Authorship Verification Based on CoSENT

Notebook for PAN at CLEF 2023

ZhaoHao Huang, Leilei Kong*, Mingjie Huang

Foshan University, Foshan, China

Abstract

Authorship verification is the task of determining whether the authorship of texts is the same based on stylistic features. In this paper, we propose two different approaches to the author attribution task. The first approach employs a positive-example-based data optimization approach to reorganize the training dataset. The second method uses CoSENT, a feature-based text classification method, to accomplish the task of authorship verification. This method enables the model to have a better ability to identify whether sentence pairs are similar or dissimilar.

Keywords

Authorship Verification, Data optimization, Feature-based, CoSENT

1. Introduction

The authorship verification task is to determine whether two texts are written by the same author by analyzing various characteristics and styles of texts. By utilizing effective authorship verification techniques, we can identify authors in literature, law, journalism, and other fields, thereby helping to solve important tasks such as the authenticity of texts, copyright issues, and the identification of forged documents. This paper presents our approach for the authorship verification task [1] on PAN 2023. This task is an open-set validation task, and the test dataset contains authors who were not seen in the training dataset, so writing style models built for the authors or topics of the training dataset are not supported in the open-set validation task.

Our idea is to encode text information, get text features, and judge whether two texts are the same author by comparing the similarity of features. In order to make the model better classify texts, the characteristic text matching method CoSENT [2] is used to expand the distance of feature vectors of dissimilar texts and reduce the distance of feature vectors of similar texts. In this task, the lack of enough labeled data to train the language model is also a problem to be solved. We try to expand the training set data by restructuring text sentence pairs from the official training set data. Then use the pre-trained language model BERT [3] to complete the author verification task.

2. Datasets

The training dataset provided by the PAN@CLEF 2023 organization [4] consists of cross-discourse types of authorship verification cases using the following discourse types (DT): essays, emails, interviews, and speech transcriptions. Among the four DTs, essays and emails belong to the written discourse, and interviews and speech transcriptions belong to the spoken discourse. The dataset includes texts from approximately 56 authors who are native English speakers and have

[1]CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece
EMAIL: zhaohaohuang6@gmail.com (A. 1); kongleilei@fosu.edu.cn (A. 2) (*corresponding author); mingjiehuang007@163.com
ORCID: 0000-0003-3278-8472 (A. 1); 0000-0002-4636-3507 (A. 2); 0000-0002-0889-5027 (A. 3)

similar ages ranging from 18 to 22. The topic of the text samples is not limited or constrained. The total number of text pairs in the dataset provided by PAN@CLEF 2023 is 8,836. Each problem is composed of two texts belonging to two different DTs. All text pairs contain 886 texts from different authors. The number distribution of different discourse types is shown in Table 1.

Since the text length of texts of emails and interviews can be very small, each text belonging to these DTs is actually a concatenation of different messages. We use the <new> tag to denote the boundaries of original texts. New lines within a text are denoted with the <nl>tag. In addition, author-specific and topic-specific information, e.g., named entities, has been replaced with corresponding tags. In spoken discourse types, additional tags are used to indicate nonverbal vocalizations (e.g., cough, laugh).

Table 1
The number distribution of different discourse types.

Type	Quantity
Essays	93
Emails	450
Interviews	275
Speech transcriptions	68
Overall	886

3. Method

3.1. Data preprocessing

For deep learning methods, the more text data, the more text features the language model can learn. Of the 8,836 text pairs in the original training dataset, there are only 886 distinct texts. We try to use these different texts to scramble and combine them into a more extensive data set for model training. We combine two groups of texts belonging to the same author into a positive pair, and the total number of positive pairs is 6,945. In order to keep the ratio of positive and negative samples at 1:1, each of the 886 different texts in the original training dataset is combined with any text of other authors to form a negative pair, and each text is randomly combined with 8 negative pairs. There are a total of 7088 negative sentence pairs. A total of 14033 positive and negative sentence pairs are shuffled and divided into 75% training set and 25% test set, as shown in Table 2.

Table 2
The partition of the training dataset of PAN 2023 into a new training dataset and test dataset.

dataset	proportion	number of text pairs
New training dataset on PAN23	75%	10524
New test dataset on PAN23	25%	3509

3.2. Model settings

In this paper, we choose to use a feature-based text-matching approach for this task. The Bert model is fine-tuned during the training phase using the expanded training dataset. In the prediction stage, we send the text1 and text2 of the text pair separately to the pre-trained language model, bert, where the text is encoded respectively and the cosine similarity is calculated. Finally, whether the cosine similarity is greater than 0.5 is used as the classification standard for the sentence pair. The overall structure of the model is shown in Figure 1.

We hope to improve the classification ability of the model. The work of Su [2] suggests that a new loss function is proposed, which improves the model's ability to identify similar text and distinguish

dissimilar text to a certain extent. We record Ω_{pos} as the set of all positive sample pairs and Ω_{neg} as the set of all negative sample pairs.

In fact, we hope that for any positive sample pair $(\mathrm{i},\mathrm{j}) \in \Omega_{pos}$ and negative sample pair $(\mathrm{k},l) \in \Omega_{neg}$, both have

$$\cos(u_i, u_j) > \cos(u_k, u_l) \tag{1}$$

Where $\mathrm{u_i}$, $\mathrm{u_j}$, $\mathrm{u_k}$, $\mathrm{u_l}$ are their respective sentence vectors. Here we only hope that the similarity of positive sample pairs is greater than that of negative sample pairs, and it is up to the model to decide how much larger it is. The loss function is shown in formula (2).

$$\log\left(1 + \sum_{(i,j)\in\Omega_{pos},(k,l)\in\Omega_{neg}} e^{\lambda(\cos(u_k,u_l)-\cos(u_i,u_j))}\right) \tag{2}$$

λ is a hyperparameter, and we set it to 20 in this experiment.

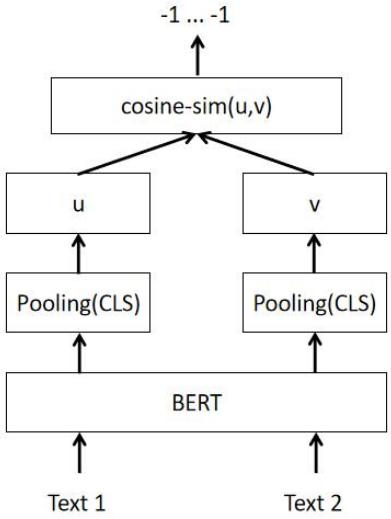

Figure 1: Architecture diagram for our model.

4. Experiment setting

In this work, $\text{BERT}_{\text{BASE}}$ (L=12, H=768, A=12, Total Parameters=110M) is chosen as the pre-trained model size, and we use Pytorch to construct BERT and fully connected network classification model. Our hyperparameters are set as follows: the batch size is 16, the maximum sequence length is 512, the initial learning rate is set to 1e-5, and 20 epochs are trained. Each training is optimized with AdamW, and the warm-up rate is set to 0.1.

5. Evaluation and results

5.1. Evaluation

To evaluate the proposed models, we use the TIRA [5] evaluation tool with the following

metrics:

AUC: The area-under-the-curve (ROC) score

F1-score: F1 score is the harmonic mean between precision and recall [6].

c@1: rewards systems that leave complicated problems unanswered [7].

F_0.5u: A measure that puts more emphasis on deciding same-author cases correctly [8].

Brier: The complement of the Brier score for evaluating the goodness of (binary) probabilistic Classifiers [9].

5.2. Results

We test the performance of our model on a new reorganized test set and test our model on the PAN23 authorship verification test dataset.The test results are shown in Table 3.

Table 3

The experimental results of the Cosent method on the new test dataset on PAN23 and the true test dataset on PAN23.

Test datasets	AUC	c@1	f_05_u	F1	Brier	Overall
New test dataset on PAN23	0.935	0.935	0.925	0.938	0.935	0.934
PAN23 authorship verification test dataset	0.563	0.563	0.55	0.511	0.563	0.55

6. Conclusion

In this paper, we present our approach to authorship verification on PAN 2022. We re-pair each dissimilar sentence in the training dataset into a new larger training dataset. By augmenting the dataset in this way, we hope the model can learn more about the similarities between texts by the same author and the differences between texts by different authors. In addition, we also introduced a special loss function in the model to make the model better learn the similarity and difference information between sentence pairs. Then use the pre-trained language model BERT to extract text features and calculate sentence cosine similarity to judge whether they are from the same author.

From the results, the method does not perform well on open datasets with unknown authors. In the follow-up work, we should use more effective methods to enhance the data and improve the classification ability of the model in the open set.

Acknowledgments

This work is supported by the National Social Science Foundation of China (No. 22BTQ101).

References

[1] 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.

[2] J. Su, CoSENT (I): A more efficient sentence vector scheme than Sentence-BERT, 2022. URL: https://spaces.ac.cn/archives/8847
[3] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., Scikit-learn: Machine learning in python, the Journal of machine Learning research 12 (2011) 2825–2830
[4] 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.
[5] 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.
[6] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., Scikit-learn: Machine learning in python, the Journal of machine Learning research 12 (2011) 2825–2830
[7] A. Peñas, A. Rodrigo, A simple measure to assess non-response (2011)
[8] J. Bevendorff, B. Stein, M. Hagen, M. Potthast, Generalizing unmasking for short texts, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 654–659
[9] G. W. Brier, et al., Verification of forecasts expressed in terms of probability, Monthly weather review 78 (1950) 1–3