Postprint: Research on Chinese Weibo Authorship Identification Based on Deep Learning

Authors: Xu Xiaolin, Cai Manchun, Lu Tianliang

Date: 2018-11-29T00:00:00+00:00

Abstract

Authorship identification has consistently played a crucial role in public security and forensic document examination. The existing process of authorial linguistic style modeling is cumbersome, and text feature engineering lacks general applicability. To address this issue, we propose the CABLSTM model for Chinese Weibo authorship identification without requiring expert feature modeling, and evaluate its accuracy on a public Weibo corpus. To maximize short text feature extraction, the model integrates the Attention mechanism into CNN while removing the pooling layer, employs bidirectional LSTM to capture contextual information, and outputs identification results through a Softmax layer. Experimental results demonstrate that, compared with traditional machine learning algorithms as well as TextCNN and LSTM algorithms on the task of Chinese Weibo authorship identification, the proposed model achieves improvements in accuracy, recall, and F-score.

Full Text

Preamble

Vol. 37 No. 1
Application Research of Computers
ChinaXiv Partner Journal

Research on Chinese Microblog Authorship Identification Based on Deep Learning

Xu Xiaolin, Cai Manchun, Lu Tianliang
(School of Information Technology and Network Security, People' s Public Security University of China, Beijing 102623, China)

Abstract: Authorship identification has always played a crucial role in public security and forensic document examination. Existing methods for model-

ing authorial linguistic style are cumbersome and lack universal applicability in text feature engineering. To address this problem, this paper proposes the CABLSTM (CNN-Attention-BiLSTM) model for Chinese microblog authorship identification without requiring expert feature engineering, and evaluates its accuracy on a public microblog corpus. To maximize feature extraction from short texts, the model integrates an attention mechanism into CNN while removing pooling layers, and employs bidirectional LSTM to capture contextual information. Identification results are output through a Softmax layer. Experimental results demonstrate that compared with traditional machine learning algorithms as well as TextCNN and LSTM, the proposed model achieves improvements in accuracy, recall, and F-measure for Chinese microblog authorship identification tasks.

Keywords: authorship identification; LSTM; CNN; automatic feature extraction

0 Introduction

Authorship identification is a critical task in forensic linguistic analysis, representing an interdisciplinary field of applied linguistics and computer science. The core idea is to quantify an author' s unconscious writing habits embedded in texts, highlighting stylistic and writing style characteristics to determine the author of anonymous texts. In public security work, harmful information author identification can also be based on textual analysis to identify suspects, and authorship identification provides analytical support for both applications.

Previous research has primarily focused on long texts, progressively developing from unigram features to multigram and multi-level text features, enabling deeper and more abstract text characteristic extraction to improve authorship identification accuracy. However, with the rapid development of the internet, massive amounts of online text have emerged, including emails, blogs, microblogs, and comments. While these short texts are abundant, long-text authorship identification methods cannot be directly applied to short texts. Current research on short texts remains limited. Qi Ruihua et al. [1, 2] extracted features from microblog short texts based on vocabulary, sentence structure, dependency relations, special symbols, and other multi-dimensional aspects to achieve authorship identification. However, this approach cannot provide a unified feature extraction method for all short texts—different short texts require different feature extraction strategies, and special symbols in microblogs significantly improve accuracy but are not universal across ordinary short texts.

Most microblog content contains fewer than 140 characters, making feature extraction extremely challenging. However, microblog posts often reflect authors' casual expressions, better representing their linguistic style. While existing feature extraction methods based on various text characteristics and microblog-specific features have achieved good results, they are tailored to specific short

text types and cannot avoid manual feature engineering by experts. Therefore, this paper proposes a deep learning-based Chinese microblog authorship identification model that automatically extracts features from short texts, eliminating the need for expert feature engineering, and tests its effectiveness on a public microblog corpus.

1 Chinese Microblog Authorship Identification Model

Deep learning possesses the capability to automatically learn and extract features. This paper leverages deep learning for authorship identification. Convolutional Neural Networks (CNN) have an effect similar to n-gram models and can extract text features through multiple convolutional layers for deeper mining. Long Short-Term Memory networks (LSTM) extract features from sequential data and effectively capture contextual information. Therefore, we combine CNN with an attention mechanism as a short-text feature extractor and integrate it with a classifier for output.

The CABLSTM model workflow is shown in [Figure 1: see original paper].

1.1 Text Preprocessing

1) Microblog Crawling
The required data consists of two categories: large-scale microblog data for building word vectors, and labeled microblog data for experiments. Since building word vectors requires substantial data, we utilized 40 GB of microblog data obtained from CSDN. Experimental data was crawled using Python' s request package and regular expressions. We manually selected candidates who met the requirements and had posted over 1,000 microblogs, then crawled data from 10 individuals, totaling 10,000 posts.

2) Text Segmentation
Training corpora were first segmented using Chinese word segmentation tools. Popular open-source Python tools include Jieba, NLPIR, and LTP. After conducting segmentation experiments on microblog corpora, we selected NLPIR, which achieved the highest accuracy.

3) Word Vector Generation
After segmentation and stop-word removal, we used Word2vec [4] with the CBOW model to establish word vectors. The input layer consists of n-1 word vectors surrounding target word x, which are summed and fed into the hidden layer. Starting from the root node, the mapping layer values undergo logistic classification along the Huffman tree while continuously adjusting intermediate vectors and word vectors, ultimately outputting the word vector for x.

1.2 CNN+Attention Mechanism for Text Feature Extraction

Due to the extreme brevity of microblog data, more abstract and high-level feature representation is necessary for effective text feature modeling. CNN

can perform convolution operations and, given its n-gram feature extraction capability, is well-suited for microblog short-text feature extraction. To enable deeper feature mining, we improved the CNN architecture as follows:

a) Removing the Max-Pooling Layer
While pooling layers reduce output vector dimensions, they also lose partial features. Therefore, we removed the Max-Pooling layer during feature extraction to maximize CNN' s convolutional feature extraction capability.

b) Adding an Attention Layer Before CNN Convolution
Traditional CNN processes each sentence through single channels, learns sentence representations, and feeds them into the classifier. This approach lacks inter-sentence connections before classification, only learning local features. Through the attention mechanism [9, 10], we construct an attention matrix between sentence pair s1 and s2. Sentence s1 multiplies with the attention matrix to obtain an attention feature map, transforming the convolutional layer input from single-channel to dual-channel. This connects sentence pairs across different CNN channels, enabling full-text feature learning and improving extraction effectiveness.

As shown in [Figure 2: see original paper], we first calculate attention matrix A, where each element represents the match_{score} between the i-th word in sentence 1 and the j-th word in sentence 2. Empirical results show that Euclidean distance works effectively as the match_{score}. The calculation formula is:

$$A_{i,j} = \text{match\_score}(F(s_{1,i}), F(s_{2,j}))$$

where match_{score} is Euclidean distance. The formulas are:

$$F(s_{1,i}) = W_0 \cdot s_{1,i} + b_0$$
$$F(s_{2,j}) = W_1 \cdot s_{2,j} + b_1$$

W_0 and W_1 are learnable parameter matrices. This paper uses shared weights (the same W) for both matrices. Multiplying s_1 with A and s_2 with the transpose of A yields two attention feature maps of the same size as the original sentence word vector matrices:

$$\text{attention\_feature\_map}_1 = s_1 \times A$$
$$\text{attention\_feature\_map}_2 = s_2 \times A^T$$

A sentence's word vector matrix and its attention feature map form two channels as CNN convolutional layer input. Using fixed window-size filters for convolution yields feature maps. To feed into LSTM, we concatenate corresponding positions from each feature map to construct the window feature sequence without connecting pooling layers, maximizing text feature mining.

1.3 Bidirectional LSTM+Softmax for Classification Output

Unidirectional LSTM at time step t only incorporates previous input information I_t, containing context from preceding text but not subsequent text. Bidirectional LSTM adds a reverse-direction LSTM, where inputs at time t represent both preceding (I_t) and following (I'_t) context information. To better extract microblog short-text features, we employ bidirectional LSTM with Softmax [11] for classification output.

As shown in [Figure 2: see original paper], CNN+Attention extracts text features and feeds the window feature sequence into bidirectional LSTM, producing two one-dimensional vectors. To preserve features effectively, we use concatenation rather than averaging to combine the vectors, preventing feature loss. Finally, fully connected layers and a Softmax layer perform classification.

The experimental workflow is shown in [Figure 3: see original paper].

2 Experiments

2.1 Experimental Data Source

The experimental data consists of Sina Weibo posts crawled for this study, comprising 10,000 microblogs from 10 public figures (1,000 posts per person). Post lengths range from 45 to 140 characters. We employed 10-fold cross-validation and evaluated authorship identification performance using average precision, recall, and F-measure across all comparison experiments.

2.2 Experimental Environment

All experiments were implemented in Python 3.6 on an Alineware machine with an i7 CPU, 16 GB RAM, Linux OS, and GTX 1070 GPU.

2.3 Microblog Word Segmentation Accuracy Comparison Experiment

We conducted comparative experiments using three popular segmentation tools: Jieba, NLPIR, and LTP. Most evaluations use the "People' s Daily corpus," which is strictly formal with low colloquialism and minimal internet slang, differing significantly from microblog corpora. To select the most accurate tool for microblog short texts, we used microblog data for comparison.

Since no standard segmented microblog corpus exists, we manually segmented 3,000 crawled microblog posts using '|'as the delimiter. A segmentation example is shown in .

The experimental workflow is shown in [Figure 3: see original paper].

We compared three datasets: 1) manual segmentation; 2) segmentation by Jieba, NLPIR, and LTP; 3) segmentation by Jieba, NLPIR, and LTP with user-defined dictionaries. The user-defined dictionaries consisted of popular microblog terms

and internet slang from the past five years. Accuracy was measured by comparing results against 3,000 manually segmented posts; processing time was tested on 100,000 posts.

Results are shown in . Key findings: (a) All three algorithms achieved over 90% accuracy on Chinese microblog segmentation, proving their effectiveness; (b) Adding user-defined dictionaries improved accuracy for all three algorithms; (c) NLPIR from the Chinese Academy of Sciences achieved the highest accuracy regardless of dictionary usage, reaching 98% with user-defined dictionaries; (d) NLPIR also had the shortest processing time—approximately one-third of the other two algorithms. Therefore, NLPIR with user-defined dictionaries offers the highest accuracy and lowest time consumption, making it optimal for word vector construction and improving model effectiveness.

2.4 Chinese Microblog Authorship Identification Algorithm Comparison Experiment

To validate the CABLSTM model' s effectiveness and superiority, we used precision (P), recall (R), and F-measure as evaluation metrics, with F1-score providing a more objective reflection of comprehensive performance.

We first verified CABLSTM' s effectiveness by comparing it with five algorithms: SVM, Decision Tree C4.5, TextCNN, LSTM, and CABLSTM. For SVM and C4.5, the Chinese microblog feature set consisted of lexical frequency features, punctuation counts, function word frequencies, and part-of-speech tagging features. Results are shown in .

Key findings: (a) All five algorithms achieved over 70% average precision, recall, and F-measure, with each author reaching over 69%; (b) TextCNN and LSTM performed similarly to traditional machine learning algorithms (SVM and C4.5) that rely on manual feature engineering; (c) The improved CABLSTM model achieved improvements across all metrics compared to the other four algorithms. Specifically, the model can more deeply mine short-text features to provide better feature representations for classification. The CNN+Attention mechanism enhances deep learning' s text extraction capability, while bidirectional LSTM with Softmax better learns features and performs classification.

Compared with traditional SVM and C4.5 algorithms, CABLSTM eliminates the manual feature engineering process based on lexical frequency, punctuation, function words, and part-of-speech features, reducing human effort while improving accuracy. Compared with TextCNN and LSTM, CABLSTM achieves higher precision, recall, and F-measure. Thus, CABLSTM can be better applied to Chinese microblog authorship identification, providing theoretical and technical support for harmful information author identification in public security and forensic document examination.

3 Conclusion

This research expands the theoretical framework and application scope of authorship identification. Considering differences between traditional long texts and online short texts in feature extraction, we addressed limitations in existing short-text feature modeling. To overcome the necessity of manual feature engineering in current Chinese microblog authorship identification, we proposed the CABLSTM model based on deep learning for automatic microblog text feature extraction and classification.

The model eliminates the manual feature engineering process in authorship identification, reducing human effort and improving efficiency. When public security agencies have databases of key individuals and their published statements, this model can analyze harmful statements of unknown authorship, providing theoretical and technical support for author identification in public security and forensic fields. Future research will focus on improving accuracy for Chinese microblog authorship identification with larger numbers of authors.

References

[1] Qi Ruihua, Yang Deli, Guo Xu, et al. Blogger identification based on multidimensional stylistic features [J]. Journal of the China Society for Scientific and Technical Information, 2015, 34(6): 628-634.

[2] Qi Ruihua, Guo Xu, Liu Caihong. Authorship attribution of Chinese microblog [J]. Journal of the China Society for Scientific and Technical Information, 2017, 36(1): 72-78.

[3] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality [J]. Advances in Neural Information Processing Systems, 2013, 26: 3111-3119.

[4] Zhang Dongwen, Xu Hua, Su Zengcai, et al. Chinese comments sentiment classification based on word2vec and SVMperf [J]. Expert Systems with Applications, 2015, 42(4): 1857-1863.

[5] Roska T, Chua L O. The CNN universal machine: an analogic array computer [J]. IEEE Trans on Circuits & Systems II Analog & Digital Signal Processing, 2015, 40(3): 163-173.

[6] Chua L O, Roska T. The CNN paradigm [J]. IEEE Trans on Circuits & Systems I Fundamental Theory & Applications, 1993, 40(3): 147-156.

[7] Gers F A, Schmidhuber J, Cummins F. Learning to forget: continual prediction with LSTM [J]. Neural Computation, 2000, 12(10): 2451-2471.

[8] Graves A. Supervised sequence labelling with recurrent neural networks [J]. Studies in Computational Intelligence, 2008, 385.

[9] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate [J]. Computer Science, 2014.

[10] Xu K, Ba J, Kiros R, et al. Show, attend and tell: neural image caption generation with visual attention [J]. Computer Science, 2015: 2048-2057.

[11] Salakhutdinov R, Hinton G E. Replicated softmax: an undirected topic model [C]// Proc of International Conference on Neural Information Processing Systems. [S. l.]: Curran Associates Inc, 2010: 1607-1614.

[12] Roska T, Chua L O. The CNN universal machine: an analogic array computer [J]. IEEE Trans on Circuits & Systems II Analog & Digital Signal Processing, 2015, 40(3): 163-173.

Note: Figure translations are in progress. See original paper for figures.