Classification of Telephone Subscriber Errors Based on Text Messages in Vietnamese Language

Phan Thi Ha, Phuong Nguyen

Abstract: ***This article describes a method for automatically classifying telecommunications subscriber errors based on text messages, using a machine learning method Support Vector Machine (SVM). The SVM method trains and tests on a set of data obtained from the text messages in Vietnamese of the actual line workers sending to the service operation centers. The results show that the proposed classification method using the SVM gives high accuracy and can be applied in practice.***

Index Terms, Text Classification, Natural Language Processing, Learning Support Vector Machine.

I. INTRODUCTION

Classification of the cause of subscriber errors (incidents) is the problem to provide the cause of the failure of subscribers in order to take diagnose of the errors for technique supports and improving the quality of telecommunications services. At each telecommunication company, this work is usually carried out by wire workers and manual labor, so there are often many errors, inconsistent terminology, non-technical expertise, resulting in difficulties for the technique service operation and management. In order to solve the above problem, it is necessary to develop a system of automated classification of subscriber incidents throughout the company to help monitoring and reporting on fluctuations of the incidents. The end goal is to help technique support and management for improving network quality and better customer service. Basically, this problem is in the field of text classification, the error incidents will be described under the observation and understanding of the wire worker or user via a text messages, based on which the system determines to which type of incident the message belongs to the list of known subscriber incident categories.

In recent years, the area of classifying texts has become more and more interested in satisfying the need to search, exploit human information as well as applying in many technique support systems. Many different problems are set up such as product evaluation, classification, sentiment analysis in the text format, etc. Text classification now has basic technical background. Text classification models are mainly using supervised machine learning methods [4, 5, 6] and non-supervised [7,8]. For Vietnamese text classification, there are also some works done on the classification of Vietnamese text by using unsupervised learning method, as described by Pham Tran Vu, Ho Chi Minh City Technological University.

Revised Manuscript Received on 03 August 2018.

Dr. Phan Thi Ha, Department of Information Technology, Posts and Telecommunications Institute of Technology, Hanoi, Vietnam (e-mail: hapt@ptit.edu.vn

Dr. Phuong Nguyen, Department of Information Technology, Posts and Telecommunications Institute of Technology, Hanoi, Vietnam (e-mail: phuongnt@ptit.edu.vn

The authors proposed to use the calculating similarities technique in text based on three aspects of the text: content, users, and whether or not they are associated with someone else. The author applied this method to calculate the similarity of the text with the previously learned by readable sample data. They also used latent semantic analysis (LSA) for record matching. In another works published in the 2011 science and technology development journal, they presented a method which do not require the use of ontology but it still have the ability to perform semantic, contextual comparisons using statistical approaches [3]. Nguyen Thi Thuy Linh National University of Vietnam in 2006 proposed a method of classification of web content regardless of languages [1]. The method enables the integration of new languages into the classifier and solves the problem of exploded features through the machine learning approach using maximized Entropy to construct the layered model and optimization of a function with many variables. Recently there has been a work on the classification of text content on some Vietnamese web pages using the SVM [2].

In this paper, we investigate the problem of automatically classifying the telephone subscriber incident causes from Vietnamese text messages by Support Vector Machine, a supervised machine learning method. The composition of the paper is as follows. In addition to Part I, Part II describes the classification of the cause of subscriber incidents from Vietnamese text messages using the SVM. Part III shows the testing and evaluation results. Finally we will give conclusion of the article and mention some future works.

II. CLASSIFICATION OF TELEPHONE SUBSCRIBE INCIDENTS FROM VIETNAMESE TEXT MESSAGES USING SVM

We used the support vector machine (SVM) to perform the training of the classifier for the problem of defect classification into six different types of categogies (labels). This is a multi-layer classification problem. The idea of a multilayered class problem is to convert it to a two-class subclass problem by constructing two classified solver.

These common multi-layered classification strategies are: One-against-One (OAO), and One-against-Rest (OAR).

In the OAR strategy, we will use the K-1 binary classifier to construct the K-class. The K class classifier is transformed into a K-1 two layer classifier. The second *i* layer is built on the *i-th* class and all the other classes. The *i-th* decision function is used to class *i* and the remaining classes are of the form:

$$y_i(x) = w_i^T(x) + b_i$$

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

The hyperplane $y_i(x) = 0$ forms the optimal partition superposition, the support vector of class *i* satisfies $y_i(x) = 1$ and the support vector of the other class satisfy $y_i(x) = -1$. If the data vector *x* satisfies the condition $y_i(x)>0$ for only one *i*, *x* will be assigned to *i-th* layer.

The OAO strategy, using K (K-1) / 2 binary classifiers, the model was constructed by pairing two classes. Thus, this strategy was also called pairwise and used as the majority method combination between these class members to determine the final classification result. The number of classifiers never exceeds K (K-1) / 2.

Compared to the OAR strategy, this strategy has the advantages of reducing the unclassified area. It also increases the accuracy of classification. The OAR strategy requires only K-1 classifiers for the K classes, while the OAO strategy requires K (K-1) / 2 classifiers. However, the number of training samples for each classifier in OAO is lower and classification model is simpler. Thus, the OAO strategy is more accurate but the cost of rebuilding the model is equivalent to that of OAR strategy. The class decisive function of class *i* for class *j* in the OAO method is:

$$y_{ij}(x) = w_{ij}^T(x) + b_{ij}$$

However, both approaches lead to ambiguous areas in the subclass (see Figure 1). We can avoid this problem by constructing a K-based linear function K of the form $y_k(x) = w_k^T x + b_{k0}$. And one point x is assigned to class C_k when khi $y_k(x) > y_j(x)$ for every $j \neq k$

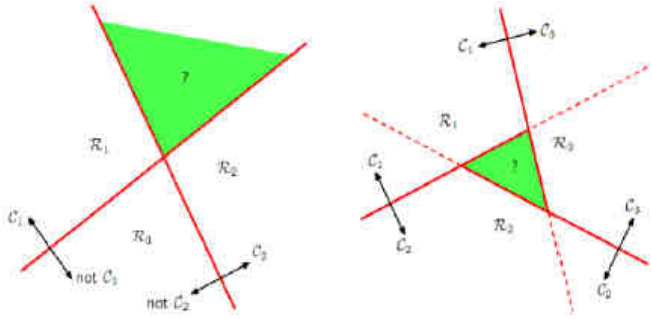

Figure 1: The Vague Region in Subclass

Specific steps in the training of the classifier:

Step1: Allows the user to enter a message or select a list of messages from the database.

Step 2: Separates words (using Vn Tokenize) and removes stop words from the text. Then the words are vector zed in the input format of the SVM algorithm.

Step3: Classify (in categories) and print the results to the screen or save the results to the database;

Express each error message in the form of a vector as follows:

<class$_i$><label$_1$>:<value$_1$><label$_2$>:<value$_2$>...<label$_n$>:<value$_n$>

where:

Class$_i$ is the classification label of each topic with i = 1 ÷ 6 (see Table 2).

Label$_j$ is the index of the j[th] feature word in the word feature space from the one that appears in the training error message with j = 1 ÷ 6.

Value$_j$ is the weight of *index$_j$* calculated by the TF.IDF formula, if value$_j$ = 0, then no such attribute is required. This format conforms to the input format of the SVM[Multiclass].

Step 4: Training the classification model based on the multilayer SVM algorithm using the OAO strategy with optimal model parameters (empirical and using some methods such as Gris Seach, Genetics, etc)

III. EXPERIMENTS AND EVALUATION

3.1. Pre-Processing of Data and Training Classification Model- Raw Training Data

The data is read from the database of the telephone subscriber's failure message system (2140 records corresponding to 2140 text messages).

- Word Separation

After "cleaning" the data by removing special chracters, numbers, date data, etc., data is extracted into tokens using the Vn Tokenizer tool [9].

- Remove the Word Stop

After the tokenization step, the stop word which is the word that appears frequently in the sentence and does not help in distinguishing the text content is removed. For example, the word "and", "then", "was" etc. usually appear in most of the text. In Vietnamese, there are about 1000 words in the word stop [9], table 1 is an example of some stop words appearing in the training data.

Table 1. The Frequent Stop Words in the Dataset

because	some	now on	go	out
use	myself	this	there	top
only	one	more	true	together
already	some	so	ok	best
segment	each	sometime	the	certain
while	new	then	what	court
head	want	occasion	hours	both
here	any	to be	between	that
for	now on	the set	should	each
lost	this	until	right	all

- Data Representation:

After word separation, the words will be weighted according to the equation (1) and then sorted in descending order of weight to make the selection of the feature words for the feature vectors representation of the text message. Each text message in the dataset is represented by an n-dimensional vector, each corresponding to a feature word. We select n dimensions corresponding to n features which have the highest weight. The expression of the error messages is described in step 3 above, while

The value *i* of the *i[th]* word in the text message vector representation *j* and the weight of the word is calculated by the formula (1)

$$\text{weight}(i,j)=\begin{cases}(1+\log(tf_{ij}))\log\left(\frac{N}{df_i}\right) & \text{If } tf_{ij} \geq 1\\ 0 & \text{If } tf_{ii} = 0\end{cases} \quad (1)$$

where:

dfi< cfi and $\sum_j tf_{ij} = cfi$

tfij (Term frequency): The number of occurrences of the word w_i in the text message d_j

dfi (Document frequency): the number of text messages containing the word w_i

cfi (Collection frequency): The number of occurrences of the word w_i in the entire dataset

If value $value_J = 0$ then no such feature is required

The data set for training and testing have a total of 2524 messages in which 2140 messages will be included in the training data set, while the remaining 384 will be the test data set. Table 2 lists the number of training and test data sets for each topic.

Table 2. Number of Training and Test Data Sets

The Incident Label	Training Dataset	Testing Dataset
Lease line	530	94
Peripheral networks	258	46
Telephone system controller, transmission	186	36
Terminals	526	93
Customers	350	63
objective reasons	290	52

3.2 Testing, Classification and Evaluation

The input of the model is text messages describing the telephone fault incidents from the wire worker, which will also be performed by separating the word, removing the stop word. After that it is represented by feature vector as described and used as input for SVM. The output of the classification is the labeled sentences; in case the model could not classify the cause then the output is labeled "Objective".

Experimental results were evaluated by means of F (F measure), which was determined by three indicators: Prec (precision), Rec (recall) and Fscore

The incident labels in these experiments use 6 basic labels e = {subscription; peripheral networks; telephone system controller, transmission; terminals; customers; objective}.

$$\text{Prec} = \frac{\text{Input}_i \cap \text{Output}_i}{\text{Output}_i} \quad (2)$$

$$\text{Rec} = \frac{\text{Input}_i \cap \text{Output}_i}{\text{Input}_i} \quad (3)$$

Where:

Input i: the number of input text messages for label *i*

Output i: the number of output text messages for label *i*

To automatically classify the errors described in the text messages, the author of the application automatically classifies the application with six main labels as above, with the 6 classes classification model was trained with the SVM algorithm.

The results of the SVM classifier evaluation on the training and test data sets are shown in Table 3 with the accuracy evaluated using the equations 2 and 3.

Table 3. Results of the Classification by Categorized Label

Number of training set / Label	2381		
	Prec	Rec	F-Score
Lease line	82,35	96,67	89,36
Peripheral networks	87,5	97,76	92,31
Telephone system controller, transmission	89,36	97,67	93,33
Terminals	89,13	95,35	92,13
Customers	70,73	67,44	69,05
Objective reasons	66.27	72.09	66.06

IV. CONCLUSION

In this paper, we have described an automated method of classification of telephone failures in Vietnamese text messages using Support Vector Machine (SVM). Testing evaluation results on a set of data obtained from the actual text messages which the phone line workers sent to the service center shows that the proposed classification method with SVM has the classification result with high accuracy in Table 3. In order to be able to apply to the systems in practice, it is necessary to build the model using training data sets which have more data. Subsequent research by the authors will aim to improve this classification method to increase the accuracy of classification and to put into practice.

REFERENCES

1. Nguyễn Thị Thùy Linh, The classification of Web pages regardless of the languages, undergraduate thesis of Hanoi National University of Technology, 2006
2. Phan Thị Hà, Hà Hải Nam: Automatic main text extraction from web pages – Journal of Science and Technology, Vietnam Academy of Science, vol. 51, no. 1, 2013.
3. Tran Vu Pham, "Dynamic Profile Representation and Matching in Distributed Scientific Networks", Journal of Science and Technology Development, Vol. 14, No. K2, 2011.
4. B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," Proceedings of the Conference on Empirical methods in natural language processing, 2002.
5. J. Martineau and T. Finin, "Delta tfidf: An improved feature space for sentiment analysis," in Proceedings of the AAAI International Conference on Weblogs and Social Media, 2009.
6. S. Aman and S. Szpakowicz, "Using roget's thesaurus for fine-grained emotion recognition," in Proceedings of the Third International Joint Conference on Natural Language Processing, 2008, pp. 296-302.
7. S. M. Kim, A. Valitutti, and R. A. Calvo, "Evaluation of unsupervised emotion models to textual affect recognition," Proceedings of the NAACL HLT Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, 2010, pp. 62-70.
8. Z. Kozareva, B. Navarro, S. V¡zquez, and A. Montoyo, "Uazbsa: a headline emotion classification through web information," in Proceedings of the 4th International Workshop on Semantic Evaluations, 2007, pp. 334-337.
9. https://www.openhub.net/p/vntokenizer ,2018.
10. http://www.cs.cornell.edu/people/tj/svm_light/svm_multiclass.html, 2018.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

11. https://github.com/stopwords/vietnamese-stopwords/blob/master/vietnamese-stopwords.txt,2017

Dr. Phan Thi Ha is currently a lecturer at the Faculty of Information Technology at Posts and Telecommunications Institute of Technology in Vietnam. She received a B.Sc.in Math & Informatics, a M.Sc. in Mathematic Guarantee for Computer Systems and a PhD. in Information Systems in 1994, 2000 and 2013, respectively. Her research interests include machine learning, natural language processing and mathematics applications. Email: hathiphan@yahoo.com, hapt@ptit.edu.vn

Dr. Phuong Nguyen is currently a lecturer and a researcher at the Faculty of Information Technology at Posts and Telecommunications Institute of Technology in Vietnam. She received a B.Sc. in Information Techology, a M.Sc. in Computer Science and a PhD. in Computer Science from University Of Maryland Baltimore County in 1998, 2008 and 2012, respectively. She is interested in parallel and distributed systems, deep learning, and natural language processing. Email: phuong3@umbc.edu, phuongnt@ptit.edu.vn