

Authorship Verification Based on SimCSE

Notebook for PAN at CLEF 2023

Yong Qiu, Haoliang Qi*, Yong Han, Kaicheng Huang

Foshan University, Foshan, Guangdong, China

Abstract

In the case of the Authorship verification task, where two texts belonging to different Discourse Types (DT) are given, the objective is to determine if they are written by the same author (cross-DT authorship verification). In this paper, we use a framework based on contrastive learning and a pre-trained language model to extract text features to solve the (authorship verification) task. Compared with traditional machine learning methods, the method based on contractive learning (SimCSE) can encode and get the comparison between the texts, enabling better handling of semantic matching. In the authorship verification task, the model can capture the author's writing style and characteristics more accurately by comparing the similarity between multiple texts of the same author, thus improving the model's classification performance. The experiment demonstrated competitive performance, achieving an accuracy of 90.76% on our test dataset, which was manually created from a PAN-provided dataset.

Keywords

Authorship Verification, Contractive Learning, Pre-trained model, Classification

1. Introduction

The goal of the PAN@CLEF 2023 Authorship verification [1][2] task is to determine whether the two texts are written by the same author by comparing their writing styles. In earlier versions of PAN [3], the effectiveness of authorship verification techniques in multiple languages and text types was investigated. In recent years, cross-domain authorship verification tasks (cross-DT authorship verification) [3] have also been successfully implemented. The authorship verification task at PAN 2023, for the first time, focuses on authorship verification across spoken discourse genres. The aim is to investigate the effectiveness and robustness of style measurement methods under more challenging and intriguing circumstances. This task provides a cross-DT author verification case using a new English corpus of a diverse sample of approximately 100 native English speakers of similar ages (18-22). The subject of the text is not restricted. Still, the forms and levels of formality vary, including essays (written discourse), e-mails (written discourse), interviews (spoken discourse), and speech transcriptions (spoken discourse).

In recent years, deep learning-based methods have been widely adopted due to their superiority in processing large-scale datasets. Neural networks have had many practices for judging text author attribution [4]. Contractive learning [5] has gained increasing attention in deep learning, such as SimCSE [6], which can use labeled and unlabeled data to train deep neural networks to alleviate the challenges caused by data scarcity or inaccurate labels. In this authorship verification task, the lack of sufficient labeled datasets is a problem that needs to be solved, and comparative learning can improve the authentication performance of authorship verification by learning the similarity between tasks, to solve these problems better. The sentence vector representation technique has always been a hot topic in NLP. In the BERT [7] era, people generally used the [CLS] vector of the BERT model to represent

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

EMAIL: Timmy_ong@163.com (A. 1); qihaoliang @fosu.edu.cn (A. 2) (*corresponding author); hanyong2005@fosu.edu.cn (A. 3); teaslate@outlook.com (A. 4)

ORCID: 0000-0003-4620-4992 (A. 1); 0000-0003-1321-5820 (A. 2); 0000-0002-9416-2398 (A. 3); 0000-0002-3473-3355 (A. 4)



© 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) Proceedings

the sentence vector by the inherent advantages of the pre-trained language model. In this paper, we apply BERT to obtain text representation, adopt contractive learning, and capture the author's writing style and characteristics by comparing the similarity between multiple texts of the same author. By experiment, we realized authorship verification.

2. Dataset

PAN@CLEF 2023 Authorship verification task belongs to the open set author verification. That is, the training data set and the test data set have the same structure and similar properties. However, there is no overlap in the author sets of training and test datasets. For the training set, there are 8836 text pairs from the Aston 100 Idiolects Corpus in English covering DTs of both written and spoken language: essays, emails, interviews, and speech transcriptions. Labels are given correspondingly, whose meaning is to judge whether the two texts are written by the same author (label 1 means yes, and vice versa is 0). To protect the author's privacy, information specific to both author and topic, such as named entities, have been substituted with corresponding tags. Moreover, for spoken discourse genres, supplementary tags are utilized to indicate nonverbal vocalizations, such as coughing and laughing.

Dataset distribution types and quantities are shown in Table 1.

Table 1

The distribution types and quantities of the PAN@CLEF 2023 Authorship verification task dataset

Type	Quantity
Essays	93
Emails	450
Interviews	275
Speech transcriptions	68

3. Method

This section will introduce our data processing approach and network framework.

3.1 Data Preprocessing

For the training set given by the task, we first separated the texts according to each author. We randomly divided the texts according to 3:1 for the training set and the test set, including 42 and 14 different authors, respectively. For the training set, we make a text triplet (" pair ": [" *text₁* ", "*text_{pos}*," "*text_{neg}*"]), where "*text_{pos}*" indicates a positive sample, and "*text_{neg}*" indicates a negative sample. The positive samples are from the texts written by the author of the first text of the triplet, while the negative ones are from other authors.

The training set made of random combinations contains 22,090 pieces of data.

3.2 Network Architecture

In this paper, the novel comparative learning framework SimCSE(Simple Contrastive Sentence Embedding Framework) is adopted, and the pre-training language model BERT is used to extract text features. And the model parameters are constantly updated with training. Precisely, as for the BERT, the stacking Transformer encoder can capture deeply bidirectional information between words in a sentence. Take the hidden of the 0th position of hidden state in the last layer of output (that is, the

hidden CLS, the vector of [CLS] tokens in the output layer) to represent the vector of the entire sentence. Directly using supervised training set for comparative learning training, the positive sentence pair can be regarded as a natural positive sample and regard other embedding in the same batch as negative samples. So, we can think of authorship verification as a binary classification problem [8]. Note that the positive example of $text_i$ for each piece of data in the same batch is the only one, while all other examples in the batch are considered negative examples. However, if the batch contains two or more texts from the same author (for example, assuming $Text_1$ and $Text_2$ in Figure 1 were written by the same author), this may cause model confusion and performance degradation. Therefore, we scrambled the dataset to avoid the occurrence of the same authors in the same batch as much as possible, and we used a lower batch size further alleviates this issue.

The core of the SimCSE model is contractive learning, which aims to better learn the representation of the data by narrowing the distance between similar data and drawing the distance between dissimilar data. This makes it more effective in text-matching tasks. For each training data (x_i, x_i^+, x_i^-) , where x_i^+ is the implied text and x_i^- is the contradictory text. The training objective of supervised learning SimCSE is:

$$l_i = -\log \frac{e^{\text{sim}(h_i, h_i^+)/\tau}}{\sum_{j=1}^N (e^{\text{sim}(h_i, h_j^+)/\tau} + e^{\text{sim}(h_i, h_j^-)/\tau})} \quad (1)$$

Where h_i and h_i^+ is the sentence vector representation of x_i and x_i^+ , N is the size of batch in the training process, $\text{sim}(h_i, h_j^+)$ is the cosine similarity of vectors h_i and h_j^+ , and τ is the temperature hyperparameter.

The network architecture of the model is shown in Figure 1.

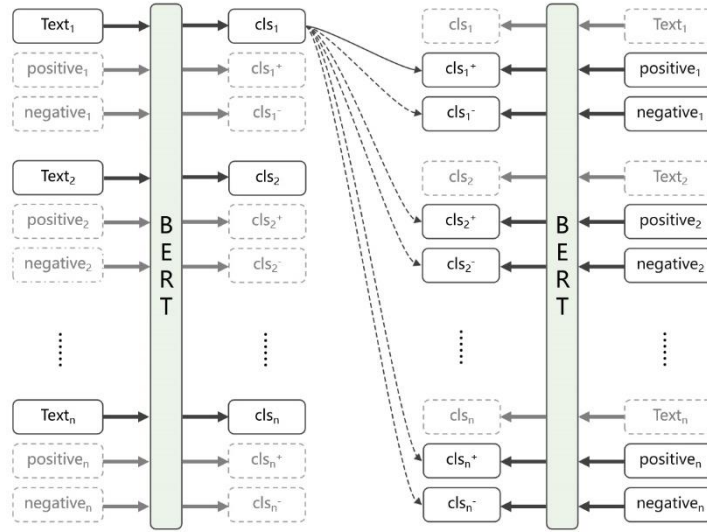


Figure 1: Network Architecture

4. Experiments and Results

4.1 Experimental Setting

In this paper, we choose BERT-base-uncased as an encoder with 12-layer, 768-hidden, 12-heads, and 110M parameters. The vocab size is 28,996. The data set was expanded to 22,090 pieces of data after partitioning. The maximum length of the encoder is set to 512. We used Adam optimizer with the learning rate set to $2e-5$. Our experiment was conducted on the A100 server. The best performance is achieved through 5 epoch models.

4.2 Evaluation

To evaluate the performance of our model, we used the evaluation platform provided by PAN, which includes the following metrics:

- AUC: the conventional area under the curve score.
- c@1: rewards systems that leave complicated problems unanswered [9].
- F_{0.5u}: focus on deciding same-author cases correctly [10].
- F1-score: harmonic way of combining the precision, and recall of the model [11].
- Brier: Brier Score evaluates the accuracy of probabilistic predictions [12].

4.3 Results

Table 2 presents the model performance. The first line shows the performance results of BERT only (without SimCSE), while the second line shows the results of additional SimCSE. By comparison, our method (BERT_{BASE}+SimCSE) performs better, which proves the effectiveness of SimCSE.

Table 2
Results on the test set

Model	AUC	c@1	f_05_u	F1	Brier	Overall
BERT _{BASE}	0.846	0.846	0.853	0.844	0.846	0.847
BERT _{BASE} +SimCSE	0.908	0.908	0.871	0.915	0.908	0.902

Table 3 demonstrates the performance of our model evaluated on the TIRA [13] environment for PAN@CLEF 2023.

Table 3
Results on pan23-authorship-verification-test

Model	AUC	c@1	f_05_u	F1	Brier	Overall
pan23-cdav-baseline	0.601	0.569	0.543	0.466	0.595	0.555
BERT _{BASE} +SimCSE	0.540	0.540	0.499	0.421	0.540	0.508

5. Conclusion

In this work, we adopted a contrastive learning framework and applied the BERT pre-trained language model to extract text features to solve the PAN 2023 Authorship verification task. The experimental results show that satisfactory results can be achieved by applying contrastive learning to the field of natural language processing, such as authorship verification task. It also demonstrates the powerful ability of the BERT model in text vector representation.

In future work, we will continue to improve our methods and strive to achieve better results in open domain authorship verification.

6. Acknowledgements

This work is supported by the National Natural Science Foundation of China (No.62276064).

7. References

- [1] Stamatatos, E., Kredens, K., Pezik, 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] Bevendorff, J., Borrego-Obrador, I., Chinea-Ríos, M., Franco-Salvador, M., Fröbe, M., Heini, A., Kredens, K., Mayerl, M., Pezik, 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.
- [3] 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.
- [4] 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.
- [5] Ye, M., Zhang, X., Yuen, P. C., & Chang, S. F. (2019). Unsupervised embedding learning via invariant and spreading instance feature. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 6210-6219).
- [6] Gao, T., Yao, X., & Chen, D. (2021). Simcse: Simple contrastive learning of sentence embeddings. arXiv preprint arXiv:2104.08821.
- [7] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
- [8] M. Koppel, J. Schler, Authorship verification as a one-class classification problem, in: C. E. Brodley (Ed.), *Machine Learning, Proceedings of the Twenty-first International Conference (ICML 2004)*, Banff, Alberta, Canada, July 4-8, 2004, volume 69 of ACM International Conference.
- [9] A. Peñas, A. Rodrigo, A simple measure to assess non-response (2011).
- [10] 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.
- [11] 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.
- [12] G. W. Brier, et al., Verification of forecasts expressed in terms of probability, *Monthly weather review* 78 (1950) 1–3.
- [13] 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.