Author Verification Of Text Fragments Based On The Bert Model

Notebook for PAN at CLEF 2023

Ji Li, Qianjin Zhang*, Mingjie Huang

Foshan University, Foshan, Guangdong, China

Abstract

Author verification is the task of determining whether two texts were written by the same author. We treat author verification as a triage task. A long text encoding method based on the pre-trained language model BERT is proposed to solve this problem. We use the BERT model, use three different preprocessing methods, obtain three different representations, and then fuse all the representations to obtain the final model. The final score of our model in the test dataset is AUC=0.5, c@ 1=0.5, f_05_u=0.55, F1=0.646, Brier=0.5, and overall=0.539.

Keywords

Author verification, pre-trained language model, BERT model

1. Introduction

Author verification is the task of determining whether two articles were written by the same author based on the writing style of the text. In the 2022 edition, each author verification case considers two texts belonging to different DTs (cross-DT authorship verification). However, all considered DTs correspond to the written language. In the PAN2023 version [1, 2, 3], the first focus is on (cross-utterance types) author verification, where both written language (i.e., papers and emails) and spoken language (i.e., interviews and speech transcriptions) represent utterance types in the set. This will provide an opportunity to study the robustness and effectiveness of stylistic measurement methods under challenging and interesting conditions. In addition, the ability of author verification methods to handle different forms of expression in written and spoken language will be emphasized.

We believe that the pre-trained language model BERT [4] is an effective text feature encoding method. Our motivation is to use three different approaches to text preprocessing using a pre-trained language BERT model. Text data is entered into the BERT model for encoding. Then we use the text feature information to determine whether the text pair comes from the same author [5].

2. Dataset

Based on the new English corpus, PAN2023 provides cross-DT author verification cases, including paper (written discourse), email (written discourse), interview (spoken discourse), and speech transcription (spoken discourse). The training dataset contains 8836 pairs of text, all written in one JSON. The training dataset comes with two UTF8-encoded newline-separated JSON files. The first file, pairs.jsonl, contains pairs of text (each pair has a unique ID) and their utterance type labels: the second file, truth.jsonl, contains the basic facts of all pairs. The ground truth consists of a Boolean flag that indicates whether the paired text is from the same author and author ID. Because the text length of the email and interview text can be very small, each text that belongs to these DTs is a concatenation of different messages. We use <new> labels to represent the boundaries of the original message. New lines

[1] CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece
EMAIL: liji151029@163.com (A. 1); 1847317374@qq.com (A. 2)(*corresponding author); mingjiehuang007@ 163.com (A. 3)
ORCID: 0009-0005-2294-5878 (A. 1);0009-0002-8067-7830 (A. 2); 0000-0002-0889-5027 (A. 3)

in the text are represented by tags <nl>. In addition, author- and topic-specific information, such as named entities, has been replaced by appropriate labels. In spoken utterance types, additional labels are used to indicate nonverbal sounds (e.g., coughing, laughing).

3. Method

3.1. Text preprocessing

Because the text length of email and interview text can be very small, each text that belongs to these DTs is actually a concatenation of different messages. Labels are used to <new>represent the boundaries of the original message. New lines in the text are represented by tags<nl>. In addition, author- and topic-specific information, such as named entities, has been replaced by appropriate labels. In spoken utterance types, additional labels are used to indicate nonverbal vocalizations. We do not believe that these labels contribute to the text feature information. So we removed those tags from the text, and we also removed all the emojis and some punctuation marks in the text [6].

Table 1
The maximum and average values of text 1 and text2 before and after pretreatment are counted.

Text pairs	Maximum (before)	Average (before)	Maximum (after)	Average (after)
Text 1	24665	4638.85	24354	4513.24
Text2	7655	2354.62	7520	2019.60

3.2. Third level heading

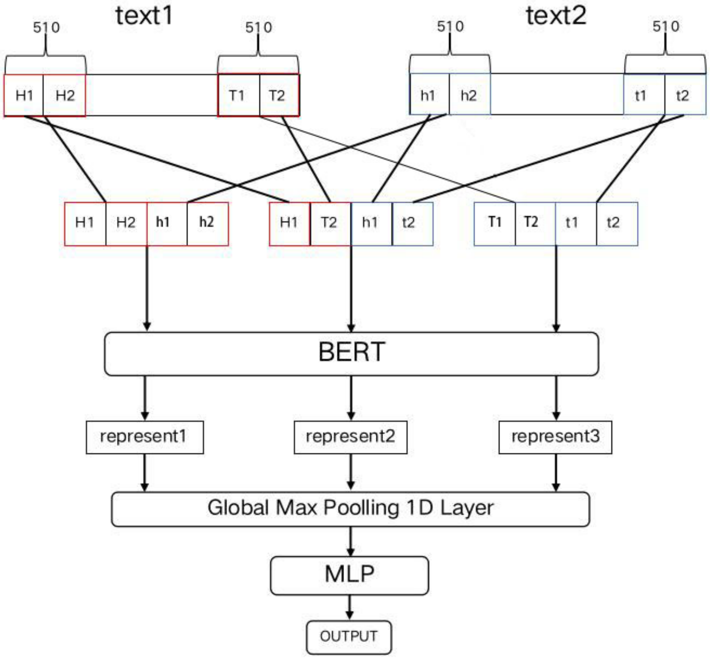

Figure 1: Model skeleton of our method.

We use three different preprocessing methods to process text pairs, first determine whether text1 and text2 are 510 characters long enough before processing, and double the text length if it does not reach 510 characters. Then start the text preprocessing: the first method takes the first 510 characters of text 1 and the first 510 characters of text2, respectively. The second method takes the last 510 characters of text 1 and the last 510 characters of text2, the third method takes the first 255 characters of text 1 plus the last 255 characters and the first 255 characters of text2 plus the last 255 characters. Then we input the reconstructed text pair of the three methods into Bert for encoding. All 8836 are processed in the above way so that we can get a representation of the text. We put the representation of the text into a global maximum pooling layer, and changed the original maximum pooling layer of 3×768 to the maximum pooling layer of 1×768 to reduce the dimension. The output of the pooling layer will be fed into a fully connected neural network to determine if the two original texts share the same author.

4. Experiments and Results

4.1. Experiment setup

We divide the 8836 text pairs of the training dataset into two parts, 6186 pairs of training data and 2650 pieces of test data. Before final submission, we use the test dataset to test and train our model.

In this work, BERTBASE (L= 12, H=768, A= 12, total parameters= 110M) is selected as the pre-trained model size, and we use Keras to build a BERT and fully connected network classification model. During the fine-tuning pre-trained model phase, our model training batch size is set to 2 and the maximum length of the encoder is set to 256. We use the Adam optimizer with the learning rate set to 1e-5, the dropout layer with the rate set to 0.2, and the sparse classification cross-entropy as the loss function [7]. We take the eigenvector and shape it into (8836, 1,768). It is shaped to (8836,768) by a global maximum pooling. The first fully connected layer output hides a size of 32, and its activation is ReLU. The other FC layer outputs a hidden size of 2, which is activated as softmax. The final FC network was trained on 500 epochs.

4.2. Results

We trained and tested with 6186 training datasets and 2650 test data before starting formal training, and then trained using official evaluation procedures, and Table 1 shows the experimental results.

Table 2

Table 2 shows the judging results on the PAN 2023 author verification task training dataset evaluated on the TIRA [8] platform.

Dataset	auc	c@ 1	f_05_u	F1	brier	overall
PAN	0.542	0.542	0.573	0.665	0.542	0.573

Table 3

Table 3 shows the judging results on the PAN 2023 author verification task training dataset evaluated on the TIRA [8] platform.

Dataset	auc	c@ 1	f_05_u	F1	brier	overall
PAN	0.5	0.5	0.55	0.646	0.5	0.539

5. Conclusions

In this paper, a pre-trained language model-based method is proposed to solve the task of PAN2023. We use the BERT model to process text information. Three different data preprocessing methods are used to input the processed data into the BERT model to obtain three different representations. The three representations are fused and then placed into a global maximum pooling layer to reduce the dimension. The output of the pool layer will be fed into a fully connected neural network, making a binary classifier to identify whether two pieces of text are the same author.

However, the model may obtain incomplete and partial information during data preprocessing. We are not sure whether it will have a significant impact on the final outcome, and we hope that someone will refine such issues in the future.

6. References

[1] E. Stamatatos, K. Kredens, P. Pezik, A. Heini, J. Bevendorff, M.Potthast, B.Stein. Overview of the Authorship Verification Task at PAN 2023. CLEF 2023 Labs and Workshops, Notebook Papers-Conference and Labs of the Evaluation Forum ,CEUR-WS.org. 2023.

[2] E. Stamatatos, M. Kestemont, K. Kredens, P. Pezik, A. Heini, J. Bevendorff, M. Potthast, B. Stein,M. Potthast. Overview of the Authorship Verification Task at PAN 2023. Working Notes of CLEF 2023:Conference and Labs of the Evaluation Forum, CEUR-WS.org(2023),Thessalonikki, Greece, sep 18-21, 2023.

[3] J. Bevendorff, I. Borrego-Obrador, M. Chinea-R{\'i}os, M. Franco-Salvador, M. Fr{\"o}be, A. Heini, K . Kredens, M. Mayerl, P. P\k{e}zik, M. Potthast, F. Rangel, P. Rosso, E. Stamatatos, B. Stein, M. Wiegmann, M. Wolska, E. Zangerle, Overview of PAN 2023: Authorship Verification, Multi-Author Writing Style Analysis, Profiling Cryptocurrency Influencers, and Trigger Detection, Lecture Notes in Computer Science, Springer, Thessalonikki, Greece, sep, 2023.

[4] J. Devlin, M.W. Chang, K. Lee, 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

[5] Pinzhakova M.,Yagel T.,Rabinovits J. Feature similarity-based regression models for authorship verification. CEUR Workshop Proceedings, 2936. 2021.

[6] Ziwang Lei, Haoliang Qi, Yong Han, Zeyang Peng, Mingjie Huang. Application of BERT in author verification task. 2022.

[7] Mingjie Huang, Kong L, Zeyang Peng, Yihui Ye, Zengyao Li, Xinyin Jiang, Zhongyuan Han . Authorship verification Based On Fully Interacted Text Segments. CLEF, 2022.

[8] M. Frobe, M. Wiegmann, N. Kolyada, et al., Continuous Integration for Reproducible Shared Tasks with TIRA.io, in: J. Kamps, L. Goeuriot, F. Crestani (Eds.), Advances in Information Retrieval. 45th European Conference on {IR} Research. Lecture Notes in Computer Science, 2023.