.

– **Tweet content and structure**. Under this category, we select features designed to capture tweet objectivity, its relevance to the topic, and its structure. We preprocess both the tweet and the topic (represented using its description). We apply the following preprocessing steps: stop words and URLs removal, expansion of hashtags by removing the # symbol and splitting the hashtag by underscores, eliminating special characters (e.g., $), removing diacritics, and finally normalizing the Arabic text to consolidate multiple spellings of the same character into a single unified form of it. The computed features are:
  - Jaccard Similarity between the topic and the tweet.
  - Count of entities identified in a tweet using a multilingual named-entity recognition tool [1].
  - Count of polarity words including positive ones (e.g., "Success") and negative words (e.g., "Corruption") identified using a large-scale multilingual sentiment lexicon [5]. We hypothesize tweets with no factual claims will include more sentiment rather than objective language.
  - Count of numbers in a tweet.
  - Count of quotes in a tweet.
  - Count of unique tokens.
  - Average of the word embedding vectors representing each token in the tweet. The embeddings were extracted from a word embedding model trained over a very large set of Arabic tweets [11]. For this feature, the tweet was preprocessed using a preprocessor provided by the model developers.

As for the classifiers, we use three classical classifiers, namely Logistic Regression, Support Vector Machine (SVM) and Random Forest, with default parameters as provided by scikit-learn Python package.[1] With leave-one-topic-out cross-validation over the training dataset, we apply a stepwise feature selection algorithm in which we greedly add the feature that results in best average performance over the folds. Eventually, we found a combination of only three features achieved the best overall performance for all three classifiers. Performance with these 3 features was superior to that achieved when using all 13 features. The features are word embeddings, *isVerified*, and count of quoted statements. We use the prediction probability of the positive class (i.e., how probable the tweet is check-worthy) as the ranking score to rank tweets in descending order per topic. We train the models using the three training topics provided by the task organizers [3,8].

2.2 Multilingual BERT

We fine-tune a Multilingual BERT (mBERT) model for the task of check-worthiness ranking. In this model, we represent the input as follows:

[CLS] + tweet text + [SEP] + topic text + [SEP]

[1] `https://scikit-learn.org/stable/`

where [CLS] is a special classification embedding and [SEP] is a token to indicate a separator between the two sentences. The topic was represented by its title concatenated with description. In order to use mBERT model for check-worthiness ranking, we add on top of it a dense layer, followed by an output Softmax classification layer to predict the probability for the two possible classes (whether the tweet is check-worthy or not). We fine tune the model in full including all layers of mBERT and the classification layer. The probability of the positive class was used as the check-worthiness score by which we rank tweets in descending order per topic.

We apply light preprocessing to both the tweet and topic by removing URLs, expanding hashtags by removing the # symbol and splitting the hashtag by underscores, eliminating special characters (e.g., $), and removing diacritics. For the model architecture specifications, we use uncased mBERT model with 12 layers and 768 hidden units. The dense layer on top of mBERT has 256 hidden units and relu activation function. We use binary cross-entropy loss for training, and set the maximum sequence length to 128 with training batch size of 32. The model was trained using the three training topics provided by the organizers.

3 Results

We submitted four runs for the task, which match exactly the models described in Section 2. Table 1 shows the best run per team for the top three teams in the task in addition to our remaining runs and the two baselines provided by the task organizers. As shown in the table, the run using mBERT achieved the best performance among all our runs measured by precision at rank 30 (P@30), which is the the official measure of the task. In fact, our team is ranked third among all participating eight teams, with a comparable performance to the second-ranked team. We find the mBERT model is our best performing model by far, which is consistent with its robust and effective performance across different ranking and classification tasks. We also observe that although all three traditional classifiers used the same features, SVM and Logistic Regression both showed superior performance over Random Forest.

We note here that our experiments on the problem are preliminary; further experiments are needed to improve and understand the results. For example, we observe that only 30% of the training data is check-worthy. Oversampling techniques of the positive class might result in better classification performance. Another future experiment is to consider integrating some of the hand-crafted features with the BERT representation in order to represent a claim with more than the content.

4 Conclusion and Future Work

Our work showed that a simple neural model using multilingual BERT had competitive performance that is superior to traditional classifiers that use many hand-crafted features for the task. However, we still need to conduct further

Table 1. Official results for best run for top three teams at Arabic Task 1 at CLEF2020 CheckThat! lab including all our runs. Our best run is boldfaced.

Run ID	**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
bigIR-svm	0.5667	0.5417	0.5472	0.4564
bigIR-logit_regression	0.5750	0.5375	0.5444	0.4525
bigIR-random_forest	0.4333	0.4542	0.4361	0.3835
baseline2	0.3500	0.3625	0.3472	0.3149
baseline1	0.3250	0.3333	0.3417	0.3244

experiments with more elaborate parameter optimization and feature selection to make more concrete conclusions. In comparison to other teams in the lab, we observe that the use of a language model pre-trained on Arabic data only can yield better performance and thus, we plan to experiment with such models next. We also hypothesize that including some of the hand-crafted features in the neural model can bring improvements to the performance and we plan to test this hypothesis in future work.

Acknowledgments

This work was made possible 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.

References

1. Al-Rfou, R., Kulkarni, V., Perozzi, B., Skiena, S.: Polyglot-NER: Massive multilingual named entity recognition. Proceedings of the 2015 SIAM International Conference on Data Mining, Vancouver, British Columbia, Canada, April 30 - May 2, 2015 (April 2015)
2. Alshaabi, T., Dewhurst, D.R., Minot, J.R., Arnold, M.V., Adams, J.L., Danforth, C.M., Dodds, P.S.: The growing echo chamber of social media: Measuring temporal and social contagion dynamics for over 150 languages on twitter for 2009–2020 (2020)
3. 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.: Overview of CheckThat! 2020: Automatic Identification and Verification of Claims in Social Media. LNCS (12260), Springer (2020)
4. Cappellato, L., Eickhoff, C., Ferro, N., Névéol, A. (eds.): CLEF 2020 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org (2020)
5. Chen, Y., Skiena, S.: Building sentiment lexicons for all major languages. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers). pp. 383–389 (2014)

6. 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, Volume 1 (Long and Short Papers). pp. 4171–4186 (2019)
7. Haouari, F., Ali, Z., Elsayed, T.: bigIR at CLEF 2019: Automatic Verification of Arabic Claims over the Web. In: Working Notes of CLEF 2019 – Conference and Labs of the Evaluation Forum (2019)
8. Hasanain, M., Haouari, F., Suwaileh, R., Ali, Z., Hamdan, B., Elsayed, T., Barrón-Cedeño, A., Da San Martino, G., Nakov, P.: Overview of CheckThat! 2020 Arabic: Automatic identification and verification of claims in social media. In: Cappellato et al. [4]
9. Kartal, Y.S., Guvenen, B., Kutlu, M.: Too many claims to fact-check: Prioritizing political claims based on check-worthiness. arXiv preprint arXiv:2004.08166 (2020)
10. Meng, K., Jimenez, D., Arslan, F., Devasier, J.D., Obembe, D., Li, C.: Gradient-based adversarial training on transformer networks for detecting check-worthy factual claims. arXiv preprint arXiv:2002.07725 (2020)
11. Soliman, A.B., Eissa, K., El-Beltagy, S.R.: Aravec: A set of arabic word embedding models for use in arabic nlp. Procedia Computer Science **117**, 256 – 265 (2017). https://doi.org/https://doi.org/10.1016/j.procs.2017.10.117, `http://www.sciencedirect.com/science/article/pii/S1877050917321749`, arabic Computational Linguistics
12. 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. [4]
13. Yasser, K., Kutlu, M., Elsayed, T.: bigIR at CLEF 2018: Detection and Verification of Check-Worthy Political Claims. In: Working Notes of CLEF 2018 – Conference and Labs of the Evaluation Forum (2018)