The Selection and Optimization of a Negative comment Identification Method

To cite this article: Haibo Luo et al 2019 J. Phys.: Conf. Ser. 1325 012100

View the article online for updates and enhancements.
The Selection and Optimization of a Negative comment Identification Method

Haibo Luo*, Jianhu Dong1 and Xianlu Luo1

1 Network Engineering Department, Neusoft Institute Guangdong, 528225, China
*Corresponding author’s e-mail: luohaibo@nuit.edu.cn

Abstract. This paper introduces the research background and significance of negative comments, and analyzes the present situation and development trend of negative comments identification at home and abroad. Several popular classification algorithms of naive Bayes, support vector machines and neural network are introduced. The word bag model, TF-IDF model and Word2Vec/Doc2Vec model feature model are also introduced. The data of feature extraction and classification algorithm in IMDB movie review website are simulated, and the results are compared and analyzed, and an optimal combination of feature extraction and classification algorithm is obtained.

1. Introduction

Although the total amount of information continues to grow, it is mixed with a large number of false and useless information. The identification of negative comments can be seen as a special text classification problem, whose technology benefits from a lot of achievements in the field of natural language processing in mathematical models. The key areas of solution can overcome the ambiguity caused by context and can meet the "standard" reference classification thesaurus. More specifically, such as the backwardness of dictionary construction, the pertinence of specific scene models, the inaccuracy of negative recognition and so on, all of which directly lead to a great difference between model training and expectations, and even a wide range of misjudgment, which affects the actual use of the effect.

This paper introduces the research background and significance of negative comments, the basic concept, characteristics and classification of negative comments, and analyzes the present situation and development trend of negative comments identification at home and abroad. Several popular classification algorithms of naive Bayes, support vector machines and neural network are introduced. The characteristics of the word-bag TF-IDF model, the vocabulary model and the Word2Vec, Doc2Vec model are introduced. The data of feature extraction and classification algorithm in IMDB movie review website are simulated, and the results are compared and analyzed, and an optimal combination of feature extraction and classification algorithm is obtained.

2. Preparation

Classification is very important for negative comments. The common classification algorithms are naive Bays, decision tree, Support Vector Machine, K nearest neighbor, logical regression and convolution neural network. In this paper, naive Bayes, Support Vector Machine and convolution neural network are used as negative comment recognition and verification algorithms.
2.1. Naive Bayes
Naive Bayes method is a classification method based on Bayes theorem and independent hypothesis of characteristic conditions. For a given training data set, first, the combined probability distribution of the input/output is learned independently based on the characteristic condition. Then, based on this model, for a given input $x$, the output with the largest posterior probability is obtained by using Bayes theorem. Naive Bayes method is a common method because of its simple implementation and high efficiency of learning and prediction. In the Naive Bayes method, it is assumed that the input variables are conditional independent. If there is a probability dependency between them, the model becomes a Bayesian network, as described in more detail in the reference [1].

Naive Bayes has been successfully used to predict whether borrowers will default on payments and to diagnose whether a person has heart disease; for privacy protection, traffic analysis, network situation assessment, expert systems, decision-making, diagnosis, statistics, forecasting, etc. Naive Bayes is used in the identification of negative comments in fewer cases.

2.2. Support Vector Machine
Support Vector Machine (SVM) is a two-class classification model. Its basic model is a linear classifier that defines the largest spacing on the feature space; Support vector machine also includes kernel technique, which makes it a nonlinear classifier in essence. The learning strategy of support vector machine (SVM) is to maximize the interval and formalize it into a problem for solving convex quadratic programming, which is also equivalent to the problem of minimizing the regular closed loss function. The learning algorithm of the support vector machine is the optimization algorithm for solving the convex quadratic programming[2].

The SVM has unique advantages in solving the problems of small sample, regression analysis, function estimation, time series prediction, nonlinear and high-dimensional pattern recognition; it is widely used in text recognition, handwritten recognition, face image recognition, gene classification and time series prediction etc [3-5]. SVM has been successfully used in spam text message recognition [6]. There are not many cases where the SVM is used for negative comment recognition.

2.3. Convolution Neural Network
Convolution Neural Network (CNN), which was first proposed by Le Cun Y[7], is a feed forward neural network which is specially used to process data with similar network structure. CNN directly inputs the original data into the network, then implicitly makes the network study from the training data, avoids the manual extraction of the features, thus leading to the characteristic of error accumulation, and the whole classification is automatic.

In recent years, CNN has been widely used in the field of image processing, face recognition, audio retrieval, clinical diagnosis signs of cardiovascular diseases, time series classification, time series measurement learning, short text clustering, visual tracking, image fusion, etc[8-9].

Although CNN has been widely used in various fields, its advantages do not mean that previous problems can be solved or improved. For example, it’s rarely used in the case of negative comment identification.

3. Model Description
The negative review identifies the following frame, as shown in Figure 1.
3.1. Negative comment feature extraction
The feature extraction methods used to identify negative comments are word bag and TF-IDF model, vocabulary model and Word2Vec/Doc2Vec model.

3.2. Word bag and TF-IDF Model
Bow (Bag-of-words, word bag) was originally used in the field of natural language processing and information retrieval to classify and identify documents. The basic principle is that the document is regarded as a set of disordered keywords, and the vector representation of each keyword is carried out by counting the frequency of each keyword appearing in a single document, so as to classify the document.

TF-IDF (Term Frequency-Inverse Document Frequency, word frequency-inverse document) was proposed by Salton in 1988. Jia Qiang et al proposed an improved TF-IDF text feature word extraction algorithm. Through the recognition of the long words in the document, the document words are normalized, the accuracy of the feature word expression is strengthened, and more effective results are obtained. The relevant formulas are as follows.

The weighted value of TF-IDF: \( W_i \), is calculated as follows:

\[
W_i = TF_i \times IDF_i
\]  
(1)
TF (Term Frequency) indicates how often a word appears in the text. For example, if a word appears in the text is WC, The total number of words in this text is TWC, and then the formula is:

$$TF = \frac{WC}{TWC} \quad (2)$$

When the text is too long, in order to avoid statistical deviation, TF value needs to be standardized. For example, the number of times a word appears in the text is WC, The highest number of word occurrences in this text is HWC, and the formula is as follows:

$$TF = \frac{WC}{HWC} \quad (3)$$

IDF (Inverse Document Frequency) represents the frequency at which a word appears in all texts. For example, the total number of text is TT If the number of text containing a word is WT (WT cannot be 0), the formula is:

$$IDF = \log \frac{TT}{WT} \quad (4)$$

According to the formula of IDF, in order to prevent the denominator 0 when a word does not appear in all texts, the denominator WT in IDF is treated with 1. The formula for the processed IDF is:

$$IDF = \log \frac{TT}{WT + 1} \quad (5)$$

The process of word bag and TF-IDF is as follows:
Step 1: store the original file in a folder named Neg;
Step 2: generate a loadfile () function that treats a comment file as a complete string;
Step 3: generate a loaddir () function, call the loadfile function, load all the files in the Neg directory, and generate the string collection X.
Step 4: load the training set Training Set, to mark the positive comment as 0 and the negative comment as 1;
Step 5: load the test set Test Set, to mark the positive comment as 0 and the negative comment as 1;
Step 6: use the CountVectorizer object of Scikit-Learn to process the word bag, and save the extracted vocabulary for the word bag test data set.
Step 7: use TF-IDF to count the dataset generated by the word bag, and get the importance of the related words in the dataset.

3.3. Vocabulary model
BoW has a good effect on which words are made up of words in the presentation text, but it cannot express the relationship between words.

The vocabulary model draws lessons from the idea of BoW and uses the generated vocabulary to code the original sentences one by one according to the words. The tflearn.data_utils.VocabularyProcessor (max_document_length, min_frequency=0, vocabulary=None, tokenizer_fn=None) function is used for processing.

The maximum length of the document is max_document_length, if the length of the text is greater than the maximum length, it will be cut, otherwise, it will be filled with 0; the minimum value of word frequency is min_frequency. If the frequency of occurrence is less than the minimum word frequency, it will not be included in the vocabulary. Vocabulary is the CategoricalVocabulary object; and Tokenizer_fn is word segmentation function.

The vocabulary model processing process is as follows:
Step 1: store the original file in a folder named Neg;
Step 2: generate a loadfile () function that treats a comment file as a complete string;
Step 3: generate a loaddir () function, call the loadfile function, load all the files in the Neg directory, and generate the string collection X.
Step 4: Format the word list;
Step 5: Code the word;  
Step 6: generate a sequence of words.

3.4. Word2Vec/Doc2Vec model

Word2Vec can represent words as real numerical vectors. Through training, Word2Vec can simplify the processing of text content to vector operation in K-dimensional vector space, and the similarity in vector space can be used to represent the semantic similarity of text. Doc2Vec and Word2Vec have the same principle.

The word vector output by Word2Vec/Doc2Vec can be used to do a lot of natural language processing related work, such as clustering, synonym finding, and part of speech analysis and so on.

The Word2Vec process is as follows:
Step 1: install gensim;
Step 2: train the data and get the corresponding vector representation according to the word.
Step 3: adjust Word2Vec parameters and control training speed and quality. For example, min_count is used for dictionary truncation, the default is 5, for performance reasons, it is recommended to set it to 10; Size represents the number of units in the hidden layer of the neural network, with recommended values ranging from tens to hundreds. After testing, it is recommended that it be set to 200. At this time, the dimension of the word vector is also 200. Window is the window of word processing, it is recommended to set it to 8; Iter represents the number of iterations and is recommended to be set to 10; the number of concurrent threads should be the same as the number of CPU on the current computer. Epochs indicates the number of training, the default is 5, it is recommended to set to 10, so as to improve the quality of the generated model. Doc2Vec will have one more tag tags than Word2Vec as a unique identity.
Step 4: output the results. Save it in a dictionary in a model object.

Training Word2Vec/Doc2Vec requires a lot of resources and time, and the classification algorithm will be changed frequently and the parameters of the classification algorithm will be modified in the debugging stage. In order to improve the efficiency, after the training, the Word2Vec/Doc2Vec model file corresponding to the word or paragraph will be saved in the model object, stored on the hard disk, and can be accessed directly like a dictionary. For example, the way to get the Word2vec for paragraph "I love you" is as follows: model ['I love you'].

4. Simulation experiment

4.1. Experimental data

The test data are from the IMDB (Internet Movie Database). A total of 100000 records were recorded in the entire dataset, 50,000 were marked and 50,000 were not marked. 50,000 marked data sets are randomly assigned into training datasets and test datasets.

4.2. Evaluation indicators and results

TP stands for real cases, TN for true and negative examples, FP for false positive cases, FN for false negative cases, and it means missing the report. Accuracy is an index used to evaluate the classification model and the percentage of correct results predicted by the model. The F1 value is a measure of the classification problem[10].

The results based on the combination of BoW and CNN are the best, the accuracy is 86.41%, but the consumption time is the longest. By comparing the two dimensions of FN and FP, it is proved that the word bag model cannot get rid of pure mathematical statistics under the action of Bayesian, and the misjudgment rate is much higher than that of Doc2vec algorithm, which uses hidden layer recognition context semantics. Under the support of neural network, the recognition rate of word bag can be improved by nearly 20% through self-training. As shown in Table 1.
Table 1. Identification, evaluation and results

| Feature | BoW | TF-IDF | Doc2vec | TF-IDF | BoW |
|---------|-----|--------|---------|--------|-----|
| Algorithm | NB | NB | NB | SVM | CNN |
| TP | 6255 | 8663 | 9208 | 9697 |
| FN | 6245 | 3837 | 3292 | 2803 |
| FP | 1595 | 2191 | 2232 | 2271 |
| TN | 10905 | 10309 | 10268 | 10229 |
| Acc | 68.64% | 75.89% | 77.90% | 79.70% | 86.41% |
| Time | 4min | 5min | 1h | 20min | 5h |

5. Conclusion

This paper introduces the research background and significance of negative comments, the basic concept, characteristics and classification of negative comments, and analyzes the present situation and development trend of negative comments identification at home and abroad. Several popular classification naive Bayes, support vector machines and neural network algorithms are introduced. The word bag model, TF-IDF model, Word2Vec/Doc2Vec model and PV-DM model are also introduced. The simulation experiment of feature extraction and classification algorithm on IMDB movie review website is carried out, and the results are compared and analyzed. Although there are still many problems to be solved in the application of deep learning technology to cyberspace security, it does not affect the development and application of cyberspace security in the future, and it will still be a hot spot in people's research for a long time in the future.

References

[1] Bishop C. (2006) Pattern Recognition and Machine Learning. Springer, Berlin.
[2] Li Hang. (2012) Statistical learning methods. Tsinghua University, Beijing.
[3] Vapnik Vladimir N. (1995) The Nature of Statistical Learning Theory. Springer, Berlin.
[4] Zhang Xuegong. (2000) On statistical learning theory and support vector machine. Journal of Automation, 26(1):1-2.
[5] Ding Shifei, Qi Bingjuan, Tan Hongyan. (2011) A review of the theory and algorithm of support vector machine. Journal of the University of Electronic Science and Technology, 40(1): 2-10.
[6] Yang Dandan. (2016) Design and implementation of spam short message monitoring system based on support vector machine. Jilin: master's degree thesis of Jilin University, 2-3.
[7] Le Cun Y. (1989) Generalization and network design strategies. Connectionism in perspective. Amsterdam: Elsevier, 19-20.
[8] Kadi I, Idri A,Fernandez-Aleman J L. (2017) Knowledge discovery in cardiology: A systematic literature review. International Journal of Medical Informatics, 97:12-32.
[9] Li Hong,Liu Fang,Yang Shu-Yuan, et al. (2016) Remote sensing image fusion based on support value learning networks. Chinese Journal of Computer, 39(8):1583-1596.
[10] FelixFuu.(2018) Calculation of TP, TN, FP, FN, Recall, Miss Rate, MCC, F1 Score et al. https://blog.csdn.net/u011956147/article/details/78967145.