Vietnamese Online Hotel Reviews Classification Bases on Term Features Selection

Tran Sy Bang[a], Choochart Haruechaiyasak[b] and Virach Sornlertlamvanich[a]
[a] *School of ICT, Sirindhorn International Institute of Technology, Thammasat University, Pathum Thani 12121, Thailand*
[b] *Speech and Audio Technology Laboratory (SPT) National Electronics and Computer Technology Center (NECTEC) Thailand Science Park, Klong Luang, Pathumthani 12120, Thailand*

Abstract. This paper aims to present the improved techniques to classify the user's feedbacks on hotel service qualities. The data were mainly collected from online feedback sources by PHP program. The training set was manually tagged as: NEGATIVE, POSITIVE, and NEUTRAL. In total, 2969 Vietnamese language terms were successfully collected. In the first part, the common machine learning techniques like K-Nearest Neighbor algorithm (KNN), Decision Tree, Naive Bayes (NB) and Support Vector Machines (SVM) were applying for classification. In the second part, we enhanced the efficiency of the text categorization by applying feature selection techniques, $\chi 2$ (CHI). At the end of the paper, we concluded that the overall performance of general machine learning techniques was significantly improved by applying feature selection.

Keywords. Vietnamese sentiment analysis, feature selection, text categorization, machine learning.

1. Introduction

Opinion mining (OM) is a recent sub discipline at the crossroads of information retrieval and computational linguistics which is concerned not with the topic a document is about, but with the opinions it expresses. OM has huge domain of applications, ranging from tracking users' opinions about products or about political candidates as expressed in online forums, to customer relationship management [1]. Since the online booking services are quite popular in Viet Nam, it is essential that the hotel business holders keep updating feedbacks from customers. Customer opinion acknowledgement helps the business runner to adjust their service patterns to meet user's expectation. In turn, they could predict customer's needs in advance so that they prepare better planning and competitive strategies in the market.

Opinion classification has widely researched in many languages such as Chinese, France, and Japanese .etc. The domains are closely related to our research field such as restaurant evaluation, and costumer services evaluation. However, online hotel's service review has less attractive for conducting research in Viet Nam due to lack of training corpus. Recently, online hotel booking service and discussion are expanding rapidly in

Viet Nam and increasing of availability of corpuses, it is sound practicality for attempting a research.

Vietnamese language structure has complicated phonetic structure that it contains 4 different kind of tone marks such as rising tone " ′ ", falling tone "`", and the sentence structure is also different from other languages. That makes it is difficult to apply common studies from other languages for Vietnamese text classification. For instance, we cannot apply the Tokenizer that is designed for other languages such as English, or Japanese because it segmented the sentence differently. Therefore, we attempted to include Vietnamese Tokenizer for the purposes of the study. After that, we performed the main contribution of the paper, the comparative study on text categorization algorithms by several kinds of feature selection techniques for Vietnamese texts.

The term features are extracted from the collected Vietnamese corpus which was automatically retrieved from Agoda[1]. The size of the features could reach to hundred thousand terms depending on training corpus. In term of proving the efficiency of our technique, we reduced the number of features to few thousand terms. The expected outcome of the model is that it can classify the text on three predefined categories such as "POSITIVE", "NEGATIVE", and "NEUTRAL".

The remainder of the paper is organized as following structure. Section 2 provides a survey on the text categorization algorithms and feature selection techniques. Section 3 describes the implementation of our methodology. Section 4 presents the experiments results, and discuss some further works.

2. Key Techniques in Text Classification

2.1. Text Classification Technique

Text classification has widely applied in many contexts, ranging from document indexing based on a controlled vocabulary, to document filtering, automated metadata generation , word sense disambiguation, population of hierarchical catalogues of Web resources, and in general any application requiring document organization or selective and adaptive document dispatching [2].

There are many available text classification techniques for conducting research, including regression models, K-Nearest Neighbor classifiers, Decision Tress, Bayesian classifiers, Support Vector Machines and Neural Networks. Yang and Liu [4] have conducted a research on those mentioned methods, and they have concluded that SVM method ranked as the best one in term of accuracy. The NB technique has lower performance on the data collected from Reuter[2]. In what follows we will describe Decision Tree, Naïve Bayes, and SVM.

[1] http://www.agoda.com/
[2] http://www.reuters.com/

2.1.1. Support Vector Machines

SVMs are starting to enjoy increasing adoption in the machine learning and computer vision research communities [11]. It is only applicable for binary classification tasks, meaning that, using this method text classification have to be treated as a series of dichotomous classification problems [5].
The SVM classifies a vector d to either –l or 1 using:

$$s = \sum_{i=1}^{N} \alpha_i y_i K(d, d_i) + b \tag{1}$$

2.1.2. K-Nearest Neighbor

K Nearest Neighbor (KNN) algorithms ranks the document's neighbors among the training document vectors based on their similarity which can be measured by for example the Euclidean distance or the cosine between the two document vectors. KNN is instance-based learning, or lazy learning that does not have an off-line training phase. Therefore, it is considered as simplest technique among other machine learning methods. The kNN algorithm is quite simple: given a test document, the system finds the k nearest neighbors among the training documents, and uses the categories of the k neighbors to weight the category candidates [9]. The kNN can be written as:

$$y(\vec{x}, c_j) = \sum_{\vec{d_i} \in kNN} sim(\vec{x}, \vec{d_i}) y(\vec{d_i}, c_j) - b_j \tag{2}$$

where $y(\overrightarrow{di}, cj) \in \{0,1\}$ is the classification for document $\overrightarrow{di}$ with respect to category c_j (y = 1 for YES, and y = 0 for NO); sim($\vec{x}, di$) is the similarity between the test document $\vec{x}$ and the training document $\overrightarrow{di}$; and b_j is the category specific threshold for the binary decisions.

2.1.3. Naive Bayes

The Naive Bayes (NB) algorithm was first proposed and used for text categorization task by D. Lewis (1998) [6]. It is flexible that requires a number of parameters linear in the number of variables (features/predictors) in a learning problem. It used the Bayes' theorem, the conditional probability can be decomposed as:

$$p(C_k|\mathbf{x}) = \frac{p(C_k)\ p(\mathbf{x}|C_k)}{p(\mathbf{x})}. \tag{3}$$

2.1.4. Decision Tree (J48)

The decision tree rebuilds the manual categorization of training documents by constructing well-defined true/false-queries in the form of a tree structure. In the decision tree structure, leaves represent the corresponding category of documents and branches represent conjunctions of features that lead to those categories [9]. J48 is an open source

Java implementation of the C4.5 algorithm in the Weka[3] data mining tool. C4.5 is a program that creates a decision tree based on a set of labeled input data. This algorithm was developed by Ross Quinlan [10].

2.2. Feature Selection Techniques

Feature selection can be sub divided into two areas which are supervised unsupervised. The supervised method will be involved with human supported in text data labeling. In the other hand, unsupervised method will be conducted without the interference of human supports. In supervised feature selection, labeled training set of data will be modeled into the desired form. In the next step, the unlabeled test set will be analyzed to predict the outcomes. In contrast unsupervised feature selection method does not require a pre labeled dataset. But, heuristics learning algorithms are used for evaluation of the features [3].

2.2.1. Information gain

Information Gain measures the number of bits of information obtained for category prediction by knowing the presence or absence of a word in at document [7].Let c_1, L, c_k denote the set of possible categories. The information gain of a word *w* is defined to be:

$$\begin{aligned} IG(w) &= -\sum_{j=1}^{K} P(c_j)\log P(c_j) \\ &+ P(w)\sum_{j=1}^{K} P(c_j|w)\log P(c_j|w) \\ &+ P(\overline{w})\sum_{j=1}^{K} P(c_j|\overline{w})\log P(c_j|\overline{w}) \end{aligned} \tag{4}$$

2.2.2. χ2 (CHI)

χ2 (CHI): CHI is based on the statistical theory. It is useful in determining the statistical significance level of association rules. CHI is a normalized value and can be compared across the terms in the same category. CHI score between a term t and a class c is defined as:

$$\chi^2(t,c) = \frac{N \times [P(t,c) \times P(\bar{t},\bar{c}) - P(t,\bar{c}) \times P(\bar{t},c)]^2}{P(t) \times P(\bar{t}) \times P(c) \times P(\bar{c})} \tag{5}$$

3. Data and Experimental Results

3.1. Data Set

In this study we separated the work into training phrase and testing phrase. The simple PHP program was developed to retrieve the user opinions on hotel service from Agoda website. We extracted reviews in Vietnamese for 50 hotels, which are located in Vietnam

[3] http://www.cs.waikato.ac.nz/ml/weka/

(mainly in Hanoi, Ho Chi Minh City, Da Nang, and Nha Trang). The raw data are stored in the XML format. The data consisted of *Score* tag that shows the rating score from the reviewer, and the *Id* tag shows the index of review or the index of the comments. The corpus was preprocessed by Sentence Detection[4], Word Segmentation[5], and Part-of-Speech Tagging[6]. The sentences were refined by removing the sentences that contain the abnormal characters, and the sentences that their structures are not in Vietnamese standard, and we also eliminated sentences without tone mark. These sentence processing tools were developed by Phuong L.H which can process very large text data. Figure 1 shows the structure preprocessed training corpus. Each sentence is normally annotated by several capital letters including V, N, and A etc. that denoted for pronounce, noun, collocation respectively. Each word in the sentence was distinguished by the white space and the word that contains more than one syllables is connected by the underline "_" character. The system is based on a maximum entropy model. The training procedure requires no hand-crafted rules, lexicon, or domain-specific information. Given a corpus annotated with sentence boundaries, the model learns to classify each occurrence of potential end-of-sentence punctuations as either a valid or invalid sentence boundary [12].

The detail of the data set volume is shown in the Table 1. Totally, 2182 comments were successfully collected, and there were 2969 keywords in extracting process. From 2182 comments, there are 1005 NEGATIVE terms, 501 NEGATIVE terms, and 676 NEUTRAL terms. The annotation process was performed by human labor. In the initial step, all the sentences were labeled by an appropriated annotator by human sense judgment. In the next step, the sentence that consists of two contradicted terms were re exanimated to conclude the final label. The annotator performs a full disambiguation of two tagged versions of the same sentences. Chance of c_A and c_B agreeing on category k: $P(cA|k) \cdot P(cB|k)$. A_e is then the chance of the coders agreeing on any k [8]:

$$A_e = \sum_{k \in K} P(c_A|k) \cdot P(c_B|k) \tag{6}$$

The Cohen's kappa coefficient of our corpus was 0.89, which can be interpreted as almost perfect agreement.

</review> -<review Score="4,3" id="0"> <sentence Id="1" Class="NEGATIVE">Phòng/N nào/P cũng/R có/V muỗi/N và/CC kiến/N ./. </sentence> <sentence Id="2" Class="NEGATIVE">Rất/R nhiều/A muỗi/N ?/? </sentence> <sentence Id="3" Class="NEGATIVE">Đồ_ăn/A thì/C dở/A ./. </sentence> </review>
Translation
</review> -<review Score="4,3" id="0"> <sentence Id="1" Class="NEGATIVE">Every rom has mosquito and ant. </sentence> <sentence Id="2" Class="NEGATIVE">A lots of mosquito </sentence> <sentence Id="3" Class="NEGATIVE">The food tastes bad </sentence> </review>

Figure 1. The preprocessed reviews format.

[4] http://mim.hus.vnu.edu.vn/phuonglh/softwares/vnSentDetector
[5] http://mim.hus.vnu.edu.vn/phuonglh/softwares/vnTokenizer
[6] http://mim.hus.vnu.edu.vn/phuonglh/softwares/vnTagger

Table 1. The corpus volume metric.

Positive Terms	Negative Terms	Neutral Terms
1005	501	676

3.2. Method used

The Decision Tree (J48) classifier, Naïve Bayes Support Vector Machines (SVM) are used in the classification work. In the 1st experiment we conducted those native methods without applying feature selection techniques. Those methods were available in the Weka machine learning software, and 5-fold cross validation test was used to conducting experiment.

In the second experiment, information gain (IG), and CHI square ($\chi 2$) feature selections were performed. The standard measures of recall, precision, and F-Score are used to evaluate the system's performance. When we are comparing two annotations X and Y, these are:

$$recall(X,Y) = \frac{\text{number of identical nodes in } X \text{ and } Y}{\text{number of nodes in } X} \quad (7)$$

$$\text{precision}(X,Y) = \frac{\text{number of identical nodes in } X \text{ and } Y}{\text{number of nodes in } Y} \quad (8)$$

F-Score is the harmonic mean of both:

$$\text{F} = \frac{2PR}{P+R} \quad (9)$$

The accuracy of the system was measure by:

$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn} \quad (10)$$

where *tp* is True Positive, *tn* is True Negative, *fn* is False Negative, and *fp* is False Positive.

3.3. Result and Analysis

The Table 2 shows the result of sentiment classification by Decision Tree, Naïve Bayes, and SVM. In overall, we obtained that Naïve Bayes delivered a highest performance in all classes, the highest result is 91.8 % in "POSITIVE" class based on Recall. The highest result of SVM method was 87.8% in "POSITIVE" class based on Recall. The highest result for Decision Tree method was 82.3% based on Recall.

Table 2. The result of sentiment classification by Decision Tree, Naive Bayes, and SVM

Methods	Precision	Recall	F-Measure
Decision Tree			
POSITIVE	0,712	**0,823**	0,764
NEGATIVE	0,5	0,441	0,469

NEUTRAL	0,67	0,574	0,618
Weighted Average	0,65	0,658	0,651
Naïve Bayes			
POSITIVE	0,698	**0,918**	0,793
NEGATIVE	0,52	0,411	0,459
NEUTRAL	0,765	0,525	0,623
Weighted Average	0,678	0,68	0,664
SVM			
POSITIVE	0,725	**0,878**	0,794
NEGATIVE	0,628	0,481	0,545
NEUTRAL	0,67	0,577	0,62
Weighted Average	0,686	0,693	0,683

Figure 2 shows the average result of three selected method. The SVM got highest place that its accuracy was 69.3%, Naïve Bayes success rate was 68%, and Decision Tree has lowest performance as its accuracy was 65.8%.

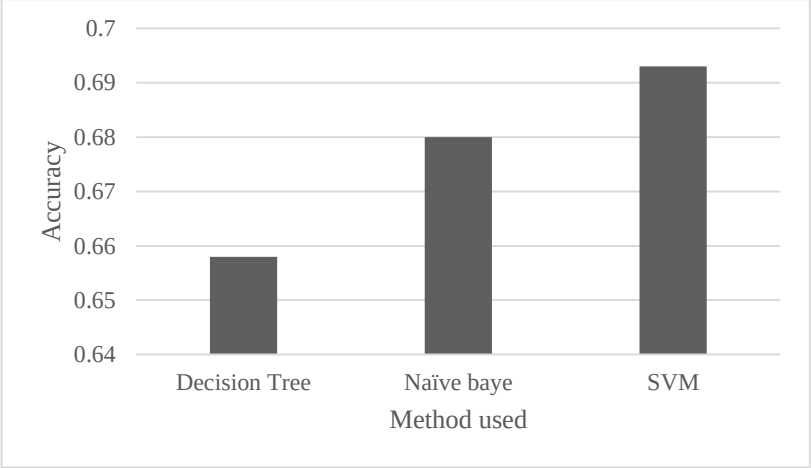

Figure 2: The average performance of the Decision Tree, Naïve Bayes, and SVM

The feature selection technique was handled by the Weka, the data mining software that allowed us to adjust the number of attributes in the preprocessing data. Information gain (IG) and $\chi 2$ (CHI) were applying in preprocess phrase in Weka. The number of attributes were selected from 2969 keywords. We run the test case on different number of attributes ranging from 240 to 1200.

Figure 3 shows the result of sentiment classification when we applied information gain. In overall, SVM delivered the best result in comparison with Naïve Bayes and Decision Tree. In Precision measurement the highest accuracy of SVM was 71.4% while Naïve Bayes was 68.8%, and Decision Tree was 65.4% respectively. The same scenarios happened in Recall measurement when SVM got the highest performance with 71% of accuracy. Naïve Bayes and Decision Tree had 68.5% and 65.8% accordingly. Lastly, in F-Score measurement SVM had its highest accuracy of 69.3%, Naïve Bayes had the second place with accuracy of 66.6%, and the lowest was Decision Tree with accuracy of 65.1%.

Figure 4 present the result of sentiment classification with application of CHI square feature selection. As we observed, the overall performance was slightly improved when we applied Information Gain feature selection technique. Decision Tree delivered the best result of 78.4 % in F-Score measurement with number of attributes are 1200. SVM has second highest accuracy which were 71.4%, 71%, 69.3% in Precision, Recall, and F-Score measurement respectively. While, Naïve Bayes has 69.7%, 69.3%, and 67.3% separately. Finally, Decision Tree has lowest performance when the accuracies were 65.6%, 65% in Precision and Recall. Although, we witnessed that Decision Tree has highest performance but throughout whole process with different number of attributes Decision Tree has lowest performance. This result confirmed that our experiment achievement agreed with other studies from other languages.

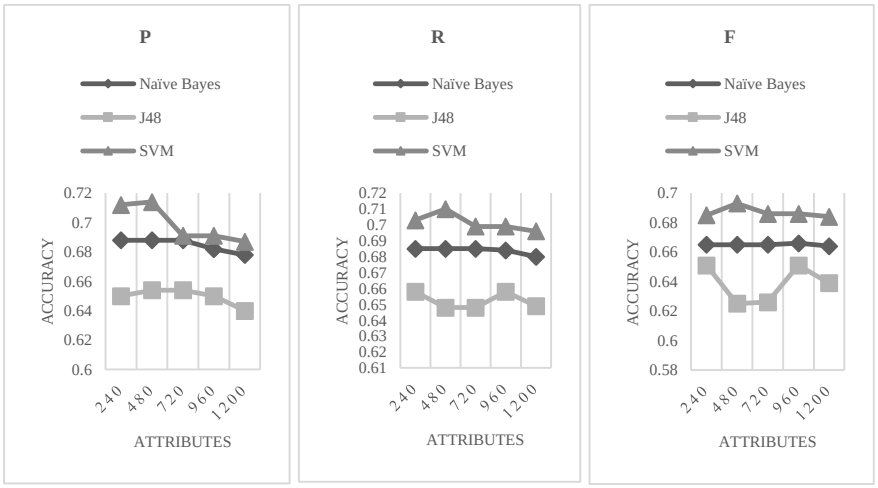

Figure 3: The result of sentiment classification by Decision Tree, Naive Bayes, and SVM with Information Gain feature selection

During the experiment, the recorded data showed that the accuracy POSITIVE term usually had the highest accuracy. For example, I was 94.9% in IG feature selection, and 95.2% in CHI feaure selection. This can be explained that the number of positive samples (1005) is dominant term in comparision with number of negative samples (501) in the corpus. Another reason may be positive sentences are usually stated clearly, while negative sentences are often stated implicitly.

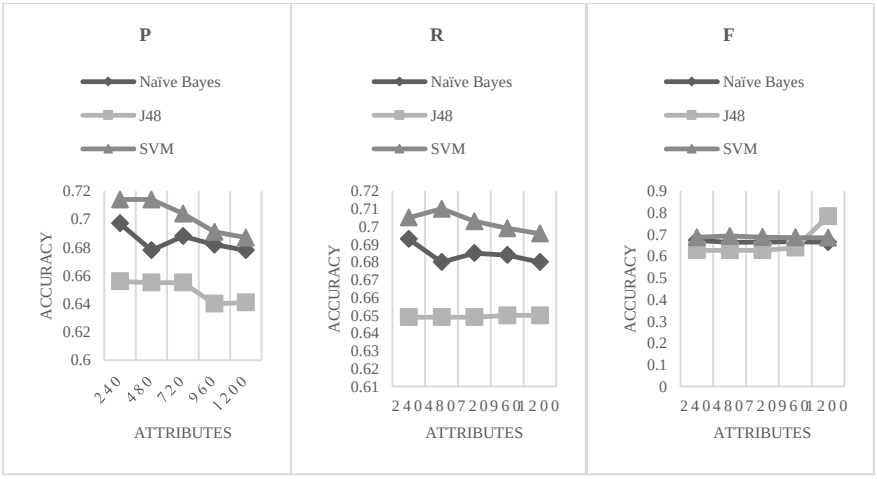

Figure 4: The result of sentiment classification by Decision Tree, Naive Bayes, and SVM with CHI square feature selection.

4. Conclusion

The purpose of text classification is to build systems which are able to automatically classify online feedbacks on hotel service domain into categories including POSITVIE, NEGATIVE, and NEUTRAL. In this paper we review the key text classification techniques including text model, feature selection methods and text classification algorithms in building a text classification system. Also we give an implementation of a text classification system based on Decision Tree and Naïve Bayes algorithm and Support Vector Machine. Our experimental results show that CHI square, and Information Gain are capable on enhancing the performance of proposed text classification algorithms. Concerning with other works, some papers that conducted experimental research on other language were slightly performed better result. However, these result are not directly comparable since their model are trained and tested on different corpus.

In further survey, we found that there were some Vietnamese Dependency parsing models have developed for structuring the sentiments. The accuracy of the model promisingly guarantee for better result of sentiment classification. This is bright track to follow in the next research.

References

[1] Andrea Esuli and Fabrizio Sebastiani, *SentiWordNet: A High-Coverage Lexical Resource for Opinion Mining.* Kluwer Academic Publishers, 2006.

[2] F.Sebastiani.*Machine Learning in Automated Text Categorization ACM Computing Survey*, 34(1): 1 - 47, 2002.

[3] H. Liu, M. Motoda, L. Yu, *"Feature Extraction, Selection, and Construction"*. In N. Ye (eds.): The Handbook of Data Mining, Lawrence Erlbaum Associates, Inc. Publishers, pp. 409-423, 2003.

[4] Y. Yang and X. Liu. A re-examination of text categorization. In *Proc. of the 22nd Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval*. Morgan Kaufmann, pp. 42-49 (1999).

[5] Thorsten Joachims.*Text Categorization with Support Vector*, 2002.
[6] D. Lewis. Naive bayes at forty: The independence assumption in information retrieval. Proc. of European Conf. on Machine Learning, pages 4–15, 1998.
[7] F.Sebastiani.Machine *Learning in Automated Text Categorization*. ACM Computing Survey, 2002, 34(1): 1 - 47.
[8] Raquel Fernández . *"Assessing the Reliability of an Annotation Scheme for Indefinites"*. Institute for Logic, Language & Computation University of Amsterdam, page 12, MoL Project Jan 2011.
[9] Menaka S, Radha N. "*Text Classification using Keyword Extraction Technique"*. International Journal of Advanced Research in Computer Science and Software Engineering. Volume 3, Issue 12, December 2013.
[10] Jay Gholap. *"Performance tuning of j48 algorithm for prediction of soil fertility".* Dept. of Computer Engineering College of Engineering, Pune, Maharashtra, India.
[11] John C. Platt. *Fast Training of Support Vector Machines using Sequential Minimal Optimization*. Microsoft Research 1 Microsoft Way, Redmond, WA 98052, USA.
[12] Le-Hong, P., and T V. Ho. "A Maximum Entropy Approach to Sentence Boundary Detection of Vietnamese Texts". RIVF 2008. Vietnam