Team bingezzzleep at PAN: A Writing Style Change Analysis Model Based on RoBERTa Encoding and Contrastive Learning for Multi-Author Writing Style Analysis

Notebook for the PAN Lab at CLEF 2024

Qida Wu[1], Leilei Kong[1,†] and Zhanhong Ye[1]

[1]Foshan University, Foshan, Guangdong, China

Abstract

The objective of writing style change detection is to identify the positions in multi-author documents where the author's writing style shifts. This paper elaborates on the strategy of using a comparative learning method adopted in the PAN 24 shared task and demonstrates the research progress in handling author style changes by utilizing RoBERTa as a pre-trained model encoder. By integrating comparative learning into the training of the encoder, its ability to capture subtle features between text pairs is enhanced. Additionally, data augmentation techniques were employed to expand the training set, thereby enhancing the model's generalization capability. On the official dataset across three difficulty levels, our method achieved F1 scores of 0.985, 0.818, and 0.807, respectively, where two exceeded the baseline method using comparative learning from the previous year, and one was on par with it.

Keywords

Style Change Detection, Contrastive Learning, Sentence Representation

1. Introduction

The task of multi-author writing style analysis requires participants to identify all changes in writing style at the paragraph level within a given text. For each pair of consecutive paragraphs, the goal is to assess whether there is a change in style between them [1]. This analysis is crucial for applications such as plagiarism detection, authorship attribution, and is instrumental in revealing the true identity of the author, preventing academic misconduct and copyright infringement, and upholding academic integrity and intellectual property protection.

To effectively tackle this challenge, traditional methods often rely on text features such as vocabulary, syntax, and structure to capture the author's writing style [2]. More recently, with the advancement of pre-trained language models, new approaches have emerged that leverage these models for style analysis [3], fine-tuning them on specific datasets to meet the task's needs.

In PAN 24, the task focuses more on addressing intrinsic style change detection, rather than relying too much on thematic information as a signal for style change [4]. For this reason, the competition provides datasets at varying levels of complexity, with progressively less thematic information at each level, encouraging participants to focus more on the subtle intrinsic changes in style rather than heavily on thematic information.

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

† corresponding author

1362990744wuqida@gmail.com (Q. Wu); kongleilei@fosu.edu.cn (L. Kong); chinwang.yip@gmail.com (Z. Ye)

0009-0002-9532-937X (Q. Wu); 0000-0002-4636-3507 (L. Kong); 0009-0001-4094-006X (Z. Ye)

2. Related work

Pre-trained Language Models (PLMs) are an important technology in the field of Natural Language Processing (NLP). They learn rich linguistic representations by pre-training on large-scale text datasets and can then be fine-tuned for specific downstream tasks to improve performance. In past style change detection tasks, Zhang et al. [5] used a pre-trained BERT [6] model to estimate the similarity of writing styles, and Lin et al. [7] achieved the best results in 2022 by using three BERT-type pre-trained models through ensemble learning.

However, the task of writing style change detection is not a semantic matching problem, but a problem of capturing writing style features of different texts. Therefore, contrastive learning can be used, which can effectively capture and compare stylistic differences between different texts, thereby identifying changes in the author's writing style. In PAN 23, Chen et al. [8] applied a contrastive learning-based method combined with the pre-trained DeBERTa model. Ye et al. [9] trained using a specialized contrastive learning loss and achieved the best results on hard datasets that year. In this study, the contrastive learning method of CoSENT [10] was adopted, which effectively guides the model to learn a more accurate and distinguishable semantic representation space, thereby enhancing the model's ability to capture subtle semantic differences in text and improving its understanding of complex linguistic phenomena.

3. Method

In this study, we employed the method of contrastive learning to optimize the training process of the RoBERTa [11] encoder. The goal is to utilize the cosine distance between two sentences to implement these text classification tasks. The core feature is to bring the embeddings of paragraphs written by the same author closer together in vector space, while pushing the embeddings of paragraphs by different authors further apart.

First, given a dataset D, the method of contrastive learning can categorize the data into a set of positive examples and a set of negative examples, where the set of positive examples is defined as Ω_{pp} and the set of negative examples is defined as Ω_{nn}. We aim to have the embeddings of the samples in the Ω_{pp} set be as close as possible in vector space, and conversely for the Ω_{nn} set.

In the testing phase, we obtain the embeddings of the two paragraphs in the vector space of the paragraph pair to be tested, and map them through a linear layer to the classifier's output space to complete the classification task.

Therefore, our model consists of three parts. The first part is the encoder block, where we use the RoBERTa-base model to encode the inputs Ω_{pp} and Ω_{nn} respectively. Next is the contrastive learning part, where during the training phase we apply L to optimize the outputs of Ω_{pp} and Ω_{nn} from the encoder block. Finally, there is the classification part, where in the testing phase we use a connection layer to complete the classification. The details of the contrastive learning training are in section 2.1.

3.1. RoBERTa Encoder Training

In the training stage, we convert the dataset D into two sets containing positive and negative examples. These are defined as $\Omega_{pp} = \{S_1, S_2, \ldots, S_p\}$ and $\Omega_{nn} = \{B_1, B_2, \ldots, B_n\}$, where Ω_{pp} is the set of positive examples, with each S_p representing a paragraph pair (i, j) with an unchanging author's style, and Ω_{nn} is the set of negative examples, with each B_n representing a paragraph pair (k, l) with a changing author's style.

Subsequently, each example is input into the encoder for encoding. For the positive examples S_p in the paragraph pair (i, j), we obtain the embedding representations $\mathbf{u}_i$ and $\mathbf{u}_j$, and the same applies to the negative examples B_n, in order to obtain the embedding representations $\mathbf{u}_k$ and $\mathbf{u}_l$ for the paragraph pair (k, l). Then, the similarity between $\mathbf{u}_i$ and $\mathbf{u}_j$ is calculated using the cosine distance.

To quantify the difference in similarity between positive and negative examples, we define a loss function based on the comparison of cosine similarities. The loss function L is represented as follows:

$$L = - \sum_{(i,j)\in\Omega_{pp},(k,l)\in\Omega_{nn}} \log\left(1 + \lambda \cdot \cos(\mathbf{u}_k, \mathbf{u}_l) - \cos(\mathbf{u}_i, \mathbf{u}_j)\right) \quad (1)$$

Here, Ω_{pp} denotes the set of positive instance pairs, Ω_{nn} denotes the set of negative instance pairs, λ is a hyperparameter greater than zero, used to balance the similarity differences between positive and negative instances. $\mathbf{u}_x$ represents the embedding representation of paragraph $\mathbf{x}$. This loss function encourages the model to reduce the cosine distance of positive pairs while increasing that of negative pairs.

Section 2.2's Algorithm 1 provides a detailed description of the training process.

In the test stage, for a test paragraph pair $\mathbf{x}$, we pass it through the encoder using the method from the training phase to obtain the embedding representation $\mathbf{u}_x$. Then, we feed $\mathbf{u}_x$ into a fully connected linear layer for linear transformation, resulting in a real-valued vector $\mathbf{B}$. Then, we apply the sigmoid activation function to convert it into a probability value that lies between 0 and 1. Paragraphs with a probability value greater than 0.5 are classified as positive examples, and those with a probability value less than 0.5 are classified as negative examples. These classifications are compared with the true labels, and the F1 score is used for evaluation.

3.2. Encoder Training Algorithm

The algorithm's input includes: a set of positive instances Ω_{pp} and a set of negative instances Ω_{nn} as training data, λ is a hyperparameter greater than zero, used to balance the similarity differences between positive and negative instances, the number of training cycles $EPOCHS$, and the encoder model E that needs to be trained. The output is the trained model E.

The intermediate variables include: paragraph pairs (i, j) and (k, l) from the sets of positive and negative instances, as well as their corresponding embedding representations $(\mathbf{u}_i, \mathbf{u}_j)$ and $(\mathbf{u}_k, \mathbf{u}_l)$. The specific process is as follows:

4. Experiments

4.1. Data Pre-processing

PAN 24 provided three datasets of varying difficulty levels for this task, each dataset consists of a training set (70%), a validation set (15%), and a test set (15%).

Task 1 (easy): The paragraphs of the document cover various topics, allowing methods that utilize thematic information to detect changes in authorship.

Task 2 (medium): There is minimal thematic diversity within the document (though still present), forcing methods to focus more on style to effectively address the detection task.

Task 3 (hard): All paragraphs in the document are related to the same topic.

For each difficulty level of the dataset, the documents are first read, then divided into paragraph pairs, and the labels indicating whether the author's style has changed are recorded. Subsequently, we obtain a large number of high-quality paragraph pairs for training the encoder based on a special data augmentation method [8].

4.2. Experimental Setup

In this experiment, the hyperparameters for RoBERTa are set as follows: the batch size is set to 32, the maximum sequence length is set to 512, and any excess will be truncated. The initial learning rate is set to 1e-5, and the number of training epochs is set to 10.

The baseline method selects the test set scores of Chen et al. [8] from last year, who used DeBERTaBASE as the pre-trained encoder. It also includes two simple baseline scores released by the official, one that always predicts 1, and the other that always predicts 0 [12]. The evaluation metric

Algorithm 1: Encoder Training for Contrastive Learning

Input: Positive instances set Ωpp
Negative instances set Ωnn
Hyperparameter λ > 0
Number of training epochs EPOCHS
Trained encoder model E
Output: Trained encoder model E
Initialize the encoder model E with pre-trained weights.
for each epoch in 1 to EPOCHS **do**
    Shuffle the training dataset to create batches
    for each batch (i, j) in Ωpp **do**
        for each batch (k, l) in Ωnn **do**
            Send (i, j) and (k, l) to E to obtain the embeddings (ui, uj) and (uk, ul)
            Calculate cosine similarity between (ui, uj) and (uk, ul)
            Calculate the contrastive loss L using the equation (1)
            Perform backpropagation to update the weights of the encoder E.
        end for
    end for
end for
Return: the trained encoder model E.

chosen is the F1 score because it takes into account both precision and recall, and strikes a balance between them, providing a more comprehensive assessment of the model.

4.3. Experimental Results

The model is submitted to TIRA [13] for execution to obtain the final metrics of the model. Table 1 provides the scores achieved by the model presented in this paper on the official test set. Table 2 demonstrates the scores of the method presented in this paper on the validation set.

Table 1
Overview of the F1 accuracy on test set for the multi-author writing style task in detecting at which positions the author changes for task 1, task 2, task 3.

Approach	Task 1	Task 2	Task 3
Chen et al.	0.915	0.820	0.676
RoBERTa	0.985	0.818	0.807
Baseline Predict 1	0.466	0.343	0.320
Baseline Predict 0	0.112	0.323	0.346

In terms of experimental results, the RoBERTa model shows a significant improvement over the baseline method on Task 1 and Task 3, and it performs roughly on par with the baseline method on Task 2. This indicates that the RoBERTa model can exhibit superior performance when dealing with documents that have a high or low degree of topic diversity. The RoBERTa model is capable of adapting

Table 2
Overview of the F1 accuracy on validation set for the multi-author writing style task in detecting at which positions the author changes for task 1, task 2, task 3.

Approach	Task 1	Task 2	Task 3
RoBERTa	0.982	0.825	0.811

well to style transfer detection tasks of varying difficulties.

5. Conclusion

This paper briefly introduces the work results of finetuning the RoBERTa encoder and linear classifier using a contrastive learning method based on CoSENT, applied to the writing style change analysis in the PAN 2024 shared task. Comparative experiments were conducted on datasets of varying difficulty levels against baseline methods. The results indicate that the writing style change analysis model, which leverages RoBERTa encoding and contrastive learning, performs exceptionally well in the multi-author writing style analysis task. It can accurately identify shifts in author style, especially with the aid of contrastive learning and data augmentation techniques.

Acknowledgments

This research is supported by the Social Science Foundation of Guangdong Province, China(No.GD24CZY02)

References

[1] J. Bevendorff, I. Borrego-Obrador, M. Chinea-Ríos, M. Franco-Salvador, M. Fröbe, A. Heini, K. Kredens, M. Mayerl, P. Pęzik, M. Potthast, et al., Overview of pan 2023: Authorship verification, multi-author writing style analysis, profiling cryptocurrency influencers, and trigger detection: Condensed lab overview, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2023, pp. 459–481.

[2] D. Zlatkova, D. Kopev, K. Mitov, A. Atanasov, M. Hardalov, I. Koychev, P. Nakov, An ensemble-rich multi-aspect approach for robust style change detection, in: CLEF 2018 Evaluation Labs and Workshop–Working Notes Papers, CEUR-WS. org, 2018.

[3] A. Iyer, S. Vosoughi, Style change detection using bert., CLEF (Working Notes) 93 (2020) 106.

[4] E. Zangerle, M. Mayerl, M. Potthast, B. Stein, Overview of the Multi-Author Writing Style Analysis Task at PAN 2024, in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2024.

[5] Z. Zhang, Z. Han, L. Kong, X. Miao, Z. Peng, J. Zeng, H. Cao, J. Zhang, Z. Xiao, X. Peng, Style change detection based on writing style similarity, Training 11 (1970) 17–051.

[6] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[7] T.-M. Lin, C.-Y. Chen, Y.-W. Tzeng, L.-H. Lee, Ensemble pre-trained transformer models for writing style change detection., in: CLEF (Working Notes), 2022, pp. 2565–2573.

[8] H. Chen, Z. Han, Z. Li, Y. Han, A writing style embedding based on contrastive learning for multi-author writing style analysis, Working Notes of CLEF (2023).

[9] Z. Ye, C. Zhong, H. Qi, Y. Han, Supervised contrastive learning for multi-author writing style analysis, in: Conference and Labs of the Evaluation Forum, 2023. URL: https://api.semanticscholar.org/CorpusID:264441623.

[10] J. SU, Cosent (1): A more efficient sentence vector scheme than sentence-bert, 2022.

[11] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).

[12] J. Bevendorff, X. B. Casals, B. Chulvi, D. Dementieva, A. Elnagar, D. Freitag, M. Fröbe, D. Korenčić, M. Mayerl, A. Mukherjee, et al., Overview of pan 2024: multi-author writing style analysis, multilingual text detoxification, oppositional thinking analysis, and generative ai authorship verification, in: European Conference on Information Retrieval, Springer, 2024, pp. 3–10.

[13] M. Fröbe, M. Wiegmann, N. Kolyada, B. Grahm, T. Elstner, F. Loebe, M. Hagen, B. Stein, M. Potthast, Continuous integration for reproducible shared tasks with tira. io, in: European Conference on Information Retrieval, Springer, 2023, pp. 236–241.