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

Toan Pham Van
Framgia Vietnam Viblo R&D Team
pham.van.toan@framgia.com

Thanh Ta Minh
Le Quy Don Technical University
ta.minh.thanh@framgia.com

***Abstract*—Text classification (TC) is one of the main applications of natural language processing. Actually, we have a lot of researches in classifying text documents, such as Random Forest, Support Vector Machine, Naive Bayes. However, most of them are applied for English documents. Therefore, the text classification researches on Vietnamese still are limited. Based on a Vietnamese news corpus, we propose some methods to solve Vietnamese news classification problems. By using Bag of Words - BOW with keywords extraction and Neural Network approaches, we trained a machine learning model that could archive an average of ≈99.75% accuracy. Additionally, we also analyzed the advantages and disadvantages of each method to find out the best of them to solve this problem.**

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

I. Introduction

Text classification - TC (or 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 $\boldsymbol{C}$ is:

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

and we will build a system to automatically label each incoming news story with a topic in $\boldsymbol{C}$. Nowadays, with the availability of more powerful hardware, many machine learning architecture is easily implemented and ***TC*** became a major subfield of the natural language processing systems. With many advantages, ***TC*** is used in many information system 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 compare accuracy with our custom Neural Network. To the best of our knowledge 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 and we have recognized that two languages have many different points in processing. The most obvious point of difference between Vietnamese and English is the word boundary identification. Not same as English, Vietnamese word boundary are not always is a space character. Vietnamese words include *single words* , *compound words, duplicative words* and *fortuitous concurrence words* [3] and the words are usually composed of special linguistic units called ***morpho-syllable***. It's maybe a morpheme or a word or neither of them [4] and the problem to recognize it called word segmentation. For example with a Vietnamese sentence as belows:

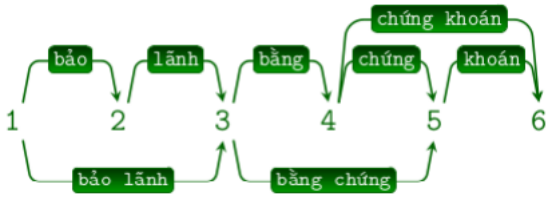

Fig. 1: *The ambiguous in Vietnamese word segmentation*

In this example, there are more than one way to understanding this sentence corresponding to two word 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 because if we segment words in the first way, we may classify it in "*finance market*". However, if we segment words in the second way, we may classify it to other category. Because our approach in this paper 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 after that.

After the keywords extraction phase, we have a dictionary of keywords. We have used it to train new model for text classification.

The rest of this paper is organized as follows: we discuss the related works in the next section. Section 3 then presents some Machine Learning methods for Text Classification in Vietnamese news data. Section 4 gives the results of experiments we conducted and Section 5 reports our conclusions and future works.

II. Related works

A. Text Classification

Text classification is the process of assigning text documents to one or more predefined categories or classes. This is not a new problem. Actually, as early as the 1800s, Knowledge Engineering (KF) techniques are used to create the automatic document classifiers in their manual construction. Nowadays, when Machine Learning (***ML***) becomes a trending vision, ***ML*** methods are used in a wide variety of domains for the purposes of classification. Of course, it's can apply to solve ***TC***. In ***ML*** we can consider this problem with a multiclass classification problem. In basically, automatic text classification uses a corpus and we extract some kind of features for each of the texts. Then we apply a mathematical model, a classifier, which somehow estimates the similarities between different texts based on the features, and guesses this category. We have some methods to approach and many of them can be directly applied to news classification as long as there exists a good training corpus [5, 6]. The most of them are *Naive Bayes* (***NB***) [9], *Support Vector Machine* (***SVM***) [8] and Convolutional Neural Network (***CNN***) [10] is a state-of-the-art for English processing. The ***TC*** process is simulate in ***Fig 2*** below:

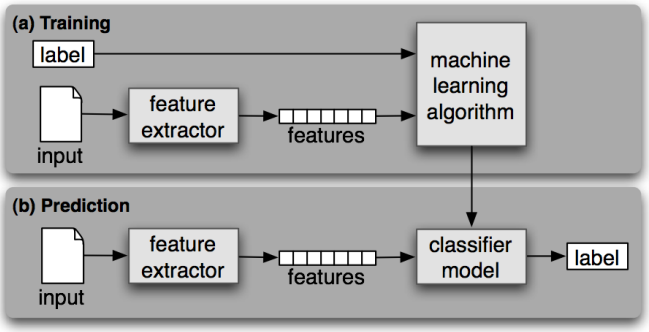

Fig. 2: *Text Classification Process*

B. Vietnamese Corpus

Some research in English ***TC*** has generally achieved satisfactory results with the results on some standard corpora such as Reuters and 20 Newsgroups ranging from 80 to 93% of accuracy [12]. However, the Vietnamese datasets are very restricted and small (from 50 to 100 files per topic) which are not available publicly for independent research [13]. Really fortunately in the research of Vu Cong Duy and colleagues [11] had constructed a Vietnamese corpus which satisfies the conditions of sufficiency, objectiveness and balance. We had used this corpus for research in current paper. Below is the detailed description of the corpus.

This corpus based on the four largest circulation Vietnamese online newspapers: VnExpress[1], TuoiTre Online[2], Thanh Nien Online[3], Nguoi Lao Dong Online[4] The collected texts are automatically preprocessed (removing the HTML tags, spelling normalization) by Teleport software and manual correction by linguists who reviewed and adjusted the documents which are classified to the wrong topics. Finally, they obtained a relatively large and sufficient corpus includes top categories. Level 1 of this corpus contains about 33,759 documents for training and 50,373 documents for testing. Two part of the dataset is shown on ***Fig 3*** and ***Fig 4***.

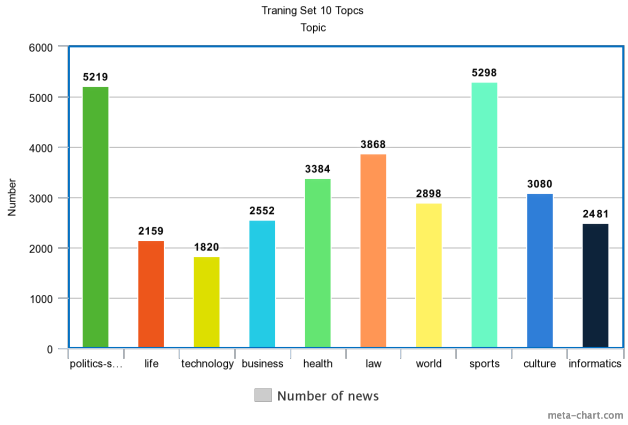

Fig. 3: *Training set in 10 topic corpus*

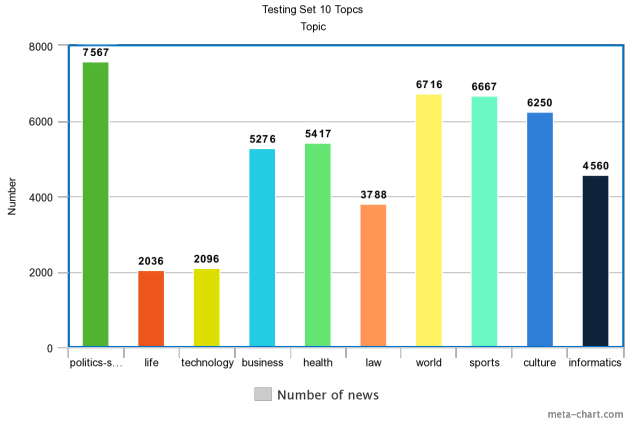

Fig. 4: *Testing set in 10 topic corpus*

Another variant of this dataset have 27 topics. The topics which are child topics of the corpus above. The division in the Level 1 is very vague meanwhile and they need to find a specific topic to experiment for TC [11]. We used both two level of dataset in this paper to training the Vietnamese text classifier.

C. Keyword Extraction

Keyword extraction is an important technique for document retrieval, text classification, document clustering, text summarization, etc. By extracting appropriate keywords, we can choose easily which document to read or learn the relation among documents and we can apply this method for reduce the demension of input space in our ***TC*** problem. Extract a keywords set in a text document is finding unique words. The unique words are the words that has no duplication and not included in stop words list and keywords and their frequency

[1] www.vnexpress.net
[2] www.tuoitre.com.vn
[3] www.thanhnien.com.vn
[4] www.nld.com.vn

are ordered by descending weight. We take ten top keywords to be calculated *Keyword Score* by using equation below:

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

Where k_i that will be calculated and $|k_i|$ is the frequency of keywords occurrence in the text, and $|W|$ is the number of unique keywords in the text. We extract top keywords of a article after each keyword has 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, a important task is that of document representation and feature selection. Actually, we have two ways can be used to represent a text document are *Bag of Words - BoW* and represent text directly as strings. Most text classification methods use the ***BoW*** representation because of its simplicity for classification purposes. In ***BoW*** method, a document is represented as a set of words, together with their associated frequency in the document. Such a representation is essentially independent of the sequence of words in the collection. Words can be made from one morpho-syllable, or many morpho-syllables.

2) Word Segmentation: So in this approach, a robust solution to document classification requires a good Vietnamese word segmentation module. We use ***vnTokenizer*** [15] - the state-of-the-art word segmentation program in the ***BoW*** approach. Text documents are segmented into words or tokens before create a dictionary in preprocessing.

3) Stop-words Removal: The most common feature selection is that of ***stop-words*** removal and stemming. In stop-words removal, we determine the common words in the documents which 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 Bag of Words and the original categories for each feature vector. We can apply some supervised learning algorithms to solve the text classification. In this paper, we consider 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 *Random Forest* 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 and each tree casts a unit vote for the most popular class at input x [16]. This classifier use averaging to improve the predictive accuracy and control overfitting. Actually, **RF** is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset.

For classification problems, given a set of simple trees and a set of random predictor variables, the **RF** method defines a margin function that 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 is the indicator function. This measure provides us not only with a convenient way of making predictions, but also with a way of associating a confidence measure with those predictions.

The predictions of the Random Forest are taken to be the average of the predictions of the trees:

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

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

RF can flexibly incorporate missing data in the predictor variables. When missing data are encountered for a particular observation during model building, the prediction made for that case is based on the last preceding node in the respective tree.

B. Support Vector Machines

Support Vector Machines - SVMs were first proposed in [17 18] for numerical data. The main principle of SVMs is to determine separators in the search space which can best separate the different classes. For example, consider the example illustrated in ***Fig 5*** below

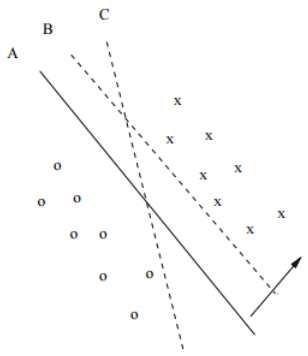

Fig. 5: *Best separating hyperplane selection in SVM*

In which we have two classes denoted by '**x**' and '**o**' respectively. We have denoted three different separating hyperplanes, which are denoted by **A**, **B,** and **C** respectively. It is evident that the hyperplane **A** provides the best separation between the different classes, because the normal distance of any of the data points from it is the largest. Therefore, the hyperplane

A represents the maximum margin of separation. We note that the normal vector to this hyperplane (represented by the arrow in the figure) is a direction in the feature space along which we have the maximum discrimination.

In this problem, numbers of classes is more than two - *multiclass problem*. Assume that we have a set of *m* training example $S = \{(x_1, y_1), .., (x_m, y_m)\}$ with x_i is the feature vector i^{th} and y_i is respective label. We assume that each example x_i is drawn from a domain $X \subseteq R^n$ and that each label y_i i is an integer from the set $Y = \{1..k\}$ with k is the numbers of classes. A ***multiclass*** classifier is a fucntion $H : X \to Y$ that maps an instance x - in this problem is the *BoW feature vector* to an label y in Y [19]:

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

where M is the matrix of size $k \times n$ over R and M_r is r^{th} of M. We interchangeably call 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. With the definition above, the predicted label is the index of the row attaining the highest similarity score with x.Our problem have $k \geq 3$ in which we maintain k prototypes $M_1, M_2, ..., M_k$ and set the label of a new input instance by choosing the index of the most similar row of M.

C. Neural Network

The basic idea of neural network is a *neuron* in which each neurol receives a set of inputs denoted by vector $\overline{X}_i$. In this case, $\overline{X}_i$ correspond to the *BoW feature vector* in the i^{th} document. Each neuron is also associated with a set of weights *W*, which are used in order to compute a function *f(·)* of its inputs. The sign of the predicted function p_i yields the class label of vector $\overline{X}_i$ A typical function which is often used in the neural network is the *linear function* as follows:

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

For the multi-class problem, the neural-network can be described formally as follows. With *d* is the lenght of dictionary after *BoW* preprocessing. For a given *d*-dimensional feature space X, the training dataset denoted by X_{tr} and each element $\overline{x} \in X_{tr}$ is associated with a class label y_i of the label Y. A neural network system F can be trained on S_{tr} such that for any given feature vector $\overline{x} \in X$ and $F(\overline{x}) \in Y$. F can be a system of neural networks or a single neural network whose weights. In this paper, we use a multi-layered feed forward neural network. We denote the input and output at a hidden node *j* 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 $\overline{x}$, and w_{ji}^h is the weight associated with the input x_i to the j^{th} hidden node. H is the number of hidden node. We applied a activation fucntion denoted by $g^h(.)$ in the hidden layer. Some of activation fucntion as *Sigmoid* [21], *ReLU* [22], *Softmax* [22]... The output value from the j^{th} hidden unit denoted by $z_j = g^h(f_i h)$. In the output layer, each node O_k has the input and output as follows:

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

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

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

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

In this paper, 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***

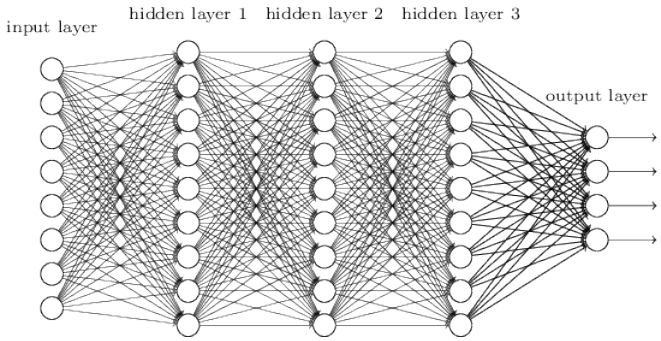

Fig. 6: *Simulation of Neural Network architecture*

IV. Result

Two recall and precision parameters are used to evaluate 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 $Precision$ as

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

In that:

- The $d_{TrueModel}$ is the number of documents classified by the model correctly.
- The $d_{AllModel}$ is the number of documents classified by the model.
- The $d_{Practice}$ is the number of documents 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 introducted in the research of Vu Cong Duy [11], and difference Machine Learning algorithms as ***SVMs multiclass,***

Random Forest, SVC. Additionally, the total accuracy is calculated from the average accuracy of all categories for each experiment. The result is present with some figures ***Fig 7, Fig 8*** and ***Fig 9***

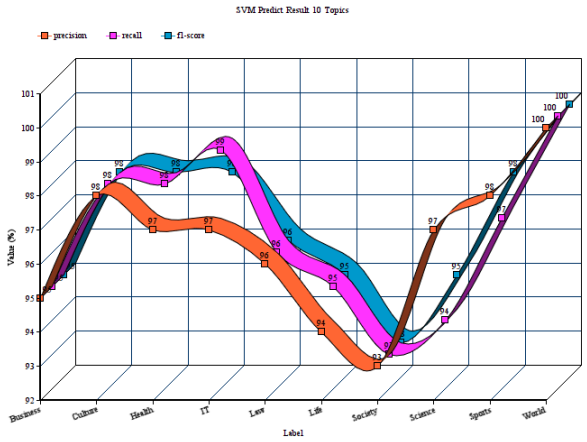

Fig. 7: *Prediction Result with SVM 10 Topics dataset*

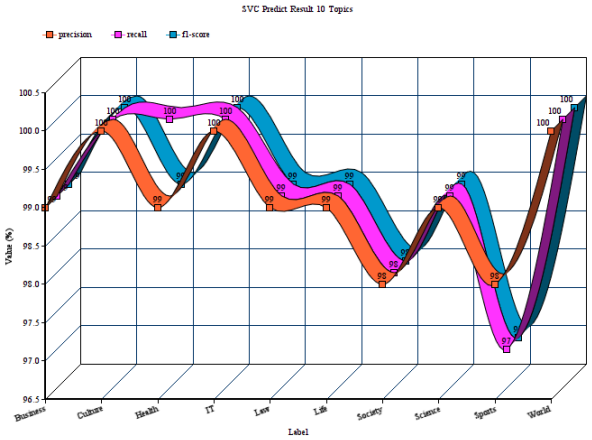

Fig. 8: *Prediction Result with SVC 10 Topics dataset*

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.

However, we are 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 1***.

V. Conclusion and Future Works

With the difference between Vietnamese and other languages, the research to find a feasible approach for Vietnamese

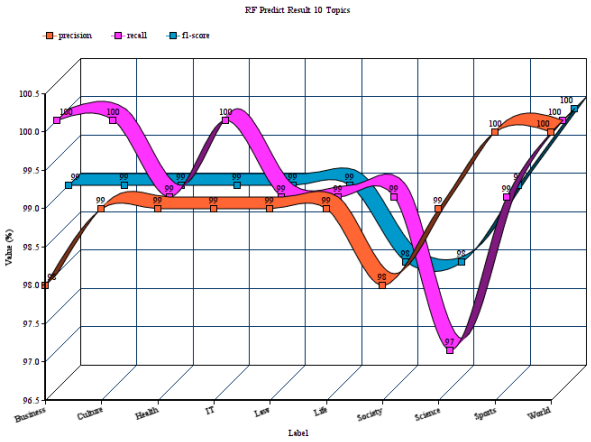

Fig. 9: *Prediction Result with Random Forest 10 Topics dataset*

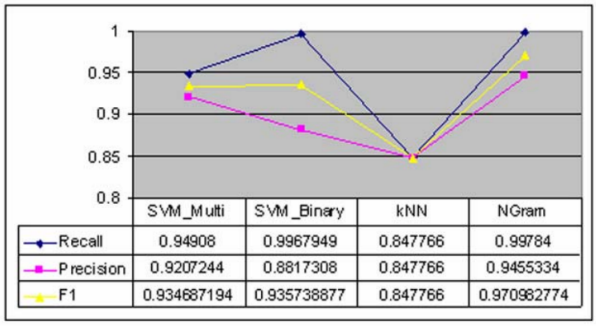

Fig. 10: *Best prediction result in other paper [11]*

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 BoW have more effective other features selection methods.

However, we also recognize that these approaches for Vietnamese TC occur some errors such as:

1) The stopwords 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.

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

APPLICATION OF RESEARCH

The research result was applied in Viblo post automatic [5] - a free service for technical knowledge sharing of ***Framgia Vietnam***[6]

ACKNOWLEDGMENT

This research was partially supported by ***Framgia Vietnam***. We are thankful to our colleagues who provided expertise that greatly assisted the research, although they may not agree with all of the interpretations provided in this paper.

REFERENCES

[1] BLOM, ALEXANDER, and SOFIE THORSEN. "A sentiment-based chat bot." (2013).

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

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

[4] D.Dien, H.Kiem, and N.V.Toan, "Vietnamese Word Segmentation". 2001. Proceedings of NLPRS'01. The 6th Natural Language Processing Pacific Rim Symposium, Tokyo, Japan, 11/2001, pp.749-756, 2001

[5] Y. Yang and X. Liu. A re-examination of text categorization methods. In 22nd Annual International SIGIR, pages 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, Pisa, IT, 1999. Revised version, 2001

[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. In C. Nedellec and C. Rouveirol, editors, Proceedings of ECML-98, 10th European Conference on Machine Learning, number 1398, pages 137—142.

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

[10] Zhang, Xiang, Junbo Zhao, and Yann LeCun. "Character-level convolutional networks for text classification." Advances in neural information processing systems. 2015.

[11] Hoang, Vu Cong Duy, et al. "A comparative study on vietnamese text classification methods." Research, Innovation and Vision for the Future, 2007 IEEE International Conference on. IEEE, 2007.

[12] Sebastiani, Fabrizio. "Machine learning in automated text categorization." ACM computing surveys (CSUR) 34.1 (2002): 1-47.

[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 (RIVF'06). Ho Chi Minh City, Vietnam , Feb 12-16, 2006

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

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

[16] Breiman, Leo. "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] Crammer, Koby, and Yoram Singer. "On the algorithmic implementation of multiclass kernel-based vector machines." Journal of machine learning research 2.Dec (2001): 265-292.

[20] Ou, Guobin, and Yi Lu Murphey. "Multi-class pattern classification using neural networks." Pattern Recognition 40.1 (2007): 4-18.

[21] Yin, Xinyou, et al. "A flexible sigmoid function of determinate growth." Annals of botany 91.3 (2003): 361-371.

[22] Glorot, Xavier, Antoine Bordes, and Yoshua Bengio. "Deep sparse rectifier neural networks." Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics. 2011.

[23] Bottou, Léon. "Large-scale machine learning with stochastic gradient descent." Proceedings of COMPSTAT'2010. Physica-Verlag HD, 2010. 177-186.

[24] Karlik, Bekir, and A. Vehbi Olgac. "Performance analysis of various activation functions in generalized MLP architectures of neural networks." International Journal of Artificial Intelligence and Expert Systems 1.4 (2011): 111-122.

[25] Fabrizio Sebastiani. 2002. Machine Learning in Automated Text Categorization. ACM Computing Surveys, Vol. 34, No. 1, March 2002, pp.1-47

[26] Salih, A. M., et al. "Modified extraction 2-thiobarbituric acid method for measuring lipid oxidation in poultry." Poultry Science 66.9 (1987): 1483-1488.

[5] www.viblo.asia

[6] www.recruit.framgia.vn