Trigger Warning Labeling with RoBERTa and Resampling for Distressing Content Detection

Notebook for PAN at CLEF 2023

Haojie Cao[1],Zhongyuan Han*[1],Guiyuan Cao[1],Ruihao Zhu[1],Yongqi Liang[1],Siman Liu[1] Minhua Huang[1] and Haihao Yu[2]

1 Foshan University,China
2 Heilongjiang Institute of Technology,China

Abstract

The objective of the "Trigger Detection 2023" task is to identify labels that contain potentially inappropriate or distressing content and trigger warnings accordingly. The trigger detector is approached as a multi-label document classification task . This paper presents a methodology based on the utilization of the RoBERTa model in conjunction with a classifier. During the model training process, the classifier and loss function are iteratively adjusted to optimize the parameter accuracy. Additionally, resampling and undersampling techniques are employed to enhance the overall precision. Finally, in the data testing phase, the accuracy of assigning warning triggers to each document is evaluated.

Keywords

Trigger Detection, RoBERTa, Resampling and Undersampling, Classifier

1. Introduction

Documents often contain content that may be distressing or inappropriate, requiring the provision of trigger warnings. The "Trigger Detection 2023"[1] focuses on assigning appropriate trigger warnings within documents. The primary goal is to determine whether the author is disseminating inappropriate content and subsequently trigger a warning mechanism to prevent such behavior. We propose a classifier based on the RoBERTa model to address this task. The model is trained using RoBERTa to assess the probability of a given label triggering an alert. To enhance the model's accuracy, techniques such as resampling and undersampling are introduced. These techniques revaluate labels with excessively high or low probabilities. These methods aim to mitigate any imbalance and improve the overall accuracy of the model.

2. Methodology

This submission introduces a voting-based transformer with a focus on recall rather than precision, achieving the third-highest macro F1 score.Our training was conducted using the NVIDIA A800 training equipment, with a duration of 10 hours per round. The prediction time, on the other hand, took approximately 30 minutes.

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece
EMAIL: caohaojie0322@163.com (H. Cao); hanzhongyuan@gmail.com (Z. Han) (*corresponding author); caoguiyuan2020@163.com(G. Cao);brotherblaozhu281@gmail.com (R. Zhu); liangy76789@outlook.com(Y. Liang); liusiman648@163.com(S. Liu); minhuahuang2023 @126.com(M. Huang);
ORCID: 0000-0002-8365-168X (H. Cao); 0000-0001-8960-9872 (Z. Han); 0009-0008-3561-851X (G. Cao); 0009-0000-7521-946X (R. Zhu); 0009-0006-1697-0900 (Y. Liang); 0009-0009-1859-4429 (S. Liu); 0009-0003-4000-3062 (M. Huang)

The proposed approach involves dividing the training documents into smaller segments or chunks. Each chunk is assigned the labels from its source document. A single classifier based on RoBERTa is then trained on each chunk. If a label is assigned to more than three tenths of the chunks, it is then assigned to the entire document.To ensure a balanced training dataset, dynamic over- and under-sampling techniques were applied. The class of "pornography" was under-sampled to 5,000 instances, while other labels were reduced to 2,000 instances. Furthermore, examples with rare labels were replicated either 2 or 8 times to enhance their representation in the training data.

2.1 Model

The model using RoBERTa and multi-model voting performed the task (the architecture is described below in Figure 1). It was obtained after each data was sliced based on a specific length and fed into RoBERTa for training. Finally, we employed the model to analyze the data and determine whether the documents contain certain types of potentially uncomfortable or distressing content and label them accordingly.

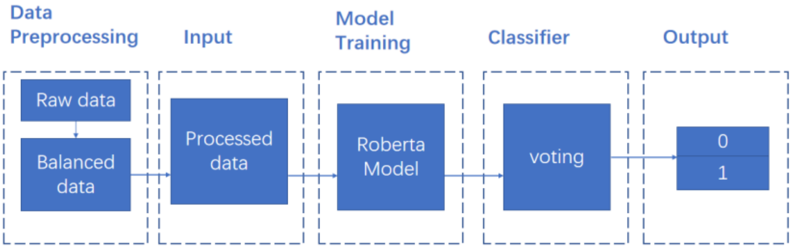

Figure 1: Architecture of the whole model.

2.2 Data processing

The dataset is provided by PAN'23 and the dataset is a jsonl file. The training dataset contains 307,102 examples, of which 17,104 are used for validation and 17,040 for test splitting.

Statistics were performed on the labels in the dataset, which revealed a large variation in the numb er of samples. To address this issue, a dual combination of resampling and undersampling was selecte d for the dataset. Based on the statistical analysis, the dataset implemented specific resampling rules to address the issue of large variation in the number of samples for different labels. Specifically, labels with a total number greater than 2000 underwent undersampling, while labels with a total number less than 100 were subject to resampling

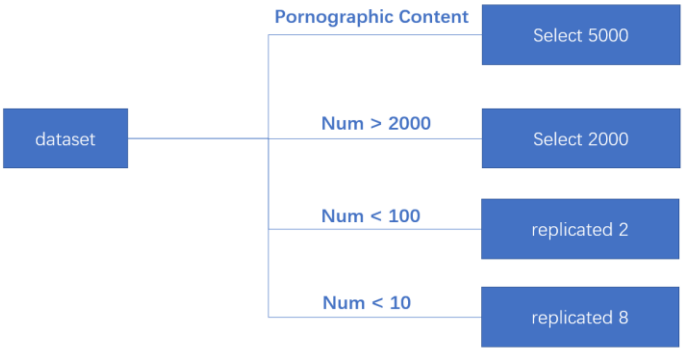

Figure 2: Resampling and undersampling

For labels with a sample size of less than 10, the samples are replicated 8 times, and for labels with a sample size between 10 and 100, the samples are replicated twice. Although these sample sizes are relatively large, they are still not large enough to match the sample sizes of the other categories, so some of the samples are replicated to increase their weight in the training process.

Next, the labels with too many total numbers (labels with high sample sizes) are undersampled in order to maintain a relative balance with the sample sizes of other categories. The specific undersampling rules are as follows:

For labels with more than 2000 samples, 2000 samples are randomly selected as undersampling. By randomly selecting some samples for undersampling, the number of these samples can be reduced to be close to the number of samples in other categories.

For the samples labeled "pornographic-content", 5000 samples were randomly selected as undersampled. This is because pornographic content is a sensitive category that usually requires more attention and processing, and by undersampling it can reduce its proportion in the dataset.

This double combination of resampling and undersampling can solve the problem of imbalance in the dataset to a certain extent and make the number of samples in each category more balanced, thus improving the performance of the model in training and testing.

In the data provided by PAN'23, we took a series of pre-processing steps to prepare the dataset. First, we performed a removal operation for the HTML tags in the text to reduce the interference of HTML tags. The processed dataset was then sliced by length 512, and the sliced data were fed into RoBERTa's model for analysis based on this rule.

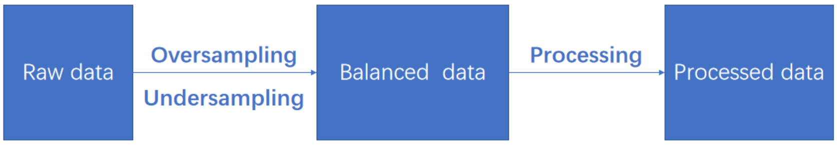

Figure 3: The whole process of Data processing

2.3 Input and Model Training

The sliced data was fed to RoBERTa for training, with a Linear() function used to output 32 neuron nodes. This involved outputting 32 neuron nodes, i.e. accessing a 32-dimensional fully connected layer to transform the 768-dimensional vector into probabilities that corresponded to the 32 labels. bertMLCF was utilized for classification.

Because the number of samples is very large, if a very high learning rate is used, a large echos will lead to overfitting, so a learning rate of 1e-5 and two rounds of epochs are used here, and the batch_size is set to 64 in order to improve the training speed.

2.4 Classifier and Output

The sliced text was combined for a summation and averaging operation. A threshold was set, and if the result was greater than the threshold, the label was considered present. If it was less than the threshold, the label was considered absent.

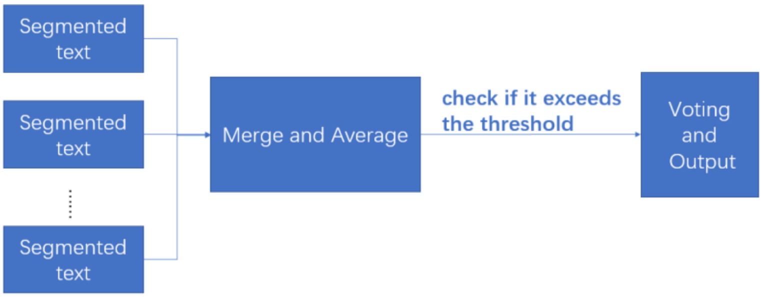

Figure 4: Merge and average

After experimentation, the threshold was set to 0.3

Initially, with the threshold set to 0.5, the macro_f1 score on the validation set was 0.212. Adjustments were made to the threshold by increasing or decreasing it with a precision of 0.1.

Eventually, it was determined that the highest macro_f1 score was achieved at a threshold of 0.3, resulting in a score of 0.245 on the validation set.

3. Results

Finally, the results of the competition are presented in the following table. Our team's mac_f1, mic_f1 and sub_acc reached 0.228, 0.557 and 0.183 respectively, ranking third.

This result is attributed to two things; first, the data from the competition was resampled and undersampled to adjust the data scaling; second, the learning rate was adjusted to determine the label type more accurately.

Table 1:
Results comparison

POS	TEAM	MAC_F1	MIC_F1	SUB_ACC
1	pan23-transformers	0.352	0.737	0.589
2	pan23-supergirl	0.35	0.753	0.622
3	baseline	0.301	0.689	0.531
4	ourTeam	0.228	0.557	0.183

4. Conclusion

This paper describes our team's assignment of trigger warning labels for resolving documents that c ontain potentially uncomfortable or distressing (triggering) content.This task is considered as a multi-label document classification task by our team. In our experiments, the model is trained using RoBER Ta to determine the probability of it being a trigger warning label or not. To enhance the model's accu racy, the data from the competition is resampled and undersampled, and the learning rate is adjusted.

5. Acknowledgements

This work is supported by Heilongjiang Province Philosophy and Social Sciences Research Planning Project (No. 21TQB117)

6. References

[1] M. Wiegmann, M. Wolska, C. Schr¨oder, O. Borchardt, B. Stein, and M. Potthast, “Pan23 trigger detection,” Feb. 2023. [Online]. Available: https://doi.org/10.5281/zenodo.7612628

[2] M. Wiegmann, M. Wolska, M. Potthast, and B. Stein, “Overview of the Trigger Detection Task at PAN 2023,” in Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, M. Aliannejadi, G. Faggioli, N. Ferro, and M. Vlachos, Eds. CEUR-WS, Sep. 2023.

[3] M. Wiegmann, M. Wolska, C. Schr¨oder, O. Borchardt, B. Stein, and M. Potthast, “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). Toronto, Canada: Association for Computational Linguistics, Jul. 2023.

[4] D. Zhang, T. Li, H. Zhang, and B. Yin, “On data augmentation for extreme multi-label classification,”CoRR,vol.abs/2009.10778,2020.[Online].Available:https://arxiv.org/abs/2009.10778

[5] Wolska, C. Schr¨oder, O. Borchardt, B. Stein, and M. Potthast, “Trigger warnings: Bootstrapping a violence detector for fanfiction,” CoRR, vol. abs/2209.04409, 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2209.04409

[6] M. Fr¨obe, M. Wiegmann, N. Kolyada, e tal. “Continuous Integration for Reproducible Shared Tasks with TIRA.io,” in Advances in Information Retrieval. 45th European Conference on IR Research (ECIR 2023), ser. Lecture Notes in Computer Science, J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, and A. Caputo, Eds. Berlin Heidelberg New York: Springer, Apr. 2023, pp. 236–241.

[7] J. Bevendorff, I. Borrego-Obrador, M. Chinea-R´ıos,e tal. “Overview of PAN 2023: Authorship Verification, MultiAuthor Writing Style Analysis, Profiling Cryptocurrency Influencers, and Trigger Detection,” in Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2023), ser. Lecture Notes in Computer Science, Springer, Sep. 2023 pp.518 526.