Vietnamese Sentiment Analysis Based on Term Feature Selection Approach

Conference Paper · November 2015

CITATIONS
9

READS
1,460

3 authors:

Bang Tran
California State University, Sacramento
21 PUBLICATIONS **420** CITATIONS
SEE PROFILE

Choochart Haruechaiyasak
National Electronics and Computer Technology Center
103 PUBLICATIONS **1,334** CITATIONS
SEE PROFILE

Virach Sornlertlamvanich
Musashino University
225 PUBLICATIONS **1,381** CITATIONS
SEE PROFILE

Vietnamese Sentiment Analysis Based on Term Feature Selection Approach

Tran Sy Bang[1] and Choochart Haruechaiyasak[2] and Virach Sornlertlamvanich[3]

[1]School of ICT, Sirindhorn International
Institute of Technology,
Thammasat University, Thailand
bangtran365@gmail.com
[2]Speech and Audio Technology Laboratory (SPT)
National Electronics and Computer Technology Center (NECTEC)
Thailand Science Park, Klong Luang, Pathumthani 12120, Thailand
choochart.haruechaiyasak@nectec.or.th
[3]School of ICT, Sirindhorn International Institute of Technology, Thammasat University,
Pathum Thani 12121, Thailand
virach@siit.tu.ac.th

Abstract. We propose an improved technique to analyze sentiment for Vietnamese texts based on term feature selection approach. The sentiment analysis task is to classify a sentence into one of the following predefined categories: positive, negative, and neutral. In order to analyze the sentiment, we compare three different text categorization algorithms including Decision Tree, Naive Bayes (NB) and Support Vector Machines (SVM). Furthermore, we enhance the efficiency of the text categorization by applying feature selection technique, $\chi 2$ (CHI). The evaluation was conducted on 1,650 hotel reviews written in Vietnamese languages. The experimental results showed that applying term feature selection could significantly improve the performance of the sentiment analysis.

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

1 Introduction

Online business is quite popular in Viet Nam with the support of Internet infrastructure improvement and the launching of many online business sites. The customer could easily find suitable products with preferred price at home . Moreover, they can also provide their feedbacks and reviews on the certain kinds of selected items. The Vietnamese analysis technique focuses on the opinioned texts from the customers regarding their opinions, beliefs and rating. From the analyzing technique, the classified results could help the customers make their own decision on selecting the suitable products such as the hotel room to stay or the restaurant for the family dining.

adfa, p. 1, 2011.

Sentiment classification has been performing extensively on the English language. There were many kinds of models created to serve this purpose. Similarly, other languages such as Japanese, France, Romania, and Chinese have a very good concern for conducting research. Previously, there were not so many attention on this topic for Vietnamese language due to the limitation of corpus, and unpopularity of Vietnamese online business sites.

In this paper we present the text categorization technique which is considered as classification approach for machine learning. The term features are extracted from the collected Vietnamese corpus. The size of the features could reach to hundred thousand terms depending on the size of the corpus. In term of modeling, we reduce 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".

Vietnamese sentences are normally having complicated structures that consist of many two syllable words such as "rộng rãi" (large), "sang trọng" (luxury). Moreover, the words are usually containing the tone marks such as rising tone "ˊ", falling tone "ˋ". Therefore, we cannot apply the tokenizer that is created for other languages such as English, or Japanese. The training corpus has to be segmented into the series of terms. After that, we can perform the main contribution of the paper, the comparative study on text categorization algorithms several kind feature selection techniques for Vietnamese text.

The rest of this paper is organized in the following structure. Section 2 provides the literature review on the text categorization algorithms and feature selection technique. Section 3 describes the implementation of our proposed methodology. Section 4 presents the experiments results, and discuss some further works.

2 Related works.

2.1 Text Classification Techniques.

Text classification is used to manage data that helps to make predictions about new data. There are several text classifications techniques have developed to serve this purpose which is including decision trees[3], regression model [2], kNN (k-Nearest Neighbor) classification [6], Bayesian methods[4], SVM (Support Vector Machines) [5]. Yang and Liu (1999) have conducted a research on those mentioned method, and they have pointed out that the SVM method was capable of delivering the best performance. The NB technique has lower performance on the data collected from Reuter[1]. This paper we decided to apply three techniques on Vietnamese text classifica-

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

tion including decision tree, Naive Bayes, and SVM to perform the comparative study.

Decision tree are a well-known method of classification [Apte and Weiss, 1997]. The C4.5 is the classification algorithm that is developed base on the decision tree. J48 is an open source Java implementation of the C4.5 algorithm in the Weka[2] data mining tool. Naive Bayes approaches applies joint probabilities of words and categories to estimate the probabilities of categories given a document (Yang and Liu, 1999). Support Vector Machines (SVM) is a relatively new learning approach introduced by Vapnik in 1995 for solving two-class pattern recognition problems[7]

2.2 Feature Selection Technique

Feature selection for methods for text classification are available for conducting the experiment. Document frequency (DF) thresholding, information gain (IG), mutual information (MI), $\chi 2$ statistic (CHI) are preferred to apply in text classification. Yang and Pedersen [9] conducted the comparative study on five feature selection methods to improve the classification techniques and concluded that DF, IG, and CHI have the best performance, respectively.

$\chi 2$ (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. In this paper we will present the feature selection technique based on CHI test.

3 Data Set and Experimental Setup

3.1 Data Set

The training corpus data was retrieved from the hotel booking online service named Agoda[3]. We selected different hotel locations in Viet Nam to ensure the diversity of data and prevent the duplication of information. The locations were mainly focus on big cities like Ha Noi, Nha Trang, Da Nang. The collected corpus has to undergo some preprocessing tasks such as sentence detection by the system called vnSentDetector[4,] vnTokenizer[5] to perform word segmentation. The next step was labeling the sentences in to three categories including "positive", "negative", and "neutral". Table 1 shows the statistics of sentences in difference categories. There were 1651sentences for conducting the experiment, including 1005 positive sentenc-

[2]http://www.cs.waikato.ac.nz/ml/weka/
[3] http://www.agoda.com
[4] http://mim.hus.vnu.edu.vn/phuonglh/softwares/vnSentDetector
[5] http://mim.hus.vnu.edu.vn/phuonglh/softwares/vnTokenizer

es, 371 negative sentence, and 275 neutral sentences. Also, the number of extracted keywords were 2434.

Class	Number of sentences
Positve	1005
Negative	371
Neutral	275
Total	1651

Table 1. The statistical information of the corpus

Figure 1 shows the customer reviews of the certain hotel in Viet Nam, we create the program by Java to retrieve the data. The expected data should contain only the opinioned text which is located in the right hand side column. Moreover, we also collect the rating score which is indicated by the bold blue number in the upper right hand side corner in each comment box. This score was not used in our experiment, but it could be useful in the further study.

Fig. 1. Customer reviews sample from Agoda

Figure 2 show the overall framework of the Vietnamese sentiment analysis system. The system is technically divided the system into 4 components.

- **Preprocessing components:** This part has the function of retrieving the raw data from the online sources such as online forums, online discussion board. After that, the raw data has to be refined to get our concerned information. We eliminated the unrelated information but kept the opinioned text from the user only.

- **Tokenizing components:** We implement the Vietnamese word segmentation by the vnTagger [6] tool that is developed by Le-Hong, P., T M H. Nguyen, A.

[6] http://mim.hus.vnu.edu.vn/phuonglh/softwares/vnTokenizer

Roussanaly, and T V. Ho [12]. This tool was created in Java programming, and it is very powerful in word segmentation with the success rate up to 98%.

- **Classification components:** This is the main contribution of the paper, it received the preprocessing text and extract the concerned features, and classify the sentences based on three categories including "NEGATIVE", "POSITIVE", and "NEUTRAL". The result achieved from this part will be used in the next part.

- **Recommendation, and suggestion:** The result that produced in the previous part will be used to evaluate the service. The evaluation information will be shown to the customers if they search for their needed service. This information will help the customers to choose the suitable services, or production.

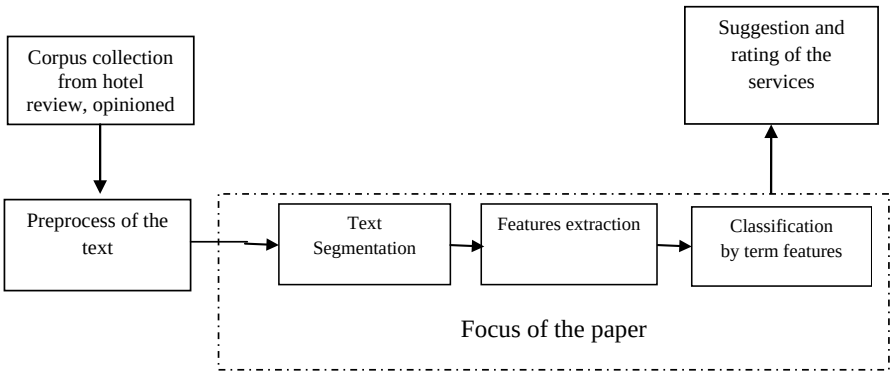

Fig. 2. The framework of the Vietnamese sentiment analysis systems

Figure 3 shows the format of the annotated training corpus with the specific classes, and the segmented words. The sample sentence has the "POSTIVE" class. The next step, we extract the keywords by using normalized Term Frequency-Inverse Document Frequency (TF-IDF) [8]. This task was manually performed, the person in charge determined the type of sentence and label them as "POSITIVE", "NEGATIVE", or "NEUTRAL" respectively. The next step, if the paragraph contains the contradicted opinioned sentences such as positive sentence vs. negative sentence , we have to solve this problem by applying Cohen's kappa coefficient that is presented as following formula:

$$\kappa = \frac{p_o - p_e}{1 - p_e},$$

where p_o is the relative observed agreement between the tagged sentences, and p_e is the hypothetical probability of chance agreement.

```
<text>
<category>
positive
</category>
<content>
Tôi rất hài_lòng về chuyến công_tác này vì đã tìm được một
khách_sạn đáp_ứng nhu_cầu của mình .
</content>
</text>
```

Translation

I'm satisfied with the work trip because I've found the suitable hotel.

Fig. 3. The training corpus format.

3.2 Experimental Setup

The experiment was performed using Weka machine learning tool by applying 5-fold cross validation test. At the first stage without applying feature selection, we deployed three learning algorithms including Decision Tree, NB, and SVM. The performance was measured based on Precision (P), Recall (R), and F-Measure.

Table 2 shows the result of text classification by applying three techniques including Decision Tree, Naive Bayes, and SVM. In general we can observe that the SVM deliver the highest performance in all classes, the highest recorded result is 88.9% in positive class based on Recall. The highest result for Naive Bayes method was 78.8% in positive class based on Precision. The highest result for Decision Tree method was 85.2% based on Recall.

In overall, the Figure 4 shows that the method NB and SVM have higher efficiency in text classification than Decision Tree. The biggest difference in those methods was in R measurement, the SVM method got the efficiency of 67.4% while NB and Decision Tree got the efficiency of 63.1% and 64.1% respectively. Therefore, we decided to eliminate the Decision Tree in the next step. The next step, we improved the efficiency of the NB and SVM by applying CHI feature selection technique.

The feature selection technique was handled by the Weka, the data mining software that allowed us to select the type of the attribute in the preprocessing data. We also run the test on different percentage of keywords. The percentage of keyword was increased by 10% on the total of 2434 keywords. The Figure 5 show the statistical indication of the performance of the NB and SVM with feature selection. The best performance was in the range of 10% to 40% of keywords. The highest archived result of SVM was 69.41% with 10% of keywords, while the best result of the NB was 67.17% with 30% of keywords.

Method		P	R	F
Decision Tree	Positive	0.722	0.852	0.782
	Negative	0.46	0.388	0.421
	Neutral	0.386	0.215	0.276
	Average	0.351	0.607	**0.641**
Naïve Bayes	Positive	0.788	0.776	0.782
	Negative	0.481	0.469	0.475
	Neutral	0.291	0.316	0.303
	Average	0.636	0.631	**0.633**
SVM	Positive	0.735	0.889	0.805
	Negative	0.594	0.445	0.508
	Neutral	0.342	0.196	0.249
	Average	0.638	0.674	**0.646**

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

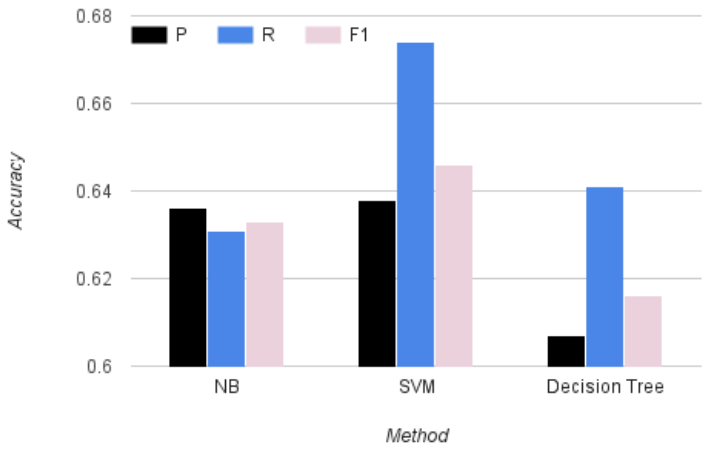

Fig. 4. The overall performance of Decision Tree, NB and SVM.

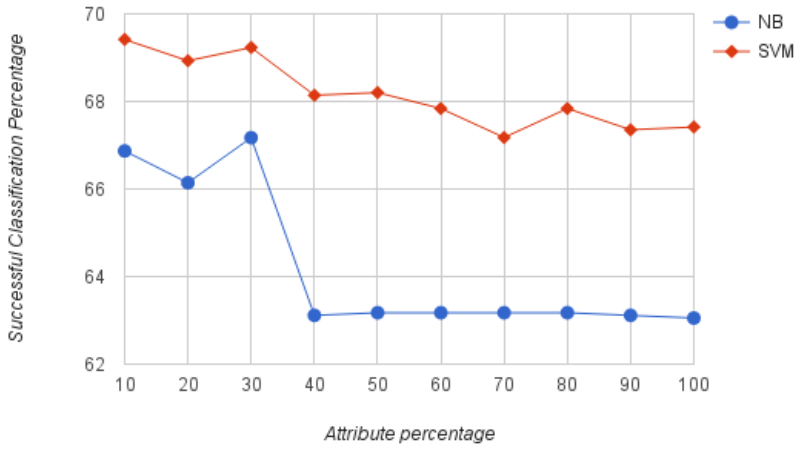

Fig. 5. Comparison between NB and SVM with CHI square attribute selection.

4 Conclusion, Discussion and Future Works

In this paper we proposed an approach to classify the Vietnamese by apply some common categorization algorithm such as Decision Tree, Naive Bayes, and Support Vector Machine. The best achieved result was 64.6 % generated by SVM. After applying CHI feature selection technique, we can increase the precision to 69.4%. This is quite impressive improvement, however, this result is still low in comparison with other language such as 82.34% from the C. A. Murthy and Tanmay Basu (IEEE, 2012) . After conducting the experiment we pointed out some problem that may cause the classification result did not get to high level as we expected. In the further experiment, we will prepare large size of training corpus to cover more Vietnamese words, the corpus will be not only focus on hotel service, but other fields. The number of class sentences should be equivalent. We also plan to deploy more feature selection technique to build better Vietnamese sentiment analysis system.

References

1. C. Apte and S.Weiss. Data Mining with Decision Trees and Decision Rules. Future Generation Computer Systems, pp. 197-210 (1997).
2. Y. Yang and C. G. Chute. An example-based mapping method for text categorization and retrieval. *ACM Transaction on Information Systems*, pp.252–277 (1994).
3. C. Apte, F. Damerau, and S. Weiss. Text mining with decision rules and decision trees. In *Proc. of the Conf. on Automated Learning and Discovery, Workshop 6: Learning from Text and the Web*, (1998).

4. D. D. Lewis. Naive (bayes) at forty: The independence assumption in information retrieval. In *Proc. of the 10th European Conf. on Machine Learning*, pp. 4–15 (1998).
5. T. Joachims. Text categorization with support vector machines: learning with many relevant features. In *Proc. of the 10th European Conf. on Machine Learning*, pp. 137-142 (1998).
6. 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).
7. V. Vapnic. The Nature of Statistical Learning Theory. Springer, New York, 1995.
8. Y. Yang and J. P. Pedersen. A comparative study on feature selection in text categorization. *Proc. of the Fourteenth Int. Conf. on Machine Learning*, pp. 412–420 (1997).
9. Yiming Yang and Jan O. Pedersen. A Comparative Study on Feature Selection in Text Categorization. In Proceedings of the Fourteenth International Conference on Machine Learning (ICML 97), pp. 412-420 (1997).
10. Tanmay Basu and C. A. Murthy. Effective Text Classification by a Supervised Feature Selection Approach. In 2012 IEEE 12th International Conference on Data Mining Workshops, pp. 918-925 (2012).
11. Nguyen Thi Duyen, Ngo Xuan Bach, and Tu Minh Phuong. An Empirical Study on Sentiment Analysis for Vietnamese. In The 2014 International Conference on Advanced Technologies for Communications (ATC'14), pp. 310-314 (2014).
12. Le-Hong, P., T M H. Nguyen, A. Roussanaly, and T V. Ho. A hybrid approach to word segmentation of Vietnamese texts. In Proceedings of the 2nd International Conference on Language and Automata Theory and Applications, pp. 240-249 (2008)