TOBB ETU at CheckThat! 2020: Prioritizing English and Arabic Claims Based on Check-Worthiness

Yavuz Selim Kartal and Mucahid Kutlu

TOBB University of Economics and Technology, Ankara, Turkey
{ykartal, m.kutlu}@etu.edu.tr

Abstract. Misinformation has many negative consequences on our daily life. While spread of misinformation is very fast, investigating veracity of claims is slow. Therefore, we urgently need systems helping human fact checkers in the combat against misinformation. In this paper, we present our participation in check-worthiness tasks (i.e., Task 1 and Task 5) of CLEF-2020 Check That! Lab. For English Task 1, we use logistic regression with fined-tuned BERT predictions, POS tags, controversial topics and a hand-crafted word list as features. For English Task 5, we again use logistic regression with fined-tuned BERT predictions and word embeddings as features. For Arabic Task 1, we use a hybrid approach of fined-tuned BERT model with the model used for English Task 5. For the Arabic task, we use AraBert as our Bert model. In the official evaluation of primary submissions, our primary models a) ranked 3^{rd} in Arabic Task 1 based on P@30 and shared the 1^{st} rank with another group based on P@5, b) ranked 5^{th} in English Task 1 based on average precision and shared the 1^{st} rank with five other groups based on reciprocal rank, P@1, P@3 and P@5 metrics, and c) ranked 3^{rd} in Task 5 based on average precision.

1 Introduction

Social media platforms provide an incredibly easy way to share information with others. Any information, including misinformation, can reach millions of people in a very short time. Unfortunately, misinformation spread over Internet cause many unpleasant incidents such as huge changes in stock prices[1]. Since the start of on-going Covid-19 pandemic, we have also witnessed how misinformation can cause unhealthy, potentially deadly, practices such as gargling with bleach to prevent Covid-19[2].

[1] https://www.reuters.com/article/net-us-usa-whitehouse-ap/hackers-send-fake-market-moving-ap-tweet-on-white-house-explosions-idUSBRE93M12Y20130423

[2] https://www.reuters.com/article/us-health-coronavirus-disinfectants/gargling-with-bleach-americans-misusing-disinfectants-to-prevent-coronavirus-survey-finds-idUSKBN23C2P2

In order to combat misinformation, there are many fact-checking websites which manually investigate veracity of claims and share their findings with public. However, misinformation spread much faster than true information [20] and investigating veracity of claims are extremely time consuming, taking around one day for a single claim [12]. Considering the vast amount of claims spread on the Internet and high cost of fact-checking, we urgently need systems helping fact-checkers to detect check-worthy claims, enabling them to focus on the important claims instead of spending their precious time for less important claims.

CLEF 2020 Check That! Lab [5] has organized two different shared-tasks (Task 1 and Task 5) for detecting check worthy claims. Task 1 has two different datasets, consisting of Arabic and English tweets while Task 5 has English political debates and transcribed speeches. In this paper, we present our methods developed for both Task 1 and Task 5.

In our study, we use two different ranking methodologies including logistic regression model and a hybrid combination of fined-tuned BERT model with logistic regression. We also investigate many different features including word-embeddings, presence of comparative and superlative adjectives, hand-crafted word list, domain-specific controversial topics, POS tags, metadata of tweets, and predictions of fined-tuned BERT models. Based on our experiments on training data, we use the following primary models for each task.

- **Arabic Task 1:** Hybrid model using word embeddings and BERT predictions as features for logistic regression model.
- **English Task 1:** Logistic regression model with POS tags, controversial topics, comparative and superlative adjectives, and BERT predictions as features.
- **English Task 5:** Logistic regression model with word embeddings and BERT predictions as features.

CLEF 2020 Check That! Lab used precision@30 (P@30) for Arabic Task 1 and average precision (AP) for English Task 1 and Task 5, as official evaluation metrics. Based on official metrics, our primary models for Arabic Task 1, English Task 1, and Task 5 ranked 3^{rd} (out of 8 groups), 5^{th} (out of 12 groups), and 3^{rd} (out of 3 groups), respectively. However, based on other metrics, our models shared the first rank with others in many cases. In particular, our primary model for Arabic Task 1 shared the 1^{st} rank with another group based on P@5. In addition, our primary model in Task 1 English shared the 1^{st} rank with 5 other groups on RR, P@1, P@3 and P@5.

2 Related Work

There are a number of studies in check-worthy claim detection. Hassan et al. [12] develop ClaimBuster which is one of the first check-worthy claim detection models. ClaimBuster uses many features including part-of-speech (POS) tags, named entities, sentiment, and TF-IDF representations of claims. TATHYA [17] uses

topics detected in all presidential debates from 1976 to 2016, POS tuples, entity history, and bag-of-words as features.

Gencheva et al. [7] propose a neural network model with a long list of sentence level and contextual features including sentiment, named entities, word embeddings, topics, contradictions, and others. Jaradat et al. [13] use similar features with Gencheva et al. [7] but extend the model for Arabic. In its followup work, Vasileva et al. [19] propose a multi-task learning model to detect whether a claim will be fact-checked by at least five of 9 reputable fact-checking organizations.

In 2018, Check That! Lab (CTL) has been organized for the first time in English and Arabic with participation of seven teams [3]. The participants investigated many learning models such as recurrent neural network (RNN) [10], multilayer perceptron [23], random forest (RF) [1], k-nearest neighbor (kNN) [8] and gradient boosting [22] with different sets of features such as bag-of-words [23], character n-gram [8], POS tags [10, 22, 23], verbal forms [23], named entities [22, 23], syntactic dependencies [23, 10], and word embeddings [10, 22, 23]. On English dataset, Prise de Fer [23] team achieved the best MAP scores using bag-of-words, POS tags, named entities, verbal forms, negations, sentiment, clauses, syntactic dependency and word embeddings with SVM-Multilayer perceptron learning. On Arabic dataset, the model of Yasser et al. [22] outperformed the others using POS tags, named entities, sentiment, topics, and word embeddings.

In CTL'19, chekc-worthiness task has been organized for only English [4]. 11 teams participated in the task and used varying models such as LSTM, SVM, naive bayes, and logistic regression (LR) with many features including readability of sentences and their context. Copenhagen team [11] achieved the best overall performance using syntactic dependency and word embeddings with weakly supervised LSTM model.

The labeled datasets provided by CTL enabled further studies in this task. Lespagnol et al. [15] explore using SVM, LR, and Random Forests with a long list of features including word-embeddings, POS tags, syntactic dependency tags, entities, and "information nutritional" features which represent factuality, emotion, controversy, credibility, and technicality of statements. Kartal et al. [14] use logistic regression utilizing BERT model with additional features including word embeddings, controversial topics, hand-crafted list of words, POS tags, presence of comparative and superlative adjectives, and adverbs. They achieve the highest AP scores on both CTL'18 and CTL'19 English datasets. In CTL'20, we use features adapted from Kartal et al. [14]'s model. However, we also explore additional features such as tweet meta data features. We also investigate a hybrid combination of fine-tuned BERT model with logistic regression.

3 Proposed Approach

In this section, we explain the features we investigate (Section 3.1) and models to prioritize claims (Section 3.2).

3.1 Features

BERT: We first remove mentions and URLs from tweets. For Arabic tweets, we also apply spelling correction using Farasa[3]. Then we fine tune BERT models using respective training data sets. We use multilingual uncased-large BERT model [6] for English tasks and Ara-BERT model [2] for the Arabic task. The prediction value of the fined tuned BERT model is used as a feature in the logistic regression model.

Word Embeddings (WE): Word embeddings are able to catch semantic and syntactic features of words. Thus, we use word embeddings to capture similarities between claims. Specifically, we represent each sentence as the average vector of its words. We use word2vec models pre-trained on Google news [16] in Task 5. For Task 1, we use fastText models pre-trained on Wikipedia [9]. Both word embedding models provide a vector size of 300. We exclude out-of-vocabulary words when we use word2vec.

Controversial Topics (CT): We use controversial topics feature defined by Kartal et al. [14]. In this feature, 11 major controversial topics in current US politics (e.g., immigration, gun policy, racism, abortion) are defined. Each topic is represented by the average word embeddings of hand-crafted related words (e.g., "immigrants", "illegal", "borders", "Mexican", "Latino", and "Hispanic" for the immigration topic). We also represent sentences/tweets to be ranked as the average word embeddings excluding stopwords of NLTK[4]. Subsequently, we calculate cosine similarity between sentences/tweets and each topic using their vector representation. This feature is used only for English datasets because this feature is valid only for claims about US politics.

Handcrafted Word List (HW): We use handcrafted word list feature defined by Kartal et al. [14]. In this feature, firstly, 66 words which might be correlated to check-worthy claims are identified (e.g., unemployment). Then, we check whether there is an overlap between lemmas of selected words and lemmas of words in the respective sentence/tweet.

Part-of-speech (POS) Tags: Informative words can make a sentence/tweet more likely to be check-worthy. Thus, in this feature set, we use the number of nouns, verbs, adverbs and adjectives in order to catch information load of sentences/tweets.

Comparative & Superlative (CS): In this feature, we use the number of comparative and superlative adjectives and adverbs in sentences/tweets, as defined by Kartal et al. [14].

Tweet Meta Data (TMD): Meta data of tweets might be an indicator for check-worthy claims. For instance, if a tweet is retweeted a lot or shared by an influential people, it might be check-worthy because it reaches to many people and affect people. Specifically, in this feature group, we use the following information about tweets: 1) whether the account is a verified one, 2) whether the tweet is flagged as sensitive content, 3) whether the tweet is quoting another tweet, 4)

[3] http://qatsdemo.cloudapp.net/farasa/

[4] https://www.nltk.org/

presence of a URL, 5) presence of a hashtag, 6) whether a user is mentioned, 7) retweet counts, and 8) favorite counts.

3.2 Ranking Methodology

We use two different approaches to prioritize claims based on their check-worthiness using features defined above.

Logistic Regression (LR): LR is commonly used in state-of-the-art check-worthy detection models [14, 15]. Thus, we also train a LR model with features defined above. Then we rank claims based on their predicted probabilities of being check-worthy.

Hybrid In this model, we apply a hybrid approach combining logistic regression model and BERT model. We first fine tune BERT model as explained above and rank claims using the fine-tuned BERT model. We keep the rankings of the top 10 claims as they are, but re-rank the other claims using logistic regression with word embeddings and BERT features explained above. For Arabic Task 1, we use Ara-Bert as our BERT model.

4 Experiments

4.1 Implementation

We use ktrain[5] and huggingface transformers [21] to fine-tune BERT models with 1 cycle learning rate policy and maximum learning rate of 2e-5 [18]. We use SpaCy[6] for all syntactic and semantic analyses. We use Scikit toolkit[7] for the implementation of LR. We use default parameters for LR.

4.2 Experimental Setup

Our experiments are in two steps. We first evaluate different models using training datasets. Subsequently, we report results for our models participated in the shared task on the test data. In evaluation of different models with the training data, we use different cases for each task and language because the data formats and sizes are different for each of them. In particular, in Arabic Task 1 , we use 5-fold cross validation. In English Task 1, both training and validation data sets are provided in the development phase of the shared task. Thus, we use the same setup. In English Task 5, transcripts of 50 political debates and speeches are provided. Following the suggestion of the shared task organizers, we use the first 40 files (i.e., debates) as training and remaining 10 files for evaluating different models in the development phase of the shared task.

We evaluate the models with the following metrics: average precision (AP), precision@1 (P@1), precision@5 (P@5), precision@10 (P@10) and precision@30 (P@30). The official metrics are P@30 for Arabic and AP for English tasks.

[5] https://pypi.org/project/ktrain/
[6] https://spacy.io/
[7] https://scikit-learn.org

4.3 Experimental Results

Experiments on Training Data. We first compare performance of different models on Arabic training dataset using 5-fold cross validation. In particular, we use fine-tuned Multilingual BERT (M-BERT) [6], Ara-BERT, logistic regression with different combinations of BERT, WE and TMD features defined in Section 2, and our hybrid model.

The results are shown in Table 1. Our observations are as follows. Firstly, Ara-BERT outperforms M-BERT, showing superior performance of language specific models compared to multilingual models. Secondly, TMD features do not yield higher prediction accuracy. Lastly, hybrid model outperforms all other models based on all metrics. Thus, we choose hybrid model as our primary model for Arabic Task 1. We also choose the second best model which is LR with Ara-BERT and WE, as our contrastive submission (C1).

Table 1. Evaluation results for different models on the training data for Arabic Task 1 using 5-fold cross validation.

Model	AP	P@1	P@5	P@10	P@30
M-BERT	.690	.76	.87	.876	.838
Ara-BERT	.750	1	1	.986	.886
LR w/ TMD	.497	.600	.692	.639	.591
LR / WE	.720	.980	.934	.898	.874
LR w/ {BERT+TMD}	.708	.810	.870	.893	.827
LR w/ {BERT+WE}	.752	.980	.932	.916	.905
LR w/ {WE+TMD}	.700	.970	.880	.883	.852
LR w/ {BERT+WE+TMD}	.750	.970	.956	.942	.892
Hybrid	.762	1	1	.988	**.909**

Next, we compare different models using training data provided for English Task 1. In particular, we investigate performance of fined tuned BERT model, logistic regression with different sets of features defined in Section 2, and our hybrid model. In this set of experiments, we also use two different word embedding models, word2vec and fastText (FT) for WE features.

The results are shown in Table 2. Our observations based on the results are as follows. Firstly, word2vec yields higher AP scores than fastText in our logistic regression model (0.625 vs. 0.573). However, we observe the opposite case in our hybrid model such that fastText yields slightly higher results than word2vec (0.799 vs. 0.805). Secondly, using only BERT outperforms all models that do not use BERT. Thirdly, we achieve our best AP scores results when we use logistic regression with our BERT, POS, CT, and HW features together. Lastly, replacing HW with CS yields slightly lower AP (0.817 vs. 0.821) but higher P@30 (0.867 vs. 0.833). Based on these results, we choose logistic regression with BERT, POS, CT, and HW features, as our primary model.

Table 2. Evaluation results for different models on the training data for Task 1 English

Model	AP	P@1	P@5	P@10	P@30
BERT	.807	1	1	.800	.833
LR w/ word2vec	.625	1	.800	.600	.600
LR w/ fastText	.573	0	.400	.400	.600
LR w/ {BERT+fastText}	.797	1	1	.800	.867
LR w/ {POS+CT+CS+BERT}	.817	1	1	1	.867
LR w/ {POS+CT+HW+BERT}	.821	1	1	1	.833
Hybrid w/ word2vec	.799	1	1	.800	.833
Hybrid w/ fastText	.805	1	1	.800	.867

For Task 5, we investigate performance of fined tuned BERT model, logistic regression with different sets of features defined in Section 2, and our hybrid model. The results are shown in Table 3. The primary model for English Task 1 (i.e., LR with POS, CT, HW and BERT features) achieve the best P@30 scores while hybrid model (i.e., primary model for Arabic Task 1) is inferior to other models. Logistic regression with BERT and WE features achieve the best AP scores. Thus, we select this model as our primary model for Task 5.

Table 3. Evaluation results for different models on the training data for Task 5

Model	AP	P@1	P@5	P@10	P@30
BERT	0.101	0.0	0.1	0.11	0.067
LR w/ {BERT+WE}	0.124	0.2	0.1	0.1	0.06
LR w/ {POS+CT+HW+BERT}	0.113	0.1	0.12	0.09	0.07
Hybrid	0.096	0.0	0.1	0.11	0.057

Experiments on Test Data. We train our primary and contrastive models using training data provided in the development phase of the shared task. The results are shown in Table 4. In Arabic Task 1, our best run (C1) is ranked 2^{nd} among all best runs per team based on official metric P@30. Our primary model also shares the first rank with another group based on P@5 metric. Considering all runs submitted for Arabic Task 1, our contrastive and primary models are ranked 5^{th} and 7^{th} among 28 participants, respectively, based on P@30.

In English Task 1, our primary model is ranked 5^{th} among all primary models. However, our primary model and second contrastive model (C2) share the first rank with nine other models based on P@1 and P@5 metrics. Our second contrastive model actually outperforms our primary model and shares the first rank with five other models based on P@10.

In English Task 5, all our models unfortunately show poor performance on the test dataset. Our primary model ranked third among three primary models.

Table 4. AP, P@1, P@5, P@10 and P@30 scores of our primary and contrastive models on the test data for each task. Official metric results are written in bold. (P) indicates that the respective model is our primary model. (C1) and (C2) represent our first and second contrastive models

Task	Model	AP	P@1	P@5	P@10	P@30
Task 1 Arabic	**(P)** Hybrid	.589	-	.733	.683	**.636**
	(C1) LR w/ {BERT+WE}	.582	-	.700	.700	**.644**
Task 1 English	**(P)** LR w/ {POS+CT+HW+BERT}	**.706**	1	1	.900	.660
	(C1) Hybrid w/ fastText	**.564**	0	0	.300	.660
	(C2) LR w/ {POS+CT+CS+BERT}	**.710**	1	1	1	.680
Task 5 English	**(P)** LR w/ {BERT+WE}	**.018**	0	0	.300	.660
	(C1) LR w/ {POS+CT+HW+BERT}	**.042**	.050	.030	.015	.018

5 Conclusion

In this paper, we present our participation in Task 1 and Task 5 of CLEF-2020 Check That! Lab. We use three different models for Arabic Task 1, English Task 1, and Task 5 as our primary models. For Arabic Task 1, we propose a hybrid model which uses a fine-tuned BERT model for the top ten claims and then use logistic regression model with BERT and word embedding features to re-rank the remaining claims. For English Task 1, we rank claims using a logistic regression with features including domain-specific controversial topics, prediction of fined-tuned BERT model, a handcrafted word list, and POS tags. For English Task 5, we use logistic regression with BERT and word embedding features.

Our primary models for Arabic Task 1, English Task 1, and Task 5 ranked 3^{rd} (out of 8 groups), 5^{th} (out of 12 groups), and 3^{rd} (out of 3 groups), respectively, based on official evaluation metric of each task. Our models also share the first rank with other groups in Arabic Task 1 and English Task 1 based on various evaluation metrics.

We believe that misinformation is a global problem. Therefore, we plan to work on different languages and build a multilingual check-worthy claim detection model in the future. Furthermore, the limited number of annotated datasets is one of the main obstacles to develop effective systems. Thus, we also plan to explore weak supervision methods and develop deep learning models for this task.

References

1. R. Agez, C. Bosc, C. Lespagnol, N. Petitcol, and J. Mothe. IRIT at checkthat! 2018. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018*, 2018.
2. W. Antoun, F. Baly, and H. Hajj. Arabert: Transformer-based model for arabic language understanding. In *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, page 9.
3. P. Atanasova, A. Barron-Cedeno, T. Elsayed, R. Suwaileh, W. Zaghouani, S. Kyuchukov, G. D. S. Martino, and P. Nakov. Overview of the clef-2018 checkthat! lab on automatic identification and verification of political claims. task 1: Check-worthiness. *arXiv preprint arXiv:1808.05542*, 2018.
4. P. Atanasova, P. Nakov, G. Karadzhov, M. Mohtarami, and G. Da San Martino. Overview of the clef-2019 checkthat! lab on automatic identification and verification of claims. task 1: Check-worthiness. In *CEUR Workshop Proceedings*, 2019.
5. A. Barrón-Cedeño, T. Elsayed, P. Nakov, G. Da San Martino, M. Hasanain, R. Suwaileh, F. Haouari, N. Babulkov, B. Hamdan, A. Nikolov, S. Shaar, and Z. Sheikh Ali. Overview of CheckThat! 2020: Automatic identification and verification of claims in social media.
6. J. Devlin, M.-W. Chang, K. Lee, and 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)*, pages 4171–4186, 2019.
7. P. Gencheva, P. Nakov, L. Màrquez, A. Barrón-Cedeño, and 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 2017*, pages 267–276, 2017.
8. B. Ghanem, M. Montes-y-Gómez, F. M. R. Pardo, and P. Rosso. UPV-INAOE - check that: Preliminary approach for checking worthiness of claims. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018*, 2018.
9. E. Grave, P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.
10. C. Hansen, C. Hansen, J. G. Simonsen, and C. Lioma. The copenhagen team participation in the check-worthiness task of the competition of automatic identification and verification of claims in political debates of the clef-2018 checkthat! lab. In *CLEF*, 2018.
11. C. Hansen, C. Hansen, J. G. Simonsen, and C. Lioma. Neural weakly supervised fact check-worthiness detection with contrastive sampling-based ranking loss. In *Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9-12, 2019*, 2019.
12. N. Hassan, G. Zhang, F. Arslan, J. Caraballo, D. Jimenez, S. Gawsane, S. Hasan, M. Joseph, A. Kulkarni, A. K. Nayak, V. Sable, C. Li, and M. Tremayne. Claimbuster: The first-ever end-to-end fact-checking system. *PVLDB*, 10:1945–1948, 2017.
13. I. Jaradat, P. Gencheva, A. Barrón-Cedeño, L. Màrquez, and 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*, pages 26–30, 2018.

14. Y. S. Kartal, B. Guvenen, and M. Kutlu. Too many claims to fact-check: Prioritizing political claims based on check-worthiness. *ArXiv*, abs/2004.08166, 2020.
15. C. Lespagnol, J. Mothe, and M. Z. Ullah. Information nutritional label and word embedding to estimate information check-worthiness. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 941–944. ACM, 2019.
16. T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.
17. A. Patwari, D. Goldwasser, and S. Bagchi. Tathya: A multi-classifier system for detecting check-worthy statements in political debates. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 2259–2262. ACM, 2017.
18. L. N. Smith. A disciplined approach to neural network hyper-parameters: Part 1 - learning rate, batch size, momentum, and weight decay. *ArXiv*, abs/1803.09820, 2018.
19. S. Vasileva, P. Atanasova, L. Màrquez, A. Barrón-Cedeño, and P. Nakov. It takes nine to smell a rat: Neural multi-task learning for check-worthiness prediction. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, 2019.
20. S. Vosoughi, D. Roy, and S. Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018.
21. T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Brew. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771, 2019.
22. K. Yasser, M. Kutlu, and T. Elsayed. 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.
23. C. Zuo, A. Karakas, and R. Banerjee. A hybrid recognition system for check-worthy claims using heuristics and supervised learning. In *CLEF*, 2018.