Analysis of Emotional Tendency Based on Chinese Sugar-water Shop Evaluation Text

Xueqi Feng

University of Science and Technology Liaoning
e-mail: FengXueqi303@outlook.com

Abstract

The catering market has developed rapidly, and under the influence of the Internet and the epidemic, online consumption has become increasingly strong. Food is the life of the people, and sweets are an important factor in improving happiness in life. For the sugar water store, combined with online and offline fine operation, grasp the user evaluation, in order to stand out in the competition. This research uses the evaluation data of the eight sugar water shops in Guangzhou to extract keywords using TextRank for the texts with long evaluation data, so that the evaluation texts are controlled within 200 words, and then machine learning algorithms are used to analyze, mine and classify them. The experimental results show that this method can solve the binary classification problem of positive emotion and negative emotion in short time.

Keywords: Sentiment analysis Chinese text classification TextRank Natural language processing

1. Introduction

With the rapid development of the country and the gradual improvement of people's living standards, people's pursuit of quality of life is becoming more and more extreme, and food is an important standard for the pursuit of happiness index of life. As the saying goes, food is the priority of the people, people all over the world have the pursuit and love for food. In today's Internet era, most people tend to look at the reviews and ratings of major food review software before making a choice. For catering enterprises, combined with online and offline fine operation, grasp the user evaluation, in order to stand out in the competition.

In recent years, with the deepening of the Internet application, the text data containing important information has exploded. As for the catering industry, as an important part of the tertiary industry, most enterprises are actively combining the development trend of network platforms, and more and more tourism consumers can share the experience brought by food through review websites and obtain evaluation information. These evaluation information contain a lot of effective information about food market preferences, and it is of great significance to extract effective information from consumer evaluation content and adjust it according to consumer demands for the development of food recommendation. In all kinds of websites at present, food and beverage reviews are mainly based on numerical scores and text reviews. For text reviews containing more real feedback, manual judgment of text content is very inefficient. Therefore, it is necessary to classify them by automatic algorithm and put forward corresponding improvement measures.

The advent of machine learning algorithms has made text classification and data mining more efficient. Traditional machine learning algorithms mainly include Support Vector Machine[1], Decision Tree[2], Random Forest[3], etc. These network models need to rely on manual completion and are semi-automatic models. Deep learning, which came later, has even greater advantages.

At present, there are more and more researches on the mining and analysis of text information at home and abroad, which are widely used in many fields such as product recommendation, catering and social media. Zhang Jianhua[4] firstly extracted emotion words from product review texts, used LDA model to establish topic distribution for the extracted texts, and conducted positive and negative binary classification according to the characteristics of comment topic distribution, finally verifying the superiority of this algorithm, but this algorithm relies heavily on emotion dictionary. Chen Z[5] reduces the dimension of the feature matrix corresponding to the text according to the characteristics of the words themselves, thus speeding up the training speed of the Convolutional Neural Network model without reducing its accuracy. Feng J[6] extracted emotion words with dependency syntax and Word2Vec, and applied the semi-supervised classification method to the evaluation of catering field with good results. Yan S[7] designed an extended method of clustering semi-supervised topics to get users' attention. Then, he used a special sentiment dictionary to conduct sentiment analysis on the topics, obtain users' preferences, and make personalized recommendations according to their preferences. Guan P F[8] constructed a parallel network model of word vector attention mechanism and BiLSTM, which showed superior performance. Deng H[9] added the attention mechanism into the network model integrated by CNN and BiLSTM to classify and predict positive and negative emotions in social media comments, and verified the high recognition accuracy of this method.

Inspired by the above researches, this research studies text classification algorithms for the evaluation of sugar and water stores based on machine learning algorithms, and the workflow is shown in Figure 1.

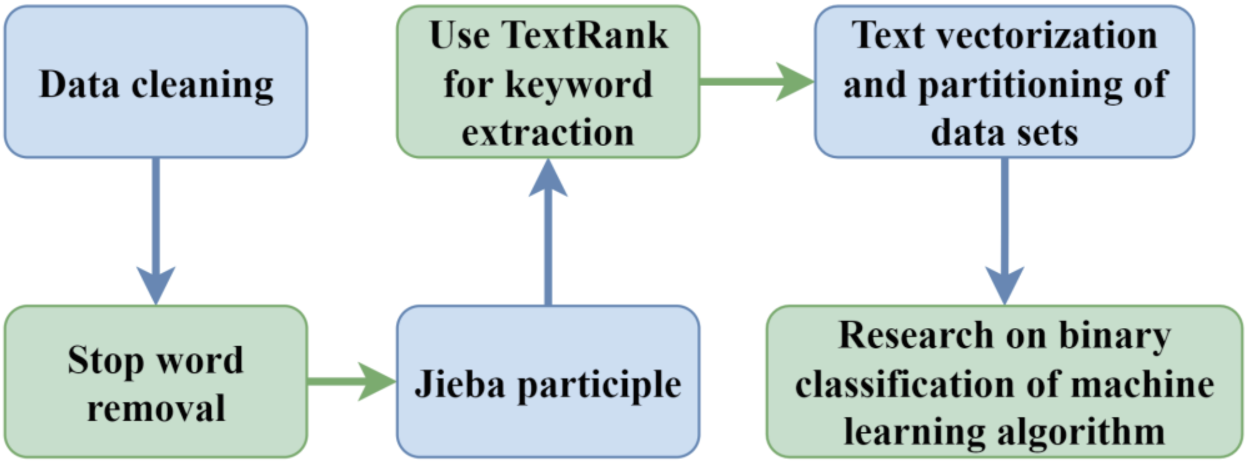
Data cleaning
Use TextRank for keyword extraction
Text vectorization and partitioning of data sets
Stop word removal
Jieba participle
Research on binary classification of machine learning algorithm

Figure 1 Work flow chart of this research

2. Data

2.1 Data Preparation

In this research, we obtain from the website1 the public comments data of eight popular sugar-water stores in Guangzhou from July 2005 to September 2018, including 32,485 comment texts and the corresponding customer ID, comment time, rating of taste, environment and service, and store ID.

The data is cleaned to remove 54 reviews that do not write reviews or rate taste, environment and service, leaving 32,429 reviews.

2.2 Data Preprocessing

The data set is analyzed according to the characteristics of the mining data, and the corresponding solution is proposed. In the stage of data set analysis, this research finds that there are extremely high frequency words and extremely low frequency words in the data set, and it is necessary to do word segmentation processing for the evaluation content.

In natural language, only a small number of words are used frequently, and a large number of words are used relatively infrequently, some of which are used very infrequently. And some punctuation marks and English letters are meaningless for judgment statements. Therefore, when preprocessing text, it is necessary to remove punctuation marks and extremely high and low frequency and unimportant words from the feature set, such as words that have no practical significance for emotion classification (stop word). The significance of this is reflected in two aspects: it can not only narrow the feature space to reduce the training cost of the model, but also increase the weight of meaningful content words in the text, so that the model can focus on learning them. Eliminate stops that don’t make much sense for emotional classification. The stop word table used in this article is the Chinese stop word table. A sentence is divided by using a Jieba segmentation, and a Jieba is a Chinese word segmentation device.

Six of the evaluation contents after removing the stop words and word segmentation are empty, and the final dataset contains 32,421 evaluations. As shown in Figure 2, this research makes statistics on the number of words in these evaluations. There are 2320 comments with more than 200 words, and the maximum number of words is 1478, which will affect the accuracy of the model. Therefore, the text with more than 200 words is extracted with the TextRank algorithm, and the keywords are extracted within 200 words. It can retain key information to the maximum extent, which helps to improve the accuracy of classification.

1 https://tianchi.aliyun.com/dataset/4366

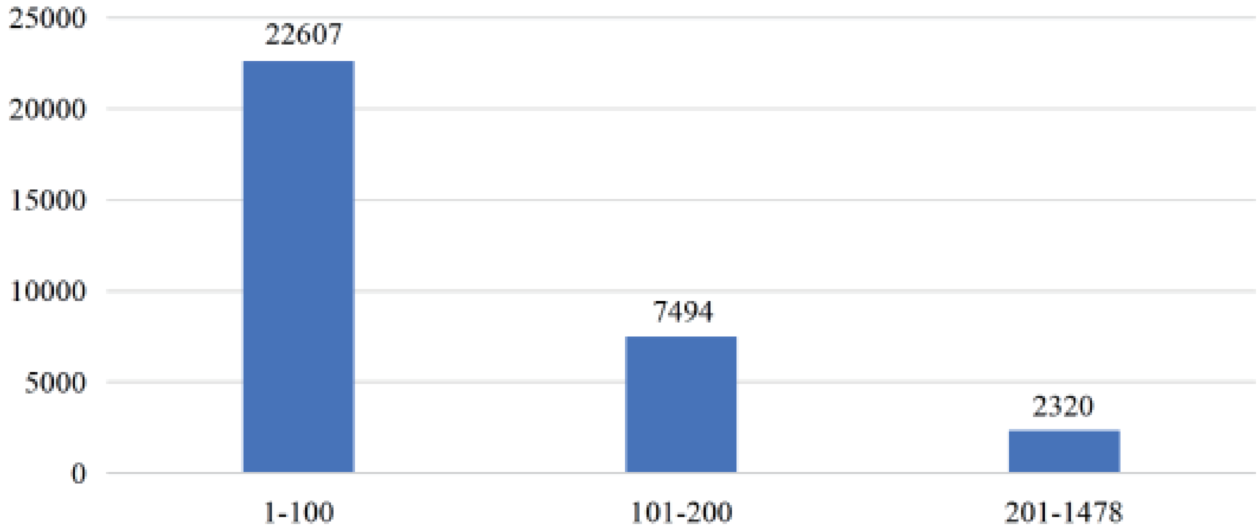

Figure 2 Evaluation word count

2.3 Text Vectorization and Partitioning of Data Sets

The text data faced by natural language processing is unstructured and disorganized text data, while the data processed by machine learning algorithms is often fixed-length input and output. Therefore, machine learning cannot directly process raw text data. Textual data must be converted into a form that can be recognized by computer numbers, such as vectors. Therefore, text vectorization is required, which is a necessary step to achieve text classification.

The main text vectorization models are One-hot representation and Bag of Words (BOW) model. One-hot is almost the first method used to extract text features, reducing text directly to a collection of words, each of which is independent regardless of its grammatical and word-order relationships. Word2Vec and word2id are equivalent to two methods of the bag of words model. Word2Vec is mainly to use context information to extract the feature vector of the word, the method is to predict the word; The word2id package is equivalent to giving each word a computer-recognized ID, standardizing the words. The data in this research is a large number of words after word segmentation, so the word bag model is more suitable for the experiment in this research.

The Decision Tree(DT) and Logistic Regression(LR)[10] classifiers use the Word2vec package. The Text Convolutional Neural Network(TextCNN)[11] uses the word2id package. In all experiments, the training set and the test set are divided by 8:2. The training set includes 25,936 review data and the test set included 6485 review data.

3. Methodology

Kim[11] earlier applied CNN to text classification and built TextCNN model. The short text classification model based on CNN usually includes seven parts: input layer, convolution layer, pooling layer, Dropout, full connection layer and output layer. Convolution layer and pooling layer are the most critical feature extraction links. The convolution layer extracts text features by constructing one-dimensional convolution kernel, moving it up and down, and performing convolution operations with the text representation matrix in the convolution window. The pooling layer selects the extracted features to screen out the most significant features, while reducing the feature dimension and preventing overfitting. In general, in short text classification, the convolutional layer and pooling layer need to be alternately superimposed, and the feature information can be obtained from multiple angles through multiple feature extraction and feature selection. Then, enter the full connection layer, integrate the feature information, and display the results in the output layer. Figure 3 shows the CNN structure of this research. The convolution layer contains three convolution layers, and the convolution kernel size is 3,4,5.

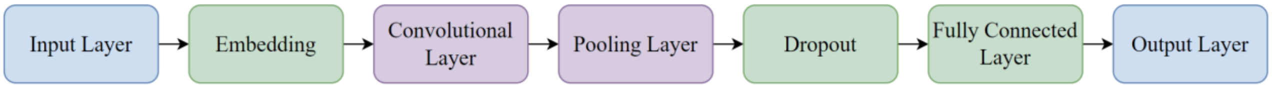

Figure 3 CNN structure chart

Although the convolution and pooling constructed by TextCNN model are single-layer and the parameter adjustment is simple, it has a good classification effect in experimental verification, which shows the effectiveness of CNN for short text classification to a certain extent. Therefore, this research chooses this model to do experiments.

4. Experiments

4.1 Performance Evaluation Index of Algorithm

The criteria included precision, recall, F1 and support, macro avg, and weighted avg. Support can be defined as the corresponding number of samples in each class of target values.

Table 1 Evaluation standard index

		Predicting results	
		Positive	Negative
Real results	True	TP	FN
	False	FP	TN

$$precision = \frac{TP}{TP+FP} \tag{1}$$

$$recall = \frac{TP}{TP+FN} \tag{2}$$

$$F1 = \frac{2 \cdot precision \cdot recall}{precision + recall} \tag{3}$$

$$accuracy = \frac{TP+TN}{TP+FN+FP+TN} \tag{4}$$

Formula 1 is the calculation formula of precision rate, formula 2 is the calculation formula of recall rate, formula 3 is the calculation formula of F1, also known as the harmonic average of accuracy rate and recall rate, and formula 4 is the calculation formula of accuracy rate. TP represents the actual positive emotion and is judged to be positive emotion; FN is actually positive emotion but judged negative emotion; TN indicates the actual negative emotion and is judged to be negative emotion; FP is actually negative emotion but judged positive emotion.

Macro avg: The average accuracy for all categories -- formula 5; Weighted avg: is an improvement over macro averaging that takes into account the exact weighting of the number of samples in each category as a percentage of the total sample -- equation 6.

$$macroavg = (precision\_0 + precision\_1)/2 \tag{5}$$

$$weightedavg = precision\_1 \cdot \left(\frac{\sup port\_1}{\sup port\_all}\right) + precision\_0 \cdot \left(\frac{\sup port\_0}{\sup port\_all}\right) \tag{6}$$

4.2 Experimental Environment

The CPU is 11thGenIntel (R) Core (TM) i7-11370H@3.30GHz, the RAM is 16GB, and the operating environment is Window10. In the experiment of Decision Tree and Logistic Regression, python3.9 and pytorch1.13.1 are used, while Text Convolutional Neural Network is used in different environments, specifically, tensorflow2.1.0, pytorch1.10.2, keras2.3.1. python3.6 version.

4.3 Comparison of experimental results

Table 2 Comparison of DT and LR experiment results

		precision	recall	F1	support
DT	positive	0.81	0.84	0.82	5111
	negative	0.30	0.26	0.28	1376
	accuracy			0.71	6487
	macro avg	0.55	0.55	0.55	6487
	weighted avg	0.70	0.71	0.71	6487
LR	positive	0.97	0.84	0.90	6136
	negative	0.16	0.56	0.25	351
	accuracy			0.82	6487
	macro avg	0.57	0.70	0.58	6487
	weighted avg	0.93	0.82	0.86	6487