Can Multilingual Transformers Fight the COVID-19 Infodemic?

Lasitha Uyangodage[‡], **Tharindu Ranasinghe**[§], **Hansi Hettiarachchi**[♡],
[‡]University of Münster, [§]University of Wolverhampton,[♡]Birmingham City University
luyangod@uni-muenster.de

Abstract

The massive spread of false information on social media has become a global risk especially in a global pandemic situation like COVID-19. False information detection has thus become a surging research topic in recent months. In recent years, supervised machine learning models have been used to automatically identify false information in social media. However, most of these machine learning models focus only on the language they were trained on. Given the fact that social media platforms are being used in different languages, managing machine learning models for each and every language separately would be chaotic. In this research, we experiment with multilingual models to identify false information in social media by using two recently released multilingual false information detection datasets. We show that multilingual models perform on par with the monolingual models and sometimes even better than the monolingual models to detect false information in social media making them more useful in real-world scenarios.

1 Introduction

By June 2021, the coronavirus(COVID-19) pandemic has affected 219 nations around the world with 176 million total cases and 3.81 million deaths. The nature of the virus caused many governments to implement lockdown in their countries. As a result, many people started spending more time at home during the pandemic and started using social media more, providing an unexpected boost to engagement on these platforms (Hettiarachchi and Ranasinghe, 2020b).

As a drawback of these exponential growths, social media has become a conduit for spreading both rumours and deliberate misinformation, and many perpetrators are deploying sites such as Facebook, Twitter, YouTube, and WhatsApp to create a sense of panic and confusion. On the other hand, the general public can not completely ignore the information seen in social media due to the fact that the Centers for Disease Control and Prevention, the World Health Organisation (WHO), numerous journals, government and other health care organisations are regularly posting guidance across a host of platforms. Therefore, rather than completely disregarding information seeing in social media, accurate identification of false information is crucial (Nguyen et al., 2020).

Considering the high data generation in social media, manual approaches to filter false information require significant human efforts. Therefore an automated technique to tackle this problem will be invaluable to the community. In the light of this many shared task has been organised to tackle the false information detection in social media (Shaar et al., 2021; Nakov et al., 2021a) leading to implement various machine learning models which can identify false information automatically (Uyangodage et al., 2021; Tziafas et al., 2021). However, most of these approaches build language-specific models trained specifically on a particular language. Given the fact that most of the social media platforms are massively multilingual, maintaining fake news identification models for each language would not be feasible. One machine learning model that can work across many languages would be invaluable to the community.

In this research, we explore multilingual models for false information detection. We experiment with two recently created datasets that target two different aspects in false information detection which also covers 5 languages; Arabic, Bulgarian, English, Spanish and Turkish. We show that multilingual models based on pretrained transformer models perform on par with the language-specific models trained for each language on both aspects in false information detection making them more feasible in real-world applications.

Proceedings of Recent Advances in Natural Language Processing, pages 1432–1437
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_160

2 Related Work

False Information Detection Identifying false information in social media has been a major research topic in recent years. According to literature, mainly, there are two types of methods for false information detection as Social Context-based methods and Content-based methods (Guo et al., 2020). Social Context-based methods use different properties in user profiles such as user's credibility (Li et al., 2019) or stances (Mohammad et al., 2017) while the Content-based methods use different features in the content of posts such as certain keywords, number of URLs and the length of textual content to detect false information. However, due to ethical considerations, most of the social media platforms do not allow to release datasets with details which can be used to identify users of the posts. Therefore Social Context-based methods have not been popular in recent research as Content-based methods. Due to the nature of datasets we use for this research, we also focused on Content-based methods.

Content-based methods mainly focus on different features of post contents. For example, Castillo et al. (2011) found that highly credible tweets have more URLs and lengthy textual contents than low credible tweets. Also, many studies utilise lexical and syntactic components of the content as useful features. For instance, Qazvinian et al. (2011) found part of speech (POS) as a distinguishable feature for false information detection. Similarly, Kwon et al. (2013) found that some types of sentiments including positive words (e.g. love, nice, sweet), negating words (e.g. no, not, never), cognitive action words (e.g. cause, know) and inferring action words (e.g. maybe, perhaps) as apparent features for a periodic time-series model to identify key linguistic differences between true and fake tweets. With the recent popularity gained by embedding and deep learning-based approaches in natural language processing, there was a tendency to use deep neural networks powered by content embeddings to perform false information classification too (Ma et al., 2016). Later, with the introduction of transformers (Devlin et al., 2019; Conneau et al., 2020), there was a tendency to involve large pretrained transformer models also (Uyangodage et al., 2021; Tziafas et al., 2021; Qarqaz et al., 2021). However, all of these models were trained specifically on a single language making them less useful in real scenarios where we need to process multilingual data.

Multilingual models Multilingual models allow training a single model to perform a task on multiple languages. These types of models have been used by many tasks such as offensive language identification (Ranasinghe and Zampieri, 2020, 2021a,b) and machine translation (Nguyen and Chiang, 2017; Aharoni et al., 2019). All of these studies train one machine learning model on all the languages which the training data is available and show that the multilingual models perform on par with or sometimes even better than monolingual models. The recently released multilingual transformer models that support more than 100 languages like BERT (Devlin et al., 2019), XLM-RoBERTa (Conneau et al., 2020) have made multilingual research easier. Even though these multilingual models improve the feasibility of the research to be applied on a real-world application, to the best of our knowledge, no prior work has been done for multilingual false information identification focused by this paper.

3 Data

For this research, we used two recently released datasets on false information identification. We mainly considered two factors when selecting datasets; the dataset should be annotated in multiple languages and it should have been annotated very recently.

The first dataset (NLP4IF) which was released for NLP4IF shared task; Fighting the COVID-19 Infodemic is about predicting several binary properties of a tweet on COVID-19 such as whether it is harmful, whether it contains a verifiable claim, whether it may be of interest to the general public and whether it appears to contain false information (Shaar et al., 2021). The data has been released for three languages; English, Arabic and Bulgarian[1]. Seven labels were targeted by this dataset. The first label was **Verifiable Factual Claim:** *Does the tweet contain a verifiable factual claim?*. We only considered this label for our research as this is directly related to false information detection and this label had the most annotated data out of the seven labels. False information detection using this label can be considered as a binary text classification task.

[1]The dataset can be downloaded from https://gitlab.com/NLP4IF/nlp4if-2021

The second dataset (CLEF2021) that we considered was released for CLEF2021 CheckThat-Lab (Nakov et al., 2021a) Task 1: Check-Worthiness Estimation of the tweets (Nakov et al., 2021b). Given a tweet, the participants need to predict whether it is worth fact-checking. This task is directly related to the first label of the NLP4IF dataset. However contrast to the binary classification in the previous task, the models in this task need to predict a continuous value between 0-1 that reflects the worthiness to perform fact-checking. The dataset has been annotated in five languages; Arabic, Bulgarian, English, Spanish and Turkish (Nakov et al., 2021b) promoting multilingual research.

4 Architecture

The main motivation for our architecture is the recent success that the transformer models had in various natural language processing tasks including text classification (Ranasinghe and Zampieri, 2020, 2021a,b), word sense disambiguation (Hettiarachchi and Ranasinghe, 2020a, 2021), language identification (Jauhiainen et al., 2021) etc. Apart from providing strong results compared to RNN based architectures (Ranasinghe et al., 2019), transformer models like BERT (Devlin et al., 2019) provide pretrained multilingual language models that support more than 100 languages which will solve the multilingual issues of these tasks (Ranasinghe and Zampieri, 2020).

Transformer models take an input of a sequence and output the representations of the sequence. There can be one or two segments in a sequence which are separated by a special token [SEP] (Devlin et al., 2019). In this approach we considered a tweet as a sequence and no [SEP] token is used. Another special token [CLS] is used as the first token of the sequence which contains a special classification embedding. For text classification tasks, transformer models take the final hidden state **h** of the [CLS] token as the representation of the whole sequence (Sun et al., 2019). A simple softmax classifier is added to the top of the transformer model to predict the probability of a class. For the text regression tasks, a fully-connected layer is added on top of the [CLS] token. The fully-connected layer will have a single output neuron which predicts the target. For both tasks, all the parameters of the transformer model as well as W are fine-tuned jointly by maximising the log-probability of the gold truth.

5 Experimental Setup

We trained a transformer model for each dataset mentioned in Section 3. Given the very unbalanced nature of the datasets, the transformer models tend to overfit and predict only the majority class. Therefore, for each label, we took the number of instances in the training set for the minority class and undersampled the majority class to have the same number of instances as the minority class.

We then divided this undersampled dataset into a training set and a validation set using the 0.8:0.2 split. We mainly fine-tuned the learning rate and the number of epochs of the classification model manually to obtain the best results for the validation set. We obtained $1e^{-5}$ as the best value for the learning rate and 3 as the best value for the number of epochs for both datasets. The other configurations of the transformer model were set to a constant value over all the experiments in order to ensure consistency between them. We used a batch size of 8, Adam optimiser and a linear learning rate warm-up over 10% of the training data. The models were trained using only training data. We performed early stopping if the evaluation loss did not improve over 10 evaluation rounds. The implementation was done using HuggingFace transformer implementation (Wolf et al., 2020). A summary of hyperparameters and their values used to obtain the reported results are mentioned in Table 1. The optimised hyperparameters are marked with ‡ and their optimal values are reported.

Parameter	Value
learning rate‡	$1e^{-5}$
number of epochs‡	3
adam epsilon	$1e^{-8}$
warmup ration	0.1
warmup steps	0
max grad norm	1.0
max seq. length	120
gradient accumulation steps	1

Table 1: Hyperparameter specifications

For monolingual experiments, we trained language-specific transformer models on that particular language only. As pretrained transformer models, we used Arabert (Antoun et al., 2020) for Arabic, bert-base-cased (Devlin et al., 2019) for English, BETO: Spanish BERT for Spanish (Cañete et al., 2020) and BERTurk for Turkish. Unfortunately for Bulgarian, we could not find a suitable

pretrained transformer model. Therefore, for Bulgarian, we used the bert-multilingual-cased (Devlin et al., 2019) model.

For multilingual experiments, we first combined data instances from all the languages of each task which left us with two large multilingual false information identification datasets. Then we trained the transformer models on that combined datasets. As the multilingual pretrained transformer model, we used the bert-multilingual-cased (Devlin et al., 2019) model.

6 Results

In Table 2 we show the results we got for the test set of the NLP4IF dataset. We used the same evaluation metric as the organisers of the task; Macro F1 inorder to compare our approach with the baselines and the best systems submitted.

Language	Model	Macro F1
Arabic	Monolingual	0.852
	Qarqaz et al. (2021)	0.843
	Multilingual	0.802
	Random Baseline	0.552
	Ngram Baseline	0.510
Bulgarian	Multilingual	0.956
	Ngram Baseline	0.909
	Shaar et al. (2021)	0.887
	Monolingual	0.647
	Random Baseline	0.594
English	Multilingual	0.842
	Tziafas et al. (2021)	0.835
	Monolingual	0.819
	Ngram Baseline	0.647
	Random Baseline	0.552

Table 2: Results ordered by Macro F1 for Arabic, Bulgarian and English languages in NLP4IF dataset. Monolingual implies the results of the monolingual models and Multilingual implies the results of the multilingual model for each language. Additionally, we report Ngram and Random baselines, and best systems submitted for the shared task.

As can be seen in the results, for the NLP4IF dataset, multilingual models perform better than the monolingual models and the best systems in Bulgarian and English while performing on par in Arabic. Please note that these best systems (Qarqaz et al., 2021; Tziafas et al., 2021) have been trained specifically on those language pairs using language specific natural language processing pipelines, yet the multilingual models outperform them in English and Bulgarian.

The results for CLEF2021 dataset is shown in Table 3. For this dataset also we used the same evaluation metric that the organisers used - Mean Average Precision (MAP) (Nakov et al., 2021b) [2].

Language	Model	MAP
Arabic	Best System	0.658
	Multilingual	0.651
	Monolingual	0.647
	Ngram Baseline	0.428
Bulgarian	Best System	0.737
	Multilingual	0.711
	Monolingual	0.700
	Ngram Baseline	0.588
English	Best System	0.224
	Monolingual	0.196
	Multilingual	0.188
	Ngram Baseline	0.052
Spanish	Best System	0.537
	Multilingual	0.522
	Monolingual	0.508
	Ngram Baseline	0.450
Turkish	Best System	0.581
	Multilingual	0.565
	Monolingual	0.555
	Ngram Baseline	0.354

Table 3: Results ordered by Mean Average Precision (MAP) for Arabic, Bulgarian, English, Spanish and Turkish languages in CLEF2021 dataset. Monolingual implies the results of the monolingual models and Multilingual implies the results of the multilingual model for each language. Best system denotes the results of the best system submitted to the language. Additionally, we report the Ngram baseline.

As can be seen in the results multilingual models outperformed monolingual models in Arabic, Bulgarian, Spanish and Turkish languages while performing on par with English. Similar to the results of the previous dataset, these multilingual models are very competitive with the best systems submitted to each of the languages.

7 Conclusion

In this paper, we explored multilingual models for false information identification using two recently created datasets. In our experiments, we observed that multilingual models built using powerful pretrained multilingual transformers perform on par or sometimes even better than the monolingual models. These results are consistent with

[2]The results are extracted from https://gitlab.com/checkthat_lab/clef2021-checkthat-lab

both datasets and across five languages. Findings in this paper would be valuable when building real-world applications for false information identification where maintaining separate machine learning models for each language would be more expensive and chaotic.

As future work, we would like to expand this research into more transformer models and more languages. We would like to experiment with how the multilingual transformer models with the cross-lingual concepts like XLM-RoBERTa would perform in multilingual false information identification. Furthermore, we would explore zero-shot and few-shot learning with multilingual models which would be beneficial to low resource language where the training data is scarce.

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. 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 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.

Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, page 675–684, New York, NY, USA. Association for Computing Machinery.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, Minneapolis, Minnesota. Association for Computational Linguistics.

Bin Guo, Yasan Ding, Lina Yao, Yunji Liang, and Zhiwen Yu. 2020. The future of false information detection on social media: New perspectives and trends. *ACM Comput. Surv.*, 53(4).

Hansi Hettiarachchi and Tharindu Ranasinghe. 2020a. BRUMS at SemEval-2020 task 3: Contextualised embeddings for predicting the (graded) effect of context in word similarity. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 142–149, Barcelona (online). International Committee for Computational Linguistics.

Hansi Hettiarachchi and Tharindu Ranasinghe. 2020b. InfoMiner at WNUT-2020 task 2: Transformer-based covid-19 informative tweet extraction. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 359–365, Online. Association for Computational Linguistics.

Hansi Hettiarachchi and Tharindu Ranasinghe. 2021. TransWiC at SemEval-2021 task 2: Transformer-based multilingual and cross-lingual word-in-context disambiguation. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 771–779, Online. Association for Computational Linguistics.

Tommi Jauhiainen, Tharindu Ranasinghe, and Marcos Zampieri. 2021. Comparing approaches to Dravidian language identification. In *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 120–127, Kiyv, Ukraine. Association for Computational Linguistics.

S. Kwon, M. Cha, K. Jung, W. Chen, and Y. Wang. 2013. Prominent features of rumor propagation in online social media. In *2013 IEEE 13th International Conference on Data Mining*, pages 1103–1108.

Quanzhi Li, Qiong Zhang, and Luo Si. 2019. Rumor detection by exploiting user credibility information, attention and multi-task learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1173–1179, Florence, Italy. Association for Computational Linguistics.

Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J. Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, page 3818–3824. AAAI Press.

Saif M. Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. 2017. Stance and sentiment in tweets. *ACM Trans. Internet Technol.*, 17(3).

Preslav Nakov, Giovanni Da San Martino, Tamer Elsayed, Alberto Barrón-Cedeño, Rubén Míguez, Shaden Shaar, Firoj Alam, Fatima Haouari, Maram Hasanain, Nikolay Babulkov, Alex Nikolov, Gautam Kishore Shahi, Julia Maria Struß, and Thomas Mandl. 2021a. The clef-2021 checkthat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news. In *Advances in Information Retrieval*, pages 639–649, Cham. Springer International Publishing.

Preslav Nakov, Giovanni Da San Martino, Tamer Elsayed, Alberto Barrón-Cedeño, Rubén Míguez, Shaden Shaar, Firoj Alam, Fatima Haouari, Maram Hasanain, Watheq Mansour, Bayan Hamdan, Zien Sheikh Ali, Nikolay Babulkov, Alex Nikolov, Gautam Kishore Shahi, Julia Maria Struß, Thomas Mandl, Mucahid Kutlu, and Yavuz Selim Kartal. 2021b. 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).

Dat Quoc Nguyen, Thanh Vu, Afshin Rahimi, Mai Hoang Dao, Linh The Nguyen, and Long Doan. 2020. WNUT-2020 task 2: Identification of informative COVID-19 English tweets. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 314–318, Online. Association for Computational Linguistics.

Toan Q. Nguyen and David Chiang. 2017. Transfer learning across low-resource, related languages for neural machine translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 296–301, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Ahmed Qarqaz, Dia Abujaber, and Malak Abdullah. 2021. R00 at NLP4IF-2021 fighting COVID-19 infodemic with transformers and more transformers. In *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 104–109, Online. Association for Computational Linguistics.

Vahed Qazvinian, Emily Rosengren, Dragomir R. Radev, and Qiaozhu Mei. 2011. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1589–1599, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Tharindu Ranasinghe and Marcos Zampieri. 2020. Multilingual offensive language identification with cross-lingual embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5838–5844, Online. Association for Computational Linguistics.

Tharindu Ranasinghe and Marcos Zampieri. 2021a. An evaluation of multilingual offensive language identification methods for the languages of india. *Information*, 12(8).

Tharindu Ranasinghe and Marcos Zampieri. 2021b. MUDES: Multilingual detection of offensive spans. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 144–152, Online. Association for Computational Linguistics.

Tharindu Ranasinghe, Marcos Zampieri, and Hansi Hettiarachchi. 2019. BRUMS at HASOC 2019: Deep learning models for multilingual hate speech and offensive language identification. In *Proceedings of the 11th annual meeting of the Forum for Information Retrieval Evaluation*.

Shaden Shaar, Firoj Alam, Giovanni Da San Martino, Alex Nikolov, Wajdi Zaghouani, Preslav Nakov, and Anna Feldman. 2021. Findings of the NLP4IF-2021 shared tasks on fighting the COVID-19 infodemic and censorship detection. In *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 82–92, Online. Association for Computational Linguistics.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? In *Chinese Computational Linguistics*, pages 194–206, Cham. Springer International Publishing.

Georgios Tziafas, Konstantinos Kogkalidis, and Tommaso Caselli. 2021. Fighting the COVID-19 infodemic with a holistic BERT ensemble. In *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 119–124, Online. Association for Computational Linguistics.

Lasitha Uyangodage, Tharindu Ranasinghe, and Hansi Hettiarachchi. 2021. Transformers to fight the COVID-19 infodemic. In *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 130–135, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.