Authorship verification Based On Fully Interacted Text Segments

Notebook for PAN at CLEF 2022

Mingjie Huang, Leilei Kong*, Zeyang Peng, Yihui Ye, Zengyao Li, Xinyin Jiang, Zhongyuan Han

Foshan University, Foshan, China

Abstract

Authorship verification is the task of deciding whether two texts have been written by the same author based on comparing the texts' writing styles. We propose a method to extract the interactive relationship between text pairs with texts separated into segments and combined in a specific order. The features of text pairs are extracted with a pre-trained model. The experiment in this paper is based on the open set, where part of authors doesn't appear in training dataset.

Keywords [1]

Authorship verification, Pre-trained model, Long text, Open set

1. Introduction

The authorship verification task in this paper is one of the sharing tasks at PAN 2022 [1]. The purpose of the authorship verification task is to determine whether two texts share the same author [3]. The authorship verification task this year is defined as cross-DT task which means two texts within each text pair belong to different discourse types (DT) [2]. The authorship verification task at PAN 2022 is based on an open set. Specifically, the author sets in training and test datas sets are not overlapping. This means that the author's writing styles learned in the training set may not be sufficient for prediction in the test set. Therefore, we are more inclined to analyze the texts' writing styles rather than specific authors' writing styles.

In recent years, neural networks have had many practices for judging text author attribution [4]. Since pre-trained model BERT [5] have excellent performance in text classification, we use BERT to extract stylistic similarity between texts. In this work, we divided the text into segments and let the text segments fully interact. We take this approach for two reasons. Firstly, we try to use this method to allow the model to learn the features between texts without discarding any segment fully. Secondly, this task is based on cross-DT text pairs, where some texts of message or email type are originally spliced from concise segments. So we try to extract the features behind short texts with this fully interactive method and use these features to infer the authorship of the two spliced texts.

2. Datasets

The authorship verification task at PAN 2022 is based on an open set of texts written by 100 different authors, and some authors do not appear in the training set. In general, there are four DTs: essays, emails, text messages, and business memos. In order to protect the author's privacy, author-specific and topic-

[1]CLEF 2022 – Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy

EMAIL: mingjiehuang007@163.com (A. 1); kongleilei@fosu.edu.cn (A. 2) (*corresponding author); pengzeyang008@163.com (A. 3); oldsport996@gmail.com (A. 4)

ORCID: 0000-0002-0889-5027 (A. 1); 0000-0002-4636-3507 (A. 2); 0000-0002-8605-4426 (A. 3); 0000-0002-7369-7537 (A. 4);

specific information in the text has been replaced with special tags. In addition, emojis or expressions consisting of punctuation marks are also preserved in email, message, and memo types.

There are 12,264 pieces of data in the training dataset containing ids, authors, text pairs, DTs, and labels. The lengths of the texts taken from essays are mostly more extended than that of the other three DTs. Specifically, the lengths of the text of essay type are mostly more than 1,000 characters, and the lengths of other three types are the opposite. Apart from essay type, texts of the other three DTs are composed of multiple short-length texts, separated by <new> tag. This means that the semantics of essay types of texts will be more coherent, while the semantics of the other three types of texts are more scattered.

The types and quantities of special symbols and emoticons in all texts of training datasets are listed in the table below.

Table 1
Types and quantities of special symbols and emojis

symbol type	examples	types	quantities
author-specific and topic-specific	<email_address>	372	3,153,444
emojis	😂 🥺	132	68,962

3. Method

3.1. Text Preprocessing

The emojis in the text can be used as a style feature of the text, but not all emojis can be encoded using the pre-trained model BERT as those special symbols. As cleaning of the data, these symbols and emojis are removed from the texts.

Texts vary in length for different discourse types. Specifically, the texts of the essay type are much longer than other three types. Therefore, the task can be divided into two parts, one part is the discrimination between long text and short text, and the other part is the discrimination between two short texts. We divide the texts into short segments and this approach has two benefits. First of all, after the texts divided into shorter segments with less than 510 characters, these two parts of tasks can be handled in the same way, i.e., author identification between text segments. Secondly, There is a fuller and more efficient interaction between text pairs, specifically, taking 90% part of training datasets as an example. The original datasets consisting of 11,000 text pairs have been expanded to 301,764 shorter text pairs. At the same time, the ratio of positive and negative samples is kept at about 1:1.

Table 2
Quantities of 11,000 text pairs before and after division, the ratio (number of true samples/number of false samples), and the total count of all samples

	false (label = 0)	true (label = 1)	rate	total
before	5528	5472	1 : 1.01	11,000
after	147482	154282	1 : 0.96	301,764

For each text pair, denoted as $text_1$ and $text_2$, suppose $text_1 = \{t_{11}, t_{12}, \dots, t_{1m}\}$ and $text_2 = \{t_{21}, t_{22}, \dots, t_{2n}\}$, where t_{11} is the first segment of $text_1$ and m is the maximum amount of segments that $text_1$ can be devided into. Then m segments of $text_1$ and n segments of $text_2$ are combined in pairs to form $m * n$ new data sets. In this way, all segments of $text_1$ can fully interact with those of $text_2$.

3.2. Neural Network Architecture

Limit maximum length to 510 characters, and all texts are splited into segments. Assuming that $text_1$ is divided into m segments and $text_2$ is divided into n segments, then $m * n$ new text pairs formed by

pairwise combination will be fed into a pre-trained model BERT, where each text pair would be encoded and computed. During this process, each segment of $text_1$ has an interaction with all segments of $text_2$. Encoded with BERT, original text pairs turn into $m * n$ features in the form of $features = \{f_{t_{11}t_{21}}, \dots, f_{t_{1i}t_{2j}}, \dots, f_{t_{1m}t_{2n}}\}$, where $f_{t_{1i}t_{2j}}$ represents the feature of pair $t_{1i}t_{2j}$. Then we will do average pooling over all features and get the represent of original text pairs. In this process, a feature array in shape of (1, mn, 768) will be averaged and compressed to a new array in shape of (1, 768). Finally, the represent will be fed into a fully connected neural network to determine whether this two original texts share the same author.

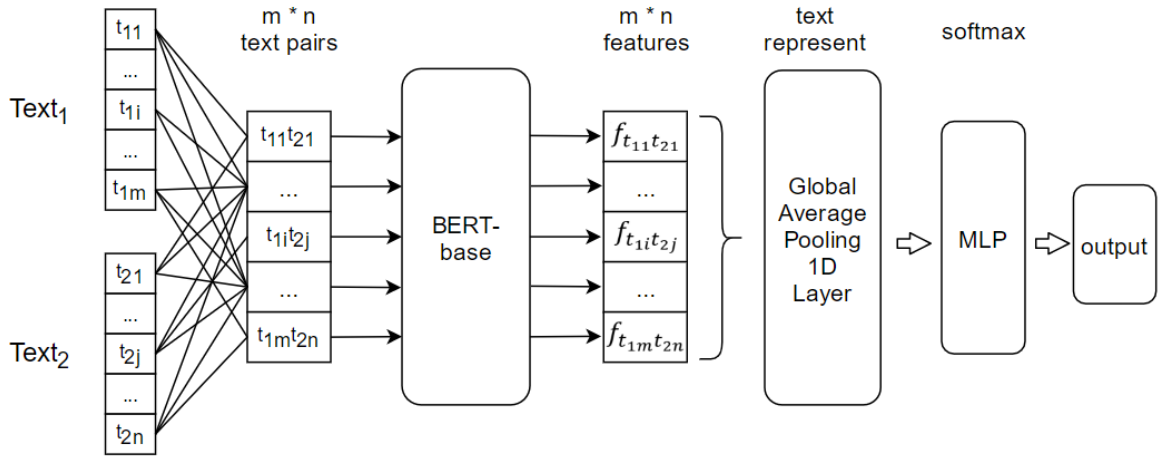

Figure 1: Neural Network Architecture

4. Experiments and Results

4.1. Data Partition

In order to simulate the test environment of the open set, we randomly selected six authors from the data set and took their corresponding texts as the test set. After that, we randomly select 70% of the remaining dataset as the training set and 30% of the remaining dataset as the validation set. The training dataset and valid dataset contain text pairs written by 50 authors.

Table 3
The partition of the training dataset into a new training dataset, valid datasets, and text dataset. The fourth line is the true test dataset of PAN 22.

dataset	proportion	number of authors	number of text pairs
training dataset	60%	---	7,381
valid dataset	26%	---	3,164
test dataset	14%	6	1,719
PAN22 authorship verification test dataset	---	44	10,478

4.2. Experiment setup

In this paper, we choose BERT-base cased as an encoder with 12-layer, 768-hidden, 12-heads, and 110M parameters. The vocab size is 28,996. After division, 7381 text pairs are extended to 50805 pairs. The training batch size is set to 16, and the maximum length of the encoder is set to 256. We use Adam optimizer, learning rate set to 2e-5, and dropout layer, the rate set to 0.5, to avoid overfitting during fine tuning. One epoch is enough to fit the model to fine tune on the training dataset.

Vectors of the last layer of BERT except for cls and last terminator are extracted and average pooled into a new vector of 768 dimensions. In other words, the CLS embedding (of BERT's output) is not used to represent the text segment pair of the input. Instead, all token embeddings except CLS and SEP are average pooled. As we use BERT an encoder, we believe the described method could obtain more comprehensive sentences features than taking CLS embedding [6]. All these 50805 new vectors will be average pooled and reshaped into a numpy array with a shape of (7381, 1, 768). Then this array will be fed to two fully connected layers. There are 16 units in the first dense layer with activation of ReLU and two units in the second dense layer with activation of softmax. Sparse categorical cross-entropy is used as the loss function for our model [7]. We set 500 epochs to fit the MLP to the training set.

4.3. Results

We test the performance of our model on 1,719 text pairs split from the training dataset and also test our model on the PAN22 authorship verification test dataset. There are 10,478 text pairs in this test dataset, including texts that are written by 44 authors who do not appear in training dataset.

The models are tested on TIRA [8] and evaluated on five measures: area under the ROC curve (AUC), F1-score, c@1 (a variant of the F1-score, which rewards systems that leave complex problems unanswered [9]), F_0.5u (a measure that puts more emphasis on deciding same-author cases correctly [10]), and the complement of the Brier score [11][12]. The results are shown in the table below.

Table 4

Line 1 is the results of 1,719 text pairs split from the training dataset. Line 2 is the results of 10,478 text pairs of the PAN22 authorship verification test dataset.

dataset	auc	c@1	f_05_u	F1	brier	overall
1719 text pairs	0.56	0.694	0.408	0.253	0.694	0.522
PAN22 authorship verification test dataset	0.519	0.519	0.328	0.196	0.519	0.416

From the data above, we can observe that the more text pairs in the test dataset and the more unknown authors, the worse our model performance will be. This may indicate that models trained on closed sets may not be powerful enough to capture textual features on open sets.

5. Conclusion

In this paper, we present our approach for authorship verification at PAN 2022. We split the text into shorter segments and let these segments interact in pairs. In this way, we hope to be able to augment the data and make the text pairs sufficiently interactive. Then the features between text pairs will be extracted with the pre-trained language model BERT. Finally, we will integrate these features to determine whether two texts belong to the same author. As seen from the results, this method does not perform well on an open dataset containing unknown authors.

In the follow-up work, we should use a more effective method to extract the author's style characteristics in the text, and it is not enough to make the text interact between fragments simply by the way of manual combination.

6. Acknowledgment

This research was supported by the Natural Science Foundation of Guangdong Province, China (No. 2022A1515011544).

7. References

[1] J. Bevendorff, B. Chulvi, E. Fersini, A. Heini, M. Kestemont, K. Kredens, M. Mayerl, R. Ortega-Bueno, P. Pezik, M. Potthast, F. Rangel, P. Rosso, E. Stamatatos, B. Stein, M. Wiegmann, M. Wolska, E. Zangerle. Overview of PAN 2022: Authorship Verification, Profiling Irony and Stereotype Spreaders, and Style Change Detection, in: A. B. Cedeno, G. D. S. Martino, M. D. Esposti, F. Sebastiani, C. Macdonald, G. Pasi, A. Hanbury, M. Potthast, G. Faggioli, N. Ferro (Eds.), Proceedings of the Thirteenth International Conference of the CLEF Association (CLEF 2022), Springer, 2022.

[2] E. Stamatatos, M. Kestemont, K. Kredens, P. Pezik, A. Heini, J. Bevendorff, M. Potthast, B. Stein. Overview of the Authorship Verification Task at PAN 2022. Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (2022)

[3] Koppel M, Winter Y. Determining if two documents are written by the same author. Journal of the Association for Information Science and Technology, 2014, 65(1): 178-187.

[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] 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] Youngjune Gwon, Hyunjin Choi, Judong Kim and Seongho Joe. “Evaluation of BERT and ALBERT Sentence Embedding Performance on Downstream NLP Tasks” International Conference on Pattern Recognition (2021)

[7] Zeyang Peng, Leilei Kong, Zhijie Zhang, Zhongyuan Han, Xu Sun. Encoding Text Information By Pre-trained Model For Authorship Verification. CLEF (Working Notes) 2021: 2103-2107

[8] M. Potthast, T. Gollub, M, Wiegmann, B. Stein: TIRA Integrated Research Architecture, in: Information Retrieval Evaluation in a Changing World, ser. The Information Retrieval Series, N. Ferro, C. Peters, Berlin Heidelberg New York: Springer, Sep. 2019.

[9] A. Peñas, Álvaro Rodrigo, A simple measure to assess non-response, in: ACL, 2011, pp. 1415–1424. URL: http://www.aclweb.org/anthology/P11-1142.

[10] J. Bevendorff, B. Stein, M. Hagen, M. Potthast, Generalizing unmasking for short texts, in: Pro- ceedings of t he 2019 Conference of the North American Chapter of the Association for Com- putational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), As- sociation for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 654–659. URL: https: //www.aclweb.org/anthology/N19-1068. doi:10.18653/v1/N19-1068.

[11] G. W. Brier, Verification of forecasts expressed in terms of probability, Monthly weather review 78 (1950) 1–3.

[12] Weerasinghe, J., Singh, R., & Greenstadt, R. (2021). Feature vector difference based authorship verification for open-world settings. CEUR Workshop Proceedings, 2936, 2201-2207.