Sentiment Analysis of Online Catering User Comments Based on Random Forest Feature Extraction

Yanqiu Liu, Fuming Ye
Computer and Information Engineering
Guizhou University of Commerce
Guiyang, China
651220506@qq.com

2022 IEEE 10th International Conference on Information, Communication and Networks (ICICN) | 978-1-6654-9082-5/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICICN56848.2022.10006565

***Abstract*—With the development of the Internet economy, online ordering modes have gradually been widely accepted, and even become an indispensable part of the lives of office workers. But, the online catering platform has many difficulties, such as low entry threshold for merchants, no restriction of online catering industry standards, and unable to effectively and long-term supervise a large number of merchants. In order to realize the effective supervision of online catering and take into account the characteristics of Chinese, this paper first obtains the user generated content of catering businesses through the Internet and uses the Jieba word segmentation tool for Chinese word segmentation, then carries out data cleaning and uses the Random Forest model to screen out five features that have great impact, and finally uses the SVM, Decision Tree and AdaBoost algorithm for data modeling, prediction and analysis. The experimental results show the better performance of our method in online catering user comments data.**

Keywords—random forest, sentiment analysis, online catering

I. INTRODUCTION

With the further development of the Internet and the impact of the epidemic in recent years. In order to reduce unnecessary contact with strangers, more and more people shop on the Internet and make comments, resulting in the prevalence of online catering. However, the low access threshold of online catering platform merchants and the absence of online catering industry standard constraints make it difficult for users to choose food that meets their own wishes in a short time. How to quickly identify the advantages and disadvantages of catering businesses is particularly important. In order to further attract more consumers and improve their dining experience, many catering platforms allow consumers to comment on various delicacies and shops. It is precisely because of the rapid growth in the number of user comments that we have the possibility to use emotional analysis to distinguish the advantages and disadvantages of catering businesses.The emotional tendency of Chinese comments includes positive and negative. Different from the general topic based classification method, sentiment analysis is more complex to a certain extent, involving more implicit information in comments. Topic based methods can be classified only through keyword recognition, while emotion analysis is more about extracting the hidden information in comments [1]. Compared with the sentiment analysis of English comments, the sentiment analysis of Chinese increases the difficulty of analysis to a certain extent due to the structural diversity of sentences. At present, the common Chinese sentiment analysis methods mainly include the method based on emotion dictionary [2] and the method based on machine learning [3].

Among them, typical methods based on emotional dictionary include HowNet emotional dictionary [4] and emotional analysis method based on domain dictionary [5]. Machine learning based affective analysis methods include semantic rule-based methods (Yang H et al. [6]) and PMI based methods (Zhang Y et al. [7]). Some researchers believe that the emotional polarity of comments is mainly judged by adjectives in comments. It is proposed to establish an emotional dictionary based on adjectives, and then calculate the emotional polarity of comments according to certain rules. However, this method relies too much on the quality of the emotional dictionary, requires certain industry experience and knowledge, and has poor generalization ability. (Hu et al. [8]). Therefore, the effect of the method based on emotional dictionary or machine learning is not good. Therefore, this paper proposes a method of feature selection based on random forest model and emotional analysis using multiple classification models. The experimental results show that this method performs better than the method based on emotion dictionary in emotion analysis.

The organizational structure of this paper is as follows. Section 2 mainly deals with the related work of Chinese comment emotion analysis, including the method based on emotion dictionary and the method based on machine learning. Section 3 describes data preprocessing. Section 4 mainly deals with the research methods of this paper. Section 5 describes the model selection and analysis of comparative experimental results. Finally, section 6 describes the research conclusions and possible future work.

II. Related Work

In this section, we give an brief introduction to the related work on the sentiment dictionary-based methods and machine learning-based methods for sentiment classification.

A. Sentiment Classification Based on Sentiment Dictionary

The main research method based on affective dictionary is to match the affective tendencies of calculation comments by constructing affective dictionaries in corresponding fields. (Sazzed S [9]) considers the influence of the distance between degree words, negative words and emotional words on the emotional tendency of criticism. But, the emotional expression of some comments is more obscure, and using this method to identify will cause some errors. In addition, the emotional tendencies in user reviews change, which makes the analysis difficult.

B. Sentiment Classification Based on Supervised Machine Learning

The main research method based on data mining is to transform the sentiment analysis of comments into a binary classification supervised learning, that is, to classify comments into positive sentiment and negative sentiment(Such as SVM,etc).By selecting representative features through a specific algorithm model and using feature coding to transform features into feature vectors, we can achieve the classification of Chinese comments.(Aurachman R et al. [10]) propose a system and programming script that can predict currency values using sentiment analysis.

III. Data Preprocessing

In this section, we give an brief introduction to the work on Data preprocessing.

A. Data Acquisition

The data set used in this paper is from www.dianping.com, and the crawler method is used to randomly obtain a total of 200000 user comments from 20 stores.After removing repeated comments and comments with less than 20 words, there are 100000 comments in total.

B. Data Cleaning

The comment data includes the following types: store name, user ID, user comment, user score and likes. The collected comment data should be filtered according to the research content of this paper. Before the emotional analysis of the data, the word segmentation and part of speech tagging of the comment data are necessary. This paper selects jieba word segmentation as a natural language processing tool to complete the data segmentation.

C. Emotional Word Tagging

Negative words in comments can directly change the emotional attitude expressed by comments, while adverbs of degree can strengthen the emotion expressed by comments to a certain extent. So, we can more accurately judge the emotional tendency of comments according to the negative words and degree adverbs near the emotional words. Therefore, this paper specially constructs the relevant negative words and degree adverbs table, as shown in Table 1 and Table 2 below:

TABLE I. Negative Words

common negative words	Bu4, Mei2, Fei1,Fou3,Wu4,Bu2Shi4 etc.

TABLE II. Degree Words and Weights

Level	Degree words	Weight
6	Bai3Fen1Bai3,Fei1Chang2,te4bie2	1.0
5	Guo4Du4,Chao1E2,He2Zhi3	0.8
4	Duo1Me,Fen4Wai4,Ge2Wai4	0.6
3	Geng4Wei2,Hai2,Hai2Yao4,Jiao4	0.4
2	Shao1Xu3,Yi4Dian3,shao1Wei1	0.2
1	Ruo4,Si1Hao2,Bu4Ding1Dian3	0.1

IV. Methodology

This paper mainly focuses on the risk early warning of online catering enterprises based on user generated content. Firstly, the random forest model is used to extract features, reconstruct comment data, and use SVM, decision tree and AdaBoost algorithm to train the model.The following is the overall process of this method:

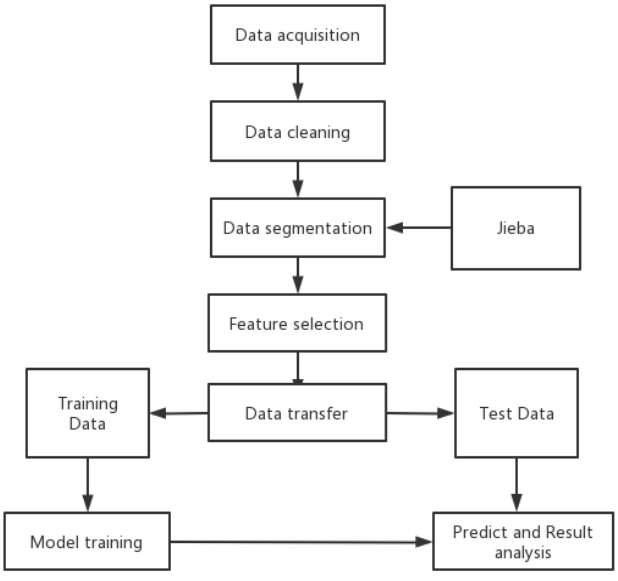

Fig. 1. The general framework of our work

A. Feature Selection and Data Conversion

A key point of emotion analysis based on machine learning is feature selection, which is related to the accuracy of emotion classification. This paper first selects 20 nouns that appear

most frequently in all comments as the initial data features; Then, find the emotional words closest to the feature, and assign values according to the above emotional word weight table to form a new data representation.

Algorithm: Feature selection and data conversion

1: Data segmentation
2: Part-of-speech tagging
3: Build weight table **a**⟵Emotional word
4: Build feature set **b** ⟵top 20 nouns
5: **for** comment **in** comments data set:
6: **for** word **in** comment:
7: **if** word in **b :**
8: find emotional word **e** and nagetive word **n** near word
9: get weight(e) in table a
10: dec = {word : weight(e)*$(-1)^{\text{count(n)}}$}
11: find all dec in comment
12:find all dec in comments data set
13:Data reconstruction

B. Feature Dimensionality Reduction

However, the features selected by this method are not accurate enough; Therefore, in this paper, we choose the Random Forest model which has the function of screening features to further reduce the number of our features.On the basis of considering both the model training time and the model accuracy, after several model training, we selected five features that have the greatest impact on the accuracy as the final features.

C. Over-sampling SMOTE

Because the data set used in this paper is unbalanced in the process of collecting positive and negative sample data, it may have a certain impact on the training of the model.So we use the SMOTE method to solve the problem of data imbalance.

V. Experiment and Analysis

This section mainly introduces the basic information of the data used in the experiment, the evaluation indicators of the model and the comparative experimental resultsAuthors and Affiliations

A. Data Sets

According to the research needs of this paper, we divide the comment data set into positive comments and negative comments according to users' ratings. After smote processing, 100000 positive comments and 100000 negative comments were obtained respectively. We divide the positive and negative comments into five subsets. Each time, 4 subsets are selected for training, and the remaining 1 subset is tested. A brief summary of data sets is presented in Table III.

TABLE III. A Brief Summary of Data Sets

Comments Data	Positive	Negative
Total	100000	100000
Training	80000	80000
Testing	20000	20000

B. Evaluation Criteria

Likes previous method,we evaluate the experimental results with precision, recall and F1-Measure. These three classic values are used for measuring the performance of positive comments and negative comments. Accuracy is used to evaluate the overall performance of sentiment analysis.

C. Experimental Results

In this section evaluates the results through comparative experiments of multiple models. The main evaluation indicators are precision, recall, F1 score and accuracy.

1) Model

In this paper, we use AdaBoost, decision tree and support vector machine to train the model. Among them, AdaBoost model, as an integrated model that continuously optimizes the accuracy of the model, can solve the over fitting problem to a certain extent.

2) Performance of our method

In this subsection, We will compare the performance of the method proposed in this paper and the method based onHowNet sentiment dictionary on the emotional analysis of online catering users' comments.

TABLE IV. Feature Extraction Using Random Forest

Method	Precision(%)		Recall(%)		F1-Measure(%)		Accuracy(%)
	Pos	Neg	Pos	Neg	Pos	Neg	
SVM	89.03	87.37	88.62	86.17	88.82	86.76	88.83
Decision Tree	91.31	90.73	91.52	94.21	91.41	92.44	91.33
Adaboost	90.19	91.53	92.62	89.17	91.39	90.33	90.07

TABLE V. Based on the HowNet Emotion Dictionary

Method	Precision(%)		Recall(%)		F1-Measure(%)		Accuracy(%)
	Pos	Neg	Pos	Neg	Pos	Neg	
SVM	86.27	87.45	87.65	86.05	86.95	86.74	86.85
Decision Tree	91.83	90.87	91.75	94.52	92.93	92.47	90.77
Adaboost	93.1	89.25	92.76	91.71	93.32	91.30	90.29

The performance of our method in three model is listed in Table IV, compare with the performance of HowNet emotion dictionary-based method in Table V. The method that we proposed shows the better performance. The reason may be that the method based on HowNet affective dictionary is more rough in feature selection, and the generalization ability of features is not as good as that of random forest. A more intuitive representation is shown in the figure below.

D. *Analysis*

From the above experiments, we can see that the method based on HowNet emotional dictionary and the method based on random forest feature extraction can get good results. However, the overall performance of the method used in this paper is better. Then, we discuss the advantages of this method.

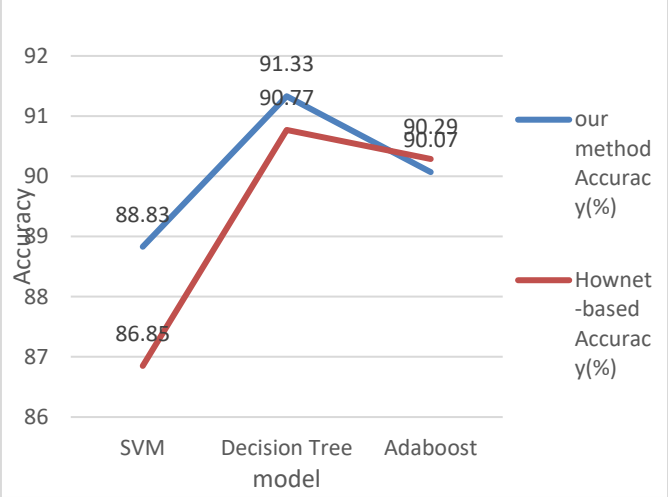

Fig. 2. The performance of two method in three model.

Compared with the method based on HowNet affective dictionary, our method has the following advantages. First, we use the relationship between negative words, degree words and features to obtain more accurate emotional tendencies of user comments. Secondly, we use the random forest model to further screen the features to reduce the interference of non key features on the results. To sum up, the performance of the method proposed in this paper is slightly better than the method based on HowNet emotional dictionary.

VI. Conclusion and Future Work

This paper studies the emotion classification of online restaurant users' comments, and proposes a comment emotion classification method based on random forest feature extraction. The performance of the three models is better than that of HowNet emotion dictionary alone, but there are still some shortcomings. For example, the processing of data imbalance is not detailed enough; There is still room for improvement in setting the weight of some emotional words; There is no analysis on the regional differences of online catering users. In the future research, we will further improve the above improvements to make the research in this field richer and more valuable.

Acknowledgment

The research is supported by "Guizhou Provincial Department of Education Youth Science and Technology Talent Growth Project, China (Project No. KY [2021]270)"

References

[1] Zhai Z, Bing L, Xu H, et al. Grouping product features using semi-supervised learning with soft-constraints.[C]// The. 2010.

[2] Taboada M, Brooke J, Tofiloski M, et al. Lexicon-based methods for sentiment analysis[J]. Computational Linguistics, 2011, 37(2):267-307.

[3] Bobicev V , Sokolova M . Inter-Annotator Agreement in Sentiment Analysis: Machine Learning Perspective[C]// RANLP 2017 - Recent Advances in Natural Language Processing Meet Deep Learning. 2017.

[4] Fellbaum C, Miller G. WordNet:An Electronic Lexical Database[M]. 1998.

[5] Ding, X. , B. Liu , and P. S. Yu . "A holistic lexicon-based approach to opinion mining." Proceedings of the International Conference on Web Search and Web Data Mining, WSDM 2008, Palo Alto, California, USA, February 11-12, 2008 ACM, 2008.

[6] Yang H , Chen J , Wang C , et al. Intelligent planning of product assembly sequences based on spatio-temporal semantic knowledge[J]. Assembly Automation, 2020, ahead-of-print(ahead-of-print).

[7] Zhang Y , Huang J F , University S , et al. Sentiment Analysis on Movie Comments Text From Douban Based on PMI Algorithm[J]. Modern Computer, 2019.

[8] Hu M , Liu B . Mining and summarizing customer reviews[C]// Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, USA, August 22-25, 2004. ACM, 2004.

[9] Sazzed S . Development of Sentiment Lexicon in Bengali utilizing Corpus and Cross-lingual Resources[C]// 2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science (IRI). IEEE, 2020.

[10] Aurachman R , Ramadani L , Utomo N S . Machine Learning Sentiment Analysis in Detection System for Rupiah Currency Value Using SysML Language[J]. Journal of Physics Conference Series, 2021, 1764(1):012177.