UPV at CheckThat! 2021: Mitigating Cultural Differences for Identifying Multilingual Check-worthy Claims

Ipek Baris Schlicht, Angel Felipe Magnossão de Paula and Paolo Rosso

Universitat Politècnica de València, Spain

Abstract

Identifying check-worthy claims is often the first step of automated fact-checking systems. Tackling this task in a multilingual setting has been understudied. Encoding inputs with multilingual text representations could be one approach to solve the multilingual check-worthiness detection. However, this approach could suffer if cultural bias exists within the communities on determining what is check-worthy. In this paper, we propose a language identification task as an auxiliary task to mitigate unintended bias. With this purpose, we experiment joint training by using the datasets from CLEF-2021 CheckThat!, that contain tweets in English, Arabic, Bulgarian, Spanish and Turkish. Our results show that joint training of language identification and check-worthy claim detection tasks can provide performance gains for some of the selected languages.

Keywords

Check-worthy Claim Detection, Language Identification, Sentence Transformers, Multilingual, Joint Training, Bias

1. Introduction

The number of fact-checking initiatives worldwide has increased to fight misinformation. Manual fact-checking is a labor-intensive and time-consuming task that cannot cope up with the dissemination of misinformation [1]. Therefore, the automation of fact-checking steps is required to speed up the process.

Check-worthy claim detection is a crucial step of an automated fact-checking pipeline [2, 1, 3] to prioritize what is needed to be fact-checked by fact-checkers or journalists. There has been an ongoing effort to address the claim-detection task by different research communities. Prior studies rely on machine learning methods that use statistical features with bag of words [4, 5, 6]. Additionally, CLEF CheckThat! Lab (CTL) has organized shared tasks to tackle this problem in political debates [7, 8] and social media [9]. This year, CTL 2021 [10] organized the shared task in English, Turkish, Bulgarian, Spanish and Arabic where the task datasets are collected from social media [11]. The task's input is a tweet and the output is a score indicating the check-worthiness of the tweet.

Multilingual language models have been widely used in natural language understanding tasks with low-resourced languages (e.g. comment moderation [12], fake news detection [13]).

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

ibarsch@doctor.upv.es (I. B. Schlicht); adepau@doctor.upv.es (A. F. M. d. Paula); prosso@dsic.upv.es (P. Rosso)

© 2021 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)

CEUR-WS.org/Vol-2936/paper-36.pdf

However, the exhibition of cultural differences is inevitable in tasks in which cultural context is required [14]. This issue could harm the transfer of knowledge across languages. Fact-checking is one of such tasks where disagreements could exist on credibility assessments even among the domain experts [15, 16]. Furthermore, exposure of global claims and their credibility (e.g. Covid-19) could vary by country [17].

With this motivation, in this paper, we present a unified framework that processes the input in different languages and uses a multilingual sentence transformer trained on the mixed language training set to learn representations for the low-resourced languages. To mitigate the bias in the sentence representations, we introduce a language identification task and train the model jointly for check-worthiness detection (CWD) and language identification (LI) tasks.

Our contributions can be summarized as follows:

1. We introduce a framework whose aim is to be aware of cultural bias. We conduct an extensive analysis on its performance.
2. We employ joint learning to reduce unintended bias. To the best of our knowledge, a similar method has not been applied to reduce bias in multilingual fact-checking tasks.
3. Our framework could be extended with various multilingual transformer models in Huggingface [18]. The source code and the trained models are publicly available [1].

2. Related Work

ClaimBuster is the first study to address the check-worthy claim detection task. The component of ClaimBuster [4, 5] that detects check-worthy claims is trained with a Support Vector Machine (SVM) classifier using tf-idf bag of words, named entity types, POS tags, sentiment, and sentence length as a feature set. [6] proposed a fully connected neural network model trained on claims and their related political debate content. Last year, CTL 2020 [9] organized a shared CWD task in English and Arabic for the claims in social media. In this shared task, multilingual transformer models performed well on the Arabic dataset [19]. However, for the English datasets, the participants did not utilize the multilingual transformer model. In our approach, we fine-tune the multilingual sentence transformers [20], which is computationally less expensive than the BERT models, on the mixed language of the training dataset. We trained one model and employed this for all languages.

The multi-task learning approach has been a proven method to mitigate unintended bias. Das et al. [21] applied multi-task learning on a face recognition task using Convolution Neural Network. As a related example in the Natural Language Processing (NLP) domain, Vaidya et al. [22] mitigate the identity bias in toxic comment detection. Their model encodes the inputs with a Bidirectional Long Short-Term Memory Network (BiLSTM). However, our approach and the tasks we deal with are different from those studies.

[1] https://github.com/isspek/Cross_Lingual_Checkworthy_Detection

3. Methodology

In this section, we introduce our framework, which is depicted in Figure 1. The input of the framework is a Twitter post. The input is tokenized with a sentence transformer encoder in order to be fed into the transformer layer. After obtaining the shared text representation from a sentence transformer, the framework fine-tunes the shared representation and the classification layers for CWD and LI tasks by minimizing a joint loss. In the following subsections, we give more details about the sentence transformer and the joint training.

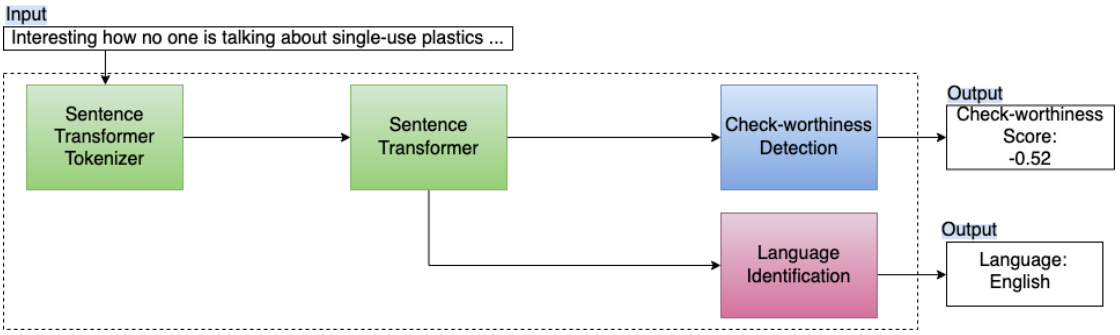

Figure 1: Our proposed framework (QDMSBERT$_{joint}$) for mitigating unintended bias.

3.1. Sentence Transformer

The framework uses a Sentence-BERT (SBERT) transformer [23], which is a modified BERT that uses a siamese and a triplet network. The SBERT can provide semantically more meaningful sentence embeddings than the BERT models. To support multilingualism in our framework and to enable fine-tuning with a small GPU, we use a pre-trained SBERT that was obtained by applying knowledge distillation [20] and that was trained on a multilingual corpus from a community-driven Q&A website[2]. We refer to it as QDMSBERT.

We apply mean pooling on the output of QDMSBERT to obtain sentence embeddings. We set the maximum length of the tokens as 128 by padding shorter texts and truncating longer texts.

3.2. Joint Learning

The framework contains two task layers: one is for the CWD task and the other is for the LI task. The input of the task layers are shared QDMSBERT embeddings. Both task layers use the same neural network structure, consisting of two fully-connected layers followed by a softmax layer that outputs the probabilities of task classes. During the training, the weighted loss of the CWD and the LI task are summed up to compute the joint loss as seen in Equation 1 where α is a probability indicating the importance of the tasks. Lastly, the joined loss is minimized by optimizing the weights of the transformer network and the tasks' classification layers.

$$J_{joint} = \alpha J_{CWD} + (1 - \alpha) J_{LI} \tag{1}$$

[2]https://www.quora.com/

Table 1
Topic, class distribution and average tokens in the CheckThat! dataset. Pos-Class means check-worthy, Neg-Class means not check-worthy.

Properties	English	Turkish	Bulgarian	Arabic	Spanish
Topic	Covid-19	Miscellaneous	Covid-19	Miscellaneous	Politics
Pos-Class (Train)	290	729	392	921	200
Neg-Class (Train)	532	1170	2608	2798	2295
Pos-Class (Dev)	60	146	62	107	109
Neg-Class (Dev)	80	242	288	279	1138
Pos-Class (Test)	19	183	76	242	120
Neg-Class (Test)	331	830	281	358	1128
Avg. Tokens (Train)	31.69	19.11	20.27	27.85	36.73
Avg. Tokens (Dev)	34.71	18.22	16.66	36.68	36.19
Avg. Tokens (Test)	35.33	23.72	17.02	23.47	36.21

4. Experiments

In this section, we give the details of the CLEF 2021 CheckThat! dataset, explain the baselines and the systems that we compared, and present the experimental settings.

4.1. Dataset

The CLEF 2021 CheckThat! offers datasets in English, Spanish, Arabic, Turkish, and Bulgarian for the CWD task. The statistics of the datasets are given in Table 1. The class distribution of the datasets for each language is highly imbalanced, which reflects the real-world cases. Check-worthy samples (Pos-Class) are the minority. English and Bulgarian datasets contain only COVID topic. The Turkish dataset covers miscellaneous topics, and the Spanish dataset has only samples about politics. The topics of the Arabic dataset are mainly COVID related.

4.2. Baselines

We compare the proposed model (QDMSBERT$_{joint}$) against the following models and systems:

- **SVM:** It encodes the texts with unigrams.
- **Monolingual Models and Mk-Bg-BERT:** We use a distilled SBERT [23] model[3] for the English samples. We couldn't find any monolingual SBERTs for Arabic, Turkish and Spanish; therefore, we use popular BERT [24] variants that are trained on monolingual corpora. TrBERT [4] is the model for Turkish samples, BETO [25] for Spanish, and lastly AraBERT [26] for the tweets in Arabic. For Bulgarian tweets, we leverage a BERT model (Mk-Bg-BERT) trained on Macedonian and Bulgarian corpora [5].

[3]https://huggingface.co/sentence-transformers/distilbert-base-nli-stsb-mean-tokens
[4]https://huggingface.co/dbmdz/bert-base-turkish-cased
[5]https://huggingface.co/anon-submission-mk/bert-base-macedonian-bulgarian-cased

- **CLEF-2021:** Submissions for the CLEF-2021 CWD task [11] that support all languages, namely Accenture, BigIR and TOBB ETU[6].
- **QDMSBERT:** QDMSBERT$_{joint}$ where the weights are only optimized for the CWD task.

4.3. Experimental Settings and Environment

We split the training dataset randomly into five chunks and thus train five different QDMSBERT models with the epochs of 3, weighted decay Adam optimizer [27], and in batches of 16. The mean of each model's predictions represents the final score. We use the GPU of Google Colab[7] for training the models.

5. Results

Table 2 presents the results of each model. We report the test results in official metrics of the shared task: Mean Average Precision (MAP), precision scores at 1-50 (P@1-P@50), R-Precision (R-Prec), and R-Rank. We first compare QDMSBERT$_{joint}$ with the SVM and QDMSBERT. QDMSBERT$_{joint}$ outperforms QDMSBERT in many metrics across the languages except for Arabic, also QDMSBERT$_{joint}$ underperforms the SVM in Spanish. We see performance gains on the English, Bulgarian and Turkish samples. The results indicate that QDMSBERT$_{joint}$ presents better results on the examples in COVID-19, but is generalized less to other topics in Spanish and in Arabic.

Among the results by the teams who submitted runs in all languages (group CLEF-2021), the performance of QDMSBERT$_{joint}$ is the best in English and the second in Bulgarian which is promising for a low-resource language.

Monolingual BERT models outperformed our model and the other teams' submissions in English and Spanish. TrBERT and AraBERT also show better results than our approach. Although we improve our outcome compared to QDMSBERT by mitigating differences across the languages, the performance of the monolingual embeddings is still unsurpassed in this task.

The presented results of QDMSBERT$_{joint}$ were accomplished by using a contribution of task loss (α) of 0.6. This initial value was choose heuristically. As an ablation study, we change the α values of the tasks' loss and train QDMSBERT$_{joint}$ for each α value to understand its influence on the CWD learning. The optimal alpha value is 0.8. In Bulgarian samples, the lower alpha values could also yield good performance for the CWD task.

Lastly, we analyze the feature representations of QDMSBERT and QDMSBERT$_{joint}$. We visualize the feature representations by applying t-distributed stochastic neighbor embedding nonlinear dimensionality reduction (T-SNE) [28]. As depicted in Figure 2, the features that QDMSBERT$_{joint}$ produces are more clearly separated. For instance, the cluster with English samples (lower right region of Figure 2a) in the T-SNE for QDMSBERT overlaps with both the cluster of Turkish and the cluster of Bulgarian samples. In contrast, the T-SNE for QDMSBERT$_{joint}$

[6]At the time of the writing the paper, we didn't know the system descriptions of their models
[7]https://colab.research.google.com/

Table 2
The results of the models on the test set. Our submission is **QDMSBERT**$_{joint}$.

Language	Models	MAP	R-Rank	R-Pr	P@1	P@3	P@5	P@10	P@20	P@50
English	SVM	0.052	0.020	0.000	0.000	0.000	0.000	0.000	0.000	0.020
	SBERT	**0.198**	**1.000**	**0.211**	**1.000**	**0.333**	**0.200**	**0.300**	**0.200**	**0.160**
	Accenture	0.101	0.143	0.158	0.000	0.000	0.000	0.200	0.200	0.100
	BigIR	0.136	0.500	0.105	0.000	0.333	0.200	0.100	0.100	0.120
	TOBB ETU	0.081	0.077	0.053	0.000	0.000	0.000	0.000	0.050	0.080
	QDMSBERT	0.114	0.500	0.105	0.00	0.333	0.200	0.100	0.100	0.100
	QDMSBERT$_{joint}$	0.149	**1.000**	0.105	**1.000**	0.333	0.200	0.200	0.100	0.120
Turkish	SVM	0.354	**1.000**	0.311	**1.000**	0.667	0.600	0.700	0.600	0.460
	TrBERT	0.563	**1.000**	0.530	**1.000**	**1.000**	**1.000**	0.800	0.850	**0.780**
	Accenture	0.402	0.250	0.415	0.000	0.000	0.400	0.400	0.650	0.660
	BigIR	0.525	**1.000**	0.503	**1.000**	**1.000**	**1.000**	0.800	0.700	0.720
	TOBB ETU	**0.581**	**1.000**	**0.585**	**1.000**	**1.000**	0.800	0.700	0.750	0.660
	QDMSBERT	0.549	**1.000**	0.579	**1.000**	0.333	0.600	0.700	0.650	0.680
	QDMSBERT$_{joint}$	0.517	**1.000**	0.508	**1.000**	**1.000**	**1.000**	**1.000**	**0.850**	0.700
Bulgarian	SVM	0.588	**1.000**	0.474	**1.000**	**1.000**	**1.000**	0.900	0.750	0.640
	Mk-Bg-BERT	0.661	**1.000**	0.645	**1.000**	**1.000**	**1.000**	0.900	0.700	0.700
	Accenture	0.497	**1.000**	0.474	**1.000**	**1.000**	0.800	0.700	0.600	0.440
	BigIR	**0.737**	**1.000**	0.632	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**0.800**
	TOBB ETU	0.149	0.143	0.039	0.000	0.000	0.000	0.200	0.100	0.060
	QDMSBERT	0.667	**1.000**	0.566	**1.000**	**1.000**	**1.000**	**1.000**	0.900	0.720
	QDMSBERT$_{joint}$	0.673	**1.000**	0.605	**1.000**	**1.000**	**1.000**	**1.000**	0.800	0.700
Arabic	SVM	0.428	0.500	0.409	0.000	0.667	0.600	0.500	0.450	0.440
	AraBERT	0.640	**1.000**	0.591	**1.000**	**1.000**	0.600	0.800	0.750	0.760
	Accenture	**0.658**	**1.000**	0.599	**1.000**	**1.000**	**1.000**	**1.000**	**0.950**	**0.840**
	BigIR	0.615	0.500	0.579	0.000	0.667	0.600	0.600	0.800	0.740
	TOBB ETU	0.575	0.333	0.574	0.000	0.333	0.400	0.400	0.500	0.680
	QDMSBERT	0.571	**1.000**	0.579	0.000	0.667	0.600	0.600	0.550	0.580
	QDMSBERT$_{joint}$	0.548	**1.000**	0.550	**1.000**	0.667	0.600	0.500	0.400	0.580
Spanish	SVM	0.450	**1.000**	0.450	**1.000**	0.667	**0.800**	0.700	0.700	0.660
	BETO	**0.569**	**1.000**	**0.533**	**1.000**	0.667	**0.800**	0.800	**0.750**	**0.720**
	Accenture	0.491	**1.000**	0.508	**1.000**	0.667	**0.800**	**0.900**	0.700	0.620
	BigIR	0.496	**1.000**	0.483	**1.000**	**1.000**	**0.800**	0.800	0.600	0.620
	TOBB ETU	0.537	**1.000**	0.525	**1.000**	**1.000**	**0.800**	0.900	0.700	0.680
	QDMSBERT	0.398	0.500	0.425	0.000	0.333	0.600	0.600	0.500	0.580
	QDMSBERT$_{joint}$	0.446	0.333	0.475	0.000	0.333	0.600	0.800	0.650	0.580

shows that only very few non-English samples fall close to the English cluster (upper region of Figure 2b).

Table 3
The performance of QDMSBERT$_{joint}$ under the different α values

Language	α	MAP	R-Rank	R-Pr	P@1	P@3	P@5	P@10	P@20	P@50
English	**0.3**	0.143	**1.000**	0.105	**1.000**	**0.333**	0.200	**0.200**	0.100	0.080
	0.4	0.145	**1.000**	0.105	**1.000**	**0.333**	0.200	**0.200**	0.100	0.080
	0.5	0.151	**1.000**	0.105	**1.000**	**0.333**	**0.400**	**0.200**	0.100	0.080
	0.6	0.149	**1.000**	0.105	**1.000**	**0.333**	0.200	**0.200**	0.100	0.120
	0.7	0.123	0.500	0.105	0.00	**0.333**	0.200	**0.200**	0.100	0.120
	0.8	**0.155**	**1.000**	**0.158**	**1.000**	**0.333**	0.200	**0.200**	**0.150**	**0.120**
	0.9	0.144	**1.000**	0.105	**1.000**	**0.333**	0.200	0.100	**0.150**	**0.120**
Turkish	**0.3**	0.520	**1.000**	0.492	**1.000**	**1.000**	**1.000**	**1.000**	0.900	0.660
	0.4	0.531	**1.000**	0.481	**1.000**	**1.000**	**1.000**	**1.000**	0.950	0.740
	0.5	0.534	**1.000**	0.492	**1.000**	**1.000**	**1.000**	**1.000**	0.950	0.740
	0.6	0.517	**1.000**	0.508	**1.000**	**1.000**	**1.000**	**1.000**	0.850	0.700
	0.7	0.528	**1.000**	0.497	**1.000**	**1.000**	0.200	0.200	0.850	0.680
	0.8	**0.588**	**1.000**	0.563	**1.000**	**1.000**	**1.000**	**1.000**	**0.950**	**0.780**
	0.9	0.582	**1.000**	**0.568**	**1.000**	**1.000**	**1.000**	**1.000**	0.850	0.740
Bulgarian	**0.3**	0.657	**1.000**	0.618	**1.000**	**1.000**	**1.000**	0.900	0.800	**0.720**
	0.4	0.666	**1.000**	0.618	**1.000**	**1.000**	**1.000**	0.900	0.800	0.700
	0.5	0.670	**1.000**	0.618	**1.000**	**1.000**	**1.000**	0.900	0.850	**0.720**
	0.6	0.673	**1.000**	0.605	**1.000**	**1.000**	**1.000**	**1.000**	0.800	0.700
	0.7	**0.677**	**1.000**	0.618	**1.000**	**1.000**	**1.000**	**1.000**	**0.850**	0.700
	0.8	0.670	**1.000**	0.592	**1.000**	**1.000**	**1.000**	**1.000**	0.800	**0.720**
	0.9	**0.677**	**1.000**	0.579	**1.000**	**1.000**	**1.000**	0.900	**0.850**	0.700
Arabic	**0.3**	0.562	**1.000**	0.558	**1.000**	0.333	0.400	**0.500**	0.400	**0.680**
	0.4	0.561	**1.000**	0.562	**1.000**	**0.667**	0.400	0.400	0.350	0.640
	0.5	0.567	**1.000**	0.562	**1.000**	**0.667**	0.400	0.400	0.400	0.660
	0.6	0.548	**1.000**	0.550	**1.000**	**0.667**	**0.600**	**0.500**	0.400	0.580
	0.7	0.561	**1.000**	0.566	**1.000**	**0.667**	0.400	**0.500**	0.400	0.620
	0.8	0.566	**1.000**	0.566	**1.000**	**0.667**	0.400	0.400	0.450	0.580
	0.9	**0.573**	**1.000**	**0.574**	**1.000**	**0.667**	0.400	**0.500**	**0.500**	0.580
Spanish	**0.3**	0.450	**0.333**	0.458	0.00	**0.333**	**0.600**	0.700	**0.750**	0.580
	0.4	0.453	**0.333**	0.475	0.00	**0.333**	**0.600**	0.700	**0.750**	0.580
	0.5	**0.456**	**0.333**	0.472	0.00	**0.333**	**0.600**	0.700	**0.750**	**0.640**
	0.6	0.446	**0.333**	0.475	0.00	**0.333**	**0.600**	**0.800**	0.650	0.580
	0.7	0.443	**0.333**	**0.483**	0.00	**0.333**	0.400	**0.600**	0.600	0.580
	0.8	0.443	**0.333**	0.475	0.00	**0.333**	0.400	0.500	0.650	0.580
	0.9	0.431	0.250	0.467	0.00	0.00	0.400	0.500	0.700	0.580

6. Conclusion

In this paper, we proposed a method to tackle the multilingual check-worthiness of claims. To mitigate bias due to cultural differences, we leveraged multilingual sentence BERTs as feature representations and trained them jointly with the language identification task. Our approach outperformed the SVM and QDMSBERT for almost all of the languages on the CLEF2021 dataset.

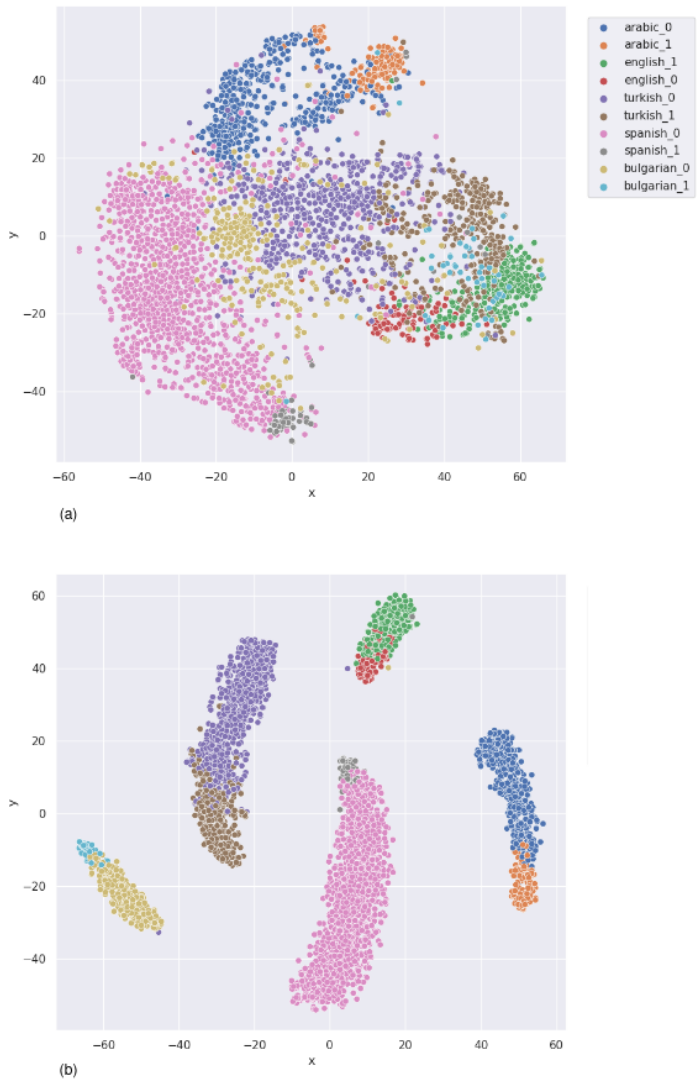

Figure 2: T-SNE visualization of QDMSBERT (a) and QDMSBERT$_{joint}$ (b). 0: not check-worthy, 1: check-worthy

Also, it became one of the top-performing approaches in Bulgarian and English datasets among the submissions that have been done for all these languages. In the future, we will investigate how the consideration of images [29] that are embedded in the tweets influence the results.

Acknowledgement

The work of P. Rosso was partially funded by the Spanish Ministry of Science and Innovation under the research project MISMIS-FAKEnHATE on MISinformation and MIScommunication in social media: FAKE news and HATE speech (PGC2018-096212-B-C31).

References

[1] L. Graves, Understanding the promise and limits of automated fact-checking, Factsheet 2 (2018) 2018–02.

[2] S. Cazalens, P. Lamarre, J. Leblay, I. Manolescu, X. Tannier, A content management perspective on fact-checking, in: WWW (Companion Volume), ACM, 2018, pp. 565–574.

[3] J. Thorne, A. Vlachos, Automated fact checking: Task formulations, methods and future directions, in: COLING, Association for Computational Linguistics, 2018, pp. 3346–3359.

[4] 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, 2015, pp. 1835–1838.

[5] 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, M. Tremayne, Claimbuster: The first-ever end-to-end fact-checking system, Proc. VLDB Endow. 10 (2017) 1945–1948.

[6] 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 2017, 2017, pp. 267–276.

[7] P. Atanasova, L. Màrquez, A. Barrón-Cedeño, T. Elsayed, R. Suwaileh, W. Zaghouani, S. Kyuchukov, G. D. S. Martino, P. Nakov, Overview of the CLEF-2018 checkthat! lab on automatic identification and verification of political claims. task 1: Check-worthiness, in: CLEF (Working Notes), volume 2125 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2018.

[8] P. Atanasova, P. Nakov, G. Karadzhov, M. Mohtarami, G. D. S. Martino, Overview of the CLEF-2019 checkthat! lab: Automatic identification and verification of claims. task 1: Check-worthiness, in: CLEF (Working Notes), volume 2380 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2019.

[9] A. Barrón-Cedeño, T. Elsayed, P. Nakov, G. D. S. Martino, M. Hasanain, R. Suwaileh, F. Haouari, N. Babulkov, B. Hamdan, A. Nikolov, S. Shaar, Z. S. Ali, Overview of checkthat! 2020: Automatic identification and verification of claims in social media, in: CLEF, volume 12260 of *Lecture Notes in Computer Science*, Springer, 2020, pp. 215–236.

[10] 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.

[11] 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.

[12] D. Korenčić, I. Baris, E. Fernandez, K. Leuschel, E. Salido, To block or not to block: Experiments with machine learning for news comment moderation, in: Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation, 2021, pp. 127–133.

[13] M. Z. Hossain, M. A. Rahman, M. S. Islam, S. Kar, BanFakeNews: A Dataset for Detecting Fake News in Bangla, in: Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020), European Language Resources Association (ELRA), 2020.

[14] B. Y. Lin, F. F. Xu, K. Q. Zhu, S. Hwang, Mining cross-cultural differences and similarities in social media, in: ACL (1), Association for Computational Linguistics, 2018, pp. 709–719.

[15] M. Mensio, H. Alani, News source credibility in the eyes of different assessors, in: TTO, 2019.

[16] D. Bountouridis, M. Makhortykh, E. Sullivan, J. Harambam, N. Tintarev, C. Hauff, Annotating credibility: Identifying and mitigating bias in credibility datasets (2019).

[17] 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, arXiv preprint arXiv:2104.10864 (2021).

[18] 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. L. Scao, S. Gugger, M. Drame, Q. Lhoest, A. M. Rush, Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 38–45. URL: https://www.aclweb.org/anthology/2020.emnlp-demos.6.

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

[20] N. Reimers, I. Gurevych, Making monolingual sentence embeddings multilingual using knowledge distillation, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2020.

[21] A. Das, A. Dantcheva, F. Bremond, Mitigating bias in gender, age and ethnicity classification: a multi-task convolution neural network approach, in: Proceedings of the European Conference on Computer Vision (ECCV) Workshops, 2018, pp. 0–0.

[22] A. Vaidya, F. Mai, Y. Ning, Empirical analysis of multi-task learning for reducing identity bias in toxic comment detection, in: Proceedings of the International AAAI Conference on Web and Social Media, volume 14, 2020, pp. 683–693.

[23] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2019.

[24] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: NAACL-HLT (1), Association for Computational

Linguistics, 2019, pp. 4171–4186.
[25] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, in: PML4DC at ICLR 2020, 2020.
[26] W. Antoun, F. Baly, H. Hajj, Arabert: Transformer-based model for arabic language understanding, in: LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020, ????, p. 9.
[27] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: ICLR (Poster), OpenReview.net, 2019.
[28] L. Van der Maaten, G. Hinton, Visualizing data using t-sne., Journal of machine learning research 9 (2008).
[29] G. S. Cheema, S. Hakimov, E. Müller-Budack, R. Ewerth, On the role of images for analyzing claims in social media, in: CLEOPATRA@WWW, volume 2829 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021, pp. 32–46.