Opinion Mining for Thai Restaurant Reviews using Neural Networks and mRMR Feature Selection

Niphat Claypo and Saichon Jaiyen

Department of Computer Science, Faculty of Science, King Mongkut's Institute of Technology Ladkrabang, Thailand

s5650805@kmitl.ac.th, kjsaicho@kmitl.ac.th

***Abstract*— Currently, Thai restaurants are popular around the world. There are tons of reviews related to foods and services in social networking websites. These tons of customer reviews make it difficult to analyze the opinions of customer toward foods and services. To help the businesses, the model of opinion mining is proposed for classifying the reviews and to analyze the attitude of customers for improving their products and services. In this research, the artificial neural network is applied to classify the positive and negative reviews. In addition, the mRMR feature selection is used to select the features of data in order to reduce the number of features in the data set. Consequently, the computational times of learning algorithms for neural networks are reduced. The experimental results show that the neural network is an effective model for classifying the Thai restaurant reviews.**

Keywords*— *Classification; Feature selection; minimal-redundancy-maximal-relevance (mRMR); multilayer perceptron (MLP); radial basis function (RBF); Support Vector Machine (SVM).

I. INTRODUCTION

The opinion mining is an approach for analyzing the user comments for products and services. The user opinions can be collected from the social media, e-commerce websites, product review sites, blogs etc. These collected comments are divided into two classes which are positive and negative comments. There are many techniques are used to classify the data in the opinion mining. The classification methods based on machine learning including Support Vector Machine (SVM), Naïve Bayesian (NB), K-nearest neighbor (KNN) are most widely used [3], [4], [5], [13]. Classifying customer reviews for online products using Sentence Weight algorithm is proposed [6]. The Sentiment fuzzy classification algorithm is proposed to predict the positive and negative comments for the movie review [7]. The Back-Propagation Neural Network (BPANN) is used for classifying the movie and hotel reviews [8]. The self-organizing map (SOM) is used to extract the feature of data vectors [9]. The opinions on Twitter micro blog data are used for classifying the positive and negative opinions [10]. The text pre-processing techniques including Tokenization, Stop-word removal (STR), Lemmatization (LM), Number replacement (NMR), Synonym recognition (SYR) and Word generalization (WG) are proposed to optimize the experiments [12]. The natural language processing techniques based on sentiment analysis are applied to online reviews to find the most influential part-of-speech [14]. The mRMR feature selection is used to select the features of omics data based on three relevance evaluation measures including MI, CC, and MIC [15]. A two-stage feature selection algorithm by combining ReliefF and mRMR are used for finding a set of genes [16].

On social networking websites, users can post comments or reviews about products and services. The opinions of products, evens, news, and so on can be found in Blogs, social media, forums and others. For online business, the user opinions are very important because they can help the business to understand the user needs and feelings. Therefore, the opinion mining is the popular tool for analyzing their customers. Furthermore, it can be applied to business intelligence, ecommerce, etc.

In this paper, we propose opinion mining method for classifying the reviews of Thai restaurant using multilayer perceptron neural network (MLP) and mRMR feature selection. The data used in the experiments are Thai restaurant reviews consisting of positive and negative reviews. The text preprocessing techniques are used for preparing the dataset. The neural network model is used to classify the positive and negative reviews and the mRMR feature selection technique is used for selecting the features of data set for reducing the number of the features of data in the data set. The experimental results are compared with Support Vector Machines (SVM) and Radial Basis Function Neural Network (RBF). The experimental results show that the multilayer perceptron neural network gives the better performance.

II. OPINION MINING

Web 2.0 technology and social media cause many opinions on the websites. There are a lot of opinions on websites toward social events, political movements, company strategies, marketing campaigns, and product preferences, and so on. The comments and reviews can be found in various blogs, forums, social media and social networking sites, virtual worlds, and tweets. The opinion mining is a discipline of web mining and computational intelligence which is used to gather the opinions in various online sources, social media comments, and other user-generated contents for extracting, classifying, and understanding. The opinion mining can be applied for sentiment analysis [1]. The opinion mining collects opinions from websites for classifying through the mining process such as SVM, MLP, decision tree, and so on. The opinion mining can help business to know the positive or negative attitude of their customers about their products and services. Moreover, the opinion mining can help them to understand the advantages and disadvantages of their products and services in order to improve their further products and services.

978-1-4799-4963-2/14/$31.00 ©2014 IEEE

III. THE PROPOSED METHOD

A. Text Preprocessing

In this research, the text preprocessing technique is used for data preparation which is divided into two steps. The first step is tokenization used to split the reviews into tokens or words. The second step is the removal of the stop words which are words that have no meaning or do not make the meaning of the sentence changes when remove these words such as "a", "an", "the", "of", "I", etc. Each document consists of many stop words. When these stop words are deleted form the document, it can make the number of words in documents lower. Consequently, the size of data is reduced and the computational time is also decreased.

B. Text Transformation

The data used in the experiments must be appropriate for the classification methods. Because the methods used in the experiments are the computational method, they must use the data in numeric format as their input. Therefore, the documents which are in the text format must be transformed to the numeric format. In this paper, the words after removing all stop words of all positive and negative comments in the training set and testing set, the words in the training set are used as a keyword for creating the input vectors. Then, each document is transformed into the input vector by calculating the frequency of keywords appearing in that document.

C. Minimum Redundancy and Maximum Relevance (mRMR) Feature Selection

The mRMR is a popular method applied to select the features by using the relationship between the feature and the target class in order to reduce the size of the dataset. The basic idea of minimum redundancy is to choose the features such that they are mutually maximally dissimilar to other features. Let S denote the subset of features and |S| is the number of features in S. The minimal redundancy (Min-Redundancy) condition is

$$\min R(S),\ R(S) = \frac{1}{|S|^2}\sum_{x_i,x_j \in S} I(x_i; x_j) \quad (1)$$

where c is the target class and $I(x_i; x_j)$ is the mutual information between the individual feature x_i and x_j.

The mutual information between individual feature x_i and the target class c is the measure of relevance of that feature. Thus, maximal relevance criterion (Max-Relevance) is to maximize the average relevance of all features in S:

$$\max D(S,c),\ D(S,c) = \frac{1}{|S|}\sum_{x_i \in S} I(x_i; c) \quad (2)$$

where $I(x_i; c)$ is the mutual information between individual feature x_i and the target class c. The mRMR feature set is obtained by optimizing the conditions in Equation (1) and (2) simultaneously.

D. Artificial Neural Network

Artificial neural network is one of disciplines of artificial intelligence (AI). Artificial neural network is the computational model that simulates the function of neurons in the human brain. In data mining neural network is another popular to do classification that can classify the data correctly and efficiently. In this paper, the opinion mining model using multilayer perceptron (MLP) neural network are proposed in order to classify the reviews of Thai restaurants. The MLP consists of three layers including input layer, hidden layer, and output layer as shown in Fig. 1. The input layer is the first layer to get data. The hidden layer is the computational layer that maps the input data in the input space into a feature space where it becomes linearly separable. The last layer is output layer used to identify the class of data.

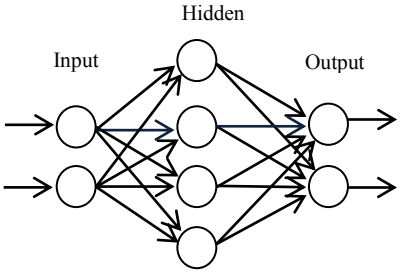

Fig. 1. Example of a simple neural network.

The MLP use the supervised learning algorithm to learn the data. The data set is divided into training and testing set. The training set consists of learning sample data and the answers called "target". The training data is fed into the network to calculate the output. Then, the output is compared to the answers to find the error for adjusting weights. These weights will be applied to other unseen data to predict the class of input data. The learning algorithm for MLP is called Back-propagation algorithm. The processes of the learning algorithm are shown in the following algorithm.

Back-propagation Algorithm

1. Input the training data to the MLP neural network and compute the network outputs.

2. For each output neuron j, calculate the local gradient δ_j defined by

$$\delta_j = e_j \varphi'(\mathrm{v_j}) \quad (3)$$

where $\varphi(\cdot)$ is the activation function, $v_j = \sum_i w_{ji} y_i$, and $y_j = \varphi(v_j)$.

3. For each hidden neuron j, calculate the local gradient δ_j defined by

$$\delta_j = \varphi'(v_j) \sum_{\mathrm{k}} \delta_{\mathrm{k}}(\mathrm{w_{kj}}) \quad (4)$$

4. Update each network weight w_{ji} by

$$w_{ji}^{new} = w_{ji}^{old} + \Delta w_{ji} \quad (5)$$

where $\Delta w_{ji} = \eta \delta_j x_i$ and η is the learning rate.

IV. EXPERIMENT

A. Data Sets

The reviews of Thai Restaurant used in this paper are collected from th.tripadvisor.com. There are 1,060 reviews randomly collected from this website. Each reviewed document include vocabularies, stop words, numbers and non-alphabet characters. Before training, the reviews documents are preprocessed by using text-preprocessing method and text transformation method as mention above and yield the results as the input vectors. These input vectors are divided into two classes including class 1 representing the positive review and class 0 representing the negative review. The dimension of the input vector is equal to the number of keywords in all comments after using text-preprocessing method. All data is divided into two set which are the training set and the testing set. In the training set, there are 510 input vectors with 1768 features and there are 550 input vectors with 1768 features in the testing set.

B. Experimental Results

In this experiment, three types of classifiers including the proposed MLP, RBF, and SVM are applied to solve this classification problem. To evaluate the performance, the experimental results of the proposed MLP method are compared with the results of RBF and SVM. This experiment is conducted using the neural network toolbox on Matlab. For the proposed MLP, we conduct the experiment with 2, 4, 8, 16 hidden neurons and the hidden neurons that give the best accuracy are chosen. After trial and error, the 8 hidden neurons that give the best accuracy are chosen. For RBF, we try to change the spread of the basis function and select the spread that give the best accuracy. After trial and error, the spread is set as 0.5 for RBF. The mRMR feather selection method is used to select the features of data. From the data set, we generate new six data sets. Each set has a number of features as 50, 100, 150, 200, 250 and 300 features.

All data sets are divided into training and testing sets. The simulations are run on all pairs of training and testing sets. The experimental results of the proposed MLP are then compared with RBF and SVM as shown in Table I. In the experiment, the comparative results from the classification by the proposed MLP, RBF, and SVM are illustrated in Table I. For all methods, reviews are classified into either positive or negative classes. The experimental results include accuracy and time obtained from six pairs of training and testing sets along three types of classification method. All three methods are simulated three times for all training and testing sets and the best accuracy are chosen. According to Table I, it has shown that the mRMR feature selection can reduce a number of features that make the data smaller and less training time in all classification methods, while it does not reduce the accuracy of all methods as well. From the comparative results, the proposed MLP method can achieve the best performance at 93.5% of accuracy, while SVM gains 90% of accuracy and RBF gains 78.9% of accuracy.

Fig. 2 shows the accuracy of MLP trained by all features and the accuracies of MLP trained by selected features using mRMR method. The accuracies of six data sets with selected features are higher than the accuracy of the data set with all features. Fig. 3 shows the accuracy of RBF trained by all features and the accuracies of RBF trained by selected features using mRMR method. The accuracies of six data sets with selected features are higher than the accuracy of the data set with all features. Fig. 4 shows the accuracy of SVM trained by all features and the accuracies of SVM trained by selected features using mRMR method. The accuracies of five data sets with selected features are higher than the accuracy of the data set with all features. However, the accuracy of SVM for the data set with 150 features is slightly less than the data set with all features. Fig. 5 illustrates the time comparison of the three methods using all features and selected features. The training times of MLP and RBF method with all features are very higher than the data sets with selected features using mRMR method. However, the training time of SVM method is almost no difference between all features and selected features because the number of all features is not sufficiently large. From the experimental results, the input vectors with 200 features give the best accuracy. Therefore, the mRMR feature selection can reduce the number of features. Enhance the accuracy and use less times for classifying reviews of Thai restaurants. From the experimental results, all three classification methods can efficiently classify the review data. However, the experimental results show that the proposed MLP is most effective for classifying reviews of Thai restaurants.

Table I. Comparison result of MLP, RBF, and SVM

Method	Non feature select		mRMR Feature Selection											
	1768		50		100		150		200		250		300	
	Time (second)	*Accuracy (%)*	*Time (second)*	*Accuracy (%)*	*Time (second)*	*Accuracy (%)*	*Time (second)*	*Accuracy (%)*	*Time (second)*	*Accuracy (%)*	*Time (second)*	*Accuracy (%)*	*Time (second)*	*Accuracy (%)*
MLP	24.31	90	0.58	91.7	0.65	92.2	0.697	92	0.8	93.5	0.58	92.2	0.47	90.9
RBF	35.3	72.5	0.24	85.3	0.24	78.9	0.24	77.5	0.24	75.7	0.43	74.4	0.65	75.9
SVM	0.54	89.7	0.2	90	0.5	88.9	0.39	86.9	0.6	90	0.33	88.2	0.3	87.1

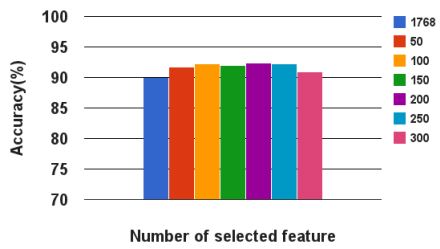

Fig. 2. Comparison of feature selection using MLP.

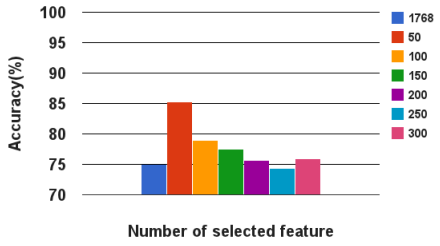

Fig. 3. Comparison of feature selection using RBF.

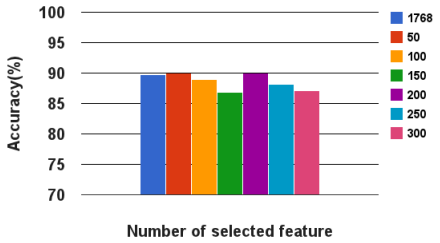

Fig. 4. Comparison of feature selection using SVM.

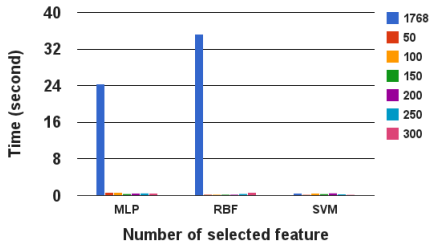

Fig. 5. Comparison time of the three methods using all features

V. CONCLUSION

In this paper, we propose the model to classify the opinions of customers for Thailand restaurants using opinion mining. For the proposed method, the mRMR feature selection is adopted to optimize the classification of Thailand restaurant reviews and the neural networks are used to classify them into positive and negative reviews. The experimental results have shown that high accuracy can be gained in all classification methods. When mRMR method is used to reduce the size of data set, it enhances the performance of the neural networks, spends less training time, and preserve high accuracy. Furthermore, an advantage of the neural networks is that they can adjust themselves to recognize patterns of the reviews and thus, they result in high accuracy in the classification. From the experimental results, they have shown that the proposed method, which utilizes the benefit of mRMR feature selection and MLP neural network, can achieve the best performance for opinion mining for Thai restaurant reviews.

VI. REFERENCES

[1] H.Chen and D.Zimbra, "AI and Opinion Mining", Intelligent Systems, Vol. 25, pp. 74 – 80, IEEE. 2010.

[2] H. Peng, F. Long and C. Ding. "Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy", Pattern Analysis and Machine Intelligence, Vol. 27, pp. 1226 - 1238, IEEE, 2005.

[3] C. Zhang, W. Zuo, T. Peng and F. He, "Sentiment Classification Reviews Using Machine Learning Methods Based on String Kernel", Convergence and Hybrid Information Technology, Vol. 2, pp. 909 – 914, IEEE, 2008.

[4] A.Khan, B.Baharudin, K.khan, "Sentence Based Sentiment Classification from Online Customer Reviews", Frontiers of Information Technology, ACM, 2010.

[5] N.Aleebrahim, M.Fathian and M.Reza Gholamian, "Sentiment Classification of Online Product Reviews Using Product Features", Data Mining and Intelligent Information Technology Applications, pp. 242 – 245, IEEE, 2010.

[6] X.Hu and B.Wu, "Classification and Summarization of Pros and Cons for Customer Reviews", Web Intelligence and Intelligent Agent Technologies, Vol. 3, pp. 73 – 76, IEEE, 2009.

[7] K.Mouthami, K.Nirmala Devi and V.Murali Bhaskaran, "Sentiment Analysis and Classification Based On Textual Reviews", Information Communication and Embedded Systems, pp. 271 – 276, IEEE, 2011.

[8] A.Sharma and S.Dey, "A Document-Level Sentiment Analysis Approach Using Artificial Neural Network and Sentiment Lexicons", ACM SIGAPP Applied Computing Review, VOL. 12, pp. 67-75, ACM, 2012.

[9] S.Nirkhi, "Potential use of Artificial Neural Network in Data Mining", Computer and Automation Engineering, Vol. 2, pp. 339 – 343, IEEE, 2010.

[10] B. Gokulakrishnan, P. Priyanthan, T. Ragavan, N.Prasath, A. Perera, "Opinion Mining and Sentiment Analysis on a Twitter Data Stream", Advances in ICT for Emerging Regions, pp. 182 – 188, IEEE, 2012

[11] G. R. Brindha and B. Santhi, "Application of Opinion Mining Technique in Talent Management", Management Issues in Emerging Economies , pp. 127 – 132, IEEE, 2012.

[12] Z. Ceska and C. Fox, "The Influence of Text Pre-processing on Plagiarism Detection", Industrial and Information Systems, pp. 376 – 380, IEEE, 2009.

[13] K. Gayathri, A. Marimuthu, "Text Document Pre-Processing with the KNN for Classification Using the SVM", Intelligent Systems and Control, pp. 453 – 457, IEEE, 2012.

[14] S. Thanangthanakij, E. Pacharawongsakda, N. Tongtep, P. Aimmanee, T. Theeramunkong, "An Empirical Study on Multi-Dimensional Sentiment Analysis from User Service Reviews", Knowledge, Information and Creativity Support Systems, pp. 58 – 65, IEEE, 2012.

[15] J. Yang, Z. Zhu, S. He and Z. Ji, "Minimal-redundancy-maximal-relevance feature selection using different relevance measures for omics data classification", Computational Intelligence in Bioinformatics and Computational Biology, pp. 246 – 251, IEEE, 2013.

[16] Y. Zhang, C. Ding and T.Li, "A Two-Stage Gene Selection Algorithm by Combining ReliefF and mRMR", Bioinformatics and Bioengineering, pp. 164 – 171, IEEE, 2007.

[17] M. Wisniewski and T. P. Zielinski, "MRMR-based feature selection for automatic asthma wheezes recognition", Signals and Electronic Systems, pp. 1 – 5, IEEE, 2010.