

A Dual-model Classification Based on RoBERTa for Trigger Detection

Notebook for PAN at CLEF 2023

Guiyuan Cao¹, Zhongyuan Han^{*1}, Haojie Cao¹, Ximin Huang¹, Zhengqiao Zeng¹, Yaozu Tan¹, Jiyin Cai¹ and Xu Sun²

¹Foshan University, China

²Heilongjiang Institute of Technology, China

Abstract

In order to predict whether a text contains much potentially uncomfortable or harmful information, our team proposed a dual-model classification method based on RoBERTa for trigger detection. After analyzing the dataset, we found that the frequency of the "pornographic-content" label was higher than the other 31 labels. To avoid this label causing unbalanced prediction results, we tried to use two models to predict the labels that this text might contain. This method involves two rounds of sampling from the original dataset. In the first round, only the label with the highest frequency is retained, forming Dataset A; in the second round, the remaining 31 labels are retained, forming Dataset B. Using these two datasets, we separately trained two RoBERTa models. The two fine-tuned models are used to predict the test documents, and the results of the two models are combined to obtain multiple labels as predictive labels. The method we used obtained a macro f1 score of 0.225 on the test set of the Trigger Detection Task at PAN 2023.

Keywords

Trigger Detection, dual-model classification, RoBERTa

1. Introduction

With the development of the Internet, various types of information have emerged online. Among this information, there are many useful ones, but also many potentially uncomfortable or harmful ones. Therefore, PAN at CLEF 2023 proposed a Trigger Detection task [1] to investigate this issue. PAN at CLEF 2023 requires participants to develop software or models to determine whether a document contains trigger content [2], with trigger content including pornographic content, violence, death, sexual assault, and 32 other labels[3]. For this task, our team proposes a dual-model classification approach for trigger detection based on RoBERTa[4], aiming to improve the model's accuracy. This approach trains two models based on RoBERTa to determine

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece

✉ caoguiyuan2020@163.com (G. Cao); hanzhongyuan@gmail.com (Z. Han^{*}); caohaojie0322@163.com (H. Cao); huangximin2020@163.com (X. Huang); zhengqiaozeng@163.com (Z. Zeng); tanyaozu2023@163.com (Y. Tan); caijiyin0904@163.com (J. Cai)

🆔 0009-0008-3561-851X (G. Cao); 0000-0001-8960-9872 (Z. Han^{*}); 0000-0002-8365-168X (H. Cao); 0009-0003-6028-7592 (X. Huang); 0009-0000-5415-349X (Z. Zeng); 0009-0003-0184-3050 (Y. Tan); 0009-0001-6418-908X (J. Cai)



© 2023 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)

whether a document fits specific labels. If a document’s score for a specific label exceeds a threshold, then the content of that document matches the label. Otherwise, the content of that document does not match the label.

Section 2 of this paper describes the specific methods used, including the data processing process, a brief introduction to the models, and the model training process. Section 3 describes the predicted results obtained using this method and analyzes those results. Section 4 presents the conclusions drawn from this experiment.

2. Method

This section provides an overview of the classification methods and specific model training used. Section 2.1 details the process of data processing; Section 2.2 describes the overall structure of the model; Section 2.3 explains the key points of model training.

2.1. Data Processing

The dataset is sourced from CLEF 2023PAN [5]. In the dataset, there is a text corpus where each text is assigned at least one warning trigger. In the original dataset provided by 2023PAN, there are many HTML tags present. Therefore, we first cleaned the HTML tags. Secondly, while analyzing the cleaned dataset, we discovered a long-tail frequency issue, where some tags have high frequency while others have low frequency. To address this problem, we recorded the most frequent tags, the least frequent tags, and the tags with frequencies in between. Next, we processed the texts that contained the most frequent tags as follows: we extracted 80,000 samples from the original dataset, where 40,000 samples matched the most frequent tags and another 40,000 did not. We then processed these sampled data by dividing long texts into subsets of no more than 512 words while maintaining the same tags as their parent sets to obtain the processed training dataset A.

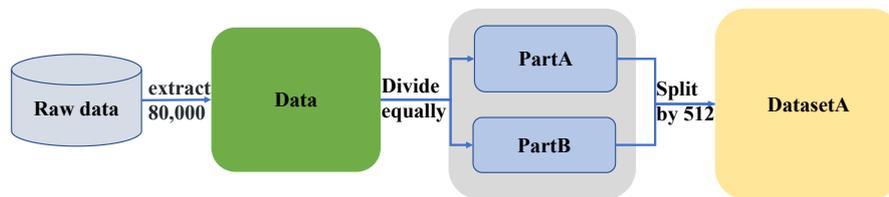


Figure 1: Process for Obtaining Dataset A

Finally, for the remaining 31 tags, we employed a sampling method of oversampling and undersampling, which balanced the occurrence frequency of each tag to some extent. The oversampling rule is as follows: if a tag appears less than 20 times in the original dataset, the data with that tag will be replicated and added to the dataset six times. If a tag appears between 20 and 200 times in the original dataset, the data with that tag will be replicated four times and added to the dataset. The undersampling rule is as follows: if a tag appears more than 3,000 times in the original dataset, only 3,000 data with that tag will be retained. After the aforementioned sampling process, we obtained the training dataset B.

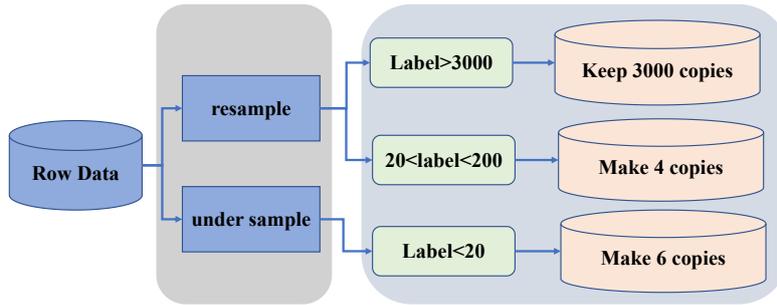


Figure 2: Process for Obtaining Dataset B

2.2. Model

Our team conducted two rounds of sampling on the original dataset. In the first round, dataset A was obtained through random sampling. In the second round, resampling and undersampling techniques were used to create a new dataset, dataset B. These two datasets were used as training data for the RoBERTa models, resulting in the training of two new models, model A and model B. In this process, we used RoBERTa, which is an improved version based on BERT. RoBERTa incorporates more model parameters, more training data, and larger batch sizes compared to BERT. It undergoes training on a scale that is an order of magnitude larger than BERT, which takes a longer time. This enables RoBERTa representations to generalize better to downstream tasks and exhibit superior performance compared to the original BERT.

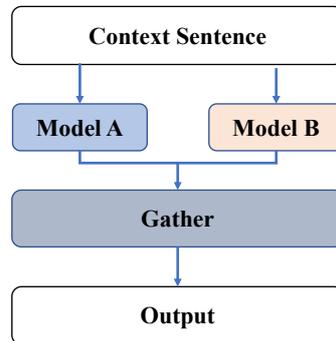


Figure 3: Overall Prediction Process

2.3. Model Training

First, on the hardware platform of Intel(R) Xeon(R) Gold 6348 CPU @ 2.60GHz + NVIDIA A800-SXM4-80GB, the batch size for each training round is set to 32, and the data is fed into RoBERTa for training. After five rounds of training, Model A and Model B are generated. For Model A, each training round takes about 10 nutes, while prediction time takes about 20 minutes. For Model B, each training round takes 10 hours, and prediction time takes about 20 minutes.

Then, based on the output results of Model A and Model B, our team has established the following criteria to evaluate the matching degree between the text and labels: "In the data processing step, since each long text has been divided into several smaller text segments, the matching degree between each long text and the label depends on the average score of all relevant small texts and labels. This average score is called the total score. If the total score of a label exceeds the threshold of 0.5, the label is assigned a value of 1; otherwise, it is assigned a value of 0." According to this criterion, if the output of the model is 1, the file corresponds to a certain label; if it is 0, the file does not correspond to a certain label.

Since Model B needs to learn multiple labels, its output results need to be connected to a fully connected layer, forming a mapping relationship between the output results of Model B and the 31 labels mentioned in Section 2.1 of the data processing part. If this mapping relationship meets the criteria of the text-label fitting mentioned above, it can be considered that the text matches some of the 31 labels.

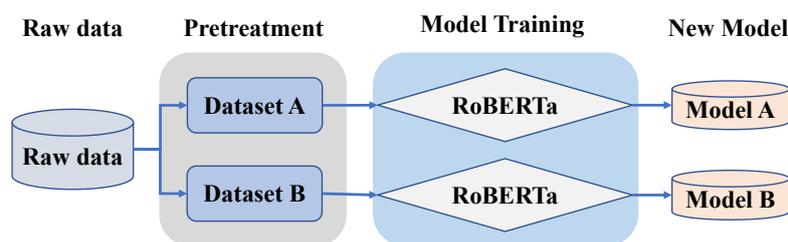


Figure 4: Model Training Process

3. Result

Based on the above experiments, we obtained the following results on the CLEF 2023PAN dataset ¹.

Table 1
Predicted results for different datasets

	mac_f1	mic_f1	sub_acc
validation set	0.223	0.623	0.323
test set	0.225	0.616	0.317

The analysis shows that the difference in macro f1 scores between the test set and the validation set is 0.002, and the difference in micro f1 scores is 0.007. Both differences are less than 0.01. Therefore, the model can predict the possible labels contained in a document stably. Although this method has high stability, there is still a lot of room for improvement in its accuracy. Therefore, in the future, we will continue to explore how to use dual models to improve the matching between documents and labels.

¹Pan23 trigger detection (1.1) [data set], 2023. URL: <https://doi.org/10.5281/zenodo.7612628>.

4. Conclusion

Regarding the question of determining whether a document contains potentially uncomfortable or harmful information, our team proposed a dual-model classification method based on RoBERTa. This method involves setting two sampling rules for the dataset based on the RoBERTa model. The training sets obtained from these two sampling methods are used to train RoBERTa, resulting in Model A and Model B. By combining the predictions of both models on the same text, we achieved a macro f1 score of 0.225 and an accuracy of 0.317 on the CLEF 2023PAN dataset. Finally, we conclude that using a dual-model approach has a positive effect on determining whether a document contains one or more harmful information.

5. Acknowledgments

This work is supported by the Heilongjiang Province Philosophy and Social Science Research Planning Project (No.20TQB065).

References

- [1] M. Wiegmann, M. Wolska, M. Potthast, B. Stein, Overview of the Trigger Detection Task at PAN 2023, in: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, CEUR-WS, 2023.
- [2] M. Wiegmann, M. Wolska, C. Schröder, et al., Trigger warning assignment as a multi-label document classification problem, in: Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023.
- [3] M. Wolska, C. Schröder, O. Borchardt, et al., Trigger warnings: Bootstrapping a violence detector for fanfiction, CoRR abs/2209.04409 (2022). URL: <https://doi.org/10.48550/arXiv.2209.04409>. doi:10.48550/arXiv.2209.04409. arXiv:2209.04409.
- [4] Y. Liu, M. Ott, N. Goyal, et al., RoBERTa: A robustly optimized BERT pretraining approach, CoRR abs/1907.11692 (2019). URL: <http://arxiv.org/abs/1907.11692>. arXiv:1907.11692.
- [5] J. Bevendorff, I. Borrego-Obrador, M. Chinea-Ríos, et al., Overview of pan 2023: Trigger detection, in: Experimental IR Meets Multilinguality, Multimodality and Interaction. Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2023), Lecture Notes in Computer Science, Springer, 2023, pp. 518–526.