A Weakly Supervised Text Classification Method Based on Vocabulary Construction

Peidong Li, Di Lin(✉), and Zijian Li

University of Electronic Science and Technology of China, Chengdu, Sichuan, China
202022090635@std.uestc.edu.cn, lindi@uestc.edu.cn

Abstract. Text classification is an important research direction in natural language processing. The computer can automatically classify and label texts according to certain classification standards through text classification technology. Traditional text classification tasks require a large amount of labeled data. However, human-labeled data is not only expensive, but also susceptible to the subjective consciousness of the labelers. Therefore, unsupervised text classification using computers becomes relevant. In most cases, the label name of each category is instructive for the classification task. In this paper, we design a weakly supervised text classification method. This method only needs to provide the label names that guide the classification to complete the automated text classification. Our method was tested on several publicly available datasets and performed well.

Keywords: Text Classification · Weak Supervision · Word Vocabulary

1 Introduction

We are in an era of information explosion, with billions of people active on the Internet every day, and more and more people expressing their views and opinions by publishing text messages. The automatic classification of a large amount of text is a very important part of natural language processing tasks. Efficient classification can provide a good dataset for other applications and is a great help in filtering out useful information on the Web. Examples include summary generation, spam detection, and sentiment analysis. To build an automatic text classification, it is generally necessary to use human-labeled documents as the basis for model's learning and training. Some deep learning-based classifiers such as CNN [1] and RNN [2] possess powerful representation learning capability and can effectively capture the semantic information of text sequences. By inputting a large amount of human-labeled texts to train the classifier, a more accurate classification can be achieved.

In June 2017, Google team proposed the classic work of NLP, Transformer, in their paper "Attention Is All You Need" [3], using a new model to completely replace RNN. The Transformer model proposed in this paper outperforms RNNs and CNNs in machine translation tasks, using only encoder-decoder and attention mechanisms. Decoder and attention mechanisms to achieve good results, and its biggest advantage is that it can

X. Jiang (Ed.): MLICOM 2022, LNICST 481, pp. 128–136, 2023.
https://doi.org/10.1007/978-3-031-30237-4_12

be efficiently parallelized. In 2018, the BERT [4] model was born. It performed well in SQuAD1.1 and recorded the best results in 11 different NLP tests. The release of the Transformer architecture creates a new baseline for deep learning text classification methods. Many new models and methods based on the Transformer architecture are created [5]. Lan et al. [6] proposed the ALBERT model to achieve cross-layer parameter sharing while retaining the original Transformer encoder architecture, and the improved model surpasses BERT in all aspects. In order to solve the problem of context fragmentation that BERT is not able to encode the whole article input at one time, Transformer-XL [7] adopts two levels of sentence and paragraph looping mechanism and relative position encoding scheme. This means that the input sequence does not need to be split into arbitrary fixed lengths, but can follow natural linguistic boundaries, such as sentences and paragraphs. This not only helps to understand the deep context of multiple sentences, paragraphs and possibly longer texts, but also enables training on larger datasets. RoBERTa [8] makes several adjustments based on BERT: 1) longer training time, larger batch size, and more training data; 2) remove the next predict loss; 3) longer training sequences; 4) use dynamic adjustment Masking mechanism on training strategy.

Weakly supervised approaches in text classification have attracted a lot of attention from researchers in recent years, as it relieves experts from the burden of annotating a large number of documents, especially for specific domains. In general, in a weakly supervised model, each category contains some tag document or seed word and other open data [9]. Although these forms are much weaker than a fully labeled data, they still require additional knowledge from language experts.

In this paper, we design a weakly supervised text classification method that only need to use the textual information of category names. we expand the description information for classes by constructing a vocabulary. Then, the text with high confidence is added to the training set, and pseudo labeled according to the weighted algorithm we designed. Finally, the pseudo labeled data set is input into the classification model for training.

We conducted experiments on some publicly available datasets for comparison between different methods, and the experimental results shows that our method is more effective than previous methods mentioned in this paper.

2 Related Work

2.1 BERT

BERT is an autoencoder language model, and it uses two tasks to pre-train the model.

The first task is the using of "Mask". In order to make the model learn the ability of bidirectional coding effectively, BERT used masking language model (MLM) in the training process, that is, masking some positions in the input sequence at random, and then predicting them through the model. Because the MLM prediction task can make the result of model coding contain the context information of the context at the same time, it is beneficial to train a deeper BERT network model.

The second task adds a sentence level continuity prediction task on the basis of the bidirectional language model, that is, two sentences were input into the model at the same time, and then the second sentence was predicted to be the next sentence of the first sentence.

Compared with RNN, which relies on the previous calculation to calculate the features of the current word, self-attention in BERT's Transformer-encoder uses context information to calculate the features of the current word. Moreover, it extracts relational features from different levels, which can solve the problem of ambiguity of a word and reflect sentence semantics more comprehensively. Due to the advantages mentioned above, the BERT model has been widely used [10] in natural language processing tasks. In our work, we use BERT for two purposes: (1) representing documents by vector and (2) training a text classifier.

2.2 Weakly Supervised Text Classification

Weakly supervised text classification aims to classify text documents based only on word-level descriptions of each category such as their class names, avoiding the dependence on any labeled documents. Weakly supervised methods can make use of external knowledge [11], such as Gabrilovich et al. who represent text with pre-defined natural concepts that are weighted and easy to understand. An important advantage is that it makes use of a large amount of human-edited knowledge in the encyclopedia. A machine learning approach is used to build a "semantic interpreter" that can shadow natural language text fragments to a weighted vector of wiki concepts. We also can make use of keyword information. For example, ConWea [12] automatically group the same word into an adaptive number of different interpretations based on the contextual representation and the seed information provided by the user, and use the contextual corpus to disambiguate the keywords. WeSTClass [13] contains two modules. One is a text generation module, which is used to generate training set with labels based on the seed information. Another one is a self-training module, which refines the model on the labeled dataset using bootstrap method. WeSHClass [14] extends WeSTClass to hierarchical labels. XClass [15] classifies the data by the idea of clustering. Specifically, a comprehensive category representation is first estimated by incrementally adding the most similar words in each category until inconsistencies arise. Following a tailored hybrid class-holding mechanism, the authors obtain document representations by a weighted average of content-oriented labeled representations. Then, the documents are clustered and the prior of each document is assigned to its closest class. Finally, the most plausible documents from each clustering class are selected to train a text classifier. The method proposed in this page is improved on the basis of LOTClass [16]. Considering the words in vocabulary have different weights for the classification guidance, we use a weighed calculation mechanism. Finally, we use Baidu's translation API to expand the dataset.

3 Method

The classification method in this paper mainly includes two parts. The first part is to predict the replaceable words through the Bert model to construct the vocabulary. In the second part, we use the vocabulary to filter out documents as training set and train the classifier.

3.1 Vector Representation of Text

Firstly, the model takes a single character in the text as the minimum unit token and performs token ID conversion word by word. Then each ID corresponds to a specific vector expression. BERT model adds CLS vector at the initial position of each text. The CLS position vector is randomly assigned at the beginning of training, there is no obvious semantic information, and the CLS vectors of each language segment are independent of each other. The purpose is to take the CLS vector as an object that integrates the global semantic information of the text. The dashed box in the figure represents the preprocessed single text data matrix x0, the first column represents CLS vector, X1 to Xn represent each character vector, and the length of each vector has 768 dimensions (Fig. 1).

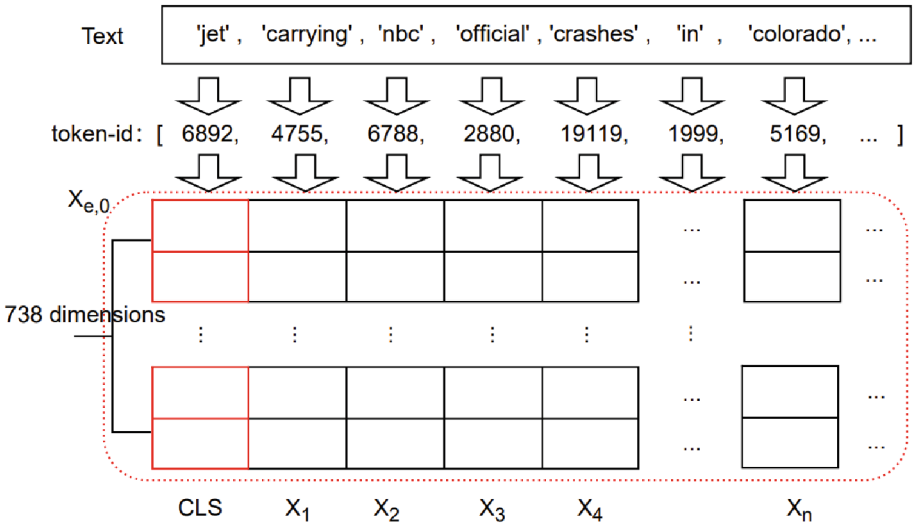

Fig. 1. Vector representation of text.

Each coding module includes multi head attention mechanism and feedforward neural network. Residual connection and layer normalization are used to solve the problem of saturation of feature extraction ability. The loaded pre training BERT model parameters include the weight matrix of each layer to replace the randomly generated values in the original BERT model. Each encoder of BERT has an output, and each output can be used to represent the text vector (Fig. 2).

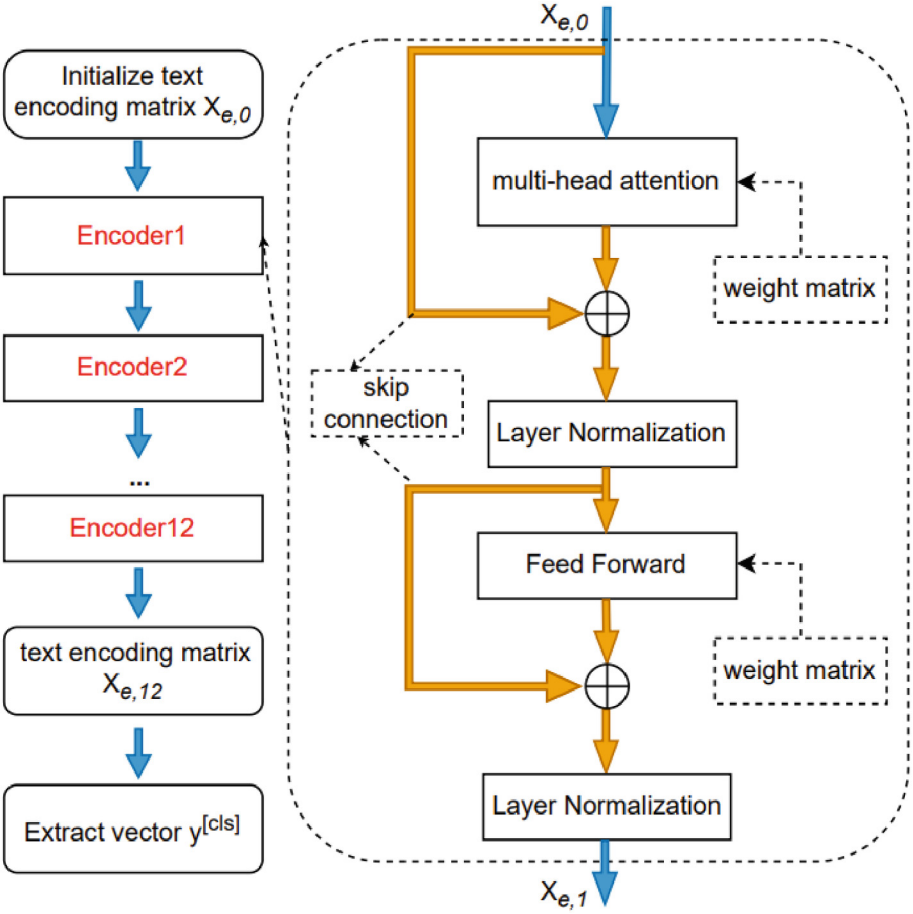

Fig. 2. Encoders' structure.

3.2 Expanding the Description of a Category by Label Name

In the original sentences, there must exist sentences containing label names. In these sentences, we do replaceable word prediction for the location of these label names and get the top-50 replaceable words for each sentence, and the replaceable words are those with similar meaning to the category labels. Each sentence is expanded with category description information using such method. Finally, each category has a certain number of words, and we filter these words to get the top-110 words in terms of frequency of occurrence. We show the predict top-50 task's result in Table 1. The result of the vocabulary for each category are shown in Table 2.

Table 1. Sentence and the predict top words.

Sentence	**Predict words**
On standby, set to fly. TORONTO -- Last week of August, the playoff race is heating up, and a Red Sox player's thoughts turn to… Football? In some ways, these guys are no different from the typical **sports** fan on the cusp of an NFL season. Sure, they're focused on the business at hand, winning a spot in their own postseason tournament…	'athletics', 'baseball', 'team', 'players', 'teams', 'athletes', 'club' …
Chain Store Sales Up 0.1 Percent - Report. NEW YORK (Reuters) - U.S. chain store sales rose in the latest week, but business continued to fluctuate from the effects of Hurricane Charley and high gasoline prices, a report said on Tuesday	'trade', 'shop', 'money', 'market', 'commerce', 'store', 'venture' …

Table 2. Category vocabulary.

Label Name	**Vocabulary**
politics	'politics', 'politicians', 'government', 'elections', 'democratic', 'party', 'state', 'leadership', 'election', 'politically', 'affairs', 'issues', 'governments', 'voters', 'debate', 'cabinet', 'congress', 'democrat', 'administration', 'president', 'religion', 'republican', 'history', 'war', 'crisis', 'legislature', 'candidates', 'governance', 'opposition', 'problems', 'relations', 'finance', 'justice', 'struggle', 'rhetoric' …
sports	'sports', 'athletics', 'game', 'national', 'news', 'athletic', 'espn', 'basketball', 'arts', 'baseball', 'tv', 'hockey', 'pro', 'press', 'team', 'red', 'home', 'bay', 'kings', 'legends', 'city', 'winning', 'miracle', 'olympic', 'go', 'giants', 'champions', 'ball', 'players', 'boxing', 'teams', 'athletes', 'tennis', 'club', 'coaches', 'gold', 'west', 'toronto', 'classic', 'pittsburgh', 'super', 'nfl', 'magic', 'key' …
business	'business', 'trade', 'commercial', 'enterprise', 'shop', 'money', 'market', 'commerce', 'corporate', 'global', 'future', 'sales', 'general', 'group', 'retail', 'companies', 'management', 'operations', 'operation', 'corporation', 'store', 'division', 'firm', 'venture', 'brand', 'contract', 'revenue', 'economic', 'branch', 'subsidiary', 'personal', 'cash', 'short', 'line', 'bank', 'customer', 'concern', 'family', 'work', 'products', 'big', 'scientific', 'virtual' …
technology	'technology', 'tech', 'software', 'technological', 'device', 'equipment', 'hardware', 'devices', 'system', 'technique', 'digital', 'technical', 'concept', 'systems', 'functionality', 'material', 'process', 'facility', 'feature', 'capability', 'content', 'security', 'ability', 'network', 'internet', 'computing', 'modern', 'communication', 'language', 'mechanism', 'computer', 'design', 'cyber', 'standard', 'tool', 'development', 'format', 'protocol', 'wireless' …

3.3 Pseudo-Tagging of Text Data by a Word List of Categories

We use the Mask method to make replaceable predictions for each word of each sentence, extracting the top 50 replaceable words at that position. For the word list of each category in the previous step, which has a size of 110, the higher ranked words appear more frequently, we weight the word list. Every ten words are divided into a group, and the weight of each group decreases from 1.5 to 0.5. By weighting calculation, we can get the weight of each sentence corresponding to different categories. If the weight is greater than 25, then we think the sentence can be pseudo-labeled as belonging to a certain category, and if there are more than one categories with weights greater than the threshold of 25, we take the category with the highest weight.

3.4 Classify Task

We use back-translation method to enhance the data set to make the classification more stable. For the problem that the number of documents in each category may vary too much, we use the resample approach to alleviate. The classifier is then trained by the BERT classification model.

4 Experiment

4.1 Dataset

We used AgNews and IMDB datasets for our experiments. The AgNews is a news article dataset, which contains four categories, politics, sports, business and technology, with 120,000 documents in the training set and 7,600 documents in the test set. IMDB is the review text data of movies, containing both categories of positive and negative, with 560,000 documents for training documents and 70,000 documents for testing.

4.2 Expirement Designing

The BERT model inserts a [CLS] symbol in front of the text and uses the output vector corresponding to this symbol as the semantic representation of the whole text. It can be understood that this symbol with no obvious semantic information will "fairly" integrate the semantic information of each word in the text compared with other words in the text. Therefore, we use the first position as the vector representation of the text, instead of the Mean approach where all positions are considered.

We use the BERT-base-cased as the basis of the experiment and leave all hyper-parameters unchanged.

4.3 Result

The Fig. 3 shows the accuracy and the variation of loss when training the classification model on the AgNews dataset.

After experiment, the results show that our method has some improvement in accuracy compared to other methods on AgNews as well as IMDB datasets. The following

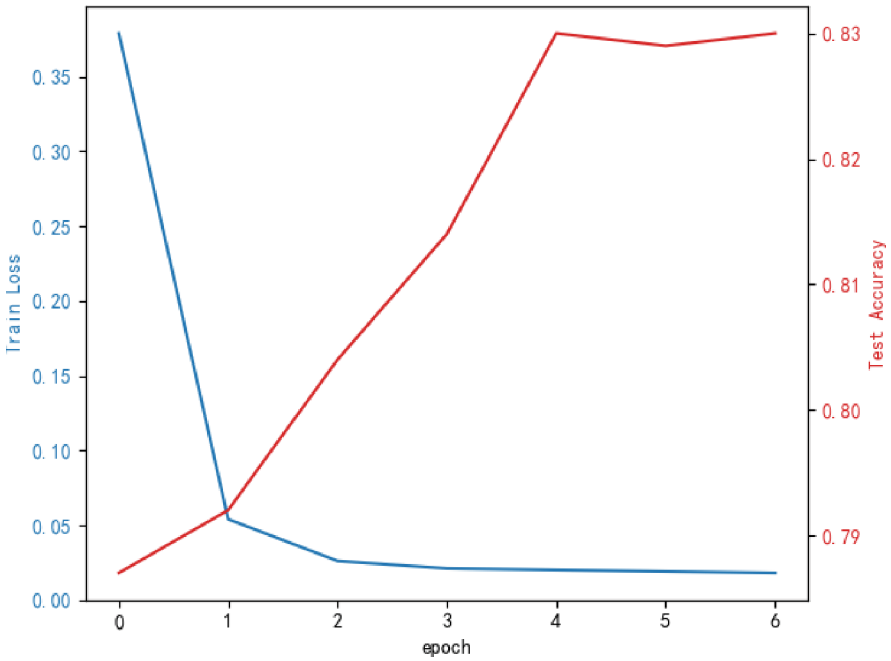

Fig. 3. Experiment's train loss and test accuracy.

data is the result of the experiment when the vocabulary size is 110. In fact, when we take a smaller vocabulary size of 80, the accuracy of pseudolabeling improves from 79.7% to 81.4% and the performance in the testing set also improves a little, although we obtain 13% less pseudo-labeled documents. Controlling the size of the vocabulary for different datasets may make the results better. Table 3 shows our experiment' results.

Table 3. Accuracy comparison.

Method	**AgNews**	**IMDB**
BERT simple match	0.752	0.677
WeSTClass	0.823	0.774
Our method	0.829	0.812

5 Conclusion

We design and implement a weakly supervised text classification method and achieved good experimental results. In our work, the Mask method is used for constructing a vocabulary of category description information and pseudo-label documents. These

pseudo-labeled text documents are used as the training set to train the classification model. Through the above methods, we achieve weak supervision. Although the accuracy of this method is only about 82%, it is still acceptable considering only relying on label names. We believe that if the relationship between sentences can be taken into account in the process of filtering out training set, a higher quality data set may be obtained. This work may make the final result better.

References

1. Zhu, X., Huang, J., Zhou, Z., Han, Y.: Chinese article classification oriented to social network based on convolutional neural Networks. In: 2016 IEEE First International Conference on Data Science in Cyberspace (DSC), pp. 33–36 (2016)
2. Xiao, J., Zhou, Z.: Research progress of RNN language model. In: 2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA), pp. 1285–1288 (2020)
3. Vaswani, A., et al.: Attention is All you Need **1706**, 03762 (2017)
4. Devlin, J., Chang, M.W., Lee, K., et al.: Bert: Pre-training of deep bidirectional transformers for language understanding **1810,** 04805 (2018)
5. Ohashi, S., Takayama, J., Kajiwara, T., et al.: Text classification with negative supervision. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 351–357 (2020)
6. Lan, Z., Chen, M., Goodman, S., et al.: Albert: A lite bert for self-supervised learning of language representations **1901**, 11942 (2019)
7. Dai, Z., Yang, Z., Yang, Y., et al.: Transformer-xl: Attentive language models beyond a fixed-length context **1901**, 02860 (2019)
8. Liu, Y., Ott, M., Goyal, N., et al.: Roberta: A robustly optimized bert pretraining approach **1907**, **11692** (2019)
9. Yin, W., Hay, J., Roth, D.: Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pp. 3912–3921 (2019)
10. Li, B., Zhou, H., He, J., et al.: On the sentence embeddings from pre-trained language models **2011**, 05864 (2020)
11. Gabrilovich, E., Markovitch, S.: Computing semantic relatedness using wikipedia-based explicit semantic analysis. In: IJCAI (2007)
12. Mekala, D., Shang, J.: Contextualized Weak Supervision for Text Classification. In: ACL (2020)
13. Yu. M., Shen, J., Zhang, C., Han, J.: Weakly-supervised neural text classification. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management, pp. 983–992 (2018)
14. Meng, Y., Shen, J., Zhang, C., Han, J.: Weakly-supervised hierarchical text classification. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 6829–6833 (2019)
15. Wang, Z., et al.: X-Class: Text Classification with Extremely Weak Supervision **2010**, 12794 (2021)
16. Meng, Y., et al.: Text classification using label names only: a language model self-training approach. In: EMNLP (2020)