International Journal of Social Science and Education Research Volume 4 Issue 4, 2021
ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202104_4(4).0053

# Method of Short Text Classification based on TF-IWF Feature Selection

Qijun Duan

School of Economics and Management, Chongqing University of Posts and Telecommunications, Chongqing, 400065, China

## Abstract

**[Objective] TF-IDF algorithm solves the problem of external corpus dependence in short text classification, but it has the problem of weight concentration and low text discrimination when calculating text features. Therefore, a short text classification method based on Chi square statistics and tf-iwf algorithm is proposed. [method] the feature words are extracted from the training data set by chi square statistics. The feature words are weighted by tf-iwf algorithm, and then classified by SVM classifier. [results] the experimental results show that the accuracy of text classification is improved by 3.1%, the recall is improved by 5.2%, and the F1 value is improved by 3.7% by combining chi square statistics and tf-iwf. [Conclusion] the method expands the range of the weight value of feature words, increases the variance of the weight value of the text set, and solves the problem of sparsity of short text content to a certain extent, so as to improve the performance of short text classification.**

## Keywords

**Short-Text; TF-IWF Algorithm; Feature Selection; Sentiment Classification.**

## 1. Introduction

With the rise of social networks and e-commerce, a large number of short text data are produced. These short texts directly reflect people's feelings and views on different events. The mining of short text information and the classification of short text have become a hot topic. Short text classification is an important branch of natural language processing, which is of great significance in search engine, automatic question answering, public opinion analysis and emotion analysis.

Due to the problems of short text content, serious colloquialism and high noise, the feature polarity is not significant in the process of emotion classification [2]. The key to accurate classification of short text content is the extraction of important content of short text. Generally, it is realized by introducing the external corpus and the content features of the text itself. The effectiveness of the external corpus expansion method depends heavily on the quality of the corpus, and the calculation is slow and inefficient [3]. The way based on the content features of short text itself is to fully mine the key information such as text semantics and word frequency to obtain the important features of the text, which has high requirements for text feature extraction.

This paper proposes a feature extraction and short text classification method based on Chi square statistics and tf-iwf algorithm, which aims to solve the problem that TF-IDF algorithm has a small range of eigenvalue weight and improve the accuracy of short text classification.

## 2. Research Status

Compared with the long text, the content of the short text is sparse, and the information unit is difficult to collect accurately. It is difficult to achieve good results by directly applying the

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202104_4(4).0053

traditional text classification methods such as support vector machine and naive Bayes classification to the short text classification [4-5]. To solve this problem, researchers propose a series of methods for short text feature extraction, so as to improve the accuracy of short text classification.Some researchers make up for the low information density of short text corpus by introducing external corpus, while others create and improve the method of extracting key information of short text to make the extracted short text features more representative.

### 2.1. Introducing External Corpus to Expand Features

External corpora usually refer to Wikipedia, HowNet and other knowledge bases that contain a large number of text content and language materials. These external corpora can supplement short text sets and increase the amount of information.Fan Yunjie [et al] combined with statistical knowledge and category information, established semantic set through external knowledge base Wikipedia.Hu Xia et al. [integrate seed words with short text features into the established hierarchical structure model, and then expand to obtain semantic information based on seed word features with the help of external corpus.Ding Lianhong et al. [used knowledge map to calculate and infer short text expansion information, and applied it to short text features.Wang Sheng et al. Calculated the upper and lower position relationship of words in short text through external database, and then used this relationship to the characteristics of the text to be tested, so as to improve the classification effect.To solve the problem of too few short text information units, the short text content expansion method based on external corpus relies on the quality of external corpus.It is difficult to play a good role for some text content which involves professional fields and uses avant-garde words.Therefore, the feature mining method based on short text itself is considered. [6] [7] [8] [9]

### 2.2. Feature Mining Method based on Short Text Itself

On the premise of not borrowing the external corpus, analyzing the content of short text, mining the potential semantic relationship of short text to build a text-based feature set, how to accurately build this feature set is the current research hotspot.Zhang Qun et al. [based on word2vec training word vector, model the short text from the granularity level, and then train LDA topic model to complete the feature extraction process.Guo Dongliang et al. [used word2vec's skip gram model to obtain short text features, then sent them to CNN to further extract high-level features, and finally put them into softmax classifier after K-MAX pooling operation to obtain classification model.Yu Zheng [et al] synthesized the existing word vector model, proposed a word vector model based on coded isa relation, and then extended it to the field of long text, constructed the semantic vector representation of short text set and long text.TF-IDF algorithm [calculates the weight of training text words, takes some words with the largest weight as feature words, and then constructs short text feature group through their weight.TF-IDF algorithm is simple and easy to implement, but there are some problems such as inaccurate feature word extraction, small variance of feature word weight, low discrimination between texts and poor classification effect.Therefore, this paper improves the original algorithm and proposes a feature extraction and short text classification method combining chi square statistics and tf-iwf algorithm to solve the above problems.Tf-iwf algorithm pays more attention to the weight difference of feature words caused by the number difference of feature words when calculating the weight of text entries. [10] [11] [12] [13]

## 3. Short Text Classification Model

### 3.1. Chi Square Statistical Method

Chi square (chi square) is usually used to calculate the difference between the distribution of data and the hypothetical distribution, which is used to measure the degree of association between an entry and its category.If the chi square value is larger, the degree of association

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202104_4(4).0053

between the entry and the category is greater, and the ability of the entry to reflect the category is stronger; if the chi square value is smaller, the degree of association between the two is smaller, and the ability of the entry to reflect the category is smaller [14].

**Table 1.** Relationship between features and class

| | Belong to $\mathbf{c_i}$ | Not belong to $\mathbf{c_i}$ | total |
|---|---|---|---|
| Text containing the word t | A | B | A+B |
| Text without the word t | C | D | C+D |
| total | A+C | B+D | N=A+B+C+D |

The calculation method based on the feature word t is shown in the formula (1), (2), (3), (4)

$$\mathbf{E_{11}} = \left(\frac{\mathbf{A+B}}{\mathbf{N}}\right) \times (A + D) \tag{1}$$

$$E_{12} = \left(\frac{A+B}{N}\right) \times (B + D) \tag{2}$$

$$E_{21} = \left(\frac{C+D}{N}\right) \times (A + C) \tag{3}$$

$$E_{22} = \left(\frac{C+D}{N}\right) \times (B + D) \tag{4}$$

According to Chi square test formula to calculate the characteristic word t and category$c_i$.The correlation degree is shown in formula (5)

$$D_{11} = \frac{(A-E_{11})^2}{E_{11}} \tag{5}$$

Find out the same reason$D_{12}$, $D_{21}$, $D_{22}$Then, the feature words T and T are brought in and solved$c_i$Of $\chi^2$The value is shown in formula (6)

$$\chi^2(t,\ c_i) = \frac{N(AD-BC)^2}{(A+B)(A+C)(B+C)(C+D)} \tag{6}$$

According to the formula (6), the word t is related to the category $c_i$When the correlation is low, $\chi^2$The closer the value is to 0. Feature word t and category $c_i$ When there is a strong correlation, $\chi^2$The higher the value is.

Steps of chi square statistical square to select feature words:

1. In $c_i$ All words are listed in the category$t_{i,1}, t_{i,2}, t_{i,3} \ldots t_{i,j}$
2. Calculate each word $t_k$ And category $c_i$ Of $A_{i,k}, B_{i,k}, C_{i,k}, D_{i,k}$
3. Computational words $t_k$ And categories $c_i$ Of $\chi^2(t_k, c_i)$value
4. Take $\chi^2$Maximum value m Three words as feature words

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202104_4(4).0053

## 3.2. Tf-iwf Algorithm

Inverse document frequency (IDF) only focuses on the difference between the number of documents, and ignores the difference in the weight of inverse document frequency caused by the difference in the number of entries in different documents.

Tf-iwf (term frequency inverse word frequency) is an algorithm used to evaluate the extent to which a word can reflect its corpus.

Term frequency (TF) is a measure of a word $t_i$In the document$d_j$The more frequent the word appears, the higher the word frequency value. The formula is shown in (7)

$$tf_{ij} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{7}$$

Where, $n_{i,j}$ Express entry $t_i$ In the document $d_j$ The number of occurrences in the, $\sum_k n_{k,j}$ Represents text $d_j$The total number of occurrences of all entries in.

Inverse word frequency (IWF) refers to words $t_i$The reciprocal of the proportion of the total number of documents in D. The function of IWF is to avoid high frequency words, but to get high weight words with little effect on documents.The mathematical formula is shown in (8)

$$iwf_i = \log \frac{\sum_m n_j t_i}{n_j t_i} \tag{8}$$

Where, $\sum_m n_j t_i$ Indicates that all documents in class m have entries $t_i$The total number of $n_j t_i$ Represents a document $d_j$Entry in $t_i$The number of.

The tf-iwf value is determined by $tf_{ij}$Value sum $iwf_i$The value is multiplied by the value $w_{i,j}$The calculation formula is shown in (9)

$$w_{i,j} = tf_{ij} iwf_i \tag{9}$$

Tf-iwf is used to filter common words and give more weight to the words that can better reflect the corpus.If the high-frequency term in a text presents low-frequency state in the text set, the tf-iwf value of the term has a high weight value.

## 3.3. Short Text Classification Process

The flow chart of feature extraction and short text classification is shown in Figure 1. Firstly, the data is cleaned, and the training text is preprocessed, including word segmentation, stop words and so on. Then, the feature is extracted, and the chi square values of all the entries and their categories are calculated. According to the chi square values, the entries in each category are arranged in order.The next step is to calculate the weight of the feature words, and calculate the weight of the extracted feature words according to the tf-iwf algorithm described above.Compared with other classification systems, SVM is a kind of linear classifier which realizes data binary classification according to supervised learning, so SVM classifier is selected as the classification model.SVM classifier is trained by short text feature vector and its label.In the process of testing, the text to be tested is extracted, and then the classification of the text to be tested is predicted by the trained classifier. [15-17]

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202104_4(4).0053

# 4. Experiments

## 4.1. Data Platform and Data Set

(1) Experimental platform

The hardware platform is based on Windows 7 operating system, and the memory is 8GB.The algorithm part is written in Python 3.6 language, and the python modules used include natural language processing library: gensim3.6.0; Machine Learning Library: sklearn0.20.2; mathematical calculation library: numpy1.15.4; progress processing library tqdm4.43.0.

(2) Experimental data set

In this paper, we collected 10000 hotel reviews from Feizhu, including 7000 good reviews and 3000 bad reviews.80% of the text is used for training and 20% of the data is used for testing.The data in training set and test set are independent of each other and there is no duplicate text.All the data are preprocessed: remove the English text, emoticons, duplicate, stop words, etc., and use stuttering segmentation to segment the data.The distribution of the data set used in the experiment is shown in Table 2.

**Table 2.** Distribution of dataset

| Text category | Training set | Test set |
|---|---|---|
| Positive | 5600 | 1400 |
| Negative | 2400 | 600 |

## 4.2. Experimental Parameter Setting

(1) Parameter setting of feature selection model based on Chi square statistics

Python is used to simulate the relationship between different feature numbers obtained by chi square statistics and text classification effect, as shown in Figure 1. Therefore, the feature number 400 when the accuracy converges to the maximum value is selected as the feature number obtained by chi square statistics.

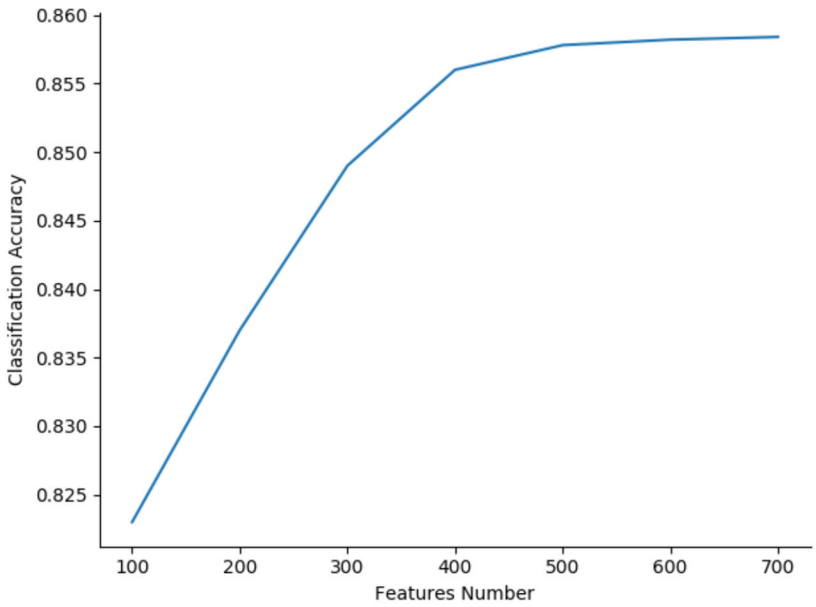


**Figure 1.** Relationship between number of features and accuracy of test set

(2) Parameter setting of SVM text classifier

The parameter setting of SVM classifier is shown in Table 3.

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202104_4(4).0053

**Table 3.** Parameter setting of SVM classifier

| Parameter name | Parameter setting | Parameter name | Parameter setting |
|---|---|---|---|
| Kernel | Linear | Class Weight | {0:1.3 1:1} |
| Penalty coefficinent | 1 | | |

## 4.3. Experimental Evaluation

Accuracy, recall, and F1 score are three evaluation indexes commonly used in classification experiments.Among them, TP represents the number of samples that belong to the positive class and are predicted to be positive class; FN represents the number of samples that belong to the positive class and are predicted to be negative class; FP represents the number of samples that belong to the negative class and are predicted to be positive class; TN represents the number of samples that belong to the negative class and are predicted to be negative class.

Accuracy refers to the proportion of the number of correctly classified samples in the classification results to the number of all classified samples, as shown in formula (10).

$$P = \frac{TP}{TP+FP} \tag{10}$$

Recall rate refers to the proportion between the number of correctly classified samples and the actual number of texts in the classification result, as shown in formula (11).

$$R = \frac{TP}{TP+FN} \tag{11}$$

F1 score is a comprehensive evaluation standard integrating accuracy and recall, as shown in formula (12).

$$F = \frac{2PR}{P+R} \tag{12}$$

## 4.4. Experimental Results and Analysis

(1) Validity verification based on feature selection

In order to verify the effectiveness of feature selection based on Chi square statistics, two groups of experiments are set up: the first group uses short text classification method combining chi square statistics and TF-IDF feature extraction, and the second group uses short text classification method based on TF-IDF feature extraction. Both groups of experiments use SVM model for classification, and the results are shown in Table 4.

**Table 4.** Contrast verification between feature choosing

| Experiment | accuracy | Recall rate | F1 value |
|---|---|---|---|
| First group | 85.7% | 82.7% | 83.9% |
| Second group | 72.5% | 69.0% | 71.3% |

(2) Validity verification based on feature extraction

In order to verify the effectiveness of the fusion of chi square statistics and tf-iwf algorithm, three groups of experiments were conducted: the first group used word2vec method for feature extraction; the second group used the fusion of chi square statistics and TF-IDF algorithm for feature extraction and short text classification; the third group used the fusion of Chi Square statistics and tf-iwf algorithm for feature extraction and short text classification.The three

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202104_4(4).0053

groups of experiments are classified by SVM classifier model, and the results are shown in Table 5.

**Table 5.** Contrast verification between feature selection

| Experience | Category | Accuracy | Recall | F1value |
|---|---|---|---|---|
| First group | Negative | 79.3% | 77.3% | 78.3% |
| | Positive | 90.0% | 91.0% | 90.5% |
| | Mean value | 84.7% | 84.2% | 84.4% |
| | Weight mean | 86.8% | 86.9% | 86.8% |
| Second group | Negative | 83.2% | 68.8% | 75.8% |
| | Positive | 87.3% | 94.1% | 90.6% |
| | Mean value | 85.3% | 81.5% | 83.2% |
| | Weightmean | 86.1% | 86.5% | 86.2% |
| | Negative | 84.7% | 78.4% | 80.9% |
| Third group | Positive | 91.4% | 93.5% | 92.2% |
| | Mean value | 88.1% | 86.0% | 86.9% |
| | Weight mean | 89.4% | 89.0% | 88.9% |

(3) Feature extraction and short text classification based on tf-iwf algorithm compared with other classification methods

In order to make a comprehensive judgment on the classification effect of the feature extraction and short text classification method based on Chi square statistics and tf-iwf algorithm, this method is compared with other text classification methods.As a traditional text vector representation method, vector space model (VSM) simplifies text content to vector operation; support vector machine (SVM) is a classical binomial classifier based on supervised learning; and k-nearest classification algorithm (k-nearest) is an effective way to classify dataNeighbor, KNN, as a classical machine learning classification algorithm, has the characteristics of simple operation and strong interpretability.Therefore, SVM classifier is used to classify and compare the text vector represented by VSM Model.The experimental results are shown in Table 6.

**Table 6.** Validation based on classification method

| Methon | Accuracy | Recall | F1 value |
|---|---|---|---|
| VSM +SVM | 74.5% | 72.1% | 73.0% |
| TF-IWF+KNN | 80.5% | 79.3% | 80.2% |
| TF-IWF+SVM | 87.2% | 86.4% | 86.9% |

It can be seen from the table that the proposed method of feature extraction and short text classification based on Chi square statistics and tf-iwf algorithm is better than the short text classification based on VSM Model.

## 5. Conclusion

Aiming at the problem that the traditional feature extraction algorithm in text classification has not strong ability to obtain feature representation, this paper proposes a feature extraction algorithm based on Chi square statistics and tf-iwf, which solves the problem of low text discrimination caused by TF-IDF's concentrated weight distribution to a certain extent,

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202104_4(4).0053

and constructs a short text classification model based on Chi square statistics and tf-iwf algorithmThe model has enhanced the ability of text feature extraction and improved the classification accuracy.

## References

[1] SHENG Cheng Cheng,ZHU Yong, LIU Tao. Public opinion analysis based on Weibo social network[J]. Intelligent Computer and Applications, 2019, 9(01): 57-59+64.

[2] Li Ding-yu,Hu Xue-gang. Cross-domain Sentiment Classification Algorithm for Short Text[J]. Journal ofChineseMini-MicroComputerSystems, 2018, 39(05): 1005-1009.

[3] SHAO Yunfei, LIU Dongsu.Classifying Short-texts with Class Feature Extension[J].Data Analysis and Knowledge Discovery.,2019, 3(09): 6067.

[4] DING Yue,WANG Xueming.Naïve Bayes classification algorithm based on improved feature weighting[J].Application Research of Computers, 2019, 36(12): 3597-3600+3627.

[5] WANG Zheng, LIU Shi-pei, PENG Yan-bing. An Essay Context Recognition Model Based on Syntax Decision Tree and SVM Algorithm [J]. Computer and Modernization, 2017(03):13-17.

[6] FAN Yun-jie, LIU Huai-liang. Research on Chinese Short Text Classification Based on Wikipedia[J]. New Technology of Library and Information Service, 2012(03): 47-52.

[7] HU X, SUN N, ZHANG C, et al. Exploiting internal and external semantics for the clustering of short texts using world knowledge [C] / /Proceedings of the 18th ACM conference on Information and knowledge management. Hong Kong: ACM, 2009: 919-928.

[8] DING lianhong, SUN Bin, ZHANG Hongwei. Short Text Classification Based on Knowledge Graph Extension[J]. TechnologyIntelligenceEngineering, 2018, 4(05):38-46.

[9] WANG Sheng, FAN Xinghua, CHEN Xianlin. Chinese short text classification based on hyponymy relation[J]. Journal of Computer Applications, 2010, 30(03): 603-606+611.

[10] ZHANG Qun, WANG Hongjun, WANG Lunwei. Classifying Short Texts with Word Embedding and LDA Model[J]. New Technology of Library and Information Service, 2016(12):27-35.

[11] GUO Dongliang, LIU Xiaoming, ZHENG Qiusheng. Internet Short-text Classification Method Based on CNNs[J]. Computer and Modernization, 2017(04): 78-81.

[12] YU Zheng. The Study and Application of Text Embeddings with Deep Learning Technique[D]. East China Normal University, 2016.

[13] YANG Bin, HAN Qingwen, LEI Min, ZHANG Yapeng, LIU Xiangguo, YANG Yaqiang, MA Xuefeng. Short Text Classification Algorithm Based on Improved TF-IDF Weight[J]. Journal of Chongqing University of Technology(Natural Science), 2016, 30(12): 108-113.

[14] LU Yun-qing. Problems in the Statistical test with Pearson's CHI square statistics[J]. Statistics & Decision, 2009(15): 32-33.

[15] LI Lingli. A Review on Classification Algorithms in Data Mining[J]. Journal of Chongqing Normal University(Natural Science), 2011, 28(04): 44-47.

[16] Kotsiantis S B. Supervised machine learning: a review of classification techniques. Proceedings of the 2007 Conference on Emerging Artificial Intelligence Applicationsbin Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies. Amsterdam, the Netherland. 2007. 3–24.

[17] Wang Yang, Xu Shanshan, Li Chang, Ai Shicheng, Zhang Weidong, Zhen Lei, Meng Dan. Classification model based on support vector machine for Chinese extremely short text[J/OL]. Application Research of Computers:1-5.https://doi.org/10.19734/j.issn.1001-3695.2018.06.0514.

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202104_4(4).0053

[18] FENG Guohe, WU Jingxue. A Literature Review on the Improvement of KNN Algorithm[J]. Library and Information Service,2012,56(21):97-100+118.

[19] ZHENG Teng, WU Yu-chuan. Research on the Classification Methods of Multiple SVM Short Texts based on LDA Feature Extension[J]. Journal of Wuhan Textile University, 2019, 32(02): 72-76.