SU-NLP at CheckThat! 2021: Check-Worthiness of Turkish Tweets

Buse Carik[1], Reyyan Yeniterzi[1]

[1]*Sabancı University, İstanbul, Turkey*

Abstract

The growth in social media usage increases the spread of misinformation on these platforms. In order to prevent this disinformation spread, automated fact checking systems that identify and verify claims are needed. The first step of such systems is the identification of whether a claim is worth-checking or not. This paper describes our participation to the check-worthiness task of CLEF 2021 CheckThat! 2021 Lab for Turkish tweets. We propose an ensemble of BERT models which ranked the second best in terms of MAP score.

Keywords

Check-Worthiness, Twitter, Turkish, BERT, Ensembles

1. Introduction

In today's digital era, social media platforms, like Twitter, became very popular among the community which increased the dissemination speed and scale of user shared content over these environments. Unfortunately, in these platforms, there is not any mechanism to check the correctness of these users' published content. These combined together cause a wide spread of misinformation, which can have serious consequences, like the one observed during the spread of COVID-19 [1].

Due to the amount of data accumulated every day, manual inspection of these claims is not possible. In order to address this issue, automatic fact checking systems should be used. Building an automatic identification and verification of claims system consists of several steps. First of all, the claims which are worth checking should be identified among all user posts. The second step is to see whether these identified claims have been fact-checked already or not. And the final phase is estimating the veracity of these claims.

CLEF CheckThat! Lab has been conducting shared tasks specifically focusing on these subtasks since 2018 [2, 3, 4]. The focus was on English and Arabic in previous years and this year [5] additional languages, like Turkish, Bulgarian and Spanish, were covered in the first subtask of the lab [6]. In this paper, we describe our approach for the Turkish check-worthiness estimation task.

In order to detect whether a Turkish tweet contains a claim worth checking, we fine-tune pretrained bidirectional encoders. Our analysis shows some variance therefore ensemble models

CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania

busecarik@sabanciuniv.edu (B. Carik); reyyan@sabanciuniv.edu (R. Yeniterzi)
0000-0002-4511-5827 (B. Carik); 0000-0002-8501-6209 (R. Yeniterzi)

are also explored. Our proposed BERT-Ensemble model ranked 2nd place in the 1A Turkish subtask. Our code is available at our Github repository[1].

2. Related Work

Automatically checking the worthiness of claims has been studied in the CLEF CheckThat! Lab in the past three years [2, 3, 4]. The CheckThat! 2020 Task 1, the identification of check-worthy claims, was carried out in English and Arabic languages [4]. The top-ranked team in the English subtask developed a model based on Roberta with additional layers [7]. The second-best team experimented with various static and contextual word embeddings using SVM and logistic regression classifier [8]. Yet they achieved the highest score with RoBERTa pre-trained model, along with several preprocessing steps. The four teams following the second-place utilized transformer-based models as well [9, 10, 11, 12]. The results on the Arabic subtask also illustrated the success of these models since the first and second-ranked teams in that task used the variation of BERT models, namely AraBERT and multilingual BERT [7, 13]. We also use BERT models in this paper as they consistently show superior performance compared to other models.

3. Task and Data Description

This year the CheckThat! Lab introduced check-worthiness estimation for Turkish. The objective of this subtask is to predict whether a provided Turkish tweet is worth to verify or not. For the training data, TrClaim-19 dataset [14] which consists of 2287 annotated Turkish tweets, was provided in two splits; training and validation. This data was collected in 2019 and contains tweets related to local events in Turkey such as local elections, earthquakes, and military operations in Syria. The data collection was labeled by 7 people (divided into 3 groups) with graded judgements. Later on, these annotations were mapped to binary values with majority voting. Among the 2287 tweets, 875 of them were labeled as worth checking while the remaining examples were annotated as not. Detailed statistics for the datasets are presented in Table 1.

For the testing, COVID-19 related 1013 tweets from 2020, which were annotated by 3 people, were used [6]. The provided training (TrClaim-19) and test (COVID-19) data were rather different than each other in terms of content, time of collection, and annotation procedure. This realistic setting made this task more challenging.

Table 1
Statistics of the Data Collection

	TrClaim-19		COVID-19
Label	Train Set	Validation Set	Test Set
Worth-Checking	729	146	-
Not Worth-Checking	1170	242	-
Total	1899	388	1013

[1] https://github.com/busecarik/Check-Worthiness-Estimation-in-Turkish-Social-Media

In our analysis of the training dataset, we noticed a possible duplication problem. 68 of the instances have at least one duplicate which makes 145 tweets in total. While there is a single copy of the 63 instances, five samples have more than one duplicate. Most of these duplicate tweets have the exact same content. In 24 of them, only the URL links at the end of the tweet are different while the rest of the content was the same. Among these 68 tweets, 15 of them had a label mismatch problem. Since these tweets were initially labeled by different groups on a graded scale, and then converted to binary format with majority voting, some duplicate tweets were mapped to different labels. For instance, the following tweet exists six times in the training set. Three of these samples were annotated as check-worthy, whereas the remaining three were labeled as not.

- Turkish: *@... Devlet borcunun milli gelire oranı dünya genelinde yüzde 80 civarında iken Türkiye'de bu oran yüzde 28 civarında. Bu oran 2002'de Türkiye'de yüzde 78'idi.*
- English: *@... While the ratio of government debt to national income is around 80 percent worldwide, this rate is around 28 percent in Turkey. This rate was 78 percent in Turkey in 2002.*

During model development, we keep the data as it is, mainly because we believe these instances are challenging cases due to ambiguity. We want our models to have that ambiguity and not be definite about these tweets being worthy or not. For test data, since all tweets were labeled by one group, there is not any label mismatch problem.

4. System Overview

In this section, we initially explain our data preprocessing steps and then describe our classification approach.

4.1. Data Pre-processing

The following preprocessing steps are applied to the tweets:

- All mentions are replaced with special token *@USER*
- All URLs are replaced with special token *URL*
- All emojis and hastags are removed

URLs can be connected to useful websites, like news websites, and this source information can be a useful feature to detect worthiness. Therefore, substituting all URLs with the same *URL* token may actually cause information loss. In order to prevent this and also to utilize more information about claims, we expand the shortened URLs and use the linked website's domain instead of the URL. For example, if a URL points to Cumhuriyet's (a newspaper in Turkey) website, the URL is replaced with the *CUMHURIYET* token. Similarly, for a URL which links to a tweet, the token *TWITTER* and the username of the tweet owner are used together to replace the URL. An example of the transformation of a tweet is illustrated below. The tweet "*It is seen in 8 percent in childhood, 6 percent in adolescence, and 4 percent in adults.*" has two URLs, one linked to the website of Cumhuriyet newspaper and the other a tweet that was shared by the official Twitter account of the Cumhuriyet newspaper. Hence, we replace these URLs with *CUMHURIYET* and *TWITTER cumhuriyetgzt* as shown.

- **Before Preprocessing:** *Çocuklukta yüzde 8, ergenlikte yüzde 6, yetişkinlerde ise yüzde 4 oranında görülüyor https://t.co/wwSocmei7t https://t.co/EhsMrLI1LI*
- **After Preprocessing:** *Çocuklukta yüzde 8, ergenlikte yüzde 6, yetişkinlerde ise yüzde 4 oranında görülüyor CUMHURIYET TWITTER cumhuriyetgzt*

4.2. Classification Models

Transformer-based models achieve superior performance in the check-worthiness estimation task for other languages [7, 8, 9, 10, 11, 12, 13]. Therefore, we also use the Bidirectional Encoder Representations (BERT) [15] model, which learns the contextual representation of a given input with masked language modeling. The following models are fine-tuned for the check-worthiness task using the provided training data.

- BERTurk[2]: BERT model pre-trained on Turkish Wikipedia dump, OSCAR[3], and OPUS[4] data sets.
- Loodos_BERT[5]: BERT model pre-trained on Turksh newspapers, e-books, online blogs, Twitter, and Wikipedia dump.
- Loodos_ALBERT[5]: pre-trained on the same dataset as Loodos_BERT.
- mBERT[6]: multilingual BERT model pre-trained on Wikipedia in 104 languages.
- XLM-RoBERTa[7]: this cross-lingual model was pretrained on the CommonCrawl corpus in 100 languages.

All BERT models above are base models, and all of them have 12 encoder layers with 768 hidden units per feed-forward layer.

Even though this is a classification task, the organizers use ranking metrics like Mean Average Precision, R-Precision, Precision@k etc., therefore, in addition to predicting whether the input is worth checking or not, we also use the returned probability to rank the tweets. Tweets which receive higher probability are ranked at top ranks.

5. Experiments

To observe the effects of expanded URLS in preprocessing, we compare two different preprocessing approaches on the BERTurk model. For each technique, five BERTurk models were fine-tuned with the same hyper-parameters using the training set. On average 0.662 MAP (Mean

[2] https://huggingface.co/dbmdz/bert-base-turkish-cased/
[3] https://oscar-corpus.com/
[4] https://opus.nlpl.eu/
[5] https://github.com/Loodos/turkish-language-models
[6] https://huggingface.co/bert-base-multilingual-cased
[7] https://huggingface.co/xlm-roberta-base

Table 2
Results of Different Transformer Models on the Validation Set

Models	MAP	RP	macro-F1
BERTurk	**0.693**	0.631	**0.672**
Loodos_BERT	0.682	**0.680**	0.669
Loodos_ALBERT	0.508	0.479	0.554
mBERT	0.640	0.618	0.627
XLM_RoBERTa	0.524	0.493	0.430

Average Precision) score is received when data was preprocessed without the URL expanding. Using URL expanding in preprocessing returns 0.684 MAP score when averaged over five models. This increase indicates that the domain name of the URL in the tweets provides useful information while identifying whether tweets are worth checking or not. This URL expansion step is used in the rest of the experiments.

5.1. Experiments with Transformers

We experimented with different transformer-based models based on the variation of the transformer model and the language of the corpus used in the pretraining. Three Turkish models, namely BERTurk, Loodos_BERT, and Loodos_ALBERT are fine-tuned using the given training data. Moreover, two multilingual models, multilingual BERT and XLM-RoBERTa are tried. The results for these models are shown in Table 2.

In Table 2, MAP stands for Mean Average Precision and RP stands for R-Precision. Even though it is not reported officially, we also use macro-F1 score to evaluate our models. The multilingual models (mBERT and XML_RoBERTa) show poor performance compared to models pretrained on Turkish, which indicates the importance of the language in the training phase of the model.

Among the Turkish BERT models, BERTurk achieved the highest MAP score, although the dataset size used in BERTurk pre-training is smaller than the data used to train Loodos_BERT model. Although both Loodos_BERT and Loodos_ALBERT were pre-trained with the same corpus, the BERT variant achieved significantly higher results compared to the ALBERT one. The difference between the performance of these two models is not surprising since the BERT version also performed better than ALBERT in other tasks such as Sentiment Analysis and NER[8].

5.2. The Ensemble Model

In our experiments on the validation set, the BERTurk model trained with the same hyperparameters showed remarkable differences for different seed values. Results obtained with 10 different seed values are presented in Table 3. The blue ones are the highest values obtained, while the red ones are the lowest.

Among the ten seeds, the highest MAP score achieved is 0.7093, while the lowest is 0.5933. The mean and the standard deviation of these ten models are 0.66 and 0.035, respectively. Since

[8] https://github.com/Loodos/turkish-language-models

Table 3
Results of BERTurk with Different Seed Values

	Seed									
	100	101	102	103	104	105	106	107	108	109
MAP	0.6720	0.6230	0.6658	0.6389	0.5933	0.6763	0.6493	0.6934	0.7093	0.6859
RP	0.6597	0.6111	0.6319	0.6041	0.5486	0.6458	0.611	0.6527	0.6458	0.6597
macro-F1	0.6813	0.6242	0.6224	0.6431	0.6344	0.6376	0.6867	0.6036	0.6441	0.6817

Table 4
Results of the Ensemble Model on Validation Set

Model	MAP	RP	macro-F1
BERTurk_1	0.7081	**0.6805**	**0.7445**
BERTurk_2	0.7074	0.6597	0.7141
BERTurk_3	0.7034	0.6736	0.7434
BERTurk_4	0.6996	0.6458	0.7033
BERTurk Ensemble	**0.7176**	0.6666	0.7342

a 0.116 difference due to random initialization is a bit unusual, the model's confidence on its predictions was examined carefully. According to our analysis over the validation set, almost half of the probability estimates generated by the model are squeezed between 0.4 and 0.6. For instance, for the following tweet,

- *"We Increased Our Hospital Bed Capacity to 240 thousand. We Increased The Number Of Our Doctors To 231 thousand, Mr. Kemal, Do You Know About These? Recep Tayyip Erdogan"*

the model with the lowest MAP score predicts not worth checking with 0.4904, while the model with the highest score predicts worth checking with 0.5366. These results show how hesitant our model is about its decisions. Hence, we create an ensemble of four models with the highest MAP scores obtained so far, in order to reduce the variance caused by the random initialization. The probabilities of these four best models are averaged in our final ensemble model. The individual performance of these four models together with the ensemble model are presented in Table 4.

As we illustrated in Table 4, the ensemble model achieved the highest MAP score compared to all individual BERTurk models. RP and macro-F1 scores of the ensemble are slightly lower than the first model. This is expected since these four models are chosen due to their high MAP performance. Since MAP is the official metric, all optimizations are performed with MAP.

The official results on the official test set are shown in Table 5. The results of two models are shown here. In addition to the ensemble model, the BERTurk_1 model from Table 5 is also shown as BERTurk. The ensemble model outperforms the individual BERTurk model here as well, this time not only in terms of MAP but also for RP and other metrics as well. The result indicates that the ensembling strategy reduces the variance caused by the random initialization and increases the performance where the model is hesitant. We could not perform further analysis on test data since we do not have gold standard labels.

Table 5
Official Results on the Test Set

Model	MAP	RP	RR	P@3	P@5	P@10
BERTurk	0.565	0.568	1.0	0.667	0.8	0.7
BERTurk Ensemble	**0.574**	0.585	1.0	1.0	1.0	0.8

6. Conclusion

In this paper, we describe our proposed system for predicting whether a tweet is worth checking or not. We use a better pre-processing approach and also experiment with ensemble models. We specifically focus on Turkish and our model ranked the second-best with 0.574 MAP score at the CLEF CheckThat! 2021 Lab. As future work, we will work with other languages to see whether the proposed preprocessing and ensembling strategy work in general for them as well.

References

[1] S. Loomba, A. de Figueiredo, S. J. Piatek, K. de Graaf, H. J. Larson, Measuring the impact of covid-19 vaccine misinformation on vaccination intent in the UK and USA, Nature human behaviour 5 (2021) 337–348.

[2] P. Nakov, A. Barrón-Cedeno, T. Elsayed, R. Suwaileh, L. Màrquez, W. Zaghouani, P. Atanasova, S. Kyuchukov, G. Da San Martino, Overview of the clef-2018 checkthat! lab on automatic identification and verification of political claims, in: International conference of the cross-language evaluation forum for european languages, Springer, 2018, pp. 372–387.

[3] P. Atanasova, P. Nakov, G. Karadzhov, M. Mohtarami, G. Da San Martino, Overview of the clef-2019 checkthat! lab: Automatic identification and verification of claims. task 1: Check-worthiness., in: CLEF (Working Notes), 2019.

[4] 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: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2020, pp. 215–236.

[5] P. Nakov, G. Da San Martino, T. Elsayed, A. Barrón-Cedeño, R. Míguez, S. Shaar, F. Alam, F. Haouari, M. Hasanain, N. Babulkov, A. Nikolov, G. K. Shahi, J. M. Struß, T. Mandl, The CLEF-2021 CheckThat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news, in: Proceedings of the 43rd European Conference on Information Retrieval, ECIR '21, Lucca, Italy, 2021, pp. 639–649. URL: https://link.springer.com/chapter/10.1007/978-3-030-72240-1_75.

[6] 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.

[7] E. Williams, P. Rodrigues, V. Novak, Accenture at checkthat! 2020: If you say so: Post-hoc fact-checking of claims using transformer-based models, arXiv preprint arXiv:2009.02431 (2020).

[8] A. Nikolov, G. D. S. Martino, I. Koychev, P. Nakov, Team alex at clef checkthat! 2020: Identifying check-worthy tweets with transformer models, arXiv preprint arXiv:2009.02931 (2020).

[9] G. S. Cheema, S. Hakimov, R. Ewerth, Check_square at checkthat! 2020: Claim detection in social media via fusion of transformer and syntactic features, arXiv preprint arXiv:2007.10534 (2020).

[10] 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, arXiv preprint arXiv:2008.13160 (2020).

[11] Y. S. Kartal, M. Kutlu, Tobb etu at checkthat! 2020: Prioritizing english and arabic claims based on check-worthiness, Cappellato et al.[10] (2020).

[12] S. Shaar, A. Nikolov, N. Babulkov, F. Alam, A. Barrón-Cedeno, T. Elsayed, M. Hasanain, R. Suwaileh, F. Haouari, G. Da San Martino, et al., Overview of checkthat! 2020 english: Automatic identification and verification of claims in social media, Cappellato et al.[10] (2020).

[13] M. Hasanain, T. Elsayed, bigir at checkthat! 2020: Multilingual bert for ranking arabic tweets by check-worthiness, Cappellato et al.[10] (2020).

[14] 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.

[15] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).