Transformers to Fight the COVID-19 Infodemic

Lasitha Uyangodage[‡], **Tharindu Ranasinghe**[§], **Hansi Hettiarachchi**[♡],
[‡]Department of Information Systems, University of Münster, Germany
[§]Research Group in Computational Linguistics, University of Wolverhampton, UK
[♡]School of Computing and Digital Technology, Birmingham City University, UK
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. NLP4IF-2021 shared task on fighting the COVID-19 infodemic has been organised to strengthen the research in false information detection where the participants are asked to predict seven different binary labels regarding false information in a tweet. The shared task has been organised in three languages; Arabic, Bulgarian and English. In this paper, we present our approach to tackle the task objective using transformers. Overall, our approach achieves a 0.707 mean F1 score in Arabic, 0.578 mean F1 score in Bulgarian and 0.864 mean F1 score in English ranking 4^{th} place in all the languages.

1 Introduction

By April 2021, coronavirus(COVID-19) pandemic has affected 219 nations around the world with 136 million total cases and 2.94 million deaths. With this pandemic situation, a rapid increase in social media usage was noticed. In measures, during 2020, 490 million new users joined indicating a more than 13% year-on-year growth (Kemp, 2021). This growth is mainly resulted due to the impacts on day-to-day activities and information sharing and gathering requirements related to the pandemic.

As a drawback of these exponential growths, the dark side of social media is further revealed during this COVID-19 infodemic (Mourad et al., 2020). The spreading of false and harmful information resulted in panic and confusions which make the pandemic situation worse. Also, the inclusion of false information reduced the usability of a huge volume of data which is generated via social media platforms with the capability of fast propagation. To handle these issues and utilise social media data effectively, accurate identification of false information is crucial. 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.

Targeting the infodemic that occurred with COVID-19, NLP4IF-2021 shared task was designed to predict several properties of a tweet including harmfulness, falseness, verifiability, interest to the general public and required attention. The participants of this task were required to predict the binary aspect of the given properties for the test sets in three languages: Arabic, Bulgarian and English provided by the organisers. Our team used recently released transformer models with the text classification architecture to make the predictions and achieved the 4^{th} place in all the languages while maintaining the simplicity and universality of the method. In this paper, we mainly present our approach, with more details about the architecture including an experimental study. We also provide our code to the community which will be freely available to everyone interested in working in this area using the same methodology[1].

2 Related Work

Identifying false information in social media has been a major research topic in recent years. False information detection methods can be mainly categorised into two main areas; Content-based methods and Social Context-based methods (Guo et al., 2020).

Content-based methods are mainly based on the different features in the content of the tweet. For example, Castillo et al. (2011) find that highly credible tweets have more URLs, and the textual content length is usually longer than that of lower credibility tweets. Many studies utilize the lexical and

[1]The GitHub repository is publicly available on https://github.com/tharindudr/infominer

Proceedings of the 4th NLP4IF Workshop on NLP for Internet Freedom, pages 130–135

syntactic features to detect false information. For instance, Qazvinian et al. (2011) find that the part of speech (POS) is a distinguishable feature for false information detection. Kwon et al. (2013) find that some types of sentiments are apparent features of machine learning classifiers, including positive sentiments 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). Then they propose a periodic time-series model to identify key linguistic differences between true tweets and fake tweets. With the word embeddings and deep learning getting popular in natural language processing, most of the fake information detection methods were based on embeddings of the content fed into a deep learning network to perform the classification (Ma et al., 2016).

Traditional content-based methods analyse the credibility of the single microblog or claim in isolation, ignoring the high correlation between different tweets and events. However, Social Context-based methods take different tweets in a user profile or an event to identify false information. Many studies detect false information by analyzing users' credibility (Li et al., 2019) or stances (Mohammad et al., 2017). Since this shared task is mainly focused on the content of the tweet to detect false information, we can identify our method as a content-based false information identification approach.

3 Data

The task is about predicting several binary properties of a tweet on COVID-19: whether it is harmful, whether it contains a verifiable claim, whether it may be of interest to the general public, whether it appears to contain false information, etc. (Shaar et al., 2021). The data has been released for three languages; English, Arabic and Bulgarian [2]. Following are the binary properties that the participants should predict for a tweet.

I **Verifiable Factual Claim**: Does the tweet contain a verifiable factual claim?

II **False Information**: To what extent does the tweet appear to contain false information?

III **Interest to General Public**: Will the tweet have an effect on or be of interest to the general public?

IV **Harmfulness**: To what extent is the tweet harmful to the society?

V **Need of Verification**: Do you think that a professional fact-checker should verify the claim in the tweet?

VI **Harmful to Society**: Is the tweet harmful for the society?

VII **Require attention**: Do you think that this tweet should get the attention of government entities?

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

4 Architecture

The main motivation for our architecture is the recent success that the transformer models had in various natural language processing tasks like sequence classification (Ranasinghe and Hettiarachchi, 2020; Ranasinghe et al., 2019; Pitenis et al., 2020), token classification (Ranasinghe and Zampieri, 2021a; Ranasinghe et al., 2021), language detection (Jauhiainen et al., 2021), word context prediction (Hettiarachchi and Ranasinghe, 2020a, 2021) question answering (Yang et al., 2019) etc. Apart from providing strong results compared to RNN based architectures (Hettiarachchi and Ranasinghe, 2019; 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 et al., 2020; Ranasinghe and Zampieri, 2021b, 2020).

Transformer models take an input of a sequence and outputs 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 c as shown in Equation 1 where W is the task-specific parameter matrix. In the classification task all the parameters from transformer as well as W are fine tuned jointly by maximising the log-probability of the correct label. The architecture of transformer-based sequence classifier is shown in Figure 1.

$$p(c|\mathbf{h}) = softmax(W\mathbf{h}) \quad (1)$$

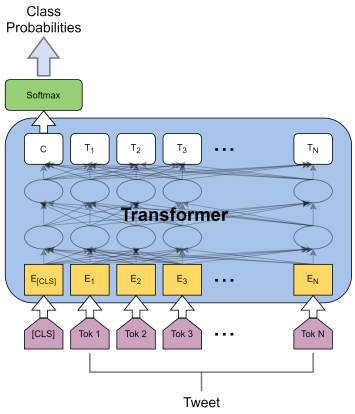

Figure 1: Text Classification Architecture

5 Experimental Setup

We considered the whole task as seven different classification problems. We trained a transformer model for each label mentioned in Section 3. This gave us the flexibility to fine-tune the classification model in to the specific label rather than the whole task. Given the very unbalanced nature of the dataset, 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 0.8:0.2 split. We mainly fine tuned the learning rate and number of epochs of the classification model manually to obtain the best results for the development set provided by organisers in each language. We obtained $1e^{-5}$ as the best value for learning rate and 3 as the best value for number of epochs for all the languages in all the labels. The other configurations of the transformer model were set to a constant value over all the languages in order to ensure consistency between the languages. We used a batch-size of eight, Adam optimiser (Kingma and Ba, 2014) 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 ten evaluation rounds. A summary of hyperparameters and their values used to obtain the reported results are mentioned in Appendix - Table 3. The optimized hyperparameters are marked with ‡ and their optimal values are reported. The rest of the hyperparameter values are kept as constants. We did not use any language specific preprocessing techniques in order to have a flexible solution between the languages. We used a Nvidia Tesla K80 GPU to train the models. All the experiments were run for five different random seeds and as the final result, we took the majority class predicted by these different random seeds as mention in Hettiarachchi and Ranasinghe (2020b). We used the following pretrained transformer models for the experiments.

bert-base-cased - Introduced in Devlin et al. (2019), the model has been trained on a Wikipedia dump of English using Masked Language Modelling (MLM) objective. There are two variants in English BERT, base model and the large model. Considering the fact that we built seven different models for each label, we decided to use the base model considering the resources and time.

roberta-base - Introduced in Liu et al. (2019), RoBERTa builds on BERT and modifies key hyperparameters, removing the next-sentence pretraining objective and training with much larger mini-batches and learning rates. RoBERTa has outperformed BERT in many NLP tasks and it motivated us to use RoBERTa in this research too. Again we only considered the base model.

bert-nultilingual-cased - Introduced in Devlin et al. (2019), the model has been trained on a Wikipedia dump of 104 languages using MLM objective. This model has shown good performance in variety of languages and tasks. Therefore, we used this model in Arabic and Bulgarian.

AraBERT Recently language-specific BERT based models have proven to be very efficient at language understanding. AraBERT (Antoun et al., 2020) is such a model built for Arabic with BERT using scraped Arabic news websites and two publicly available Arabic corpora; 1.5 billion words Arabic Corpus (El-khair, 2016) and OSIAN: the Open Source International Arabic News Corpus (Zeroual et al., 2019). Since AraBERT has outperformed multilingual bert in many NLP tasks in Arabic (Antoun et al., 2020) we used this model for Arabic in this task. There are two version in AraBERT; AraBERTv0.1 and AraBERTv1, with the difference being that AraBERTv1 uses pre-segmented text where prefixes and suffixes were

	Model	I	II	III	IV	V	VI	VII	Mean
English	roberta-base	0.822	0.393	0.821	0.681	0.461	0.235	0.251	0.523
	bert-base-cased	0.866	0.461	0.893	0.740	0.562	0.285	0.303	0.587
Arabic	bert-multilingual-cased	0.866	0.172	0.724	0.400	0.557	0.411	0.625	0.536
	arabert-v2	0.917	0.196	0.782	0.469	0.601	0.433	0.686	0.583
	arabert-v2-tokenized	0.960	0.136	0.873	0.571	0.598	0.424	0.678	0.606
Bulgarian	bert-multilingual-cased	0.845	0.098	0.516	0.199	0.467	0.303	0.196	0.375

Table 1: Macro F1 between the algorithm predictions and human annotations for development set in all the languages. Results are sorted from Mean F1 score for each language.

	Model	I	II	III	IV	V	VI	VII	Mean
English	Best System	0.835	0.913	0.978	0.873	0.882	0.908	0.889	0.897
	InfoMiner	0.819	0.886	0.946	0.841	0.803	0.884	0.867	0.864
	Random Baseline	0.552	0.480	0.457	0.473	0.423	0.563	0.526	0.496
Arabic	Best System	0.843	0.762	0.890	0.799	0.596	0.912	0.663	0.781
	InfoMiner	0.852	0.704	0.774	0.743	0.593	0.698	0.588	0.707
	Random Baseline	0.510	0.444	0.487	0.442	0.476	0.584	0.533	0.496
Bulgarian	Best System	0.887	0.955	0.980	0.834	0.819	0.678	0.706	0.837
	InfoMiner	0.786	0.749	0.419	0.599	0.556	0.303	0.631	0.578
	Random Baseline	0.594	0.502	0.470	0.480	0.399	0.498	0.528	0.496

Table 2: Macro F1 between the InfoMiner submission and human annotations for test set in all the languages. Best System is the results of the best model submitted for each language as reported by the task organisers (Shaar et al., 2021).

splitted using the Farasa Segmenter (Abdelali et al., 2016).

6 Results

When it comes to selecting the best model for each language, highest F1 score out of the evaluated models was chosen. Due to the fact that our approach uses a single model for each label, our main goal was to achieve good F1 scores using light weight models. The limitation of available resources to train several models for all seven labels itself was a very challenging task to the team but we managed to evaluate several.

As depicted in Table 1, for English, bert-base-cased model performed better than roberta-base model. For Arabic, arabert-v2-tokenized performed better than the other two models we considered. For Bulgarian, with the limited time, we could only train bert-multilingual model, therefore, we submitted the predictions from that for Bulgarian.

As shown in Table 2, our submission is very competitive with the best system submitted in each language and well above the random baseline. Our team was ranked 4^{th} in all the languages.

7 Conclusion

We have presented the system by InfoMiner team for NLP4IF-2021-Fighting the COVID-19 Infodemic. We have shown that multiple transformer models trained on different labels can be successfully applied to this task. Furthermore, we have shown that undersampling can be used to prevent the overfitting of the transformer models to the majority class in an unbalanced dataset like this. Overall, our approach is simple but can be considered as effective since it achieved 4^{th} place in the leader-board for all three languages.

One limitation in our approach is that it requires maintaining seven transformer models for the seven binary properties of this task which can be costly in a practical scenario which also restricted us from experimenting with different transformer types due to the limited time and resources. Therefore, in future work, we are interested in remodeling the task as a multilabel classification problem, where a single transformer model can be used to predict all seven labels.

Acknowledgments

We would like to thank the shared task organizers for making this interesting dataset available. We further thank the anonymous reviewers for their insightful feedback.

References

Ahmed Abdelali, Kareem Darwish, Nadir Durrani, and Hamdy Mubarak. 2016. Farasa: A fast and furious segmenter for Arabic. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 11–16, San Diego, California. 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.

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.

Ibrahim Abu El-khair. 2016. 1.5 billion words arabic corpus. In *arXiv preprint arXiv:1611.04033*.

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. 2019. Emoji powered capsule network to detect type and target of offensive posts in social media. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 474–480, Varna, Bulgaria. INCOMA Ltd.

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 Fifteenth Workshop on Semantic Evaluation*.

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.

Simon Kemp. 2021. 15.5 users join social every second (and other key stats to know). https://blog.hootsuite.com/simon-kemp-social-media/.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. In *arXiv preprint arXiv:1412.6980*.

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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. In *arXiv preprint arXiv:1907.11692*.

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).

Azzam Mourad, Ali Srour, Haidar Harmanai, Cathia Jenainati, and Mohamad Arafeh. 2020. Critical impact of social networks infodemic on defeating coronavirus covid-19 pandemic: Twitter-based study and

research directions. *IEEE Transactions on Network and Service Management*, 17(4):2145–2155.

Zesis Pitenis, Marcos Zampieri, and Tharindu Ranasinghe. 2020. Offensive language identification in Greek. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5113–5119, Marseille, France. European Language Resources Association.

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, Sarthak Gupte, Marcos Zampieri, and Ifeoma Nwogu. 2020. WLV-RIT at HASOC-Dravidian-CodeMix-FIRE2020: Offensive Language Identification in Code-switched YouTube Comments. In *Proceedings of the 12th annual meeting of the Forum for Information Retrieval Evaluation*.

Tharindu Ranasinghe and Hansi Hettiarachchi. 2020. BRUMS at SemEval-2020 task 12: Transformer based multilingual offensive language identification in social media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1906–1915, Barcelona (online). International Committee for Computational Linguistics.

Tharindu Ranasinghe, Diptanu Sarkar, Marcos Zampieri, and Alex Ororbia. 2021. WLV-RIT at SemEval-2021 Task 5: A Neural Transformer Framework for Detecting Toxic Spans. In *Proceedings of the Fifteenth Workshop on Semantic Evaluation*.

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. MUDES: Multilingual Detection of Offensive Spans. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*.

Tharindu Ranasinghe and Marcos Zampieri. 2021b. Multilingual Offensive Language Identification for Low-resource Languages. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*.

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 task on fighting the COVID-19 infodemic and censorship detection. In *Proceedings of the Fourth Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, NLP4IF@NAACL' 21, 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.

Wei Yang, Yuqing Xie, Aileen Lin, Xingyu Li, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. 2019. End-to-end open-domain question answering with BERTserini. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 72–77, Minneapolis, Minnesota. Association for Computational Linguistics.

Imad Zeroual, Dirk Goldhahn, Thomas Eckart, and Abdelhak Lakhouaja. 2019. OSIAN: Open source international Arabic news corpus - preparation and integration into the CLARIN-infrastructure. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 175–182, Florence, Italy. Association for Computational Linguistics.

A Appendix

A summary of hyperparameters and their values used to obtain the reported results are mentioned in Table 3. The optimised hyperparameters are marked with ‡ and their optimal values are reported. The rest of the hyperparameter values are kept as constants.

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 3: Hyperparameter specifications