NITK-IT_NLP at CheckThat! 2022: Window based approach for Fake News Detection using transformers

Hariharan RamakrishnaIyer LekshmiAmmal [1], Anand Kumar Madasamy[1]

[1]Department of Information Technology, National Institute of Technology Karnataka, Surathkal

Abstract

Misinformation is a severe threat to society which mainly spreads through online social media. The amount of misinformation generated and propagated is much more than authentic news. In this paper, we have proposed a model for the shared task on Fake News Classification by CLEF2022 CheckThat! Lab[1], which had mono-lingual Multi-class Fake News Detection in English and cross-lingual task for English and German. We employed a transformer-based model with overlapping window strides, which helped us to achieve 7^{th} and 2^{nd} positions out of 25 and 8 participants on the final leaderboard of the two tasks respectively. We got an F1 score of 0.2980 and 0.2245 against the top score of 0.3391 and 0.2898 for the two tasks.

Keywords

Fake news, DeBERTa, Disinformation, RoBERTa

1. Introduction

Social media can be considered the leading media through which the latest news is propagated. Even though this can be seen as the advancement of technology, it poses a severe threat to society by giving false or fake information to the users who are dependent on them. As per the literature, it's evident that the number of fake news propagates faster than authentic news [1]. During the COVID-19 pandemic, we have seen a lot of disinformation connected to the virus's cure and vaccines, which all portray the importance of combating fake news on social media.

CLEF-2022 CheckThat! [2, 3] Lab had organized a shared task named Fake News Classification[1] [4]. They had two subtasks for Fake News Detection of News Articles: a Mono-lingual task in English and a Cross-lingual task for English and German (English training data and German test data). Both tasks were to classify the articles into fake, partially fake, other, and true labels. They had provided datasets for both the tasks, from 2010 to 2022, covering several topics like elections, COVID-19, etc.

In this paper, we have used transformers-based models, which are pre-trained on a vast amount of data and are capable of understanding words based on their context. We have used the various transformer-based models like DeBERTa, RoBERTa [5, 6**?**] and used a novel window striding approach to handle the long documents which were given in the training data. This trained model was then used to get predictions for the test data.

[1]https://sites.google.com/view/clef2022-checkthat/tasks/task-3-fake-news-detection

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

hariharanrl.22@gmail.com (H. R. L.); m_anandkumar@nitk.edu.in (A. K. Madasamy)
0000-0003-2695-311X (H. R. L.); 0000-0003-0310-4510 (A. K. Madasamy)

This paper is presented as follows, section 2 about the related works, section 3 describes the dataset, section 4 explains the methodology and system description, and section 5 reports on the experiments and results. This is followed by the conclusion and some outlook on future work.

2. Related Work

The ease of social media data availability had made Fake news analysis and detection gain attention. The primary reason is that they need to reduce the spread of unauthentic news propagating through their platform. Many traditional systems are based on machine learning methods to classify fake and authentic information. Still, their performance is not so accurate because of the inability to understand the context of the data. With the advent of different deep learning methods, they have become an integral part of most designs. We have analyzed recent works incorporating deep learning models and other methods explained below.

Shahi et al. [7] have done an experimental study of COVID-19 misinformation on Twitter. They have analyzed the propagation, authors, and content of misinformation to gain early insights and categorized tweets into false, partially false, true, and other. They have also found that fake claims disseminate faster than partially false claims. In our previous work [8] we have used an ensemble of different transformer-based models which try to account for both long and normal-sized text. Shahi et al. [9] have proposed a benchmark classification dataset for fake news, which had multilingual cross-domain fact-checked news articles for COVID-19, collected from 92 fact-checking websites. Shahi [10] proposed an annotation framework of multi-modal social media data. They have presented a semi-automated framework for collecting multi-modal annotated data from social media combining machines and humans in the data compilation process. They have also implemented this framework for gathering COVID-19 misinformation. Mehta et al. [11] have proposed a transformer model for fake news classification of a specific domain dataset, including human justification and metadata for added performance. They have used multiple BERT models with shared weights between them to handle various inputs.

3. Dataset Description

There were two tasks given on the competition website[2] under Fake News Detection by CLEF2022-CheckThat! Lab, namely Multi-class Fake News Detection of news articles (English) and Cross-Lingual Task (German). We were provided with training and validation data in English (only) language, whose classwise details are given in Table 1. We have been given test data in English and German, where German was for the cross-lingual task, Table 1 provides the details of the test data.

[2]https://codalab.lisn.upsaclay.fr/competitions/4633

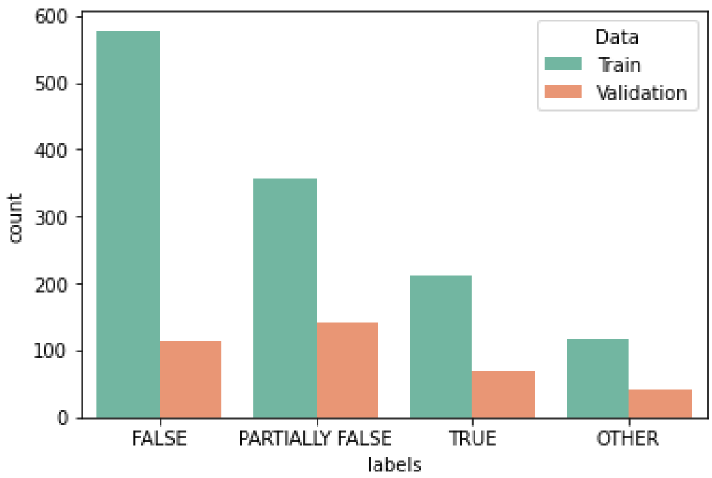

Figure 1: Distibution of Training and Validation Data

Table 1
Test Data

Language	# of articles
English	612
German	586

4. Methodology

In this shared task, we had Multi-class fake news detection of new articles in English and German, where German is a Cross-Lingual Task; both have to be classified into different labels as given in Figure 1. The proposed model had steps such as Text Preprocessing, Tokenization, and Model building, whose design is as shown in figure 2; a detailed description of each step is explained in the upcoming subsections.

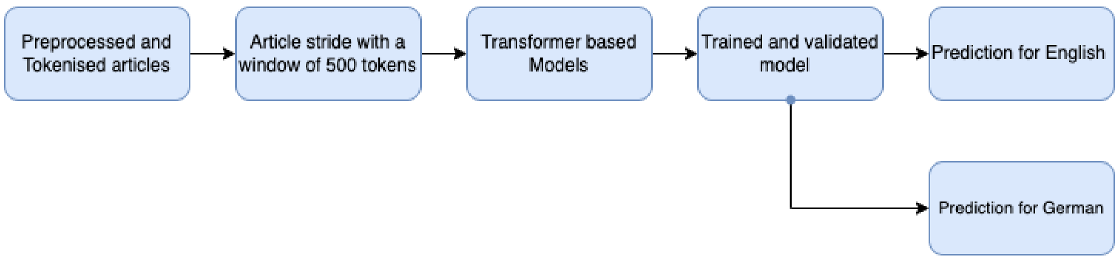

Figure 2: Model Design

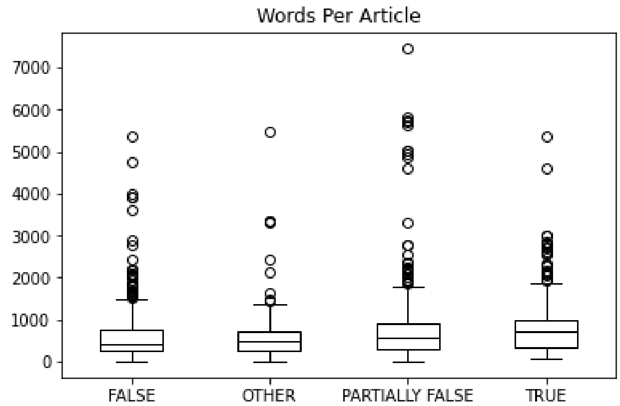

Figure 3: Box plot depicting word distribution of Training Data

4.1. Text Preprocessing and Tokenization

As the dataset was retrieved from many social media sites, preprocessing was essential before proceeding with model building. Here, we used the clean-text[3] library from python, which helped remove contents like URLs, ASCII conversions, etc. We had 1264 articles in the training data, which were mostly long articles or news. We had to convert these articles to token sequences and pass them on to the models to process them. We have used tokenization methods [4] corresponding to the particular pre-trained model being used. These tokenizers will tokenize according to the structure underlying the respective model. RoBERTa [6], BERT [12], BigBird [13], DeBERTa [?] are the different models which we have used to build our system.

4.2. Model Architecture

The transformer-based pre-trained models[5] were used for building our system. The models have been individually trained for data using the pre-trained weights, which give the probabilities for the different labels. We have finetuned the pre-trained models, which had their vocabulary and embeddings with the training data for predictions. The same model configuration is used for test data prediction. The hyperparameters were as per the standard values used for the particular pre-trained model.

4.3. Data Analysis and Modeling

We have explored different transformer-based models for building the system. As the articles were long, as shown in Figure 3, we concentrated on bringing the best out of the models, even though most models have a maximum token size of 512. The data analysis shows that the average number of tokens falls within 1000, though few articles had more than 7000. As

[3]https://pypi.org/project/clean-text/
[4]https://huggingface.co/docs/tokenizers/python/latest/
[5]http://huggingface.co/models

mentioned previously, we have used models BERT, RoBERTa, and DeBERTa, which all are capable of having a maximum token of 512. Along with them, we have tried the BigBird model, which was designed for long text.

We have used a window-based striding approach to handle the long articles. The main idea was to leverage the use of pre-trained models for the overall document without losing much information. We have divided the whole document into batches of 500 tokens, with each set of 500 tokens having the same label as the original one. We attempted different overlapping strides over the text, trained the model, and validated it, which gave good results compared to the usual approach. We tried different stride values so that the model could learn better without losing the entire context.

5. Experiments and Results

We have fine-tuned the pre-trained models for the training data using AdamW [14] and learning rate 3e-5. We used the cross-entropy loss as the loss function. The experiments were performed on Kaggle, Tesla V100 16GB GPU. The learning rate was the same for all models and trained for five epochs, with callbacks on validation loss.

5.1. Results of Different Models

This subsection will discuss the individual model results for English on the validation data. As far as the cross-lingual task is concerned, we used the best model, which gave good results on English data. The different model results are shown in Table 2. Here we can observe that DeBERTa is the best model among all the others, giving an overall F1 score of 0.9893. DeBERTa model with a stride of 0.9 means we are overlapping 90% of the tokens of different batches and combining the result as the final label for an article during training. As the DeBERTa model with stride 0.9 gave the best results, we employed the same value for different models, whose results are also included in Table 2.

Table 2
Different Model Results on Validation Data

Model	Stride	Precision	Recall	F1-Score
DeBERTa	0.5	0.9823	0.9886	0.9854
	0.7	0.9858	0.9856	0.9855
	0.9	**0.9862**	**0.9928**	**0.9893**
RoBERTa	0.9	0.9728	0.9833	0.9778
XLM-RoBERTa		0.9332	0.9454	0.9390
BigBird		0.9666	0.9699	0.9681

5.2. Result of Model on Test Data

Here we will discuss our results on test data. We have submitted the system for testing, which gave the best result on validation data to the shared task, whose results are as shown in Table 3. We got an F1 score of **0.2980**, which was placed at **7$^{\text{th}}$** position on the leaderboard with the best score of 0.3391. In the case of cross-lingual task with German data, we got an F1 score of **0.2245**, which was placed at $\mathbf{2^{nd}}$ position on the leaderboard with the best score of 0.2898. The final leaderboard for the two subtasks are published here [6]. We further analyzed the test data to improve the results after the publication of the final leaderboard, but we were not able to improve due to time constraints. The distribution of test data shows that it had more 'true' labels, which could be a problem for our model performance degradation because the distribution was different for training and validation data.

Table 3
Results on Test Data

Language	Precision	Recall	F1-Score
English	0.3804	0.3140	**0.2980** (7^{th} position)
German	0.3342	0.2785	**0.2245** (2^{nd} position)

6. Conclusion and Future Work

Nowadays, people consider social media as the primary source of news. Hence, verified and accurate information must be propagated through social media, which is not always the case. In this paper, we have concentrated on building a model that can classify long news articles from social media comprising politics, entertainment, COVID-19, etc., as fake or not. We have used a novel window-based method to fine-tune the transformer-based models like DeBERTa, RoBERTa, etc., which helped to have a model that can handle long text documents. In the future, we would like to extend our work further to improve the results so that the model can perform even on different distributions of datasets.

References

[1] S. Vosoughi, D. Roy, S. Aral, The spread of true and false news online, Science 359 (2018) 1146–1151. URL: https://science.sciencemag.org/content/359/6380/1146. doi:10.1126/science.aap9559.

[2] 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: Proceedings of the 13th International Conference of the CLEF Association: Information Access Evaluation meets Multilinguality, Multimodality, and Visualization, CLEF '2022, Bologna, Italy, 2022.

[3] 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,

[6] https://codalab.lisn.upsaclay.fr/competitions/4633#participate

K. Balog, K. Nørvåg, V. Setty (Eds.), Advances in Information Retrieval, Springer International Publishing, Cham, 2022, pp. 416–428.

[4] J. Köhler, G. K. Shahi, J. M. Struß, M. Wiegand, M. Siegel, T. Mandl, M. Schütz, Overview of the CLEF-2022 CheckThat! lab task 3 on fake news detection, in: Working Notes of CLEF 2022—Conference and Labs of the Evaluation Forum, CLEF '2022, Bologna, Italy, 2022.

[5] 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, Huggingface's transformers: State-of-the-art natural language processing, 2020. arXiv:1910.03771.

[6] 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, 2019. URL: https://github.com/pytorch/fairseq. arXiv:1907.11692.

[7] G. K. Shahi, A. Dirkson, T. A. Majchrzak, An exploratory study of covid-19 misinformation on twitter, Online Social Networks and Media 22 (2021) 100104.

[8] H. R. LekshmiAmmal, A. b Madasamy, Nitk_nlp at checkthat! 2021: Ensemble transformer model for fake news classification, in: CLEF, 2021.

[9] G. K. Shahi, D. Nandini, FakeCovid – a multilingual cross-domain fact check news dataset for covid-19, in: Workshop Proceedings of the 14th International AAAI Conference on Web and Social Media, 2020. URL: http://workshop-proceedings.icwsm.org/pdf/2020_14.pdf.

[10] G. K. Shahi, Amused: An annotation framework of multi-modal social media data, arXiv preprint arXiv:2010.00502 (2020).

[11] D. Mehta, A. Dwivedi, A. Patra, ·. M. Anand Kumar, A transformer-based architecture for fake news classification 11 (2021) 39. URL: https://doi.org/10.1007/s13278-021-00738-y. doi:10.1007/s13278-021-00738-y.

[12] M.-w. C. Kenton, L. Kristina, J. Devlin, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (2018). arXiv:arXiv:1810.04805v2.

[13] M. Zaheer, G. Guruganesh, K. A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham, A. Ravula, Q. Wang, L. Yang, et al., Big bird: Transformers for longer sequences, Advances in Neural Information Processing Systems 33 (2020).

[14] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, 2019. arXiv:1711.05101.