BERT Based Fake News Detection Model

Yang Zhang[1], Yi Shao[1], Xuan Zhang[2], Wenbo Wan[1], Jing Li[1,*], and Jiande Sun[1,*]

[1] *Shandong Normal University, China*
[2] *Shandong Police College, China*

Abstract

Fake news has become one of the global risks that endanger people's production and life. With the development of electronic information and artificial intelligence, the technology of automatic detection of fake news has also appeared. Inspired by these issues, the main goal of this article is to develop a multi-labelling multi-class detector fake news test algorithm. We use the dataset from the MediaEval 2022 fake news detection challenge, which consists of 1913 texts from Twitter. We propose a Bert-based multi-label multi-category fake news detection model to solve the text classification task in the MediaEval 2022 fake news detection challenge. We first convert the 9 categories of tweet tags into 27 categories, and then input them into the model for classification. The Accuracy Score of our model is 0.495, the Micro F1 Score is 0.926, and the Macro F1 Score is 0.538.

1 INTRODUCTION

The MediaEval Fake News Detection Challenge has been running for several years [1, 2], the purpose of this year's fake news detection challenge is to design a multi-label and multi-category model to classify Tweets [3]. The correct label may be 0 to multiple, and there are three categories of the category. The way we use is to divide the 9 categories of 3 tags into 27 categories. The tags of each classification are 0 or 1. As a result, the data set label is obtained, and the problem is transformed into a multi-label classification problem. We use BERT to obtain text features, use the full connection layer to obtain the classification results, and finally evaluate the accuracy of the model.

2 RELATED WORK

Currently news contains multiple modal data, such as text, images, and videos. We can detect the authenticity of news and classify it only through text. Ma Jing and others applied deep learning technology to fake news detection for the first time [4]. This method inputs each sentence of the news into RNN [5], and uses the hidden layer vector of the cyclic neural network to represent the news information. Then input the hidden layer information into the classifier to get the classification result. FENG et al. Using convolutional neural network modeling articles for the first time [6], the various posts of the news event were mapped to vector space, and then each POST vector was spliced into a matrix. Embedded vector into the classifier to get the classification results. Ma Jing and others applied Multi-Task Thought to fake News Test [7]. This method is to combine fake news testing tasks and position classification tasks into a multi-tasking model, and use RNN as Backbone to train two tasks. Ma Jing and others applied the idea of fighting training into the test of fake news [8]. This method uses the generator to transform the rumor into non -rumor, expand the training data, and then enter the generated news and original news to the judgment Fake news testing in the device. VAIBHAV and others transform fake news detection issues into graph classification problems [9]. This method models news articles as a sentence with a sentence with a sentence, with the similarity between sentences, and using GCN [10] fusion diagram. The information between the middle nodes obtains the node embedding vector, embeds the figure pool by the node vector, and enters into the classifier to obtain the detection result.

3 APPROACH

We use the BERT pre-trained model to obtain text features, and use the fully connected layer as a classifier. After we get the text features, this news representation is then passed through a fully connected neural network for classification.

3.1 Textual Feature Extractor

It uses Bidirectional Encoder Representations from Transformers(BERT) [11] to represent words and sentences in a way that best captures underlying semantic and contextual meaning. We use BERT-base version that has 12 encoding layers(termed as transformer blocks). It takes as input a sequence of words that keep moving up the stack. Each layer applies self-attention, and

MediaEval'22: Multimedia Evaluation Workshop, January 13–15, 2022, Bergen, Norway and Online
*Corresponding author.
2021317099@stu.sdnu.edu.cn (Y. Zhang); 2021020981@stu.sdnu.edu.cn (Y. Shao); zx@sdpc.edu.cn(X. Zhang); wanwenbo@sdnu.edu.cn(W. Wan); lijingjdsun@hotmail.com (J. Li); jiandesun@hotmail.com (J. Sun)

passes its results through a feed-forward network, and then hands it off to the next encoder. A detailed description of the textual feature extractor is shown in Figure 1.

3.2 Input and Classification

In order to facilitate data processing, we convert the labels according to the following rules, "0,0,1" means label 1, "0,1,0" means label 2, and "1,0,0" means label 3. The original 9-category labels are converted to 27 categories. The text representation and labels are obtained from this, and the fully connected layer is used as a classifier for classification.

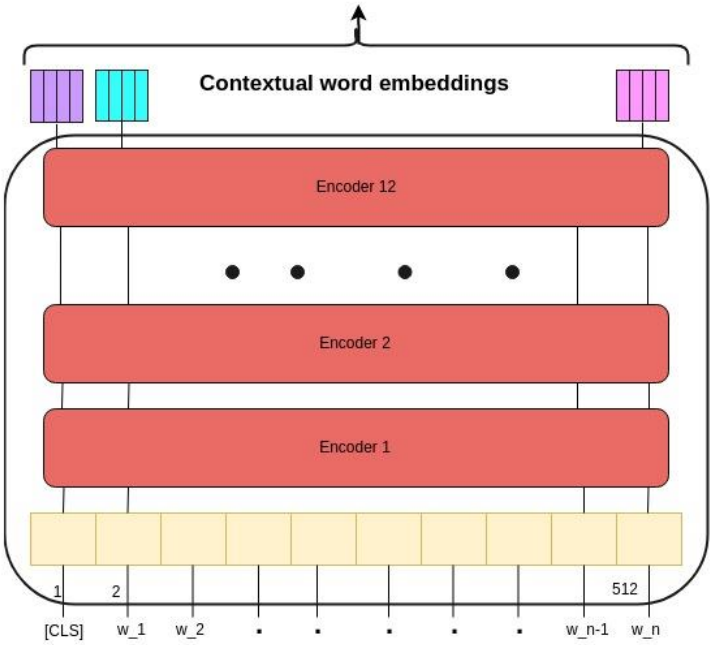

Figure 1: Textual Feature Extractor.

4 RESULTS AND ANALYSIS

4.1 Dataset

We divide the training set and test set according to the ratio of 8:2. The labels of the data set are converted according to the rules before being input into the model, and the labels are restored after the model output is obtained.

Table 1: **Dataset**

Train	Test
1530	383

4.2 Training

We use ADAM optimizers, the initial learning rate is 0.00001, and the loss function is BcewithlogitsLoss. The proposed network is trained using Intel core i5 processor, GeForce GTX950 GPU, 8GB RAM and Platform Pytorch. This model is evaluated with Accuracy Score, Micro F1 Score and Macro F1 Score.

4.3 Results

Table 2 is the overall evaluation result of the data set. Table 3 is the evaluation result of each classification of the data set. It can be seen from the results of Table 3 that the accuracy of each label classification is very different, resulting in the low accuracy of the overall classification.

Table 2: **Overall evaluation of dataset**

Accuracy Score	F1 Score (Micro)	F1 Score (Macro)
0.495	0.926	0.538

Table 3: **Evaluation of each label of dataset**

Label1	Label2	Label3	Label4	Label5	Label6	Label7	Label8	Label9
0.326	0.455	0.506	0.408	0.421	0.509	0.760	0.703	0.296

5 CONCLUSIONS

We use the classic pre- trained BERT model and full-connected layer mode for multi-label text classification, and no pre-processing operation of the data set text is not performed. In order to meet the input and output of the model, we divided the label from 9 into 27 categories, and the final Accuracy Score was 0.495, and the F1 score was 0.538. The final result failed to reach the expected goal. After merging 27 tags into 9 categories, each type of score is large, the highest ones reach 0.7 or even higher, and the lowest is less than 0.3. Therefore, we believe that the classification of a single model and multi-labelling multi-class into more labels is inappropriate for this challenge. Especially decomposing the number of labels into more quantities may cause the performance of the model to decrease.

REFERENCES

[1] Pogorelov K, Schroeder D T, Burchard L, et al. FakeNews: Corona Virus and 5G Conspiracy Task at MediaEval 2020[C]//MediaEval. 2020.

[2] Pogorelov K, Schroeder13 D T, Brenner S, et al. FakeNews: Corona Virus and Conspiracies Multimedia Analysis Task at MediaEval 2021[C]//Multimedia Benchmark Workshop. 2021: 67.

[3] Maria Authorsen and Jacques de Coauthor. Cool Task: Challenges, Dataset and Evaluation. Proc. of the MediaEval 2022 Workshop, Bergen, Norway and Online, 12-13 January 2023.

[4] Ma J, Gao W, Mitra P, et al. Detecting rumors from microblogs with recurrent neural networks[J]. 2016.

[5] Karpathy A. The unreasonable effectiveness of recurrent neural networks[J]. Andrej Karpathy blog, 2015, 21: 23.

[6] Yu F, Liu Q, Wu S, et al. A Convolutional Approach for Misinformation Identification[C]//IJCAI. 2017: 3901-3907.

[7] Ma J, Gao W, Wong K F. Detect rumor and stance jointly by neural multi-task learning[C]//Companion proceedings of the the web conference 2018. 2018: 585-593.

[8] Ma J, Gao W, Wong K F. Detect rumors on twitter by promoting information campaigns with generative adversarial learning[C]//The world wide Web conference. 2019: 3049-3055.

[9] Vaibhav V, Annasamy R M, Hovy E. Do sentence interactions matter? leveraging sentence level representations for fake news classification[J]. arXiv preprint arXiv:1910.12203, 2019.

[10] Scarselli F, Gori M, Tsoi A C, et al. The graph neural network model[J]. IEEE transactions on neural networks, 2008, 20(1): 61-80.

[11] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT:pre-training of deep bidirectional transformers for languageunderstanding," CoRR, vol. abs/1810.04805, 2018. [Online].Available: http://arxiv.org/abs/1810.04805