Application of BERT in author verification task

Notebook for PAN at CLEF 2022

Ziwang Lei, Haoliang Qi*, Han Y, Zeyang Peng, Mingjie Huang

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. Authorship verification task for the competition PAN@CLEF 2022 is that given two texts belonging to different Discourse Types (DT), determine if they are written by the same author (cross-DT authorship verification). We propose a long text encoding method based on BERT, a pre-trained language model, to solve this task. We cut $text_1$ in a text pair into five segments. And $text_2$ is reserved when it is less than 510 characters, only the first 510 characters are reserved when $text_2$ is longer than 510 characters. Then each segment of $text_1$ is combined with $text_2$ to form a new text pair and input them into BERT for encoding. Finally, a classifier is used to get the classification label. The final score of our model in the test dataset is AUC=0.539, c@1=0.539, f_05_u=0.488, F1=0.399, Brier=0.539, overall=0.501.

Keywords

Authorship Verification, Pre-trained language model, Classification task

1. Introduction

Authorship verification technology has been applied in various fields. How to improve the accuracy of authorship verification has attracted more and more attention. The task of the PAN 2022 Authorship Verification[1][2] focuses on more challenging scenarios where each author verification case considers two texts that belong to different DTs (cross-DT authorship verification). the author sets of training and test dataset do not overlap, so it is difficulty to solve the problem for this task if we only model the author's writing style. We think that the pre-trained language model BERT[3] is an effective method to encode text features. Our motivation is to use Self-attention based BERT to capture more text feature information than traditional neural networks. Because BERT input can only be up to 512 tokens, we propose a strategy of text segmentation and interaction to input text data into BERT for encoding. Then we use the text feature information to judge whether the text pair comes from the same author. Authorship verification task is a binary classification problem[4].

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

EMAIL:Leiziwang@163.com(A.1);qihaoliang@fosu.edu.cn(A.2)(*corresponding author); hanyong2005@fosu.edu.cn(A.3); pengzeyang008@163.com(A.4) ;mingjiehuang007@163.com(A.5)

ORCID:0000-0001-7626-1643 (A. 1);0000-0003-1321-5820 (A. 2); 0000-0002-9416-2398 (A. 3); 0000-0002-8605-4426(A. 4);0000-0002-8605-4426(A. 5)

2. Datasets

The PAN 2022 provides cross-DT authorship verification cases using four DTs, they are essays, emails, text messages and business memos. The datasets has 12,264 pairs of texts. The train and test dataset consist of pairs of texts belonging to two different DTs. This means that all authors in the test dataset have not appeared in the training dataset. We counted the characters of $text_1$ and $text_2$, the statistical results are shown in the table 1. The length relationship of all sentence pairs is $text_1$ greater than $text_2$ in dataset.

Table 1

Statistics on the number of characters of sentence pairs in the dataset

Datasets	max character	min character	mean character
$text_1$	22,160	230	4,353
$text_2$	6,159	230	983

Since the text length of texts of emails and text messages can be very small, each text belonging to these DTs is actually a concatenation of different messages.

3. Model framework

Since BERT can only accept 512 tokens as input at most, we propose a method of text slicing to solve the problem that the number of input text characters is out of range. Suppose $text_1$ is the first of the text pair, $text_2$ is the second of the text pair. We found that the length of $text_1$ in all text pairs is longer than that of $text_2$. According to the text length characteristics of $text_1$ and $text_2$, we use punctuation as a separator to divide $text_1$ into 5 segments to make sure each segment consists of several complete sentences. And $text_2$ is reserved when it is less than 510 characters, only the first 510 characters are reserved when $text_2$ is longer than 510 characters. Figure 1 shows the framework of our model.

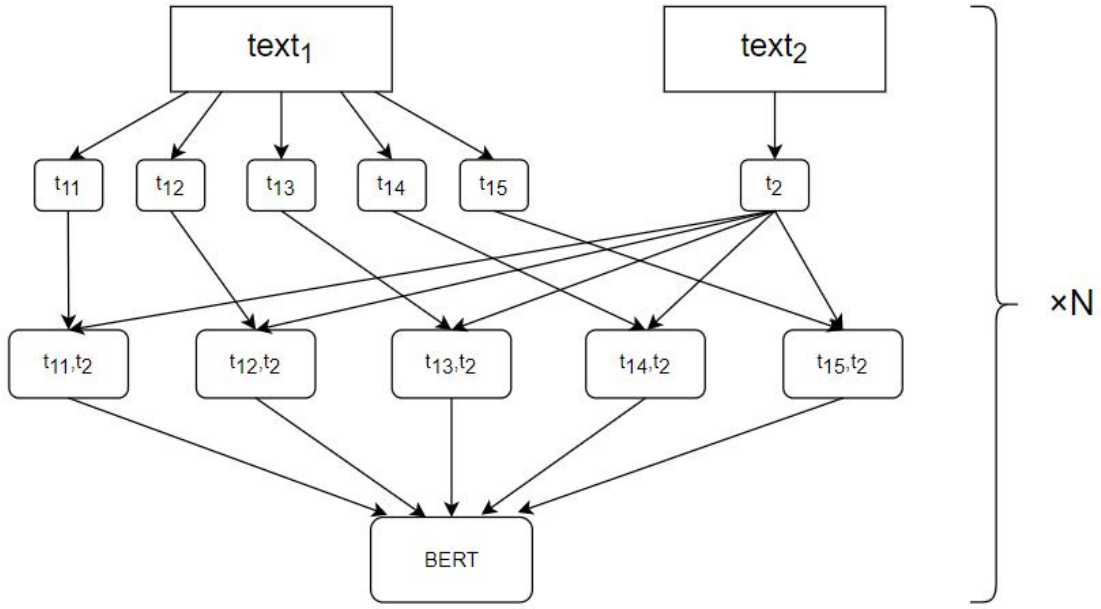

Figure 1: Model framework diagram of our method

After splitting, suppose $text_1 = \{t_{11}, t_{12}, t_{13}, t_{14}, t_{15}\}$, $text_2 = \{t_2\}$. We spliced the tokens of the five tokenized fragments with t_2 respectively, and we use the special separator <SEP> as their boundary. Then we input the restructured text pair into Bert for encoding. These five fragment pairs have the same classification tags. All N text pairs are processed in the above way, where N is 12,264. In this way, we can get the representation of the text. We put the representation of the text into a global average pooling layer to reduce the dimension. The output of pooling layer will be put into the fully connected neural network, and using softmax as the activation function to get a binary label. From this classifier, we can get the answer whether the two paragraphs of text are the same author. Inspired by the method of Peng[5]. The difference from his method is that we use different data segmentation strategies, and different ways of recombining the data after segmentation.

4. Experiments and Results

4.1 Data preprocessing

For the emails and text messages DTs, text is composed of multiple original messages through <new>tag, and new lines within a text are denoted with the <nl>tag. We think these tags have no contribution to extracting text feature information. Therefore, we removed these tags from all text. And we deleted all Emoji expressions contained in the text.

After the above preprocessing, we get the clean text. The average number of characters for $text_1$ is 4,043, the average number of characters for $text_2$ is 960, the average characters of $text_1$ and $text_2$ are 310 and 23 less than the original respectively.

4.2 Experiments

There are 12,264 text pairs in the training dataset. In order to test the effect of our model on open-set, we divide the training dataset into two parts: 11,000 pairs of training data and 1,264 pairs of test data. The authors of the text pairs of the segmented test data do not overlap with the training data. Before the final submission, we use the segmented dataset to train and test on our model.

The pretrained language model we use is $BERT_{BASE}$(L=12, H=768, A=12, Total Parameters=110M). and we use Keras to construct BERT and fully connected network classification model. We split $text_1$ into five segments, no more than 510 characters per segment, and $text_2$ is reserved when it is less than 510 characters, only the first 510 characters are reserved when $text_2$ is longer than 510 characters. We use these fragments to restructure text pairs. We obtain the feature vector and reshape it to (12264, 5, 768). Then we reshape it to (12264, 768) by a global average pooling. The final fully connected network is trained for 100 epochs. We set batch_size = 16 and the optimization method is Adam with a 2e-5 learning rate. We use sparse categorical cross-entropy as the loss function.

4.3 Results

We input the segmented training data and test data into our model for training and testing, then we use the official evaluation program to evaluate the results, The evaluation score is shown in table 2.

Table 2

Test results on dataset after segmentation, where D is the dataset after segmentation.

Datasets	AUC	c@1	f_05_u	F1	Brier	Overall
D	0.637	0.693	0.530	0.506	0.693	0.612

Table 3 shows the final evaluation results on the dataset of the PAN 2022 authorship verification task evaluated on the TIRA platform [6]. Our team name is lei22.

Table 3

Test results on dataset after segmentation, where D is the dataset after segmentation.

team	AUC	c@1	f_05_u	F1	Brier	Overall
lei22	0.539	0.539	0.488	0.399	0.539	0.501

5. Conclusions

In this paper, We propose a method based on pre-trained language model to solve the task of the PAN 2022. We use BERT to encode text information, since BERT can only receive no more than 512 characters, We split $text_1$ and $text_2$,reorganize fragment pairs, and then input them into BERT. This solves the problem that BERT cannot encode long text. Finally, the text feature information is put into a fully connected neural network, Make a binary classifier to identify whether two paragraphs of text are the same author.

However, our final experimental results are not good. One possibility is that the sentence pair loses too much information when entering BERT after splitting into fragments. Another possibility is that using BERT to encode text information is not suitable for authorship verification on open-sets.

6. Acknowledgments

This work is supported by the Social Science Foundation of Guangdong Province (No. GD20CTS02).

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] 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] 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

[5] Z. Peng, L. Kong, Z. Zhang, Z. Han, X. Sun, Encoding text information by pre-trained model for authorship verification, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2021.
[6] 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.