Siamese Networks In Trigger Detection Task

Notebook for PAN at CLEF 2023

Yunsen Su, Yong Han*, Haoliang Qi

FoshanUniversity,Foshan,China

Abstract

The trigger detection task involves providing a piece of text and determining which warning labels it belongs to. In the PAN@CLEF 2023 competition, the trigger detection task requires analyzing a text segment ranging from 50 to 6000 words and identifying which of the 32 warning labels apply to that segment. To address this task, a method based on RoBERTa-based Siamese networks and convolutional neural networks was proposed.The text is divided into two segments, with the first segment containing the first 505 words and the second segment containing the last 505 words. These segments are separately input into the Siamese RoBERTa models. The outputs of RoBERTa undergo pooling, resulting in two embeddings. These two embeddings are then subjected to one-dimensional convolutional operations. The convolutional results are fed into a classifier for multi-label classification. Using this approach, the method achieved the following results on the test dataset: mac_F1 = 0.35, mac_p = 0.544, mac_r= 0.298, mic_F1 = 0.753, mic_p = 0.798, mic_r = 0.712, and sub_acc = 0.622.

Keywords

Multi-Label Text Classification, Pre-trained language model, Siamese Networks

1. Introduction

In trigger detection, the goal is to assign trigger warning labels to documents that contain potentially distressing or painful (trigger) content [1]. The PAN@CLEF 2023 trigger detection task requires the development of software or models to determine whether a document contains trigger content. To increase the challenge, the PAN 2023 trigger detection task models the detector as a multi-label classification task, where each document is assigned all relevant trigger warnings. This task has wide applications in fields such as information retrieval [2,3], web mining, question-answering systems, and sentiment analysis. However, due to the numerous label categories, complex relevance relationships, and imbalanced sample distributions, building a simple and effective multi-label text classifier presents significant challenges [4].

For multi-label classification, there are four main methods [5]: problem transformation methods [6,7], algorithm adaptation methods [8], ensemble methods [9,10], and neural network models [11,12]. Researchers in the fields of machine learning and natural language processing have made significant efforts in developing MLTC (Multi-Label Text Classification) methods in each of these aspects [13]. Traditional machine learning algorithms for multi-label text classification can be primarily categorized into problem transformation and algorithm adaptation. The former transforms the multi-label classification problem into a series of single-label classification problems, while the latter improves existing single-label algorithms to make them applicable to multi-label data. However, traditional methods heavily rely on feature engineering and are susceptible to noise, resulting in suboptimal predictive performance [14].In recent years, with the emergence of transformer-based deep learning models, significant contributions have been made in the field of natural language

[1]CLEF 2023 – Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece
EMAIL: suyunsen2023@gmail.com (A. 1); qihaoliang@fosu.edu.cn(A. 2); hanyong2005@fosu.edu.cn (A. 3)
ORCID: 0009-0003-3609-0749 (A. 1); 0000-0003-1321-5820 (A. 2); 0000-0002-9416-2398 (A. 3)

processing. More and more researchers are utilizing transformer-based models like BERT and GPT for multi-label classification tasks and achieving promising results. BERT-BCE [15] has shown good performance in multi-label classification tasks. It utilizes the pretrained language model BERT to encode input sentences and employs Binary Cross Entropy loss for multi-label classification [13].

This article proposes improvements to BERT-BCE by addressing the issue of text length exceeding the input limit of ROBERTA. To overcome this limitation, the approach selects the first 505 words and the last 505 words of the text. This ensures that more semantic information from the text is captured.For the first 505 words and the last 505 words of the text, a pair of Siamese ROBERTA models are employed. The embeddings are obtained by taking the average of the last layer of ROBERTA. To better combine the embeddings, a one-dimensional convolutional neural network is applied to the embeddings, resulting in the final text embeddings. Finally, these embeddings are passed through a classifier for multi-label classification.By using this approach, the proposed method aims to leverage the benefits of BERT while accommodating longer texts by splitting them and using a siamese network with a convolutional layer for embedding aggregation. The final embeddings are then utilized for multi-label classification.

2. Model framework

In this section we will discuss our model.

Our base model is based on a Siamese network using ROBERTA. For a given text, we split it into two segments, which are separately input into ROBERTA. After applying pooling, we obtain embeddings for the two segments. These embeddings are then processed through convolution to generate the final embedding for the text. Finally, this embedding is fed into a classifier. The detailed diagram of the model is shown in Figure 1.

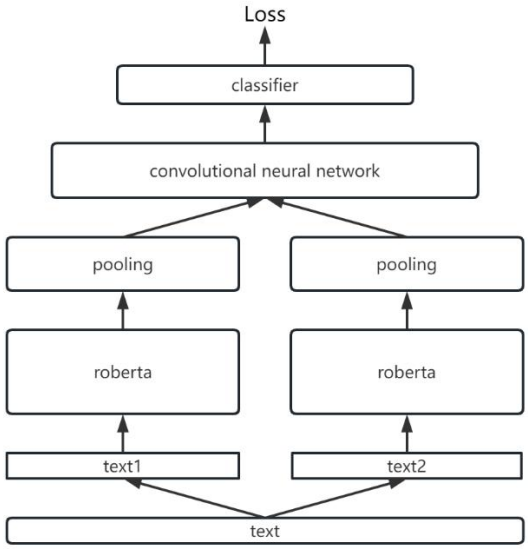

Figure 1 Model framework diagram of our method

2.1. Text Processing

For the PAN@CLEF 2023 trigger detection task, each text segment is given with a total length of 50-6000 words, and the text content is provided in HTML format. First, I clean the HTML text content by removing the HTML tags, transforming the text provided by PAN into plain text format. Since the input limit of ROBERTA is 512 tokens, and most of the given texts have a total length greater than ROBERTA's input limit, I choose to truncate the text by selecting the first 505 words and the last 505 words. It's important to note that the total length of the text can be either less than or equal to 505 words or greater than 505 words but less than 1010 words. In the case where the total length is less than or equal to 505 words, I will generate two identical segments of the text and feed

them into the Siamese network. In the case where the total length is greater than 505 words but less than 1010 words, there will be an overlap between the latter half of the first 505 words and the first half of the last 505 words, containing the same text content.

2.2. Pooling

When classifying tasks of ROBERTA, many people will directly take [cls] as the embedding of the entire text, but it may not achieve a good effect in many tasks. Therefore, this paper conducts a pooling on the output of the last layer of ROBERTA. The two most common operations of pooling are meanpooling and maxpooling. Meanpooling is adopted in this paper. Meanpooling is to calculate the average value of output corresponding to each token.

2.3. Convolutional neural network

The convolution neural network is mainly applied in the field of images, but it is also used in the field of text. Inspired by this, I did not directly add the two texts of the embedding or average of the two embedding, but convolved the two embedding into the one-dimensional convolution neural network for output. To represent the final embedding of the text.

2.4. Classifier

Our classifier is composed of several simple linear layers and activation functions. The detailed composition order is linear layer, Tanh activation function layer, dropout layer, linear layer. A total of four layers make up the classifier.

3. Experimet

In this section I will describe our experiment. For this experiment, I used the hardware provided by the school, which is a shared server. The allocated time for using the school's public server is one day, so my model was trained for only 15 hours, completing a total of 4 epochs.

3.1. Data set

The trigger detection task for PAN@CLEF 2023 is given a data set containing fan fiction retrieved from archiveofourown.org (AO3). Each piece is between 50 and 6,000 words long and is assigned one to many trigger warnings. The tag set contains 32 different trigger warnings and has a long-tail frequency distribution, meaning that some tags are very common and most tags are increasingly rare. Our training dataset contains 307,102 examples, 17,104 for validation and 17,040 for testing.

3.2. Detail of experiment

For the types of data sets given, the training set has 3070,12 examples, the verification set is 17,104 examples, and 17,040 are used for testing. Use the pre-trained ROBERTA to train on the training set and validate on the verification set.

Our code is implemented on pythorch and we use the huggingface architecture. roberta-base,the optimizer we used is the Adam optimizer. The learning rate of the optimizer is $\beta = 2e^{-5}$. The loss function is pytorch's MultiLabelSoftMarginLoss, and the batch of size is 32. The experimental results of our proposed model are shown in Table 3-1.

Table 3-1
Experimental results table

mac_f1	mac_p	mac_r	mic_f1	mic_p	mic_r	sub_acc
0.347	0.521	0.296	0.75	0.791	0.713	0.616

3.3. Ablation experiment

3.3.1. Cls as expression

To investigate whether the method used in this paper is effective for this task, it was used as a text expression for "cls", and the results are shown in Table 3-2 below.

Table 3-2
"cls" Experimental results table

mac_f1	mac_p	mac_r	mic_f1	mic_p	mic_r	sub_acc
0.265	0.427	0.208	0.701	0.808	0.623	0.568

3.3.2. One pooling

In order to verify the effectiveness of siamese networks and convolutional neural networks, we selected only the first 505 words of the text as input to the model for experimentation. The experimental results of One pooling are shown in Table 3-3.

Table 3-3
One pooling results table

mac_f1	mac_p	mac_r	mic_f1	mic_p	mic_r	sub_acc
0.269	0.504	0.206	0.711	0.816	0.63	0.578

Based on our ablation experiment results, it can be observed that without using pooling and without employing siamese networks and convolutional networks, the experimental results are inferior to the method proposed in this paper. Therefore, the method proposed in this paper is effective in improving the performance of multi-label tasks.

4. Conclusions

This paper proposes a method based on Siamese networks and convolutional neural networks to tackle the trigger detection task of PAN@CLEF 2023. We split a given text into two segments, which are then encoded by the Siamese ROBERTA network. The output from ROBERTA is fed into a one-dimensional convolutional neural network to generate the final embedding for the text. This embedding is subsequently passed through a classifier for multi-label classification.

We have achieved better results than the baseline provided by the PAN@CLEF 2023 Trigger Detection task. I believe our model performs well on the dataset provided by the PAN@CLEF 2023 Trigger Detection task. Due to limited hardware resources, my model was trained for only 4 epochs. In the future, I will explore training for more epochs to observe if my model can further improve its performance.

5. Acknowledgments

This work is supported by the National Natural Science Foundation of China (No.62276064).

6. References

[1] W, Magdalena,Schr\"oder, Christopher and Borchardt, Ole and Stein, Benno and Potthast, Martin. Trigger Warnings: Bootstrapping a Violence Detector for FanFiction.2022.
[2] Gopal, Siddharth, and Yiming Yang. Multilabel Classification with Meta-Level Features.*Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2010.
[3] Myagmar, B.; Li, J.; Kimura, S. Cross-domain sentiment classifification with bidirectional contextualized transformer language models. IEEE Access 2019, 7, 163219–163230.
[4] Duan, Lihua and You, Qi and Wu, Xinke and Sun, Jun. Multilabel Text Classification Algorithm Based on Fusion of Two-Stream Transformer. Electronics. 2022,pp.2138.dio: 10.3390/electronics11142138.
[5] Yang, Pengcheng and Sun, Xu and Li, Wei and Ma, Shuming and Wu, Wei and Wang, Houfeng. SGM: Sequence Generation Model for Multi-label Classification. International Conference on Computational Linguistics.2018.
[6] Boutell, Matthew R. and Luo, Jiebo and Shen, Xipeng and Brown, Christopher M. Learning multi-label scene classification. Pattern Recognition.doi: 10.1016/j.patcog.2004.03.009.
[7] Overview, An. Multi-Label Classification.
[8] Li, Li and Wang, Houfeng and Sun, Xu and Chang, Baobao and Zhao, Shi and Sha, Lei. Multi-label Text Categorization with Joint Learning Predictions-as-Features Metho.2015.doi: 10.18653/v1/d15-1099.
[9] Tsoumakas, Grigorios and Vlahavas, Ioannis. Random k-Labelsets: An Ensemble Method for Multilabel Classification. Machine Learning: ECML 2007,Lecture Notes in Computer Science.2007.doi: 10.1007/978-3-540-74958-5_38.
[10] Szymański, Piotr and Kajdanowicz, Tomasz and Kersting, Kristian. How is a data-driven approach better than random choice in label space division for multi-label classification?.2016,pp.282.doi: 10.3390/e18080282.
[11] Chen, Guibin and Ye, Deheng and Xing, Zhenchang and Chen, Jieshan and Cambria, Erik. Ensemble application of convolutional and recurrent neural networks for multi-label text categorization. 2017 International Joint Conference on Neural Networks (IJCNN.2017.doi: 10.1109/ijcnn.2017.7966144.
[12] Baker, Simon and Korhonen, Anna. Initializing neural networks for hierarchical multi-label text classification. BioNLP 2017.2017.doi: 10.18653/v1/w17-2339.
[13] Quanjie, Han and Xinkai, Du and Yalin, Sun and Chao, Lv. Label Dependencies-aware Set Prediction Networks for Multi-label Text Classification.2023.
[14] Duan, Lihua and You, Qi and Wu, Xinke and Sun, Jun. Multilabel Text Classification Algorithm Based on Fusion of Two-Stream Transformer. Electronics. 2022,pp.2138. doi: 10.3390/electronics11142138.
[15] Devlin, Jacob and Chang, Ming-Wei and Lee, Kenton and Toutanova, Kristina. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North. 2019. doi: 10.18653/v1/n19-1423.