Application of R-Drop in Authorship Verification

Notebook for PAN at CLEF 2023

Jiajun Lv, Yong Han* and Qian Dong

Foshan University, Foshan, China

Abstract

The text describes our method to use a pre-trained model for the PAN2023 author authentication task. Author authentication is a task to judge whether two documents are written by the same author based on comparing the writing styles of two documents. This paper uses the BERT pre-training model to extract the interaction between text pairs. The R-Drop regularization[6] method is used to improve the model's generalization performance.

Keywords

Authorship Verification, Dropout, R-Drop, Pre-trained model

1. Introduction

The text describes our approach to implementing the author authentication sharing task on PAN 2023[1]. This task is to determine whether the same author writes two texts by comparing the writing styles of two documents. In contrast to previous versions of the task, the PAN 2023 task emphasizes the ability of the author authentication method to handle different forms of expression in written and spoken language[2].

The data set[1]includes written language, which includes essays and emails, and spoken language, which includes interviews and voice transcripts. The data set consists of different discourse type forming a pair of texts and predicting whether the same author writes them.

The pre-trained language BERT[5] model has performed very well in natural language processing in recent years. We use the BERT pre-trained language model[5]to extract stylistic features between texts and use these features to judge whether the same author writes text pairs[3]. In addition, to prevent model training from overfitting, we use an R-Drop[6] method to enhance the robustness of the model to dropout[4] by adding a regularization term.

2. Datasets

The author validation task for PAN 2023[1] is based on a set of texts from an open dataset of more than 100 authors with the unrestricted subject matter and a level of formality that can vary within specific discourse types[8]. The dataset consists of two pairs of texts of different discourse types, each pair being assigned a unique identifier to distinguish between pairs of the same author and pairs of different authors. In addition, each text provides metadata about the discourse type. The training and test datasets have the same structure and have similar properties. However, their author sets are separate. Since the text length of email and interview texts can be very short, each text belonging to these discourse types concatenates different messages.

PAN2023 author verification task data is quite challenging. Since it is a text pair with intersecting types, the text length of email and interview text may be concise, and it is difficult to find similarities in writing styles between different text pairs.

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

EMAIL: lvjiajun.96@gmail.com (A. 1); hanyong2005@fosu.edu.cn (A. 2) (*corresponding author); dongqian@fosu.edu.cn (A. 3)

ORCID: 0000-0002-8755-5310 (A. 1); 0000-0002-9416-2398 (A. 2); 0000-0002-9416-2398 (A. 3)

Table 1
The number of texts corresponding to different discourse types

Type	Quantity
Essays	93
Emails	450
Interviews	275
Speech transcriptions	68

In addition, author-specific and subject-specific information has been replaced by corresponding entity tags. In spoken speech types, additional labels indicate non-verbal vocalizations (e.g., coughing, laughing). The table below tallies the types and number of entity labels and emoticons in the texts of all training datasets.

Table 2
The number of entity labels and emoticons

symbol type	types	quantities
author-specific and topic-specific	393	10,690
emoji	52	141

3. Method

3.1. Text data preprocessing

In the text of the data set, there are entity labels, which are difficult to be encoded by the pre-training model BERT. As data cleaning, the entity labels are removed. In addition, since <nl> and <new> labels represent special meanings in the data set, we believe they can be used as characteristics to distinguish the writing styles of different authors. Choose to replace it with a special placeholder symbol.

Emoji in a text can be used as a style feature[4]. Compared with the tasks of previous versions, PAN2023 authors verify that text emoji in the data set are more evenly distributed and have no apparent features. Considering that the coding of adding emoji into the pre-training language model is more complex and performance improvement is difficult to guarantee, to clean the data, These symbols and emoticons are cleared from the text.

Different types of discourse have different text lengths. Most texts of Essay class will exceed the maximum number of coding BERT of the pre-training model. In the text, the method of directly intercepting the first 256 tokens is adopted to compose the training text for training.

The PAN2023 Author recognition task dataset contains 8836 text pairs. Depending on how the data set is divided, 2,650 text pairs (30%) will be used to verify model training performance, while 6,186 text pairs (70%) will be used to train model parameters.

3.2. Neural Network Architecture

A pair of text is directly truncated, $text1$ $text2$, a pair of text no longer than 256 tokens, and is spliced together with a particular match to get $text$

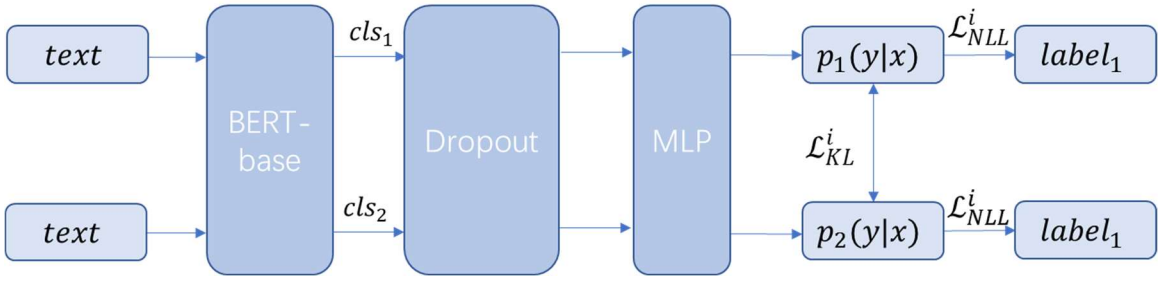

Figure 1: Neural Network Architecture

We use the BERT-base model, as shown in the figure above. An input $text$ passes through the model twice and gets two different probability distributions, $p_1(y|x)$ and $p_2(y|x)$. Since the Dropout randomly drops some neurons each time, $p_1(y|x)$ and $p_2(y|x)$ are different prediction probabilities obtained through two different subnetworks. Calculate $p_1(y|x)$ and $p_2(y|x)$

bidirectional KL divergence loss functions to make the output samples of the two models consistent with each other:

$$\mathcal{L}^i_{KL} = \frac{1}{2}(\mathcal{D}_{KL}(\mathcal{P}^w_1(y_i \mid x_i) \parallel \mathcal{P}^w_2(y_i \mid x_i)) + \mathcal{D}_{KL}(\mathcal{P}^w_2(y_i \mid x_i) \parallel \mathcal{P}^w_1(y_i \mid x_i))) \quad (1)$$

In addition, the maximum likelihood loss is:

$$\mathcal{L}^i_{NLL} = -\log \mathcal{P}^w_1(y_i \mid x_i) - \log \mathcal{P}^w_2(y_i \mid x_i) \quad (2)$$

The final training loss function is:

$$\mathcal{L}^i = \mathcal{L}^i_{NLL} + \alpha \cdot \mathcal{L}^i_{KL} \quad (3)$$

4. Experiments and Results

4.1. Experiment setup

We employ the bert-base-case pre-training model as the feature encoder, consisting of 12 layers, 768 hidden units, 12 attention heads, and 110 million parameters. The training process utilizes a batch size of 32, with a maximum encoder length of 512. We employ the Adam optimizer with a learning rate of 2e-5 and apply a dropout rate of 0.1, 0.3, and 0.5, respectively.

In order to train Multilayer perceptron and update BERT model parameters, we iterate over 30 epochs to ensure proper convergence with the training dataset. Additionally, we implement a rule that terminates the training if the epoch loss fails to decrease for five consecutive iterations.

We extract the vectors from the final layer of BERT, excluding the special tokens [CLS] and the last terminator token. An input text is fed into the pre-training BERT model twice, and two different [CLS] inserts are obtained. The [CLS] inserts are fed into the Dropout layer respectively and then connected into a fully connected layer. The probabilities $\mathcal{P}_1(y_i \mid x_i)$ and $\mathcal{P}_2(y_i \mid x_i)$ are obtained, and the bidirectional KL divergence loss L_{KL} and $\mathcal{P}_1(y_i \mid x_i)$ are calculated. $\mathcal{P}_2(y_i \mid x_i)$ is the maximum likelihood loss L_{NLL} of the real label, and finally, the two losses are combined to obtain the final loss and used for backpropagation

4.2. Evaluation

The PAN-2023 task uses five evaluation indicators[7], namely F1-score and AUC, c@1 , F_ 0.5u and Brier.

F1: The F1 score is a commonly used indicator for evaluating the performance of classification models, combining the comprehensive performance of model accuracy and recall.

auc: ROC curve is a graphical tool often used to evaluate the performance of dichotomous models. ROC curve shows the relationship between true case rate and false positive case rate under different classification thresholds.

c@1: a variant of the conventional F1-score, which rewards systems that leave difficult problems unanswered.

F_0.5u: a measure that puts more emphasis on deciding same-author cases correctly.

Brier: the complement of the well-known Brier score, for evaluating the goodness of binary classification probabilistic classifiers.

4.3. Results

We tested the model's performance in this article on 2650 text pairs separated from the training data set and our model on the PAN23[8] author authentication test data set. This test data set contains 2,650 text pairs, including text written by 56 authors.

Table 3

Evaluation scores of models trained with different dropout rates.

dropout rates	F1	auc	c@1	F_0.5u	Brier
0.1	0.843	0.842	0.842	0.85	0.842
0.3	0.815	0.825	0.824	0.851	0.824
0.5	0.786	0.786	0.786	0.795	0.786

We have submitted three versions on the TIRA platform[9]: "radioactive-copyright," "cold-rotor," and "tender-bugle." These versions represent models trained at dropout rates of 0.1, 0.3, and 0.5, respectively.

Table 4

Line 1 is the results of radioactive-copyright. Line 2 is the results of tender-bugle. Line 3 is the results of cold-rotor.

run name	dropout rates	F1	auc	c@1	F_0.5u	Brier
radioactive-copyright	0.1	0.504	0.553	0.553	0.54	0.553
cold-rotor	0.3	0.501	0.551	0.551	0.537	0.551
tender-bugle	0.5	0.465	0.55	0.55	0.524	0.55

5. Conclusion

In this paper, we introduce our method of author identification verification on PAN2023. Then, we use the pre-training language model BERT to extract the natural features of text pairs and integrate these features to judge whether the same author writes a pair of texts. We use R-Drop method to constrain the control of model freedom. In order to improve the generalization performance of the model, the results show that this method does not extract the characteristics of different authors' writing styles well, which leads to poor performance on open data sets.

6. Acknowledgement

This work is supported by the National Natural Science Foundation of China (No.62276064) and the Natural Science Foundation of Guangdong Province, China (No.2022A1515011544).

7. References

[1] Efstathios Stamatatos, Mike Kestemont, Krzysztof Kredens, Piotr Pezik, Annina Heini, Janek Wendorff, Benno Stein, and Martin Potthast. Overview of the Authorship Verification Task at PAN 2022. In Guglielmo Faggioli, Nicola Ferro, Allan Hanbury, and Martin Potthast, editors, CLEF 2023 Labs and Workshops, Notebook Papers, September 2023.

[2] Mike Kestemont, Efstathios Stamatatos, Enrique Manjavacas, Janek Bevendorff, Martin Potthast, and Benno Stein, Overview of the Cross-Domain Authorship Verification Task at PAN 2021. Working notes of CLEF 2021 - Conference and Labs of the Evaluation Forum.

[3] Adaku Uchendu, Thai Le, Kai Shu, and Dongwon Lee. 2020. Authorship Attribution for Neural Text Generation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 8384–8395, Online. Association for Computational Linguistics.

[4] Ghiasi, G., Lin, T. Y., & Le, Q. V. (2018). Dropblock: A regularization method for convolutional networks. Advances in neural information processing systems, 31.

[5] Devlin J., Chang M.W., Lee K., et al. Bert: Pre-training of deep bidirectional transformers for language understanding//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019, 1: 4171-4186

[6] Wu, L., Li, J., Wang, Y., Meng, Q., Qin, T., Chen, W., ... & Liu, T. Y. (2021). R-drop: Regularized dropout for neural networks. Advances in Neural Information Processing Systems, 34, 10890-10905.

[7] Stamatatos, E., Kredens, K., Pęzik, P., Heini, A., Bevendorff, J., Potthast, M., & Stein, B. (2023). Overview of the Authorship Verification Task at PAN 2023. In CLEF 2023 Labs and Workshops, Notebook Papers. Conference and Labs of the Evaluation Forum (CLEF 2022). CEUR-WS.org.

[8] Bevendorff, J., Borrego-Obrador, I., Chinea-Ríos, M., Franco-Salvador, M., Fröbe, M., Heini, A., Kredens, K., Mayerl, M., Pęzik, P., Potthast, M., Rangel, F., Rosso, P., Stamatatos, E., Stein, B., Wiegmann, M., Wolska, M., & Zangerle, E. (2023). Overview of PAN 2023: Authorship Verification, Multi-Author Writing Style Analysis, Profiling Cryptocurrency Influencers, and Trigger Detection. In A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, A. Giachanou, D. Li, M. Aliannejadi, M. Vlachos, G. Faggioli, & N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2023). Thessalonikki, Greece: Springer.

[9] Fröbe, M., Wiegmann, M., Kolyada, N., Grahm, B., Elstner, T., Loebe, F., Hagen, M., Stein, B., & Potthast, M. (2023). Continuous Integration for Reproducible Shared Tasks with TIRA.io. In J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, & A. Caputo (Eds.), Advances in Information Retrieval. 45th European Conference on IR Research (ECIR 2023) (pp. 236-241). Berlin Heidelberg New York: Springer.