NITK_NLP at CheckThat! 2021: Ensemble Transformer Model for Fake News Classification

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

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

Abstract

Social media has become an inevitable part of our life as we are primarily dependent on them to get most of the news around us. However, the amount of false information propagated through it is much higher than the genuine ones, thus becoming a peril to society. In this paper, we have proposed a model for Fake News Classification as a part of CLEF2021 Checkthat! Lab[1] shared task, which had Multi-class Fake News Detection and Topical Domain Classification of News Articles. We have used an ensemble model consisting of pre-trained transformer-based models that helped us achieve 4^{th} and 1^{st} positions on the leaderboard of the two tasks. We achieved an F1-score of 0.4483 against a top score of 0.8376 in one task and a score of 0.8813 in another.

Keywords

Fake news, RoBERTa, COVID-19, Ensemble

CEUR-WS.org/Vol-2936/paper-49.pdf

1. Introduction

Nowadays, social media is the primary platform for people to get the latest news and updates happening around them, either political or entertainment or even health-related. People are more dependent on reliable information during this pandemic situation, which is propagated through social media. But here, the situation is quite different as many information are fake, which spreads faster than authentic ones [1]. Also, it can be seen that there has been an increase in concern for fake news on social media [2, 3] due to various situations in the modern world. Hence tackling fake news is more important as well as challenging in this social media age. This process isn't easy because even humans can't distinguish between fake and authentic news accurately. Thus it becomes crucial to develop an automated system for fake news identification.

CLEF-2021 CheckThat! [4, 5] Lab had organized a shared task named Fake News Classification[1]. The shared task had two subtasks called Multi-Class Fake News Detection and Topical Domain Classification of News Articles. The first subtask was to classify the articles into fake, partially fake, other, true, and the second subtask was to classify into domains health, election, crime, climate, election, education. They had provided datasets for both the tasks, from 2010 to 2021, covering several topics like election, COVID-19, etc.

In this paper, we have used transformers, which is effective for text classification because of their self-attention mechanism and a better understanding of word features. We have

[1] https://sites.google.com/view/clef2021-checkthat/tasks/task-3-fake-news-detection

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

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)

© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

used the transformer-based model [6, 7] and finetuned them for the training data to get the predictions. This paper is presented as follows, section 2 about the related works, section 3 explains the dataset, section 4 explains the Methodology and System Description, section 5 about the Experiments and Results, which follows Conclusions and Future scope.

2. Related Work

Fake information analysis and detection have gained attention because of the ease of availability of the data from social media. This is because even social media needs to curb the spreading of misinformation through their platform. We have many traditional machine learning methods to classify fake and real information. However, the performance of such systems is not that accurate because of the inability to understand the data. Deep Learning is now becoming an integral part of these fake information detection systems because of the computation capability. We have analyzed some recent works which have used deep learning models and some of the newer datasets used for fake news classification, which are explained below.

Umer et al.[8] has used a CNN-LSTM deep learning architecture to detect relative stance of fake news towards it headline. They have employed PCA and Chi-Square to reduce the dimensionality of features to predict the relative stance of a news article towards its title. Their results show a 20% improvement in the F1-Score, and PCA excels Chi-Square and state-of-the-art methods. Das et al. [9] proposed an Ensemble model for COVID-19 fake news detection for the Constraint COVID19 shared task [10]. They have used a combination of pre-trained models with a heuristic algorithm based on the username handle and link-domain in tweets. Shahi et al. [11] 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. Shahi et al. [12] 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 [13] 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. [14] 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. Liao et al. [15] proposed an integrated multi-task model for fake news detection. They have considered news topics and authors as any of them can have a higher percentage of fake news. They investigate the influence of topic labels and contextual information at the same time to improve the performance on short fake news. Manouchehri et al. [16] proposed a theoretical approach to block the influence of misinformation in social networks efficiently. The main idea was to limit the spread of misinformation as much as possible.

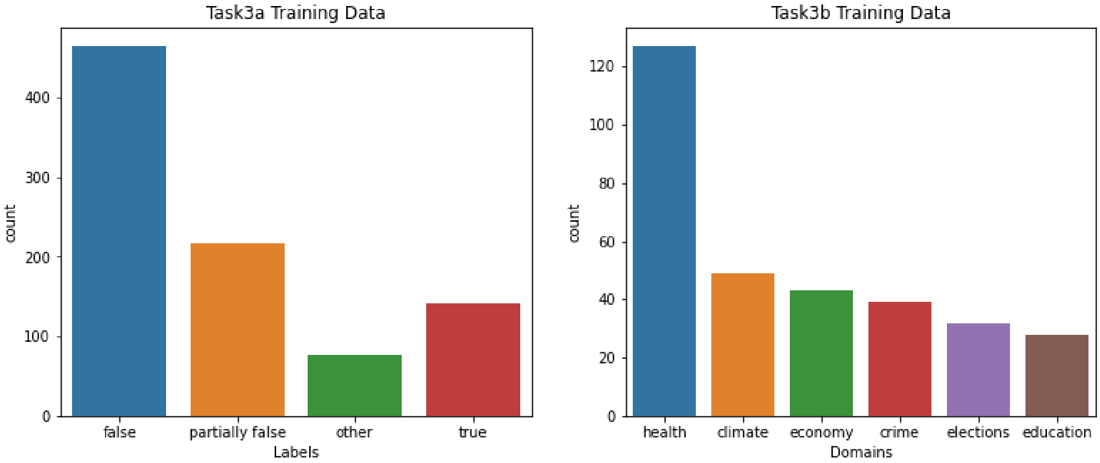

Figure 1: Dataset Distribution

3. Dataset Description

There were two tasks given on the competition website[2] under Fake News Detection by CLEF2021-CheckThat! Lab[3], namely Multi-class Fake News Detection of News Articles (Task 3a) and Topical Domain Classification of News Articles (Task 3b). The training dataset for Task3a is shown in table 1, which had 900 articles with respective labels and the testing data had 364 articles without labels. The training dataset for Task3b is shown in table 1, which had 318 articles with respective labels and the testing data had 137 articles. The label distributions of both the datasets are shown in figure 1.

Table 1
Dataset Details

Task	Data	
	Training	Testing
3a	900	364
3b	318	137

4. Methodology

We had a classification task under both Task3a and Task3b where we had to classify the articles into labels and domains as in figure 1. The proposed model had Text Preprocessing, Tokenization, Model Architecture, and Ensemble Modeling [9]. The overall design is as shown in figure 2, the

[2] https://competitions.codalab.org/competitions/31238
[3] https://gitlab.com/checkthat_lab/clef2021-checkthat-lab/-/tree/master/task3

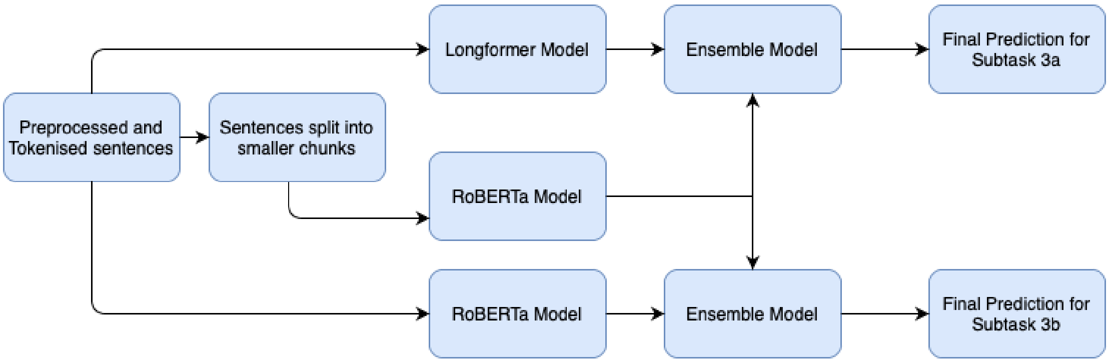
Preprocessed and Tokenised sentences
Sentences split into smaller chunks
Longformer Model
RoBERTa Model
RoBERTa Model
Ensemble Model
Ensemble Model
Final Prediction for Subtask 3a
Final Prediction for Subtask 3b

Figure 2: Overall Model Design

preprocessed sentences shown are corresponding to the subtask data. A detailed description is given in the following subsections.

4.1. Text Preprocessing

The data provided by the organizers were mainly retrieved from various news websites and articles which need to be preprocessed; for this, we used the clean-text[4] library from python, which helped in removing contents like URLs, ASCII conversions, etc.

4.2. Tokenization

We were dealing with articles that had sentences. To process the sentence data, we need to convert them to some tokens and pass them on to the model. We have used the tokenization[5] approach corresponding to the pre-trained model being used, which expects the tokens to be in an explicit structure depending upon the model. Each model will tokenize based on its structure during training on the data. We have used Longformer [17], and RoBERTa [18] models and a combination of them which are modified versions of BERT [19].

4.3. Model Architecture

We have used pre-trained models[6] as the base model for this classification task. The model has been individually trained for data using the pre-trained weights, which gives the probabilities for the different labels. As transfer learning is being used, the model has its own vocabulary and pre-trained embeddings, which is fine-tuned to get the predictions using the training data. The same tokenizer and model are used to get the predictions for test data.

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

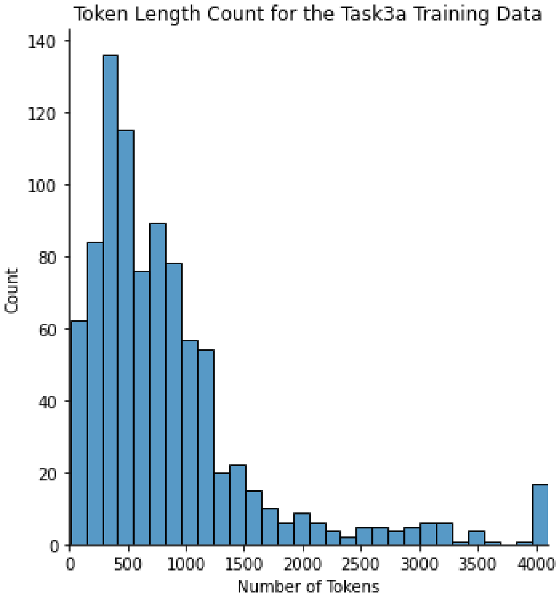

Figure 3: Token Length Distribution

4.4. Ensemble Modeling

In this method, we have used an ensemble approach to predict the labels for the testing data. We have applied RoBERTa and Longformer as the base models for our classification. Longformer was chosen as it can handle up to a maximum length of 4096, whereas RoBERTa was chosen because it outperformed BERT in most of the downstream tasks and benchmarks. We have used the model prediction vectors from these models and combined them to get the final result for our classification model. The final prediction is the sum of the individual predictions by the models and taking the maximum probability among the class labels. This helps in predicting the labels with more probability towards a particular category.

4.4.1. Data Analysis and Modeling

The longformer model was mainly chosen because as per the token distribution in the figure 3, we can see that most of the articles have tokens which are less than 3000, but we have some 20 odd number of articles having tokens of size 4096 (which is the maximum model can handle) even after truncation. Even though we lose some information, for training the longformer model, 3000 was chosen as the maximum token length for both the task as they had a similar distribution of tokens. When we tried to use the full token size, there wasn't much difference in the predictions.

The training strategy for the RoBERTa model was different as it can handle only tokens up to 512 length. To train the data with this model, we split the single article into small chunks keeping the same label for the split parts. We had split one article into 450 tokens (around 50 is left because the model splits words into subwords) each so that the model can perform well. For example, if an article had 2000 tokens, we split it into four articles having 450 tokens and one having 200 tokens with the same labels as the original one. The one drawback of this method is that while we are testing the data, articles are truncated to 450 tokens which can affect the results.

Ensemble model output is obtained using the predicted probabilities for each of the models. Instead of taking the final prediction from the models, we took the probabilities for each of the labels (Task 3a) and domains (Task 3b). These probabilities were summed and rounded up to get the final predictions. Result analysis for both the tasks will be explained in the following section.

5. Experiments and Results

We have fine-tuned the pre-trained models for the training data using AdamW [20] and learning rate 2e-5 (as recommended in the original model). We had used cross-entropy loss as the loss function. The experiments were performed on a nvidia-dgx machine with CentOS, Tesla V100 32GB GPU. The learning rate was kept the same for all the tasks, and the number of epochs varied from 10-15 with callbacks on validation loss. The same parameters were used for both the tasks as they had similar data; only classification labels were different.

5.1. Results of Individual Models

In this section, we will discuss the individual model results for both tasks. We have used a training and validation split of 0.20 for the training data, and the results of validation data are shown in the table2. The longformer model with maximum token length 3000 gave an F1-score (weighted average was taken as it will consider the proportion for each label in the dataset) of 0.60 for task 3a. Even though longformer had more tokens, it couldn't get better than that of RoBERTa. Hence we came up with an ensemble model as our proposed model whose results are explained in the following section.

While coming to task 3b, we have tried a different method because longformer could not perform well for the token length of size 3000. Here we have used the RoBERTa model (with maximum token length of 450) directly on the entire article without splitting for the long sequences whose results are shown in the table 2. Then we split the data into chunks of 450 as explained in the last section, whose results are shown as RoBERTa_SplitText in the table 2. These two models were used as an ensemble model for the task 3b classification because the model which was trained on chunks of text could have a better understanding.

Table 2
Individual Model Results

Task	Model	Accuracy	Precision	Recall	F1-Weighted
3a	Longformer	0.62	0.58	0.62	0.60
	RoBERTa	0.64	0.62	0.64	0.62
3b	RoBERTa	0.91	0.90	0.91	0.91
	RoBERTa_SplitText	0.93	0.93	0.93	0.93

5.2. Results of Ensemble Models

The final model results for test data are as shown in table 3. As discussed in the methodology section, we have used an ensemble of the fine-tuned Longformer and RoBERTa as the final prediction for the test data. We can see from the results that for task 3b ensemble model outperformed, but the Longformer performed well for task 3a. The final results of our proposed model for task 3b achieved 1^{st} position in the leaderboard of the competition (position on the leaderboard is given in brackets corresponding to the task). Eventhough the ensemble model could perform well for task 3b, we believe some mislabelled articles were causing the ensemble model to underperform in task 3a. Also, we have given F1- weighted score in the previous section because of the unbalanced class labels. Here, we have provided the F1-macro score because it was used as the final score for the leaderboard by the organizers.

Table 3
Final Model Results on Test Data

Task	Model	Accuracy	Precision	Recall	F1-Macro
3a	Longformer	0.52	0.50	0.47	**0.45(4)**
	Longformer+RoBERTa_SplitText	0.42	0.33	0.34	0.31
3b	RoBERTa	0.83	0.82	0.76	0.77
	RoBERTa+RoBERTa_SplitText	0.91	0.91	0.87	**0.88(1)**

6. Conclusions and Future Scope

In this modern social media age, we have to be vigilant in the rapid spreading of fake information, which can have immense ramifications on our day-to-day lives. In this paper, we have focussed on building a model to classify the news articles from social media comprising politics, entertainment, COVID-19, etc., as fake or not. We have fine-tuned transformer-based models Longformer and RoBERTa to predict the news articles. Moreover, our results got improved when we implemented the ensemble combination of these models. In the future, this method can be extended to learn more features with different models used in combination, and also, we would evaluate our model on generic Fake News 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] F. Monti, F. Frasca, D. Eynard, D. Mannion, M. M. Bronstein, Fake news detection on social media using geometric deep learning, 2019. arXiv:1902.06673.

[3] D. P. Calvillo, B. J. Ross, R. J. B. Garcia, T. J. Smelter, A. M. Rutchick, Political Ideology Predicts Perceptions of the Threat of COVID-19 (and Susceptibility to Fake News About It), Social Psychological and Personality Science 11 (2020) 1119–1128. URL: https://doi.org/10.1177/1948550620940539. doi:10.1177/1948550620940539.

[4] P. Nakov, G. D. S. Martino, T. Elsayed, A. Barrón-Cedeño, R. Míguez, S. Shaar, F. Alam, F. Haouari, M. Hasanain, N. Babulkov, A. Nikolov, G. K. Shahi, J. M. Struß, T. Mandl, The CLEF-2021 checkthat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news, in: D. Hiemstra, M. Moens, J. Mothe, R. Perego, M. Potthast, F. Sebastiani (Eds.), Advances in Information Retrieval - 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 - April 1, 2021, Proceedings, Part II, volume 12657 of *Lecture Notes in Computer Science*, Springer, 2021, pp. 639–649. URL: https://doi.org/10.1007/978-3-030-72240-1_75. doi:10.1007/978-3-030-72240-1_75.

[5] P. Nakov, G. Da San Martino, T. Elsayed, A. Barrón-Cedeño, R. Míguez, S. Shaar, F. Alam, F. Haouari, M. Hasanain, W. Mansour, B. Hamdan, Z. S. Ali, N. Babulkov, A. Nikolov, G. K. Shahi, J. M. Struß, T. Mandl, M. Kutlu, Y. S. Kartal, Overview of the CLEF-2021 CheckThat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news, in: Proceedings of the 12th International Conference of the CLEF Association: Information Access Evaluation Meets Multiliguality, Multimodality, and Visualization, CLEF '2021, Bucharest, Romania (online), 2021.

[6] 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.

[7] A. Vaswani, G. Brain, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention Is All You Need, Technical Report, ????

[8] M. Umer, Z. Imtiaz, S. Ullah, A. Mehmood, G. S. Choi, B. W. On, Fake news stance detection using deep learning architecture (CNN-LSTM), IEEE Access 8 (2020) 156695–156706. doi:10.1109/ACCESS.2020.3019735.

[9] S. D. Das, A. Basak, S. Dutta, A Heuristic-driven Ensemble Framework for COVID-19 Fake News Detection (2021). URL: https://competitions.codalab.org/competitions/26655http://arxiv.org/abs/2101.03545. arXiv:2101.03545.

[10] P. Patwa, S. Sharma, P. Y. Srinivas, V. Guptha, G. Kumari, M. S. Akhtar, A. Ekbal, A. Das, T. Chakraborty, Fighting an infodemic: COVID-19 fake news dataset, 2020. URL: www.boomlive.in. arXiv:2011.03327.

[11] 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.

[12] 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.

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

[14] 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.

[15] Q. Liao, H. Chai, H. Han, X. Zhang, X. Wang, W. Xia, Y. Ding, An Integrated Multi-Task Model for Fake News Detection, IEEE Transactions on Knowledge and Data Engineering 4347 (2021) 1–12. doi:10.1109/TKDE.2021.3054993.

[16] M. A. Manouchehri, M. S. Helfroush, H. Danyali, A Theoretically Guaranteed Approach to

Efficiently Block the Influence of Misinformation in Social Networks, IEEE Transactions on Computational Social Systems (2021) 1–12. doi:10.1109/TCSS.2021.3059430.

[17] I. Beltagy, M. E. Peters, A. Cohan, Longformer: The Long-Document Transformer, 2020. URL: https://github.com/allenai/longformer. arXiv:2004.05150.

[18] 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.

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

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