Emotional Classification Comparison Experiment of Weibo Comments

Yuqing Sun, Xinru Li, Zhiwei Yang, and Xiaopei Li

School of North China University of Science and Technology, Hebei 63210, China

Abstract

This article studies the Weibo topic # Do students really want to start school early #.First, using the analytical data collection using the octopus data collector, the corpus was established based on the data cleaning, from which emotional features were extracted.Secondly, in a specific experimental environment, seven comparative models such as KNN, logistic regression, random forest, decision tree, SVM, gradient promotion iterative decision tree GBDT, and integrated learning classifier Adaboost were selected for the experiments.Finally, the classification accuracy and time complexity of each model are analyzed by contrast.

Keywords

Weibo Comments, Emotional Classification, Comparative Experiment.

1. Introduction

With the increase of microblog users and the improvement of people's attention to the generated text information, the emotional analysis of applying various algorithms and models to microblogs and comments has gradually become a hot topic.At present, many studies have combined different algorithm models with the classification of microblog comments respectively, and applied various algorithms to conduct emotional analysis of microblog comments, which has achieved good results.In order to obtain the optimal model, this paper uses Section_three_way, KNN, SVM, GBDT, Adaboost and other classification methods, and the important parameters, algorithm running time, the algorithm model classification effect, with accuracy, recall and f1 value as the evaluation index, accuracy is more than 80%, get the best classification results.

2. Weibo Topic Data Pre-Processing

2.1 Data Acquisition

First of all, select a recent Weibo topic, such as: # Do students really want to start school early #, and find a blogger to comment below on this topic.

Follow only the comments and see a total of 1233 comments, collecting only its first-level comments, namely messages containing no reply.

Acquisition was performed using the octopus data collector to establish a custom task reviewing the details page (https://weibo.com/2835724503/Ix1Y1qBTX?filter=hot&root_comment_id=0 & type= comment#_rnd1588600395519) was copied into the collector and designed for the acquisition process, as shown in Figure 1:

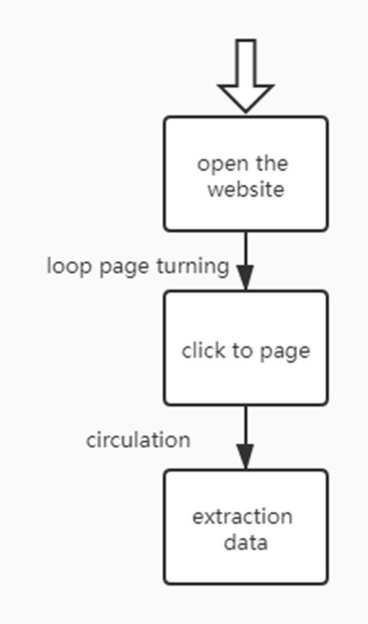

Figure 1. Acquisition process design

The principle of collection is to save and open the above copied website, and then set over the page for a large number of comments, that is, "View more" in the comment details, click around it until all the comments are completed, and then extract the data one by one from top to bottom.

Eventually extracted to 1041 level 1 comments and exported in xlsx file format.

2.2 Data Cleaning

In the collected text data, due to many noisy data, the original data needs to be cleaned, including: removing the user name, space, topic symbol # XXXXXX#, and @XXX.This part uses the built-in function and regular expression of python to achieve the purpose of data cleaning.

2.3 Establishment of Corpusww

The process diagram for building the corpus is shown in Figure 2:

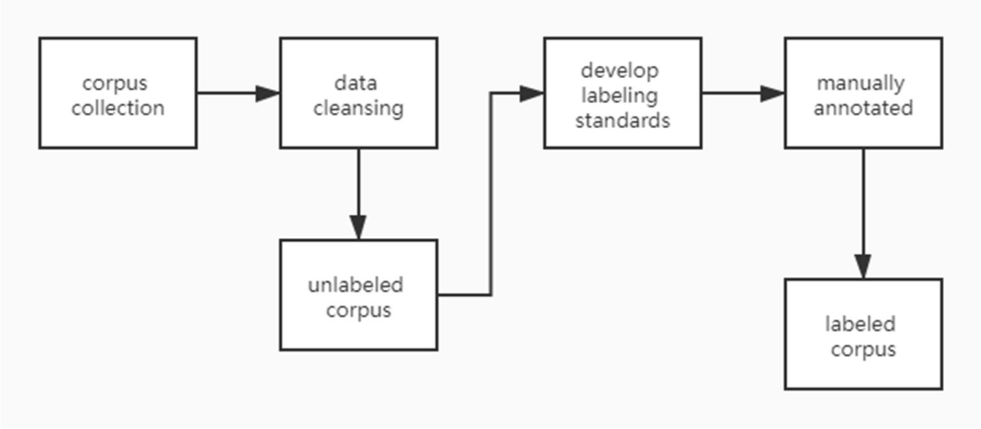

Figure 2. Process diagram of corpus

Considering the accuracy and the feasibility of the operation at the present stage, the method of manual label is positive to the topic # Do students really want to start the school early. The label is marked as a positive class, the verdict is negative, and the label is marked as 0 as a negative class.

Five hundred and 500 collected data were selected for subsequent experiments, including 250 positive and negative text each, and the annotated text is a constructed corpus.

2.4 Remove Word Segmentation and Deactivated Words

There are a variety of partitioning methods, including forward / reverse maximum matching method, word frequency statistics method, word by word traverse method, adjacency constraint method, probabilistic language partitioning method, etc.Among them, the probabilistic language participle model is currently the most reliable method to obtain the partitioning effect, because it requires training on a large-scale partitioned corpus, so the obtained partitioning effect is also better.Splciple results are generally represented by a list of multiple words.For example, the sentence "there is a person on the other side of the river", the entry list "[river, of, the other side, there, one, person]".In order to realize the text partitioning, install the python third-party library jieba, and the jieba word partitioning system is mainly divided into three modules: word segmentation, word annotation, and keyword extraction.

Stop word use has no practical significance, and only plays an auxiliary role in semantic expression, such as "in", "in", "in one", etc.A large number of stop words in the text will only interfere with the feature selection of the text, and greatly increase the difficulty of text preprocessing. In the large number of hundreds of millions of microblog comments, this problem cannot be ignored.Because the evaluation criteria of stop words vary in different fields, a universal stop word table of universality does not yet exist yet.

Due to the network language update speed is too fast, the existing online stop word list may not be complete, therefore, in this step of processing, downloaded from the network: baidu stop word table, Harbin Institute of Technology stop word table, Chinese stop word table and sichuan university machine intelligence laboratory stop word library, and the four stop word table into a stop word table.

2.5 Extraction of Emotional Features

After dividing the problem and alternative text, a collection of words can form each paragraph of text. However, due to the different length of text in each paragraph, in order to encode each text into vectors of the same dimension, in order to realize the similarity of the problem text and alternative text cosine in the subsequent algorithm, we need to extract a fixed number of key words of each text, which can best reflect the paragraph of text content, that is, the analysis text for subject modeling.TF-IDF algorithm is a widely used and effective algorithm to extract text keywords.

In the $\mathrm{TF}-\mathrm{IDF}$ methods, if the more a word appears in a certain text, the better the word reflects the content of the text. The calculation formula is as follows:

$$\omega = \mathrm{TF}\cdot\mathrm{IDF} \tag{1}$$

$$\text{word frequency(TF)} = \frac{\text{The number of times a word appears in an article}}{\text{Number of total words of the article}} \tag{2}$$

$$\text{Inverse document frequency} = \log(\frac{\text{The total number of documents in the corpus}}{\text{The number of documents that contain the word+}}) \tag{3}$$

Where, w is the weight of each word, TF is the word frequency, the frequency of a word appearing in a text, and IDF is the inverse document frequency, the reciprocal frequency of a word appearing in all texts.

3. Principles of Each Comparison Algorithm

3.1 KNN Algorithm

The KNN algorithm is a very special machine learning algorithm because it has no learning process in the general sense. It works by using the training data to partition the feature vector space using training data and using the partition results as the final algorithmic model. There is a sample data set, also known as the training sample set, and each data in the sample set has a label, where we know the correspondence of each data in the sample set to the subordinate classification.

After entering the unlabeled data, each feature of this unlabeled data is compared with the features corresponding to the data in the sample set, and then the taxonomic labels of the data (the nearest neighbor) with the most similar features in the sample are extracted.

In general, we have only select the first k most similar data in the sample dataset, which is the origin of K in the KNN algorithm, and usually k is an integer no greater than 20. Finally, the k most frequent categories in the most similar data were selected as the classification of the new data.

The classification prediction process of the KNN classification algorithm is very simple and easy to understand: for an input vector x that needs to be predicted, we only need to look for the set of k nearest vectors in the training dataset, and then predict the category of x as the largest number of categories in this k sample.

3.2 Logistic Regression

Logical regression is an algorithm very similar to linear regression, but, in essence, the type of problem that which linear regression processing is not consistent with logistic regression. Linear regression deals with numerical problems, where the last predicted result is numbers, such as house prices. Logic regression belongs to the classification algorithm, that is, the logical regression prediction result is a discrete classification, such as judging whether this email is spam, and whether the user will click on this advertisement. So logistic regression is a classical dicclassification algorithm.

In terms of implementation, logical regression only adds a Sigmoid function to the calculation of the linear regression, transforming the numerical results into the probability between 0 and 1 (the image of Sigmoid function is generally not intuitive, you just need to understand that the larger the value, the function the closer the 1, the smaller the value, the closer the function 0). Then we can predict according to this probability, such as the probability is greater than 0.5, then this mail is spam, or whether the tumor is malignant, etc.

3.3 Random Forest

The rationale of the random forest algorithm is to use Bootstrap subself-sampling to obtain different sample sets for model construction, thus increasing the difference between models and improving the ability to extrapolate predictions. For attribute selection: first, randomly select a subset of attributes including k (k, k is usually log2d, where d is the size of the attribute set) from the base decision tree junction; second, select a best attribute from this subset for division; finally, the majority outcome method, averaging method, voting method are considered to determine the final random forest algorithm.[1].

Intuitively, each decision tree is a classifier (assuming now for the classification problem), then N trees for an input sample will have N classification results. While the random forest integrates all the classified voting results, assigning the most voted category as the final output, which is one of the simplest Bagging ideas.

3.4 Decision Tree

The decision tree is a tree structure (either be a binary or non-binary tree).Each of its non-leaf nodes represents a test on a feature property, and each branch represents the output of this characteristic property on a certain domain of value, while each leaf node stores a category.

The process of using the decision tree is to start from the root node, test the corresponding feature properties in the item to be categorized, and select the output branch at its value until the leaf node, taking the category that the leaf node is stored as the decision result.

The core of the decision tree model is divided into the following parts: node and directed edges; nodes have two types: internal nodes and leaf nodes; internal nodes represent a feature, and leaf nodes represent a class.

The basic principle of decision tree algorithm is to represent decision problems with decision points, alternative schemes with scheme branches, and various possible results with probability branches. After the calculation of profit and loss values of various schemes under various outcome conditions, it provides a decision basis for decision makers.

Decision tree analysis is a common decision-making method for risk analysis. The method is a method to describe the calculation, comparison and selection of each scheme in a tree graph, and its decision is based on the expected value.People may come into several different situations about the future. Each case has a possibility, which can not be confirmed, but the probability of various natural states can be inferred from previous data.

3.5 SVM

Support vector machine is a biclassification model where the basic model is a linear classifier defined on feature space. The learning strategy of SVM is interval maximization, which can be formalized as a problem of solving convex quadratic programming and also equivalent to the minimization problem of the regularization compound page loss function.[2].

$$\min_{\omega,b}\frac{1}{2}\parallel \omega^2 \parallel + C\sum_{i=1}^{m} L_\varepsilon(y-f(x)) \tag{4}$$

L_ε is the loss function, and C is the penalty coefficient.In general, the larger the value of C is set, the higher the accuracy of the model training.However, if the C value is set too high, an overfitting problem occurs.

$$L_\varepsilon\big(y-f(x)\big)=\begin{cases}0 & |y-f(x)| \\ |y-f(x)|-\varepsilon & \text{else}\end{cases} \tag{5}$$

3.6 GBDT

The lifting (Boosting) method is a common statistical method that can be seen as an integration method by changing the weight of the training samples, learning multiple classifiers, and linear combining these classifiers and improving the performance of the model. The Boosting method mainly adopts a linear combination of basis functions with the forward distribution algorithm. The promotion method with decision tree as based function is called booting tree (BD), while Gradient boosting decision tree (GBDT) is the integration process combining the regression tree and BT and proposing to use the residual gradient to optimize the regression tree.[3].

3.7 Adaboost

AdaBoost is an abbreviation for "Adaptive Boosting" (adaptive enhancement), where the weights of the sample misclassified by the previous basic classifier increase, while the correctly classified sample decrease and again used to train the next basic classifier. Meanwhile, in each round, a new weak classifier is added and the final strong classifier is not determined until a certain predetermined sufficiently small error rate is reached or the maximum prespecified number of iterations is reached. The Adaboost algorithm can be summarized in three steps:

(1) First, it is the weight distribution D1 of the initialized training data. Assuming N training sample data, each of the training samples is given the same weight: w1=1 / N.

(2) Then, train the weak classifier hi.In the specific training process, if a training sample point is accurately classified by the weak classifier hi, its corresponding weight decreases in constructing the next training set; Instead, if a training sample point is misclassified, its weight should increase.The weight-updated sample set was used to train the next classifier, and the entire training procedure proceeds so iteratively.

(3) Finally, the weak classifier obtained from each training is combined into a strong classifier.After the training process of each weak classifier ends, the weight of the weak classifier with small classification error rate is increased, so that it plays a large decisive role in the final classification function, while reducing the weight of the weak classifier with large classification error rate, making it play a small decisive role in the final classification function.

In other words, weak classifiers with low error rates have greater weight in the final classifier, otherwise smaller.

4. Emotional Classification of Microblog Comments based on Each Classification Algorithm

4.1 Experimental Environment

CPU: Intel(R)Core(TM) i7-7500U CPU @2.70GHz 2.9.GHz;

Memory: 8.00GB;

System: windows 10 64 bit;

Data source: Octopus Data collector 8;

Experimental run: Anaconda3.

4.2 Introduction of the Comparative Experimental Algorithm Model

In the comparison experiment, KNN, logical regression, random forest, decision tree, SVM, gradient promotion iterative decision tree GBDT, and integrated learning classifier Adaboost were selected.The principle of emotion classification using these 7 classic classification models based on micro blog comments is thatfor the cleaned text, go through word separation, stop word use, feature extraction, and finally complete the weighting operation of features with tf-idf.On this basis, each classifier is defined, taking accuracy, recall and f1 value as evaluation indicators.

For selected classification algorithms such as KNN, logistic regression, decision tree, SVM, GBDT and Adaboost, classification models were constructed.The important parameter settings for the each classification model are shown in Table 1:

Table 1. Compares the experimental classification model parameters

Model name	Main parameter
knn	n_neighbors=5
logisitic_regression	penalty='l2'
decision_tree	criterion='gini', max_depth=2
gradient_boosting	n_estimators=200
svm	kernel='rbf'
adaboost	n_estimators=100

As shown in Table 1, the partial parameters of the model are selected.The n_neighbors of the KNN is the k value in the KNN algorithm, generally default 5 is most appropriate; select l2 for logical regression penalty term, that is, the sample meets Gaussian distribution; the decision tree is impure with the Gini coefficient and the optimal maximum depth of the tree is 2; the number of adjusted initial value trees in GBDT is set to 200, Adaboost is 100; SVM kernel function is really rbf radial basis function.

4.3 Analysis of the Results of the Comparative Experimental Algorithm

Organize the classification effects of the different algorithms in all instance analysis to obtain Table 2:

Table 2. Comparison of the classification effect of each algorithm model

	accuracy	precision	recall	f1
Section_three_way	0.78	0.7801	0.7765	0.7869
knn	0.76	0.7916	0.7600	0.7607
logistic_regression	0.82	0.8386	0.8200	0.8211
random_forest	0.87	0.8706	0.8700	0.8702
decision_tree	0.87	0.8706	0.8700	0.8702
svm	0.89	0.8957	0.8900	0.8906
GBDT	0.88	0.8814	0.8800	0.8803
Adaboost	0.83	0.8296	0.8300	0.8297

According to Table 2, it can be seen that for the selected dataset of Weibo comments samples: the classification effect of the three interval decision algorithms is relatively general, SVM and GBDT are the best, and KNN is the worst.

Table 3. Comparison of the algorithm running time

Algorithm	Time consuming
Section_three_way	271.01354241371155
knn	0.009972810745239258
logistic_regression	0.007976770401000977
random_forest	0.021941184997558594
decision_tree	0.007977724075317383
svm	0.11768412590026855
GBDT	0.6213383674621582
Adaboost	0.4438135623931885

Combined with Table 3 shows that the interval three decisions, although KNN in accuracy than KNN, perform much worse than K N N in running time and are the worst of all algorithms.At the same time, along with the characteristics of the algorithm itself, a lot of cycles need to be used in the implementation process, which will lead to excessive time complexity.Therefore, the model is not suitable for large-scale data, while the text of large data is more complex and has more feature properties.

5. Conclusion

This paper makes emotion analysis on Weibo comments through various classification algorithms to obtain the corresponding classification effect.Finally, the accuracy, recall, program running time and the experimental software and hardware environment are presented.

References

[1] Dong Hongyao, Wang Yidan, Li Lihong. Summary of the random forest optimization algorithm [J]. Information and Computer (Theory Edition), 2021,33 (17): 34-37.C. Li, W.Q. Yin, X.B. Feng, et al. Brushless DC motor stepless speed regulation system based on fuzzy adaptive PI controller, Journal of Mechanical & Electrical Engineering, vol. 29 (2012), 49-52.

[2] Wang Bin, Chai Huaxun, Wang Yongjian, Li Hongguang. Evaluation of Industrial Process Area Control Performance based on the GBDT + LR classifier [J]. Petrochemical Automation, 2020,56 (03): 21-26 + 30.J. Liu, E.L. Chen and Z.T. He: Journal of Shi Jia Zhuang Railway Institute (Natural Science), Vol. 22 (2009) No. 4, p.40-42.

[3] Wang Yalin, Chen Ninja. Comparison of different applications of machine learning algorithms to classification problems [J]. Heilongjiang Science, 2021,12 (04): 16-18 + 22.