Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/36/20230432

Combination of Chinese sentiment analysis datasets based on BiLSTM+Attention model

Kaiyuan Jiang

School of Computer Science and Engineering, Beihang University, Beijing, 100191, China

19373799@buaa.edu.cn

Abstract. By training the Chinese sentiment analysis model, it is found that the prediction accuracy of the model trained by one dataset is obviously low on other datasets. Considering that the existing sentiment analysis work mainly uses a single domain corpus dataset and referring to the existing data processing methods on natural language processing, this paper designs an experiment to combine Chinese datasets from different fields into a large field-imbalanced dataset, and the number of samples from different fields in this dataset is obviously different. The new dataset is used to train a comprehensive Chinese sentiment analysis model and achieves satisfactory training results. According to the results of the experiments, the model trained by the field-imbalanced dataset has high prediction accuracy for samples from various fields, and the prediction accuracy increases with the increase of the proportion of corpus in this field in the training dataset. Through the experiment in this paper, some ideas are provided for the construction of large-scale cross-domain Chinese sentiment analysis datasets in the future.

Keywords: sentiment analysis, dataset combination, natural language processing.

1. Introduction

When trying to train the Chinese sentiment analysis model, it is found that the model performs poor in cross-data universality. For example, the Long Short-Term Memory (LSTM) model trained by the movie reviews dataset has only 67% accuracy in predicting e-commerce reviews, while its prediction of takeaway reviews is even only 50%. The reason is that the number of words, vocabulary habits and language characteristics of different datasets are quite diverse.

Therefore, it is questioned that if Chinese samples from different fields together form a field-imbalanced dataset, will the model trained on the consolidated dataset have different test results on test sets in various fields?

The problem of dataset imbalance refers to the phenomenon that the sample number ratio of different categories is unbalanced. In an imbalanced dataset, fewer instances are designated as one class, which is typically the more significant class, while practically almost all instances are identified as the other class. As a result of the propensity to categorize all data into majority classes, which are frequently less significant classes, classical classifiers that seek accurate performance in all situations are obviously unsuitable for dealing with imbalanced learning tasks [1]. So when dealing with imbalanced datasets, existing work usually pays more attention to the classes with fewer samples, because the capability of the model to accurately predict the classes with fewer samples is more important than that with more

DOI: 10.54254/2755-2721/36/20230432

samples. The field-imbalanced dataset in this paper is a similar problem, that is, there are obvious differences in the number of corpus samples from different fields in a comprehensive dataset.

For the purpose of Chinese sentiment analysis, the samples may come from different fields, such as product reviews, movie reviews, etc. And similarly the existing large Chinese corpora are often composed of text information from a single field, such as film reviews or Weibo [2]. The samples collected on different platforms and topics often have different language characteristics, which could have a significant impact on the prediction effect of the trained model, especially on the cross-domain universality. Theoretically, if Chinese samples from divers fields form a field-imbalanced dataset, the trained model should be more accurate in predicting the categories that occupy a relatively large part of the dataset and easier to misjudge the categories that occupy a relatively small part. This paper will design experiments to verify this theoretical hypothesis.

2. Method

2.1. Related work

Sentiment analysis is an important mission to detect emotional polarity in texts widely used in online shopping systems, blogs and social media. Its main task is to divide documents into different polarity groups. Many preprocessing techniques are used to clean and standardize data, deny processing and strengthen processing to improve performance. In addition, data enhancement technology is proposed, which can create additional data from the original source data to supplement the existing datasets with no need for human operation [3]. In the natural language processing field, a great deal of work focuses on the processing of datasets. For example, Xuan-Phi Nguyen et al. [4] proposed Data Diversification as a straightforward approach for neural machine translation (NMT), which applies the prediction function of a plurality of forward and backward models, and subsequently integrates the predictions with the initial dataset to diversify the training data. Finally, the new merged dataset is used to train a translation model, and this work provides an experimental idea of merging and expanding training datasets.

Simultaneously, pertaining to sentiment analysis, whether in English or Chinese, the existing researches often only use datasets in a single field. For example, Purnomoputra et al. [5], Daeli et al. [6] focus on the sentiment analysis of movie reviews, Farisi et al. [7], Luo et al. [8] place a focus on the examination of emotional expressions revealed in hotel reviews, Hossain et al. [9] deploy a sentiment analysis system for online reviews of Bangladeshi restaurants, Yang et al. [10] evaluate the real book reviews of famous e-commerce websites in China, while Li et al. [11] makes an attempt to analyze the sentiment of stock comments. When it comes to imbalanced datasets, it is known that imbalanced datasets have negative effects on the model, for example, Gu et al. [12] use imbalanced datasets to test the MBGCV model and achieve competitive results, which proves the performance advantage of their work of combining BiGRU, CNN and VIB. However, just as mentioned before, existing work is common to use a single dataset and seldom involves the universality of cross-datasets and the multi-domain prediction of synthetic datasets. Inspired by experimental phenomena and existing work, in this paper, several open source datasets in different fields are merged into a large field-imbalanced dataset to train a sentiment analysis model and compare the accuracy of the model on test sets in different fields, so as to quantitatively corroborate the capacity of the trained model for making predictions in different fields.

2.2. Model

For the sake of experimental efficiency, this paper uses Attention-Based Bidirectional Long Short-Term Memory Networks (BiLSTM+Attention model) [13] to carry out the experiment, the structure design of which is apparent in Figure 1. There are five component parts in the model:

(1) Input layer: The model is provided with sentences to work;

(2) Embedding layer: Every word is mapped to a vector in the low-dimensional space;

(3) LSTM layer: BLSTM is used to gain advanced features which come from the second layer;

DOI: 10.54254/2755-2721/36/20230432

(4) Attention layer: These features at word-level of each time step are combined to generate feature vectors at sentence-level, which are realized by multiplying the previously generated weight vectors;
(5) Output layer: Feature vectors at sentence-level are used to classify emotional categories. [13]

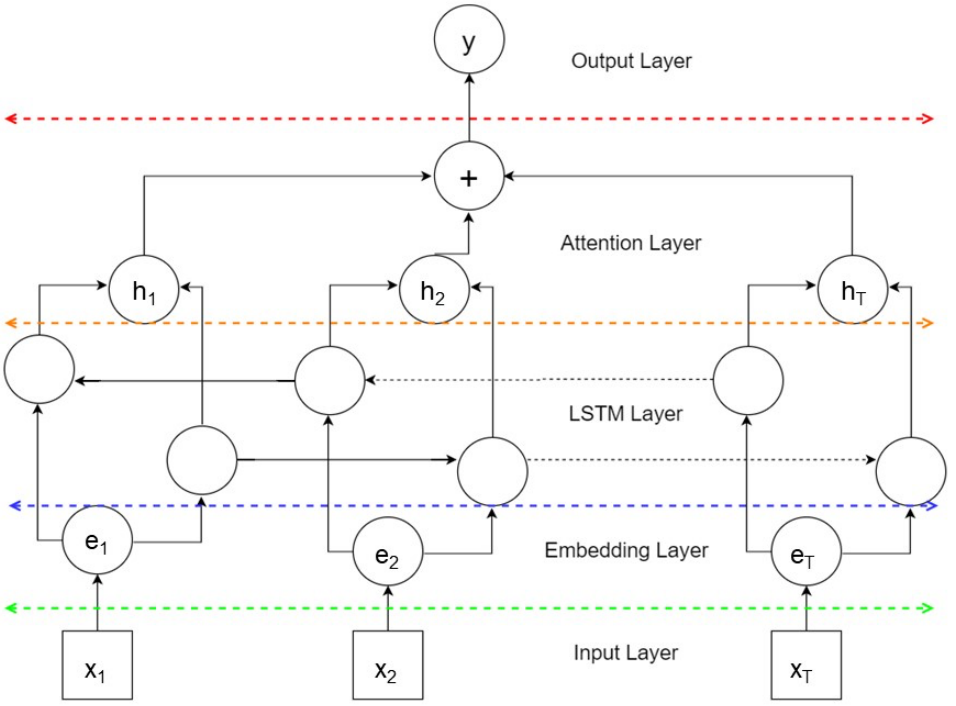

Figure 1. The model used in the experiments of this paper.

Each word in the sentences is input as xi, then transformed into its word embedding vector ei, and the network consists of two sub-networks for context relation of corresponding left and right sequences, which are forward delivery and reverse delivery respectively and can process vector ei into hi, then the final classification representation y is obtained through the attention mechanism. Model evaluation indicators include the following four parts:

$$Accuracy = \frac{TP + TN}{TP + FN + FP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1 = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{4}$$

The workflow of experiments using this model is as follows:
(1) Read the texts, separate the Chinese words in the sentences, clean the data and remove the stopped words. (2) Establish word2index, index2word table. (3) Prepare the pre-trained word embedding (using the wiki_word2vec_50.bin as tool). (4) Process Dataloader to get the sentence representation of datasets. (5) Build a BiLSTM+Attention model. (6) Configure parameters, such as n_epoch. (7) Train the model with training set. (8) Test the model with validation set. (9) Save the model for subsequent operation. In

DOI: 10.54254/2755-2721/36/20230432

order to easily overload the model framework and ensure that the parameters in the trained model have greater stability, the model in this experiment is saved in the form of state_dict, which is not affected by the versions of Pytorch and other packages.

2.3. Experimental environment

The experiment designed in this paper is carried out on a server, which is equipped with a CPU 15 vCPU Intel(R) Xeon(R) Platinum 8358P and a GPU RTX A5000(24GB) .

3. Experiments

3.1. Experimental context

The LSTM model trained by the movie review dataset has only 67% accuracy in predicting e-commerce reviews, while its prediction of takeaway reviews is even only 50%, as shown in Table 1. The reason is that the number of words, vocabulary habits and language characteristics of different datasets are quite diverse.

Table 1. Performance of sentiment analysis model of film reviews in diverse fields.

Datasets	movie reviews(as training data)	e-commerce reviews	takeaway reviews	hotel reviews
Prediction accuracy	0.81843	0.67629	0.50757	0.80743

In order to quantitatively corroborate the capacity of the trained model to make predictions in different fields, the experimental design of this paper is to combine several open source datasets from different fields into a large field-imbalanced dataset to train a sentiment analysis model, and then compare the accuracy of the model with test sets from different fields.

3.2. Datasets processing

In order to complete the experiment, this paper collect and use five open source Chinese sentiment analysis datasets from four different fields, as shown in Table 2. The positive and negative labels and file coding formats of different datasets are not consistent, such as the opposite labels, so it is necessary to deal with the datasets in advance in order for the datasets to be combined correctly.

Table 2. Details of the datasets used in this paper.

Dataset name	online_shopping_10_cats	ChnSentiCorp_htl_all	hotel_discuss2	waimai_10k	SoulDGXu_Dataset
Field	e-commerce reviews	hotel reviews	hotel reviews	takeaway reviews	movie reviews
Sample number	62775	7766	5929	11987	25996
Positive sample number	31728	5322	4758	4000	12998
Negative sample number	31047	2444	1171	7987	12998

3.3. Results

In order to verify the conjectural hypothesis about the merged field-imbalanced dataset, two rounds of experiments are carried out by combining two hotel datasets with other datasets.

DOI: 10.54254/2755-2721/36/20230432

Table 3. Division of training data.

Experimental rounds	1st Round		2nd Round	
Data division	training set	validation set	training set	validation set
Positive samples	37595	9378	41680	10422
Negative samples	38022	9527	39160	9789
Total	75617	18905	80840	20211

The first round introduces a hotel review dataset. There are 25,996 movie reviews, 11,987 takeaway reviews, 7,766 hotel reviews and 62,775 e-commerce reviews in the existing datasets. It is planned to randomly select 3,500 of these datasets as test sets, and then train the model after splicing and reorganizing the rest data, and then test the accuracy of the model in four small test sets in the fields of movie reviews, takeaway reviews, hotel reviews and e-commerce reviews respectively. After selecting the test set data, there are 95,122 remaining pieces of the four datasets. While the second round introduces two hotel review datasets. There are 25,996 movie reviews, 11,987 takeaway reviews, 14,295 hotel reviews and 62,775 e-commerce reviews in the existing datasets. It is also planned to randomly select 3,500 of these datasets as test sets, and then train the model after splicing and reorganizing the rest data, and then test the accuracy of the model in four small test sets in the fields of movie reviews, takeaway reviews, hotel reviews and e-commerce reviews respectively, and study whether the imbalance of data field affects Chinese sentiment analysis in different fields. After selecting the test set data, there are 101,051 remaining pieces of the four datasets. In two rounds of experiments, as indicated in Table 3 above, the remaining data are split into two portions that are respectively used for training and validation at a ratio of 8:2.

The results of model training during the first round of experiment could achieve an accuracy of 88.284%, which has reached a high level. The results of the testing are displayed in Table 4 for the following four test sets, each of which has 3500 samples. E-commerce reviews, which accounts for the largest proportion in the training set, reaches the highest prediction accuracy of 0.9114, while takeaway reviews accounts for 8.92%, with an accuracy of 0.8655, hotel reviews accounts for 4.48%, with an accuracy of 0.8589, which is in line with the proportional relationship.

Table 4. Results of the first round of experiment.

Field	e-commerce reviews	hotel reviews	takeaway reviews	movie reviews
Sample size in the training set	59275	4266	8487	22496
Percentage of samples in the training set	62.31%	4.48%	8.92%	23.65%
Accuracy of test set	0.9114	0.8589	0.8655	0.8034

Next, the results of model training during the second round of experiment could achieve an accuracy of 89.060%, which has achieved higher accuracy of sentiment analysis compared with the last round. The results of the testing of the following four test sets, each of which has 3500 samples, are displayed in Table 5. E-commerce reviews, which accounts for the largest proportion in the training set, still reaches the highest prediction accuracy of 0.9137. Different from the last round of experiments, the proportion of hotel reviews increases to 10.68% due to the increase of the sample number of hotel reviews, which surpasses 8.40% of takeaway reviews. Correspondingly, the prediction accuracy of hotel reviews increases to 0.9011, which surpasses 0.8626 of takeaway reviews.

DOI: 10.54254/2755-2721/36/20230432

Table 5. Results of the second round of experiment.

Field	e-commerce reviews	hotel reviews	takeaway reviews	movie reviews
Sample size in the training set	59275	10795	8487	22496
Percentage of samples in the training set	58.66%	10.68%	8.40%	22.26%
Accuracy of test set	0.9137	0.9011	0.8626	0.8171

As the number of samples in a particular field rises, the prediction accuracy of the comprehensive training model in this field is improved, which is the expected experimental phenomenon and achieves the purpose of designing two rounds of experiments.

3.4. Discussion

Referring to e-commerce reviews, hotel reviews and takeaway reviews, the results can verify the hypothesis that the prediction accuracy of the Chinese sentiment analysis model trained with the combined field-imbalanced dataset is related to the number of corpus in this field, especially after the new hotel review dataset is added into the combined field-imbalanced dataset, the prediction accuracy of the model for hotel reviews has obviously increased, surpassing the takeaway reviews.

In detail, referring to e-commerce reviews, hotel reviews and takeaway reviews, the horizontal comparison of a single experiment shows that the prediction accuracy is proportional to the proportion of training samples in the synthetic field-imbalanced dataset, and the larger category can get higher prediction accuracy in its field. By longitudinal comparison of the two experiments, after adding another hotel review dataset, its proportion in the training set increases from 4.48% to 10.68%, exceeding the proportion of takeaway reviews, and the corresponding test accuracy for hotel reviews increases from 0.8589 to 0.9011, exceeding the 0.8626 of the takeaway test, which is in line with expectations. It shows that the prediction accuracy in a specific field is improved after increasing the number of samples in this field.

As for movie reviews, which are special, they should have ranked second in proportion. It is speculated that there are many abstract expressions such as irony and metonymy in movie reviews, which affect the prediction accuracy and cross-domain universality of corpus samples in this field. Through the independent experimental verification of the movie review dataset, it is found that its training accuracy is only 0.81843, which is obviously lower than other datasets. This result shows that some Chinese sentiment analysis datasets in a certain field, such as movie reviews, has obvious unique language habits, which hinders the prediction of this field.

4. Conclusion

Through the analysis and discussion of the experimental results, this paper supports the hypothesis that the Chinese sentiment analysis model trained by a large-scale field-imbalanced dataset, which is merged by single datasets from different fields, has shown mixed results in different fields, and the prediction accuracy in a certain field is directly proportional to the proportion of samples. Besides, through the experiment in this paper, some ideas are provided for the construction of large-scale cross-domain Chinese sentiment analysis datasets in the future, such as the number of samples in different fields should be as balanced as possible, and the sample quality should be high enough so as not to affect the model prediction in this field. Considering the quantity and quality of corpus in different fields comprehensively is beneficial to build a more significative large-scale Chinese corpus, which will be beneficial to the training of large-scale Chinese language models. For example, if researchers want to build a large comprehensive Chinese corpus for sentiment analysis in the future, they should not only pay attention to the balance of the number of positive and negative labels, but also pay attention to selecting samples from multiple fields, such as Weibo reviews and e-commerce reviews, rather than a single platform. In this way, the characteristics of people's language habits in different contexts can be taken into account, and it can be better applied to sentiment analysis of new samples.

DOI: 10.54254/2755-2721/36/20230432

References

[1] Kotsiantis S, Kanellopoulos D, Pintelas P. 2006. Handling imbalanced datasets: A review. GESTS international transactions on computer science and engineering. 30(1):25–36.

[2] Peng H, Cambria E, Hussain A. 2017. A Review of Sentiment Analysis Research in Chinese Language. *Cognitive Computation*. 9(4):423–435.

[3] Duong H-T, Nguyen-Thi T-A. 2021. A review: preprocessing techniques and data augmentation for sentiment analysis. *Computational Social Networks*. 8(1):1-16.

[4] Nguyen X-P, Shafiq Joty, Wu K, Ai Ti Aw. 2020. Data Diversification: A Simple Strategy For Neural Machine Translation. *Neural Information Processing Systems*. 33:10018–10029.

[5] Purnomoputra RB, Adiwijaya A, Wisesty UN. 2019. Sentiment Analysis of Movie Review using Naïve Bayes Method with Gini Index Feature Selection. *Journal of Data Science and Its Applications*. 2(2):85–94.

[6] Daeli NOF, Adiwijaya A. 2020. Sentiment Analysis on Movie Reviews using Information Gain and K-Nearest Neighbor. *Journal of Data Science and Its Applications*. 3(1):1–7.

[7] Farisi AA, Sibaroni Y, Faraby SA. 2019. Sentiment analysis on hotel reviews using Multinomial Naïve Bayes classifier. *The 2nd International Conference on Data and Information*. 1192:012024.

[8] Luo J, Huang S (Sam), Wang R. 2020 Jun 16. A fine-grained sentiment analysis of online guest reviews of economy hotels in China. *Journal of Hospitality Marketing & Management*. 30(1): 71-95.

[9] Hossain N, Bhuiyan MdR, Tumpa ZN, Hossain SA. 2020. Sentiment Analysis of Restaurant Reviews using Combined CNN-LSTM. *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*.

[10] Yang L, Li Y, Wang J, Sherratt RS. 2020. Sentiment Analysis for E-Commerce Product Reviews in Chinese Based on Sentiment Lexicon and Deep Learning. *IEEE Access*. 8:23522–23530.

[11] Li M, Chen L, Zhao J, Li Q. 2021. Sentiment analysis of Chinese stock reviews based on BERT model. *Applied Intelligence*. 51(7):5016–5024.

[12] Gu T, Xu G, Luo J. 2020. Sentiment Analysis via Deep Multichannel Neural Networks With Variational Information Bottleneck. *IEEE Access*. 8:121014–121021.

[13] Zhou P, Shi W, Tian J, Qi Z, Li B, Hao H, Xu B. 2016. Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification. *Proceedings of the 54th annual meeting of the association for computational linguistics*. (volume 2: Short papers):207-212.