NUS-IDS at CheckThat!2022: Identifying Check-worthiness of Tweets using CheckthaT5

Mingzhe Du[1,*], Sujatha Das Gollapalli[1,2] and See-Kiong Ng[1,2]

[1]*Institute of Data Science, National University of Singapore*
[2]*Centre for Trusted Internet and Community, National University of Singapore*

Abstract

This paper describes our system CheckthaT5, which was designed in the context of the Checkthat! 2022 competition at CLEF. CheckthaT5 explores the feasibility of adapting sequence-to-sequence models for detecting check-worthy social media content in a multilingual environment (Arabic, Bulgarian, Dutch, English, Spanish and Turkish) provided in the competition. We feed all languages into CheckThaT5 uniformly, thus enabling knowledge transfer from high-resource languages to low-resource languages. Empirically, CheckthaT5 outperforms strong baselines in all low-resource languages. In addition, we incorporate tasks based on non-textual features that complement tweets and other related Checkthat! 2022 tasks into CheckthaT5 through multitask learning further improving the average classification performance by 3 per cent. With our CheckThaT5 model, we rank first in 4 out of 6 languages at Checkthat! 2022 Task 1A.[1]

Keywords

Fact-checking, Multilingual, Transformers, Social media

1. Introduction

With the rise of social media, everyone can disseminate information on the Internet, while the Internet was inherently deficient in regulatory mechanisms for information authenticity [1]. Fact-checking systems were developed, and the task of verifying whether the information has check-worthiness is the outset of fact-checking system pipelines [2].

Until recently, most fact-checking systems were designed in the context of Wikipedia [3, 4], academic papers [5], and other relatively formal domains [6, 7, 8], in which the canonical and logical writing style may help models identify factual inconsistencies [8]. Yet the real challenge of fact-checking comes from social media. Social media enables speedy dissemination of massive content, while at the same time lacks regulatory and quality control due to which the problem of fact-checking is significantly increased [9]. Thus models that are designed to sift highly influential and check-worthy content through the information flood are highly desirable.

Current state-of-the-art fact-checking models do not adequately address low-resource languages. Previous works train a dedicated model for each language [10]. However, the per-

[1]Our code and models are available at https://github.com/Elfsong/clef

CLEF 2022: Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy

*Corresponding author.
mingzhe@nus.edu.sg (M. Du); idssdg@nus.edu.sg (S. D. Gollapalli); seekiong@nus.edu.sg (S. Ng)
0000-0001-7832-0459 (M. Du); 0000-0002-4567-8937 (S. D. Gollapalli); 0000-0001-6565-7511 (S. Ng)

formance of these models usually is not comparable to English models, due to the limited corpus resources and labeled data [10]. To address this shortcoming, we seek to enhance model understanding of language-independent knowledge through multitask learning, thereby improving the model performance on low-resource languages with assistance from high-resource languages.

To this end, we propose CheckthaT5, a powerful sequence-to-sequence model based on mT5 [11]. CheckthaT5 can uniformly handle classification tasks as well as regression tasks in a multitask learning manner. In the Checkthat! 2022 task 1A, Our model achieved the best results on Arabic, Bulgarian, Dutch and Spanish datasets, ranked second among Turkish and eighth among English.

2. Task Background and Data Analysis

Checkthat! 2022 competition aims to fight the COVID-19 infodemic and detects fake news[12, 13].[1] The dataset consists of 16,000 manually annotated tweets for fine-grained disinformation analysis, which covers Arabic, Bulgarian, Dutch, English, Spanish and Turkish. Based on this dataset, the competition organisers construct multiple downstream tasks of fact-checking. In this paper, we focus on the task "check-worthiness of tweets", which predicts whether the given tweet is worth for fact-checking [14]. In this section, we describe our qualitative investigations, through which we obtain insights used to design our model.

Multilingual tasks The Checkthat! 2022 dataset includes six languages. Compared with English, the other five languages are relatively low-resource languages. Previous works typically train a dedicated model for each language, which involves fine-tuning pre-trained language models on specific datasets [15, 16, 17]. Following this approach, fact-checking models have made rapid progress in languages such as English where large-scale language models and corpora are available. However, large-scale models or corpora to train them are infeasible in low-resource languages. These languages may have available monolingual models and corresponding corpora, but their size and quantity are much less than those of mainstream languages such as English. It is challenging to train accurate classifiers using the above approach. Instead, a large, labeled dataset may be required which again comprises a roadblock for low-resource languages. Therefore, we posit that if we extract language-independent knowledge representations from multilingual language models and transfer them to low-resource language tasks, we can effectively improve the performance of these tasks. Accordingly, we adopt the multitask learning paradigm in our model.

Class Imbalance in Training Data. The competition organisers provided training, development and development-test sets for each language. Table 1 provides the label distribution of Checkthat! 2022 Task 1A datasets. As can be seen in this table, the competition datasets have significant class imbalance. Previous research has indicated that address class imbalance is crucial while designing classification models especially when the training datasets are significantly small as in low-resource settings [18]. We therefore adopt class weighted loss and two data sampling methods to alleviate this problem.

[1] https://sites.google.com/view/clef2022-checkthat

Table 1
The label distribution of Checkthat! 2022 Task 1A datasets

language	split	yes	no	total
Arabic	train	962	1,551	2,513
	dev	100	135	235
	test	266	425	691
Bulgarian	train	378	1,493	1,871
	dev	36	142	178
	test	106	413	519
Dutch	train	377	546	923
	dev	28	44	72
	test	102	150	252
English	train	447	1,675	2,122
	dev	44	151	195
	test	129	447	576
Spanish	train	1,903	3,087	4,990
	dev	305	2,195	2,500
	test	305	2,195	2,500
Turkish	train	423	1,994	2,417
	dev	45	177	222
	test	114	546	660

Table 2
Model performance difference (positive class F1) on the development split between models training with links and without links

Link settings	Arabic	Bulgarian	Dutch	English	Spanish	Turkish
with links	0.582	0.680	**0.706**	0.643	0.576	0.462
without links	**0.599**	**0.711**	0.700	**0.682**	**0.635**	**0.498**

Link Processing. Though a large number of tweets in the provided datasets have URL information, we noticed that these links are not informative and instead seem to have been generated randomly. We therefore substitute all URLs with a placeholder "[LINK]" in the preprocessing step. Experimentally, we found that the results with link replacement not only promotes the model performance but also shortens the input sequence length and thereby improves the training efficiency. Table 2 reveals the performance difference on the development set between models training with links and without links.

Tweet Meta-features. Apart from the content of tweets, tweets also contain a number of meta-features, such as the author profile, the count of tweet likes and whether the tweet topic is hot, etc. Hence, we consider the effectiveness of these features in improving check-worthiness prediction. To exploit these features, we constructed new datasets by including these tweet meta-features.

Table 3
Positive class F1 score comparison of our model and baseline models on the test split

Language	Arabic	Bulgarian	Dutch	English	Spanish	Turkish
Random	0.342	0.261	0.421	0.242	0.139	0.268
N-gram	0.378	0.377	0.531	0.515	0.489	0.148
XLM-RoBerta-Large	0.561	0.674	0.670	**0.689**	0.639	0.477
CheckthaT5(ours)	**0.610**	**0.711**	**0.721**	0.682	**0.642**	**0.498**

3. Baselines

The CLEF Checkthat! 2022 competition organizers provided three baselines, namely: Majority, Random and *N*-gram. We selected the *N*-gram baseline among them to compare with our model since it is difficult to derive task-related insights from the other two models. Furthermore, based on best performing models from the previous year's competition [10], such as Schlicht et al. [19] who employed a multilingual transformer model, we selected a strong multilingual transformer model RoBerta [20] as a baseline. We report our best results and baseline results in Table 3.

3.1. N-gram

The *N*-gram baseline uses the TF-IDF vector representation of uni-grams into a support vector classification (SVC) [21] model to predict binary labels. This method is fast and straightforward, but cannot handle unknown words in the training data and also ignores the relationship between words. In Checkthat! 2021, Martinez-Rico [22] and Touahri [23] used more sophisticated word vector techniques such as word2vec [24] and GloVe [25], which obtained more semantic information through training on a large corpus. However, these methods are inherently deficient in contextual understanding since they do not incorporate sentence-level information.

3.2. Multilingual Models

In addition to the *N*-gram baseline described in the previous section, we set up another baseline using the multilingual transformer model XLM-RoBERTa [26]. It is pre-trained on 2.5TB of filtered CommonCrawl data containing 100 languages and provides a strong baseline for comparing against our proposed CheckthaT5 model.

4. Method

4.1. Data Preprocessing

To mitigate model bias and over-fitting due to the class imbalance issue mentioned in Section 2, we use class weighted loss and two data sampling methods, scilicet upsampling and downsampling [18], to enhance model generalisation. For class weighted loss, weighing the loss computed for samples differently based on the corresponding class proportion. For downsampling, we randomly remove the preponderance class samples until the number of classes is balanced. For

upsampling, we repeatedly sample the disadvantaged class samples until the proportions of all classes are close. After experiments, we found that upsampling works best.

After the link processing step, mentioned in Section 2, data samples are divided into queries and labels in the preprocessing stage. Depending on the language and task type, we insert a prompt string to the head of queries. The specific generation method of the prompt string will be introduced in Section 4.2.[2]

4.2. Multitask Training

Based on the exploratory data analysis, we found that whether a tweet is check-worthy is also affected by factors other than the content of the tweet. For example, the number of followers, the number of retweets, and whether the related topic is popular among others. To utilise these tweet meta-features, we construct a set of auxiliary tasks.

We created a new dataset using tweets and corresponding like counts to set up an auxiliary regression task that predicts "tweet likes" counts. Similarly, we also found that other sub-tasks in Checkthat! 2022 Task 1, such as "Verifiable factual claims detection", "Attention-worthy tweet detection", can assist CheckthaT5 to improve the performance of the main task "Check-worthiness detection".

In addition, "translation" tasks are created as follows: We used the different language datasets from CheckThat! 2022 Task 1A and employed the Google translation library Rosetta [27] to get their corresponding English translations. Using each of these non-English datasets, we added five "OL to EN" translation tasks to our multitask learning setup where OL refers to a non-English language from the set AR (Arabic), BG (Bulgarian), NL (Dutch), ES (Spanish), TR (Turkish). A language prompt string is assigned for each language separately [28]. In this way by adding English translation tasks for all available languages and training them jointly, we hope to learn cross-language representation and boost performance for low-resource languages using the better resourced ones such as English.

As described above, both classification and regression tasks can uniformly added into our multitask learning set up. We list the different tasks and their prompts used in CheckThaT5 in Table 4. In Section 5.2, the effectiveness of each of these tasks for improving check-worthiness is studied experimentally.

4.3. CheckthaT5 Model

The processed datasets are afterwards fed into a point-wise model, which we call CheckthaT5. This model is based on mT5, a multilingual sequence-to-sequence transformer pretrained on the mC4 corpus, covering 101 languages [11]. All tasks can be cast as sequence-to-sequence tasks in mT5. We follow this style to convert all mentioned tasks by using the following input sequence:

$$< Prompt \| Query : query \| Label : label >$$

This core idea of CheckthaT5 is to train distinct tasks jointly, enabling the model to learn language-independent knowledge. We purpose to handle classification tasks and regression

[2]Pre-processed data can be downloaded from https://huggingface.co/datasets/Elfsong/clef_data

Table 4
Auxiliary tasks and their corresponding prompts

Task	Type	Prompt
Check-worthiness of tweets (CW)	Classification	checkworthy
Verifiable factual claims detection (FC)	Classification	claim
Harmful tweet detection (HD)	Classification	harmful
Attention-worthy tweet detection (AD)	Classification	attentionworthy
English translation (ET)	Classification	backtranslate
Tweet favorite count (TF)	Regression	favorite
Tweet retweet count (TR)	Regression	retweet

tasks uniformly through the model decoder layers. For binary classification tasks, the model is fine-tuned to produce the tokens "yes" or "no" by the particular task definition. For regression tasks, the model output should be a string of numbers. However, the original output space of language models spans the entire vocabulary. To ensure the model output only comprises of classification labels (e.g., "yes" or "no"), we ignore all logits except label words for classification tasks. Similarly, to produce numerical output predictions in regression tasks, we only consider output tokens that are digits, i.e. 0 to 9 thus ensuring that the final output is a legal number (e.g., "224" or "619").

We fine-tune the mT5-XL model[3] with a batch size of 128 for 2,000 steps. The learning rate is constant at 0.001. After fine-tuning, we pick the best checkpoint based on the development set scores. Further details of configurations are available in our code repository.

5. Results

5.1. Leaderboard Results

The official metric of Checkthat! 2022 check-worthiness was F1 score (positive class) for all languages. Table 5 lists our results. Arabic, Bulgarian, Dutch and Spanish performed the best with 0.628, 0.617, 0.642 and 0.571, respectively. It was followed by Turkish (0.173) as the second and English(0.519) as the eighth.[4] We note in the last row of Table 5, that though we rank the second, our F1 score is low.

Considering that our model achieved an F1 score of 0.498 on the dev-test split (Table 3), we conjecture that the test data distribution is slightly different from that in the data shared as part of the competition and used to train and fine-tune models (train/dev/dev-test splits). Indeed, as indicated on the leaderboard all the participating systems attain low test performance on the Turkish language with the system that ranked first obtaining an F1 score of 0.212 compared to our 0.173.

[3]Model link https://huggingface.co/google/mt5-xl
[4]The public leaderboard can be found in http://shorturl.at/nuCOS

Table 5
CheckthaT5 results from Checkthat! 2022 Task 1A public leaderboard

language	F1	Rank
Arabic	**0.628**	**1-st**
Bulgarian	**0.617**	**1-st**
Dutch	**0.642**	**1-st**
English	0.519	8-th
Spanish	**0.571**	**1-st**
Turkish	0.173	2-nd

Table 6
Multitask learning ablation study using F1 scores on the development-test splits

Language	Arabic	Bulgarian	Dutch	English	Spanish	Turkish
CW	0.582	0.680	0.706	0.643	0.576	0.462
CW + ET	**0.607**	0.675	0.701	0.663	0.637	0.489
CW + ET + TF	0.601	0.700	0.712	0.659	**0.642**	0.477
CW + ET + AD	0.600	0.709	0.710	0.665	0.618	0.491
CW + ET + FC + AD	0.596	0.698	0.704	0.656	0.605	0.494
CW + ET + FC + AD + TF	0.600	**0.711**	**0.721**	**0.682**	0.635	**0.498**

5.2. Ablation Study

We consider the following tasks in the ablation study: check-worthiness of tweets(CW), verifiable factual claims detection(FC), harmful tweet detection(HD), attention-worthy tweet detection(AD), English translation(ET), tweet favorite count(TF) and tweet retweet count(TR).

Table 6 shows the result of the ablation study using F1 scores on the dev-test splits provided as part of the competition. Compared to single-task learning models (first row), multitask learning results in improved performance for all six languages. In particular, comparing rows 1 and 2 in the table, the English translation tasks 4.2 enable the model to obtain a significant score improvement in Arabic, English, Spanish and English. In addition, we found that appending the task of predicting the number of Twitter likes further improved the F1 score demonstrating the effectiveness of the multitask learning paradigm.

5.3. Error Analysis

Although our model achieves good performance on low-resource languages and best on four out of six languages, namely Arabic, Bulgarian, Dutch and Spanish, it performs mediocre on the English dataset. We speculate that one possible reason for the lower performance on English is that we feed all language datasets into a single model, which makes the model more language-agnostic, but also loses certain language-specific information. We will explore an opportune trade-off point in further research.

6. Conclusion

In this paper, we introduced CheckthaT5, a novel pipeline for verifying multilingual fact check-worthiness. It can capture check-worthy content from the information torrent of social media. Empirically, CheckthaT5 effectively acquires language-independent knowledge through the multitask learning paradigm, which supports our system on the top of most language tracks in CLEF Checkthat! 2022 Task 1A.

Acknowledgments

This research is supported by CTIC, NUS grant number: A-0003503-05-00. Additionally, we would like to thank Google for computational resources in the form of Google Cloud credits and Google TPU Research Cloud.

References

[1] J. Y. Cuan-Baltazar, M. J. Muñoz-Perez, C. Robledo-Vega, M. F. Pérez-Zepeda, E. Soto-Vega, Misinformation of covid-19 on the internet: infodemiology study, JMIR public health and surveillance 6 (2020) e18444.

[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] J. Thorne, A. Vlachos, C. Christodoulopoulos, A. Mittal, Fever: a large-scale dataset for fact extraction and verification, arXiv preprint arXiv:1803.05355 (2018).

[4] R. Aly, Z. Guo, M. Schlichtkrull, J. Thorne, A. Vlachos, C. Christodoulopoulos, O. Cocarascu, A. Mittal, Feverous: Fact extraction and verification over unstructured and structured information, arXiv preprint arXiv:2106.05707 (2021).

[5] D. Wadden, S. Lin, K. Lo, L. L. Wang, M. van Zuylen, A. Cohan, H. Hajishirzi, Fact or fiction: Verifying scientific claims, arXiv preprint arXiv:2004.14974 (2020).

[6] C. Samarinas, W. Hsu, M. L. Lee, Improving evidence retrieval for automated explainable fact-checking, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations, 2021, pp. 84–91.

[7] W. Ostrowski, A. Arora, P. Atanasova, I. Augenstein, Multi-hop fact checking of political claims, arXiv preprint arXiv:2009.06401 (2020).

[8] Z. Guo, M. Schlichtkrull, A. Vlachos, A survey on automated fact-checking, Transactions of the Association for Computational Linguistics 10 (2022) 178–206.

[9] Y. Wang, M. McKee, A. Torbica, D. Stuckler, Systematic literature review on the spread of health-related misinformation on social media, Social science & medicine 240 (2019) 112552.

[10] S. Shaar, M. Hasanain, B. Hamdan, Z. S. Ali, F. Haouari, A. Nikolov, M. Kutlu, Y. S. Kartal, F. Alam, G. Da San Martino, et al., Overview of the clef-2021 checkthat! lab task 1 on

check-worthiness estimation in tweets and political debates, in: CLEF (Working Notes), 2021.
[11] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, C. Raffel, mt5: A massively multilingual pre-trained text-to-text transformer, arXiv preprint arXiv:2010.11934 (2020).
[12] P. Nakov, A. Barrón-Cedeño, G. Da San Martino, F. Alam, J. M. Struß, T. Mandl, R. Míguez, T. Caselli, M. Kutlu, W. Zaghouani, C. Li, S. Shaar, G. K. Shahi, H. Mubarak, A. Nikolov, N. Babulkov, Y. S. Kartal, J. Beltrán, M. Wiegand, M. Siegel, J. Köhler, Overview of the CLEF-2022 CheckThat! lab on fighting the COVID-19 infodemic and fake news detection, in: A. Barrón-Cedeño, G. Da San Martino, M. Degli Esposti, F. Sebastiani, C. Macdonald, G. Pasi, A. Hanbury, M. Potthast, G. Faggioli, F. Nicola (Eds.), Proceedings of the 13th International Conference of the CLEF Association: Information Access Evaluation meets Multilinguality, Multimodality, and Visualization, CLEF '2022, Bologna, Italy, 2022.
[13] P. Nakov, A. Barrón-Cedeño, G. Da San Martino, F. Alam, J. M. Struß, T. Mandl, R. Míguez, T. Caselli, M. Kutlu, W. Zaghouani, C. Li, S. Shaar, G. K. Shahi, H. Mubarak, A. Nikolov, N. Babulkov, Y. S. Kartal, J. Beltrán, The CLEF-2022 CheckThat! Lab on fighting the covid-19 infodemic and fake news detection, in: M. Hagen, S. Verberne, C. Macdonald, C. Seifert, K. Balog, K. Nørvåg, V. Setty (Eds.), Advances in Information Retrieval, Springer International Publishing, Cham, 2022, pp. 416–428.
[14] P. Nakov, A. Barrón-Cedeño, G. Da San Martino, F. Alam, R. Míguez, T. Caselli, M. Kutlu, W. Zaghouani, C. Li, S. Shaar, H. Mubarak, A. Nikolov, Y. S. Kartal, J. Beltrán, Overview of the CLEF-2022 CheckThat! lab task 1 on identifying relevant claims in tweets, in: N. Faggioli, Guglielmo andd Ferro, A. Hanbury, M. Potthast (Eds.), Working Notes of CLEF 2022—Conference and Labs of the Evaluation Forum, CLEF '2022, Bologna, Italy, 2022.
[15] J. Howard, S. Ruder, Universal language model fine-tuning for text classification, arXiv preprint arXiv:1801.06146 (2018).
[16] M. E. Peters, S. Ruder, N. A. Smith, To tune or not to tune? adapting pretrained representations to diverse tasks, arXiv preprint arXiv:1903.05987 (2019).
[17] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, 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] F. Thabtah, S. Hammoud, F. Kamalov, A. Gonsalves, Data imbalance in classification: Experimental evaluation, Information Sciences 513 (2020) 429–441.
[19] I. B. Schlicht, A. F. M. de Paula, P. Rosso, Upv at checkthat! 2021: Mitigating cultural differences for identifying multilingual check-worthy claims, arXiv preprint arXiv:2109.09232 (2021).
[20] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).
[21] S. R. Gunn, et al., Support vector machines for classification and regression, ISIS technical report 14 (1998) 5–16.
[22] J. R. Martinez-Rico, L. Araujo, J. Martinez-Romo, Nlp&ir@ uned at checkthat! 2020: A preliminary approach for check-worthiness and claim retrieval tasks using neural networks

and graphs. (2020).
[23] I. Touahri, A. Mazroui, Evolutionteam at checkthat! 2020: integration of linguistic and sentimental features in a fake news detection approach, Cappellato et al.[10] (2020).
[24] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781 (2013).
[25] J. Pennington, R. Socher, C. D. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.
[26] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, arXiv preprint arXiv:1911.02116 (2019).
[27] D. Mingzhe, Rosetta: A high performance translation tool powered by Google Translate, 2022.
[28] B. Lester, R. Al-Rfou, N. Constant, The power of scale for parameter-efficient prompt tuning, arXiv preprint arXiv:2104.08691 (2021).