Vietnamese News Classification based on BoW with Keywords Extraction and Neural Network

Toan Pham Van
Framgia Inc. R&D Group
13F Keangnam Landmark 72 Tower
Plot E6, Pham Hung, Nam Tu Liem, Ha Noi
pham.van.toan@framgia.com

Ta Minh Thanh
Dept. of Network Technology
Le Quy Don Technical University
236 Hoang Quoc Viet, Cau Giay, Ha Noi
thanhtm@mta.edu.vn

***Abstract*—Nowadays, text classification (TC) becomes the main applications of NLP (natural language processing). Actually, we have a lot of researches in classifying text documents, such as Random Forest, Support Vector Machines and Naive Bayes. However, most of them are applied for English documents. Therefore, the text classification researches on Vietnamese still are limited. By using a Vietnamese news corpus, we propose some methods to solve Vietnamese news classification problems. By employing the Bag of Words (BoW) with keywords extraction and Neural Network approaches, we trained a machine learning model that could achieve an average of $\approx$ 99.75% accuracy. We also analyzed the merit and demerit of each method in order to find out the best one to solve the text classification in Vietnamese news.**

***Keywords*—Vietnamese Keywords Extraction, Vietnamese News Categorization, Text Classification, Neural Network, SVM, Random Forest, Natural Language Processing.**

I. Introduction

Text classification (text categorization in other researches) is a machine learning classification problem with labeling a text document with categories from the predefined sets. For example, we have a dataset of the news denoted is:

$$N = (n_1, .., n_n)$$

documents are already labelled with a pool of categories ***C*** is:

$$C = (c_1, .., c_m)$$

and we will build a system to automatically label each incoming news story with a topic in ***C***. Nowadays, with the availability of more powerful hardware, many machine learning architecture is easily implemented and ***TC*** becomes a very important subfield of the natural language processing systems. Today, ***TC*** is also applied on various information systems such as chatbot [1], content-based recommendation [2], article auto-tagging (e.g) and build a news categorizer as the problem of this paper.

In this paper, we have applied a few popular algorithms multilabel classification for Vietnamese text classification such as Naive Bayes, Random Forest, multiclass SVM (e.g) and compared with accuracy with our custom Neural Network. This is the first time that these techniques have been used in the Vietnamese text classification problem. We have researched the similar problem, but for English. We recognize that two languages have many different points in processing. The most obvious point of difference between Vietnamese and English is the word boundaries identification. Unlike English, Vietnamese word boundaries are not always is a space character. Vietnamese words in the articles include several kind of words such as *single words, compound words, duplicative words*. Sometimes, it also includes the *fortuitous concurrence words* [3]. The words in Vietnamese are normally created by the special linguistic units called ***morpho-syllables***. Each unit can be a morpheme or a word or neither of them [4] and the process of recognizing these units is called *word segmentation*. The word segmentation problem in Vietnamese sentence is demonstrated in Fig. 1.

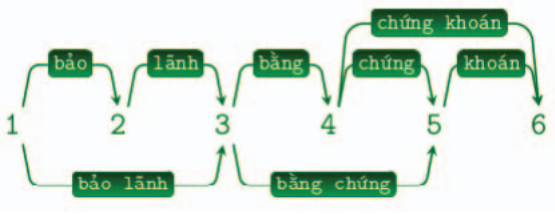


Fig. 1: *The ambiguous in Vietnamese word segmentation*

As shown in Fig. 1, there are more than one way to understanding this sentence corresponding to two words segmentation results:

1) "Bảo_lãnh bằng chứng_khoán" and this mean is "***Guarantee by stock***".
2) "Bảo_lãnh bằng_chứng khoán" is a meaningless sentence.

The segmentation is an important step in text preprocessing. If the words of sentence is segmented in the first way, we can classify the sentence in "*finance market*". However, if the words of sentence is segmented in the second way, we can classify the sentence to other category. Since our approach is based on keywords extraction, the accuracy of words segmentation progress is very important. The failure in words segmentation synonymous with low accuracy of keywords extraction leads to wrong classification.

After the keywords extraction phase, we can make a dictionary of keywords. We use it to train new model for Vietnamese text classification.

Our paper is organized as follows: we present the related works in Section 2. Section 3 then presents some Machine

978-1-5386-0743-5/17/$31.00 ©2017 IEEE

Learning methods for Text Classification in Vietnamese news data. Section 4 shows our experimental results. Section 5 presents our conclusions. The future works are given in the end of our paper.

II. Related works

A. Text Classification

TC is the method of assigning the documents to one or more predefined categories or classes. It is not a new research topic. Actually, as early as since 1800s, Knowledge Engineering techniques have been used to make the automatic document classifiers. Those are used to classify their manual construction. Nowadays, when Machine Learning (***ML***) becomes a trending vision, ***ML*** methods are used for classification purposes in various variety of domains. Of course, can be applied to solve ***TC***. In ***ML*** we can consider this problem with a multiclass classification problem. Basically, the automatic text classification methods use a predefined corpus for training and learning. We extract some kind of features for each of the text categories in the corpus. Then we apply a mathematical model, a classifier, which somehow estimates the similarities between different texts based on their features, and guesses this category. We have some methods to approach this topic. Some methods can be directly applied to classify the news article as long as there is a good training corpus [5, 6]. Most of them are implementations of *Naive Bayes* (***NB***) [9], *Support Vector Machine* (***SVM***) [8] and Convolutional Neural Network (***CNN***) [10] which are state-of-the-art for English processing. The ***TC*** process simulation is shown in Fig. 2 below:

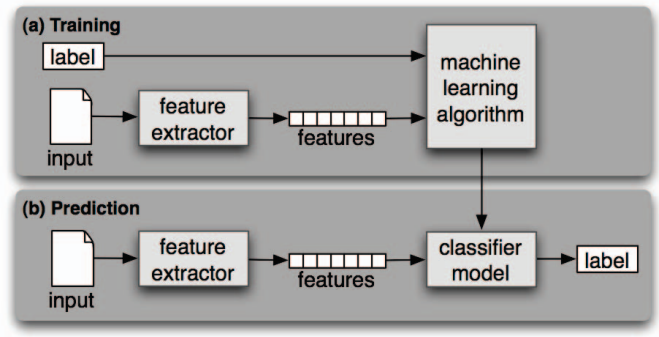


Fig. 2: *Text Classification Process*

B. Vietnamese Corpus

Some researches in English, ***TC*** has achieved satisfactory classifications by using some standard corpora such as Reuters and 20 Newsgroups. Their accuracy ranges from 80% to 93%[12]. However, the Vietnamese datasets are very restricted and small. Each topic only includes from 50 to 100 files of news article. Also, they are not available publicly for independent researches [13]. Fortunately, the research of Vu Cong Duy and colleagues [11] had created a Vietnamese corpus. That can satisfy the conditions of sufficiency, objectiveness, balance. We use their corpus for research in this paper.

Their corpus is constructed by using four well-known Vietnamese online newspapers such as VnExpress[1], Tuoi Tre Online[2], Thanh Nien Online[3], Nguoi Lao Dong Online[4]. The crawled online documents are automatically altered ($e.g$ removing the HTML tags, spelling normalization) by Teleport[5]. Afterwards, they are manually corrected by linguists. The linguists had reviewed, then edited the text documents which are classified to the wrong predefined topics. Then, they obtain a relatively large and sufficient corpus including top categories. The training documents of Level 1 of this corpus contains about 33,759 documents. Also, the testing documents of that are about 50,373 documents. Two parts of the dataset are shown in Fig. 3 and Fig. 4.

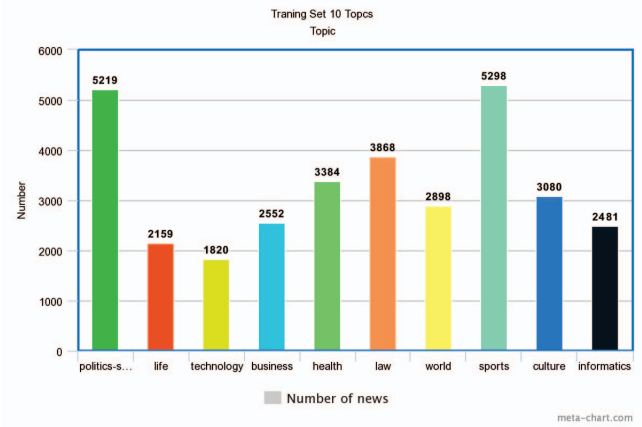


Fig. 3: *Training set in 10 topic corpus*

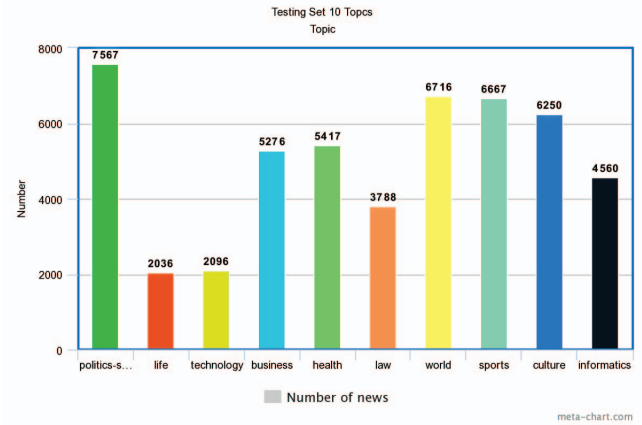


Fig. 4: *Testing set in 10 topic corpus*

Another variant of this dataset has 27 topics. Those topics are child topics of the corpus above. The division in Level 1 is very vague meanwhile and they need to find a specific topics to employ for TC [11]. We used both two levels of dataset in this paper to train the Vietnamese text classifier.

C. Keyword Extraction

Keyword extraction is an important technique for document retrieval, text classification, document clustering, text sum-

[1] www.vnexpress.net

[2] www.tuoitre.com.vn

[3] www.thanhnien.com.vn

[4] www.nld.com.vn

[5] https://teleport-pro.en.softonic.com

marization, etc. By extracting appropriate keywords, we can choose easily the appropriate document to read or learn the relation among documents and we can apply this method for reducing range of input space in our ***TC*** problem. Extracting the keywords set in a text document means finding unique words. That means those words have no duplication meaning and also are not included in the stop words list and keywords. Their frequency is ordered by descending weight. This paper takes ten top of keywords to calculate *Keyword Score* as follows:

$$KeywordScore(k_i) = 1.5\frac{|k_i|}{|W|}, \quad (1)$$

where $|k_i|$ is the frequency of keywords occurrence in an article, $|W|$ is the number of unique keywords. We extract top keywords of an article after each keyword has got its score and *build a dictionary* of keywords in all documents of our corpus.

D. Feature Selection

1) Bag of Words approach: Before any classification task, the important task is to represent document and to select the feature. Actually, we have two ways to represent text documents. Those are *Bag of Words (BoW)* and representation of text as strings. Most of TC methods use the ***BoW*** approach because of its simplicity for classification problems. In ***BoW*** method, a text document is described as a set of words with their associated frequency. This representation of BoW is essentially independent of words sequence in the collection. Words are made of one or more morpho-syllable.

2) Word Segmentation: In order to efficiently segment the words, a robust method for document classification requires a robust word segmentation method in Vietnamese. We use ***vnTokenizer*** [15] - the quire new word segmentation program in the ***BoW*** approach. ***vnTokenizer*** is used to segment the text documents into words or tokens before creating a dictionary for preprocessing.

3) Stop-words Removal: The most common feature selection is ***stop-words*** removal and stemming. In the stop-words removal, we determine the common words that are not specific or discriminatory to the different classes. Defined words (e.g., "và", "bị" and "chính là") are ignored in text processing. For this purpose, we prepared a stop-words list (about ≈2000 words, collected manually).

III. Text classification methods

After text preprocessing above, we have numeric training features from the BoW and the original categories for each feature vector. We can apply some supervised learning algorithms to solve the text classification. In the proposed method, we propose some multiclass classification algorithms and compared with our Neural Network Architecture. Some methods as ***Random Forest***, ***Support Vector Machine*** will be represented below.

A. Random Forest

Random Forest (RF) is a famous algorithm for classification in Machine Learning. A *RF* is a classifier consisting of a collection of tree-structured classifier $\{RF(x, \theta_k), k = 1, ...\}$, where θ_k are independent identically distributed random vectors. Each tree casts a unit vote for the most popular class at input vector x [16]. This classifier employs the averaging value to improve the prediction accuracy and to control the over-fitting. Actually, **RF** is a meta estimator. The function RF need to fit a number of decision tree classifiers by using various sub-samples of the experimental dataset.

With the classification problems, suppose that a set of simple trees and a set of random predictor variables are given, a margin function defined by the **RF** method measures the extent to which the average number of votes for the correct class exceeds the average vote for any other class present in the dependent variable. Given a set of classifier denoted with

$$RF_1(x), RF_2(x), ..., RF_k(x)$$

the features vector **X** and the labels vector **y**. The margin function *M* is defined as:

$$M(X, y) = avI(h_k(X) = y) - max_{j \neq y} avI(h_k(X) = j),$$

where I denotes indicator function. This measure provides the researchers not only efficient way to make their predictions, but also a way to associate a confidence measure of works with those predictions.

The predictions defined by the ***RF*** are taken to be the average value of the predictions of the trees defined as follows.

$$s = \frac{1}{N}\sum_{i=1}^{N} s_i,$$

where s_i is the prediction of tree i. The index i runs over the individual trees in the forest.

The missing data in the predictor variables can be flexibly incorporated by *RF*. When missing data are encountered during model building, the prediction made for the mentioned case is based on the last preceding node in the respective trees.

B. SVM

Support Vector Machines (SVMs) were proposed first time in [17, 18] for numerical data. In this method, determining the optimal hyperplane which can separate more efficiently the different classes is the main principle of SVMs. For instance, the example of SVM is shown in Fig. 5.

In Fig. 5, there are two classes described by '**x**' and '**o**'. Three different optimal hyperplanes are denoted by **A**, **B,** and **C**, respectively. According to Fig. 5, the hyperplane **A** gives the best separation of solution between the different classes. The reason is that the normal distance of any data points in the classes from that hyperplane is the largest. Therefore, the maximum margin of separation can be archived by using the hyperplane **A**. Note that the normal vector to the hyperplane is a direction in the feature space along.

In this problem, numbers of classes is more than two - *multiclass problem*. Assume that we have a set of *m* training example $S = \{(x_1, y_1), .., (x_m, y_m)\}$, where x_i is the feature

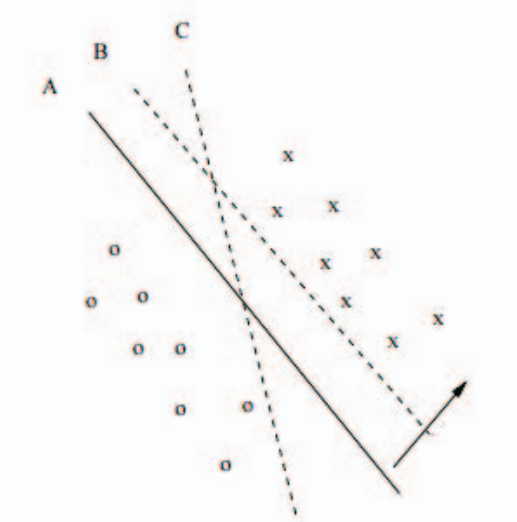


Fig. 5: *Best separating hyperplane selection in SVM*

vector i^{th} and y_i is respective label. Suppose that each example x_i is drawn from a domain $X \subseteq R^n$. Each label y_i is a value from the set $Y = \{1, .., k\}$, where k is the numbers of classes. A ***multiclass*** classifier function called as $H : X \rightarrow Y$. The function H needs to map an instance vector x - in this problem is the *BoW feature vector* - to a label y_i in Y [19]:

$$H_M(x) = arg\max_{1}^{k}(M_r.x),$$

where M is the matrix. The size of M is $k \times n$ over R. M_r is r^{th} of M. The value of the inner-product of the r^{th} row of M with the instance x the *confidence* and the *similarity score* for the r class are interchangeably called. The index of the row achieving the highest similarity score with x is assigned to the predicted label y. Since our problem have $k \geq 3$, therefore k prototypes called $M_1, \cdots, M_k$ are maintained. The labels of a new input layer are chosen by using the index of the most similar row of M. They are set as the inputs of next layers of our neural network.

C. Neural Network (NN)

The basic idea of NN is a *neuron* in which that receives a set of inputs data. That can be denoted by vector $\overline{X}_i$. $\overline{X}_i$ correspond to the *BoW feature vector* in the i^{th} text document. A set of weights, denoted by *W*, is used to express the association of each neuron. Afterwards, each neuron is computed the function *f(·)* of its inputs. The sign of the predicted function p_i yields the class label of vector $\overline{X}_i$. In general, a typical function used in the NN is the *linear function*. That can be described by,

$$p_i = W.\overline{X}_i$$

For the multi-class problem, the neural network is a little bit different. Call d is the length of dictionary after *BoW* preprocessing. Suppose that a d-dimensional feature X, the training dataset denoted by X_{tr} and each feature vector $\overline{x} \in X_{tr}$ is associated with a class y_i of the label Y. In our paper, X is *BoW* of text document, Y is categories of document. In order to classify the category, the neural network F is trained on S_{tr}. That means each feature vector $\overline{x} \in X$ then $F(\overline{x}) \in Y$. In general, F maybe a multiple neural networks or a single neural network. In our research, a multi-layered feed forward neural network is employed for classification. Suppose that the input/output at a hidden node *j* of neural network as:

$$f_j^h = \sum_i w_{ji}^h x_i,$$

with $j = 1, ..., H$, where x_i is the i^{th} input of feature vector $\overline{x}$, w_{ji}^h is the weight associated with the feature vector x_i to the j^{th} hidden node. The number of hidden node is denoted by H. We apply a activation function denoted by $g^h(.)$ in the hidden layer. Some activation functions can be used such as *Sigmoid* [21], *ReLU* [22], *Softmax* [22]... The output of the j^{th} hidden unit is denoted by $z_j = g^h(f_i h)$. Each node O_k of the output layer has the input/output as follows:

$$f_k^o = \sum_i w_{kj}^o z_j$$

Finally we have the label class associate with feature vector $\overline{x}_k$

$$y_k = g^o(j_k^o),$$

where $k = 1...M$. *M* is the number of the output nodes.

In our proposal, we create a network with 6 hidden layers with the ***tanh*** activation function [24] and used *stochastic gradient descent* [23] to optimization in this network. The input layer is the feature vector after feature seletion pharse with ***BoW*** method and the output layer is label vector of the documents. The simulation of network architecture is present in Fig. 6

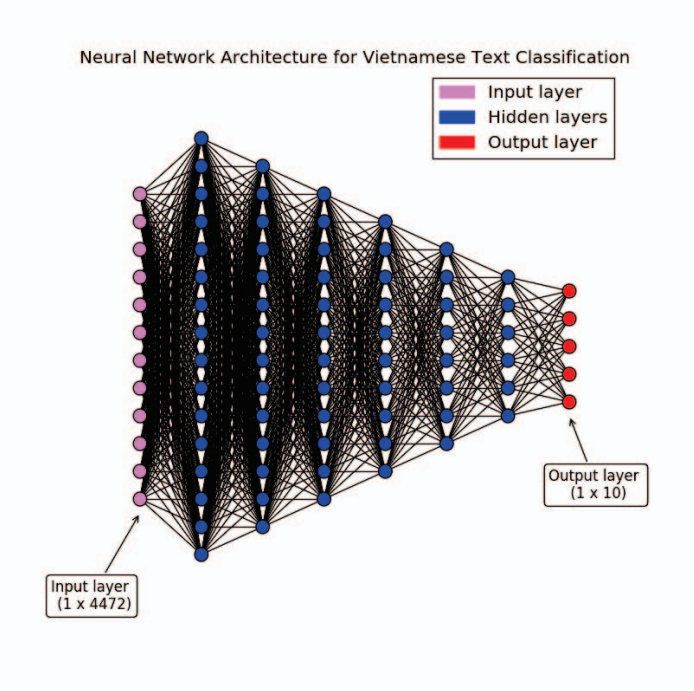


Fig. 6: *Simulation of Neural Network architecture*

IV. RESULT

We employ the recall and precision parameters for evaluating the classification models [25]. The $Recall$ is defined as below:

$$Recall = \frac{\sum_{d \in D} d_{TrueModel}}{\sum_{d \in D} d_{Practice}},$$

and the $Precision$ is computed as follows:

$$Precision = \frac{\sum_{d \in D} d_{TrueModel}}{\sum_{d \in D} d_{AllModel}},$$

In that:

- The $d_{TrueModel}$ is the number of text articles classified by the model correctly.
- The $d_{AllModel}$ is the number of text articles classified by the model.
- The $d_{Practice}$ is the number of text articles classified correctly in practice.

The F_1 score is calculated with:

$$F_1 = \frac{2.Recall.Precision}{Recall + Precision}$$

Our *Keyword extraction with BoW* method is abbreviated with ***KEBoW***. We investigate the comparison with ***N-gram*** method introduced in the research of Vu Cong Duy [11], and difference Machine Learning algorithms as ***SVMs multiclass, Random Forest, SVC***. Also, the total accuracy is needed to calculate based on the average accuracy values of all categories for each experimental dataset. The results are presented in Fig. 7, Fig. 8, and Fig. 9.

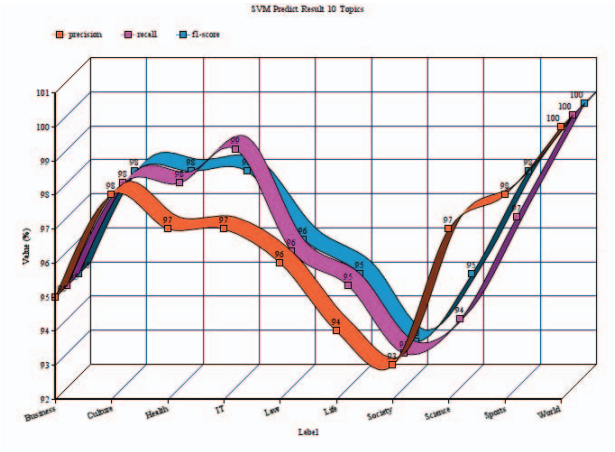


Fig. 7: *Prediction Result with SVM 10 Topics dataset*

The Fig. 7 Fig. 8 and Fig. 9 showed that the prediction result of respective algorithms with ***KEBoW*** extraction method with ***10 Topics dataset***. We can see the best result of other research [11] with current dataset in Fig. 10. Easy to see that our prediction result better than the result in Fig. 10 in the same dataset. It proves that our feature selection method with keywords extraction and BoW have more effective other features selection methods.

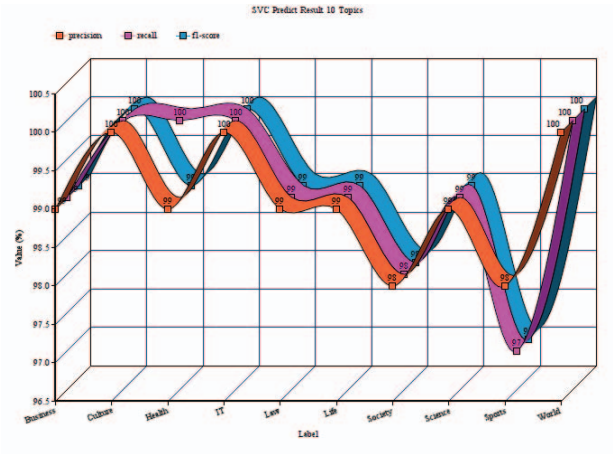


Fig. 8: *Prediction Result with SVC 10 Topics dataset*

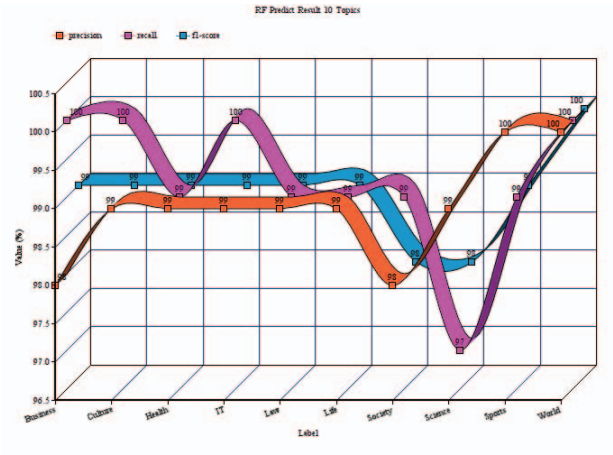


Fig. 9: *Prediction Result with Random Forest 10 Topics dataset*

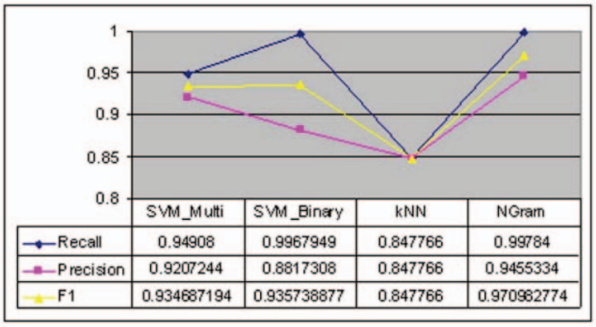
However, we not only improved the features selection pharse with ***KEBoW*** method, but also we proposed a Neural Network applied in the training pharse. The comparison of our Neural Network accuracy with some algorithms is shown in Table I.

V. CONCLUSION AND FUTURE WORKS

With the difference between Vietnamese and other languages, the research to find a feasible approach for Vietnamese text classification is the new challenge of us. By our experiments, we proposed new neural network architecture with average accuracy 99.75%. This result is much better than some methods as ***SVM***, ***Random Forest*** in the same dataset. Especially our result achieved is better than the research of Vu Cong Duy [11] with the same algorithm in the same dataset. It proves that our feature selection method with keywords extraction and

TABLE I: *Accuracy Comparsion Result*

	SVM	*Random Forest*	*SVC*	*Neural Network*
10 Topics Dataset	0.9652	0.9921	0.9922	0.9975
27 Topics Dataset	0.9780	0.9925	0.9965	0.9969



	SVM_Multi	SVM_Binary	kNN	NGram
Recall	0.94908	0.9967949	0.847766	0.99784
Precision	0.9207244	0.8817308	0.847766	0.9455334
F1	0.934687194	0.935738877	0.847766	0.970982774

Fig. 10: *Best prediction result in other paper [11]*

BoW have more effective other features selection methods.

Additionally, there are several errors occur in these approaches for Vietnamese TC. Those can be described as follows:

1) The stop-words list is builded from subjective views and it maybe not have high accuracy
2) The corpus have the ambiguities between two or many topics.
3) The segmentation is limited by third-party library.

In the future, we could improve the accuracy of our Neural Network, overcome the disvantages of preprocessing pharse and combine more semantic and contextual features in this text classification problem for Vietnamse.

Application of Research

The research result was applied in Viblo [6] - a free service for technical knowledge sharing of ***Framgia Inc.***[7] - to automatic classification the post when user publish it.

Acknowledgment

This research was partially supported by ***Framgia Inc.***.

References

[1] B. Alexander, S. Thorsen, "A sentiment-based chat bot." (2013).

[2] Mooney, J. Raymond, Roy. Loriene, "Content-based book recommending using learning for text categorization," Proc. of the 5th ACM conference on Digital libraries, ACM, 2000.

[3] D. Dinh, V. Thuy, "A maximum entropy approach for Vietnamese word segmentation." Research, Innovation and Vision for the Future, International Conference on IEEE, 2006.

[4] D. Dien, H. Kiem, N.V.Toan, "Vietnamese Word Segmentation" Proc. of the 6th Natural Language Processing Pacific Rim Symposium, Tokyo, Japan, pp.749-756, 2001.

[5] Y. Yang and X. Liu. A re-examination of text categorization methods. In 22nd Annual International SIGIR, pp. 42–49, Berkley, August 1999.

[6] F. Sebastiani. Machine learning in automated text categorisation: a survey. Technical Report IEI-B4-31-1999, Istituto di Elaborazione dell'Informazione, Consiglio Nazionale delle Ricerche, 1999.

[7] Yang, Y. 1994. Expert network: effective and efficient learning from human decisions in text categorization and retrieval. In Proceedings of SIGIR-94, 17th ACM International Conference on Research and Development in Information Retrieval (Dublin, IE, 1994), pp. 13–22.

[8] Thorsten Joachims. "Text Categorization with Support Vector Machines: Learning with Many Relevant Features." Proc. of ECML-98, 10th European Conference on Machine Learning, No. 1398, pp. 137–142.

[9] Shimodaira, Hiroshi. "Text classification using naive Bayes." Learning and Data Note 7 (2014): 1-9.

[10] Z. Xiang, J. Zhao, Y. LeCun, "Character-level convolutional networks for text classification." Advances in neural information processing systems. 2015.

[11] H. V. C. Duy, et al. "A comparative study on vietnamese text classification methods," International Conf. on Research, Innovation and Vision for the Future, 2007.

[12] S. Fabrizio. "Machine learning in automated text categorization." ACM computing surveys (CSUR), no. 34, vol. 1, pp. 1–47, 2002.

[13] Hung Nguyen, Ha Nguyen, Thuc Vu, Nghia Tran, and Kiem Hoang. 2005. Internet and Genetics Algorithm-based Text Categorization for Documents in Vietnamese. Proceedings of 4th IEEE International Conference on Computer Science - Research, Innovation and Vision of the Future, 2006.

[14] D. Gunawan, et al. "Automatic Text Summarization for Indonesian Language Using TextTeaser." IOP Conf. Series: Materials Science and Engineering, vol. 190. no. 1, 2017.

[15] L. N. Minh, et al. "VNLP: an open source framework for Vietnamese natural language processing." Proc. of the Fourth Symposium on Information and Communication Technology, 2013.

[16] L. Breiman, "Random forests." UC Berkeley TR567, 1999.

[17] V. Vapnik, "Estimations of dependencies based on statistical data," Springer, 1982.

[18] C. Cortes, V. Vapnik, "Support-vector networks. Machine Learning," 20: pp. 273–297, 1995.

[19] C.Koby, Y. Singer, "On the algorithmic implementation of multiclass kernel-based vector machines." J. of machine learning research, pp. 265–292, 2001.

[20] O. Guobin, Y. L. Murphey, "Multi-class pattern classification using neural networks." Pattern Recognition, vol. 40, no. 1, pp. 4–18, 2007.

[21] Yin, Xinyou, et al. "A flexible sigmoid function of determinate growth," Annals of botany, vol. 91, no. 3, pp. 361–371, 2003.

[22] G. Xavier, A. Bordes, Y. Bengio, "Deep sparse rectifier neural networks." Proc. of the Fourteenth International Conference on Artificial Intelligence and Statistics, 2011.

[23] B. Léon. "Large-scale machine learning with stochastic gradient descent." Proc of COMPSTAT'2010, pp. 177–186, 2010.

[24] K. Bekir, A. V. Olgac. "Performance analysis of various activation functions in generalized MLP architectures of neural networks." International J. of Artificial Intelligence and Expert Systems, vol. 1, no. 4, pp. 111-122, 2011.

[25] F. Sebastiani, "Machine Learning in Automated Text Categorization," ACM Computing Surveys, vol. 34, no. 1, pp.1–47, 2002.

[26] A. M. Salih, et al. "Modified extraction 2-thiobarbituric acid method for measuring lipid oxidation in poultry." Poultry Science, vol. 66, no. 9, pp. 1483–1488, 1987.

[6] www.viblo.asia

[7] www.recruit.framgia.vn