Encoding Text Information By Pre-trained Model For Authorship Verification

Notebook for PAN at CLEF 2021

Zeyang Peng[1], Leilei Kong[1*], Zhijie Zhang[1], Zhongyuan Han[1], Xu Sun[2]

[1]*Foshan University, Foshan, China*
[2]*Heilongjiang Institute of Technology, Haerbin, 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 present a classification method based on encoding text information by a pre-trained model for authorship verification. The proposed model achieved the highest c@1 and F1-score on the small dataset of PAN Authorship Verification datasets.

Keywords

Pre-trained model, Text information, Classification, Authorship verification

1. Introduction

There have such phenomena as telecommunication, email fraud and terrorist attacks in today's society, so it is important to verify unknown authorship. Authorship verification is an active research area of computational linguistics that can be considered as a fundamental question of stylometry, namely whether or not two texts are written by the same author [1]. Accordingly, it has become one of the staple sharing tasks at PAN. The work presented in this paper was developed as a solution to the Authorship verification task for the competition PAN @ CLEF 2021 [1]. At PAN 2021, authorship verification belongs to an open-set verification task that test dataset contains verification cases from training dataset's unseen authors and topics [2], so a writing style model has built for training dataset's author or topic is not supported on open-set verification task. Accordingly, our idea is to encode text information to get the text feature and determine whether two texts are the same author by comparing the similarity of the features.

In recent years, more and more pre-trained models, typified by BERT [3], have performed well in natural language processing. In particular, BERT, which stands for Bidirectional Encoder Representations from Transformers [4], has achieved considerable improvement in encoding text information. However, we also noticed that BERT could not encode long text efficiently. In order to encode the long text, our method is to split long texts into short texts that BERT can encode, then obtain the similarity of local text by combining two short texts from two long texts, respectively. Finally, the overall similarity of two long texts can be obtained by integrating these local similarities.

2. Datasets

[1]CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania

EMAIL: pengzeyang008@163.com (A. 1); kongleilei@fosu.edu.cn (A. 2) (*corresponding author); zhangzhijie5454@gmail.com (A. 3); hanzhongyuan@fosu.edu.cn (A. 4)
ORCID: 0000-0002-8605-4426 (A. 1); 0000-0002-4636-3507 (A. 2); 0000-0002-4854-0618 (A. 3); 0000-0001-8960-9872 (A. 4);

CEUR Workshop Proceedings (CEUR-WS.org)

Authorship verification task datasets consist of many fanfictions, which were obtained drawn from fanfiction.net, and they are writing in which fans use media narratives and pop cultural icons as inspiration for creating their own texts. The datasets include large and small training dataset, which consists of pairs of (snippets from) two different fanfics. They include 275,565 and 52,601 pairs of texts, respectively. Table 1 shows the results of data analysis over both datasets.

Table 1
The detail of Authorship verification datasets

Dataset	Samples	Positive Samples	Max Characters	Min Characters	Mean Characters
Small	52601	27834	296887	20670	21424.93
Large	275565	147778	943947	20355	21426.08

The Samples column shows the number of text pairs, and the last three columns show the Characters' statistics of texts.

3. Method

3.1. Network Architecture

Given two texts, denoted as $text_1$ and $text_2$, the task of Authorship Verification is to decide whether they are written by the same author. Suppose $text_1 = \{t_{11}, t_{12}, \ldots\ldots, t_{1N}\}$, where t_{11} is the first fragment of $text_1$ and t_{1N} is the Nth fragment of $text_1$. $Text_2 = \{t_{21}, t_{22}, \ldots\ldots, t_{2N}\}$, where t_{21} is the first fragment of $text_2$ and t_{2N} is the Nth fragment of $text_2$. The N number set to 30, so a text pair would be split into 30 short text pairs. Figure 1 shows the network architecture.

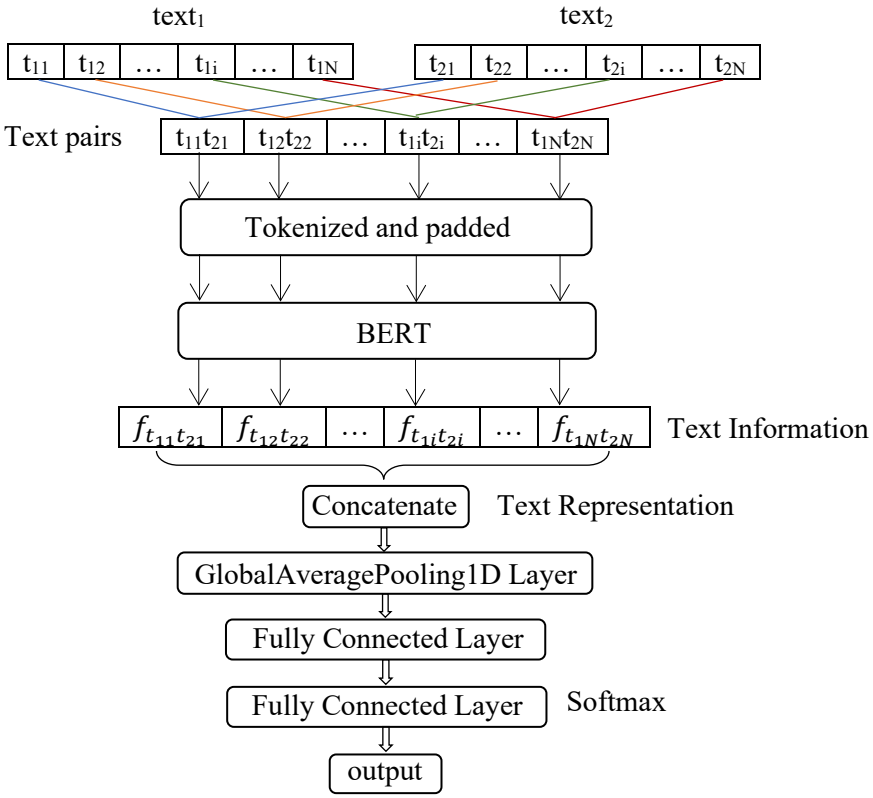

Figure 1: Architecture diagram for our model.

It can be observed, a text pair can be split into N short text pairs, and the $t_{1i}t_{2i}$ is one of the short text pairs that consist of the ith fragment of $text_1$ and $text_2$. By using BERT to encode these short text pairs,

we obtain more efficient text features, and the $f_{t_{1i}t_{2i}}$ is the text feature of the ith short text pair encoded. Then the text representation is obtained by concatenating these text features. Finally, we feed the text representation into a fully connected neural network to build a binary classification model, which determines whether two texts have been written by the same author.

3.2. Text Preprocessing

Text preprocessing corresponds to the first three steps in Figure 1. We use punctuation as a separator to intercept the first 30 fragments of each text sample. Now each train sample's text1 and text2 all have 30 fragments, and their labels are the same, so a text pair constructs 30 sub-training samples. We combine the corresponding fragments of text1 and text2, then they were tokenized and sequence padded to be a vector of max length 256.

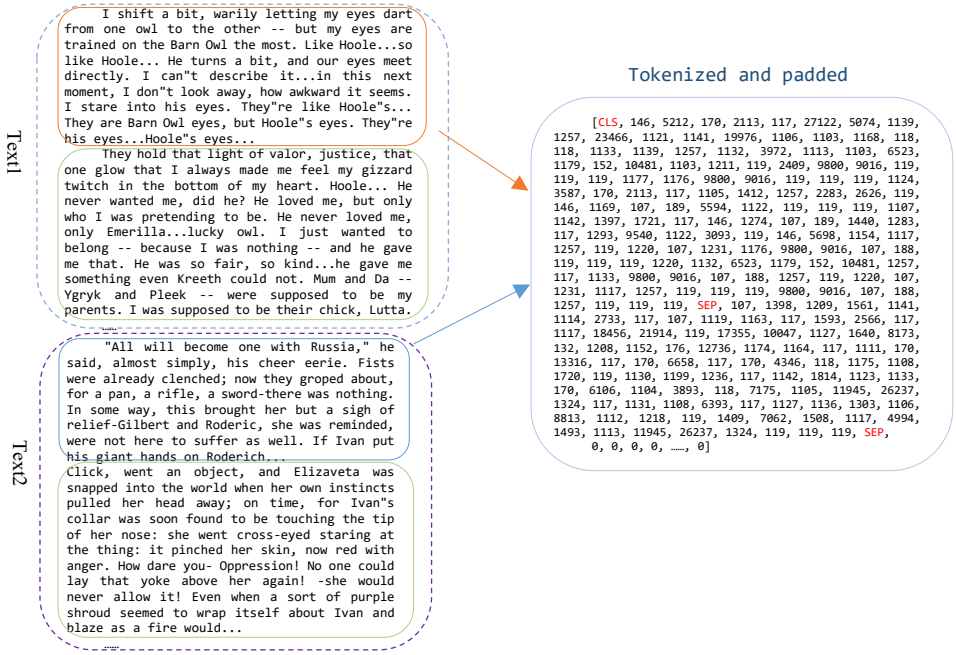

Figure 2: This is an example of how we preprocessed a text pair. The <CLS> is a special symbol added in front of every input example, and the <SEP> is a special separator token that can be used to separate two fragments in this example.

4. Experiments and Results

4.1. Experimental setting

In this work, $\text{BERT}_{\text{BASE}}$ (L=12, H=768, A=12, Total Parameters=110M) is chosen as pre-trained model size, and we use Keras to construct BERT and fully connected network classification model. A text pair is split into 30 short texts and tokenized to a vector that its shape is (30, M), and there have 52,601 such vectors on the small dataset, where the M is the max length of short texts. In the fine-tuning pre-trained model phase, we set $batch\_size = 30$ and use sparse categorical cross-entropy as the loss function, and the optimization method is Adam with a 2e-5 learning rate. Using BERT to encode text information, we obtain the feature vector and reshape it to (52601, 30, 768). After global pooling of one-dimension, its shape becomes (52601,768). The first fully connected layer output hidden size is 16, and its activation is ReLU. The other FC layer output hidden size is 2, and its activation is softmax. The final FC network is trained for 400 epochs, and its optimization is Adam.

4.2. Results

To evaluate the proposed model, two splitting strategies are adopted on the given training small dataset. The first one is 36,821 text pairs for the training dataset and 15,780 text pairs for the validation dataset, and the other is 45,000 for training dataset and 7,601 for valid dataset. The experimental results on the validation dataset are denoted as Val-1 and Val-2 separately.

Table 2
Validation results on the small dataset.

Datasets	AUC	c@1	f_05_u	F1	Brier	Overall
Val-1	0.920	0.921	0.921	0.926	0.921	0.922
Val-2	0.944	0.943	0.958	0.945	0.943	0.946

It can be observed, with the increase of the training sample, the overall score would be increased.

Table 3 shows the final evaluation results on the small datasets of the PAN 2021 authorship verification task evaluated on the TIRA platform [5]. Our model is denoted as peng21.

Table 3
Final results on the test dataset.

Team	AUC	c@1	f_05_u	F1	Brier	Overall
weerasinghe21	**0.9666**	0.9103	**0.9270**	0.9071	**0.9290**	**0.9280**
peng21	0.9172	**0.9172**	0.9200	**0.9167**	0.9172	0.9177
embarcaderoruiz21	0.9470	0.8982	0.8785	0.9040	0.9072	0.9070
menta21	0.9385	0.8662	0.8787	0.8620	0.8762	0.8843
rabinovits21	0.8129	0.8129	0.8186	0.8094	0.8129	0.8133
ikae21	0.9041	0.7586	0.7233	0.8145	0.8247	0.8050
unmasking21	0.8298	0.7707	0.7466	0.7803	0.7904	0.7836
naive21	0.7956	0.7320	0.6998	0.7856	0.7867	0.7600
compressor21	0.7896	0.7282	0.7027	0.7609	0.8094	0.7581

5. Conclusion

In this paper, we propose the method that utilizes a pre-trained model to encode text information to solve the authorship verification in the PAN@CLEF 2021. To resolve the problem of long text encoding, the method we proposed is to split long texts into short texts that a pre-trained model, BERT, can encode. As can be observed above Tabel 3, the classification model achieved the highest c@1 and F1-score on the small dataset of PAN Authorship Verification datasets. Accordingly, the approach described can encode long text information efficiently in long text pairs.

6. Acknowledgments

This work is supported by the National Natural Science Foundation of China (No.61806075 and No.61772177) and the Social Science Foundation of Heilongjiang Province (No. 210120002).

7. References

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

[2] Kestemont, M., Markov, I., Stamatatos, E., Manjavacas, E., Bevendorff, J., Potthast, M. and Stein, B.: Overview of the Authorship Verification Task at PAN 2021. Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (2021)
[3] Devlin J., Chang M.W., Lee K., et al. Bert: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019, 1: 4171-4186
[4] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, December 2017, pp:6000–6010.
[5] Potthast M, Gollub T, Wiegmann M, et al. TIRA integrated research architecture[M]//Information Retrieval Evaluation in a Changing World. Springer, Cham, 2019: 123-160.