A Comparative Study on Vietnamese Text Classification Methods

Vu Cong Duy Hoang, Dien Dinh
Faculty of Information Technology
College of Natural Sciences, Vietnam National University
Ho Chi Minh City, Vietnam
hcdvu@fit.hcmuns.edu.vn, ddien@fit.hcmuns.edu.vn

Nguyen Le Nguyen, Hung Quoc Ngo
Faculty of Computer Sciences
College of Information Technology, Vietnam National University
Ho Chi Minh City, Vietnam
nguyennl@uit.edu.vn, hungnq@uit.edu.vn

***Abstract*—Text classification concerns the problem of automatically assigning given text passages (or documents) into predefined categories (or topics). Whereas a wide range of methods have been applied to English text classification, relatively few studies have been done on Vietnamese text classification. Based on a Vietnamese news corpus, we present two different approaches for the Vietnamese text classification problem. By using the Bag Of Words – BOW and Statistical N-Gram Language Modeling – N-Gram approaches we were able to evaluate these two widely used classification approaches for our task and showed that these approaches could achieve an average of >95% accuracy with an average 79 minutes classifying time for about 14,000 documents (3 docs/sec). Additionally, we also analyze the advantages and disadvantages of each approach to find out the best method in specific circumstances.**

Keywords-text classification; text categorization; feature selection; feature extraction; language modeling;, naïve bayes; support vector machines; k-nearest neighbours

I. INTRODUCTION

Text classification - TC (or text categorization*)* has been described as the activity of labeling natural language text with thematic categories from the predefined set. TC has been built the early '60s, until the early '90s, due to increased applicative interest and to the availability of more powerful hardware, it became a major subfield of the information systems discipline. Now with many advantages, TC is used in many applicative contexts, such as: automatic document indexing, 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.

In this paper, we have applied state-of-the-art text classification techniques [1] for Vietnamese text classification problem. To the best of our knowledge this is the first time that these techniques, which have been previously evaluated on English texts, have been used for Vietnamese. The most obvious point of difference between English and Vietnamese is in word boundary identification. In Vietnamese, the boundaries between words are not always spaces as those in English and the words are usually composed of special linguistic units called "morpho-syllable". This morpho-syllable may be a morpheme or a word or neither of them [13]. For example: with a Vietnamese sentence as belows:

"Một luật gia cầm cự với tình hình hiện nay" will be understood as many different statements due to its different word segmentations (here, we use the underscore "_" to link morpho-syllables of a Vietnamese word together), e.g.:

1. "A lawyer contends with the present situation"

("Một luật_gia cầm_cự với tình_hình hiện_nay")

2. "A law poultry resists the present situation"

("Một luật gia_cầm cự với tình_hình hiện_nay")

The comparison of Vietnamese and English word segmentation is shown in the Figure 1 as below:

Vnese	Một	luật	gia	cầm	cự	với	tình	hình	hiện	nay
Vnese1	Một	luật *gia*		*cầm cự*		với	tình hình		hiện nay	
Eng1	A	lawyer		contends		with	situation		present	
Vnese2	Một	luật	*gia cầm*		*cự*	với	tình hình		hiện nay	
Eng2	A	law	poultry		resists		situation		present	

(Vnese: Vietnamese; Eng: English)

Figure 1. An ambiguous example in Vietnamese word segmentation

In this example, there are more than one way of understanding. If we segment words in way 1 (the better one in semantics), we may classify this document into category "politics-society". However, if we segment words in way 2, we may classify this document into category "Health" (Avian-Flu topics). This implies that the word segmentation is a necessary problem which affects to the topics-based document classification. This problem needs to be solved in the preprocessing step before further processing can take place.

The rest of this paper is organized as follows: in Section 2 we discuss related work, Section 3 then presents our model and the processing resources for Vietnamese, Section 4 gives the results of experiments we conducted and Section 5 reports our conclusions and future work.

II. RELATED WORK

In the '80s, in order to create the automatic document classifiers in their manual construction, knowledge engineering (KE) techniques are used. Example to build manually, an

expert system required set of manually defined rules under the following type

if *(DNF[1] Boolean formula)* **then** (*category)* **else !**(*category)*

It means that the document was classified under (category) if is satisfied (DNF Boolean formula). And the construe system, which was built by Carnegie Group for the Reuters news agency, is the typical example for this approach.

Since the early '90s, the more effective and powerful approach which has been built and replaced for the KE approach, was machine learning (ML). By extracting the characteristics of a set of documents which have been pre-classified manually under c_i by a domain expert, a general inductive process (also called the learner) automatically builds a classifier for a category c_i. The advantages of this approach are that construction of a classifier based on an automatic builder of classifier from a set of manually classified documents (learning), not of a classifier.

Two assumptions for the advantages of ML approach over KE approach:

- Assumption 1: in case the manually classified documents are already available, an organization that had already been carrying out the same categorization activity manually, and that decides to automate the process. Evidently, ML is more convenient than KE approach.

- Assumption 2: in case the manually classified documents are not available, an organization must start a categorization activity and decides to opt for an automated modality straightaway. Manually classifying a set of document and characterizing a concept extensionally is easier than building and tuning a set of rules.

Some ML techniques were use for building a classifier that achieve impressive accuracy such as Naïve Bayes (NB) [5], k-Nearest Neighbors (k-NN) [7], Neural Networks (NN) [6],…, especially Support Vector Machine (SVM) [5], a state-of-the-art English TC classifier of today.

The survey reported in [1] shows that text classification in English has generally achieved satisfactory results with the results on some standard corpora such as Reuters, Ohsumed and 20 Newsgroups[2] ranging from 80 to 93%. However, the reported results for Vietnamese are very restricted and tend to be based on small data sets (from 50 to 100 files per topic) which are not publicly available for independent analysis [15]. Unlike for English no gold standard exists for evaluation. Evaluating the performance therefore for Vietnamese is very subjective and it is difficult to identify the best methods. To overcome these problems, we propose the following methodology:

- Corpus construction: we constructed a Vietnamese corpus which satisfies the conditions of sufficiency, objectiveness and balance. A detailed description of the corpus will be discussed in the next section

- Classification model: the text classification problem usually has three main approaches:

- ✓ Bag of Words – BOW based Approach [3]
- ✓ Statistical N-Gram Language Modeling based Approach [4]
- ✓ Combining two above approaches [7]

Each approach will have an advantage/disadvantage for different languages, therefore in this paper, we concentrate on compare the effect between these approaches in the TC for Vietnamese language.

III. METHODS

A. *Preparing the Corpus*

We built a Vietnamese corpus based on the four largest circulation Vietnamese online newspapers: VnExpress[3], TuoiTre Online[4], Thanh Nien Online[5], Nguoi Lao Dong Online[6]. The collected texts are automatically preprocessed (removing the HTML tags, spelling normalization) by Teleport software and various heuristics. There followed a stage of manual correction by linguists (five master students in Linguistics of University of Social Sciences, VNU-HCM city, Vietnam) who reviewed and adjusted the documents which are classified to the wrong topics. Finally, we obtained a relatively large and sufficient corpus which includes about 100,000 documents:

❖ **Level 1**

Level 1 includes some top categories from the above popular news websites. This contains about 33,759 documents for training and 50,373 documents for testing. These documents are classified by journalists and then passed a careful preprocessing step (see above part)

TABLE I. THE TOP 10 MOST FREQUENT CATEGORIES IN THE CORPUS (LEVEL 1)

No	Topic	Train	Test
1	politics-society	5,219	7,567
2	life	2,159	2,036
3	science & technology	1,820	2,096
4	business	2,552	5,276
5	health	3,384	5,417
6	law	3868	3788
7	world news	2,898	6,716
8	sports	5,298	6,667
9	culture	3,080	6,250
10	informatics	2,481	4,560
11	**Summary**	**33,759**	**50,373**

❖ **Level 2**

Level 2 includes the topics which are child topics of the level 1. The division in the level 1 is very vague meanwhile we need to find a specific topic to experiment for TC. So the level 2 is satisfactorily used for our purpose.

[1] DNF: "disjunctive normal form"
[2] http://ai-nlp.info.uniroma2.it/moschitti/corpora.htm

[3] www.vnexpress.net
[4] www.tuoitre.com.vn
[5] www.thanhnien.com.vn
[6] www.nld.com.vn

Level 2 contains about 14375 documents for training and 12076 documents for testing. Corpus level 2 is described as follows:

TABLE II. THE DISTRIBUTION OF THE CORPUS 27 (LEVEL 2)

No	Topic	Train	Test
1	music	900	813
2	eating and drinking	265	400
3	real property	246	282
4	football	1,857	1,464
5	stock	382	320
6	bird flu - influenza	510	381
7	the life in the world	729	405
8	studying abroad	682	394
9	tourist	582	565
10	WTO	208	191
11	family	213	280
12	computer entertainment	825	707
13	education	821	707
14	sex	343	268
15	hackers and viruses	355	319
16	criminal	155	196
17	life space	134	58
18	international business	571	559
19	Beauty	776	735
20	lifestyle	223	214
21	shopping	187	84
22	fine arts	193	144
23	stage and screen	1,117	1,030
24	new computer products	770	595
25	tennis	588	283
26	young world	331	380
27	fashion	412	302
28	**Summary**	**14,375**	**12,076**

B. *Vietnamese Text Classification Model*

The general model of the TC Module is shown in the following Figure 2. This is described in detail below.

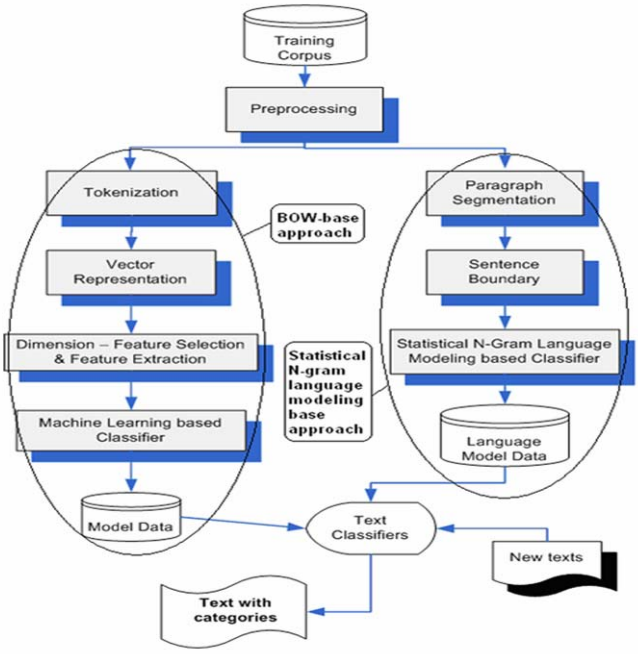


Figure 2. Our text classification model

C. *The BOW-based Approach*

In this approach the text document is transformed into a feature vector, where a feature is a single token or word. While English is an inflexional language, Asian languages such as Chinese, Thai and Vietnamese are isolating languages. With these languages, boundaries between words are not spaces as those in the flexion languages and the words are linked closely together. Words can be made from one morpho-syllable, or many morpho-syllables. So in this approach, a robust solution to document classification requires a good Vietnamese word segmentation module.

1) *Preprocessing*

❖ Tokenization

We use the state-of-the-art word segmentation program in [14] as a tokenizer in this BOW approach. All documents are segmented into words or tokens that are inputs for next steps.

❖ Removing stop words

In this phase the relevant features are extracted from documents. As usual, all words as well as numbers are considered feasible features. They are, usually called tokens [3]. After, the set of tokens is extracted it can be improved by removing features that do not bring any information. Function words (e.g., "và", "của" and "nhất là") are removed improving at least the efficiency of the target DC models. For this purpose a list of function words is prepared and used in the preprocessing phase as a stop list (about ~900 words, collected manually).

2) *Weighting Schemes*

Every text document which is input is firstly transformed into a list of words obtained by selecting only those which are not present in a list of stop words. Then the words are matched against the term dictionary. Each entry in dictionary includes current text, term frequency, the number of documents containing the term, *idf (Inverse Document Frequency)* frequency. This data structure is built during the domain learning phase and is needed here to access the *idf*, and other values for vector weighting. To weight the elements we use the standard *tf idf* product, with *tf* the term frequency in the document, and $idf=log(n/df(i))$ with n the number of documents in the collection and $df(i)$ the number of documents in the collection containing the word i, and pointers are obtained to words known to the system.

3) *Dimension Reduction – Feature Extraction and Selection*

Dimension reduction techniques can generally be classified into Feature Extraction (FE) approaches [2] and Feature Selection (FS) [8][12]. FS algorithms select a subset of the most representative features from the original feature space [16]. FE algorithms transform the original feature space to a smaller feature space to reduce the dimension. Though the FE algorithms have been proved to be very effective for dimension reduction, the high dimension of data sets in the text domain often fails many FE algorithms due to their high computational cost. Thus FS algorithms are more popular for real life text data dimension reduction problems.

In this paper, we only consider the feature selection algorithms. There has been much research done on feature selection in text classification [1] such as: MI (Mutual Information), IG (Information Gain), GSS (GSS coefficient), CHI (Chi-square), OR (Odds Ratio), DIA association factor, RS (Relevancy score). An excellent style manual for science writers is [7].

TABLE III. FORMULA OF SOME FEATURE SELECTION METHODS

Method	Formula
Information Gain	$\sum_{c\in\{c_i,\bar{c}_i\}}\sum_{t\in\{t_k,\bar{t}_k\}} P(t,c).\log\frac{P(t,c)}{P(t).P(c)}$
Mutual Information	$\log\frac{P(t_k,c_i)}{P(t_k).P(c_i)}$
Chi-square	$\frac{\lvert Tr\rvert.[P(t_k,c_i)P(\bar{t}_k,\bar{c}_i)-P(t_k,\bar{c}_i)P(\bar{t}_k,c_i)]^2}{P(t_k).P(\bar{t}_k).P(c_i).P(\bar{c}_i)}$
Odds Ratio	$\frac{P(t_k \mid c_i).(1-P(t_k \mid \bar{c}_i))}{(1-P(t_k \mid c_i).P(t_k \mid \bar{c}_i)}$
GSS Coefficient	$P(t_k,c_i).P(\bar{t}_k,\bar{c}_i)-P(t_k,\bar{c}_i).P(\bar{t}_k,c_i)$

Recently, the work in [10] has been shown that OCFS (Optimal Orthogonal Centroid Feature Selection) gives state-of-the-art performance of FS algorithms in text classification for a similar task to the one we approach.

Main ideas of OCFS:

Step 1: Calculate centroid m_i i=1,2,…,c for each category of training corpus

Step 2: Calculate centroid m for all categories of training corpus

Step 3: Calculate score for each term i-th follow by formula

$$s(i)=\sum\frac{n_j}{n}(m_j^i-m^i)^2 \quad \textbf{(1)}$$

Step 4: Choose K terms which have highest score

In this paper, we will implement six methods which are best in English text classification: MI, IG, GSS, CHI, OR, and especially OCFS. From our experiments, we will find feature selection methods which are best for Vietnamese document classification.

For the classification model we chose Support Vector Machines – SVM, a state of the art algorithm based on machine learning which has been widely applied to text classification [5].

D. Statistical N-Gram Language Modeling based Approach

1) Preprocessing

At this stage, we first pass the documents for spelling standardization which includes tone rule processing ex: hòa → hoà and letter variant processing ex: thời kỳ → thời kì. Then, they are passed the sentence and paragraph segmentation steps (will be used in afterwards probability calculation).

2) N-gram model and n-gram model based classifier

This is a new approach for text classification [4], that has been successfully applied in Chinese and Japanese languages. In this paper, we initially use this new model for Vietnamese and compare with other traditional methods (BOW approach).

N-gram model is a widely used language model. It assumes that the probability of one word in a document depends on its preceding n-1 words. Given a word sequence $s=w_1w_2\ldots w_T$, the probability of s could be calculated as follows by the chain rule of probability:

$$p(s)=\prod_{i=1}^{T} p(w_i \mid w_1\ldots w_{i-1}) \quad \textbf{(2)}$$

$p(w_i \mid w_1\ldots w_{i-1})$ can be estimated from a corpus with Maximum Likelihood criteria. That is:

$$p(w_i \mid w_{i-n+1}\ldots w_{i-1})=\frac{\#(w_{i-n+1}\ldots w_i)}{\#(w_{i-n+1}\ldots w_{i-1})} \quad \textbf{(3)}$$

Where $\#(w_{i-n+1}\ldots w_i)$ denotes the number of occurrence of the word sequence $w_{i-n+1}\ldots w_i$.

In real-world applications, $p(w_{i-n+1}\ldots w_{i-1})$ is often under-estimated due to the sparseness of training data. To solve this problem, smoothing techniques are introduced to adjust these low probabilities [4].

It is straightforward to construct text classifiers based on the n-gram model. Given an n-gram model, we can get the probability of a document being generated by this model. Therefore, after training the n-gram model on the training data of each category, we could classify test documents in the following way:

$$\begin{aligned} c^* &= \arg\max\{P(c)P(d \mid c)\} \\ &= \arg\max_{c\in C}\left\{P(c)\prod_{i=1}^{T} P(w_i \mid w_{i-n+1}\ldots w_{i-1},c)\right\} \\ &= \arg\max_{c\in C}\left\{P(c)\prod_{i=1}^{T} P_c(w_i \mid w_{i-n+1}\ldots w_{i-1})\right\} \end{aligned} \quad \textbf{(4)}$$

(C is set of categories, d is a new document, the prior *P(c)* can be estimated from training data).

In this paper, we consider text in document as a concatenated sequence of morpho-syllables instead of words. There are two main reasons:

1) We want to avoid the Vietnamese word segmentation problem which is proved to be a very difficult problem.

2) A morpho-syllable-based n-gram language model is smaller and it reduces the sparse data problem.

IV. EXPERIMENTS AND RESULTS

Two recall and precision parameters are used to evaluate the classification models [1]:

$$Recall = \frac{The\ text\ number\ classified\ by\ the\ model\ correctly}{The\ text\ number\ classified\ correctly\ in\ practice}$$

$$Precision = \frac{The\ text\ number\ classified\ by\ the\ model\ correctly}{The\ text\ number\ classified\ by\ the\ model}$$

$$F_1 = \frac{2 * recall\ *\ precision}{(recall\ +\ precision\ )} \quad \textbf{(5)}$$

In this phase, we define some following abbreviations:

- **SVM-Multi**: SVMs[7] with multi-class

- **SVM-Binary**: SVMs with binary-class

- **k-NN**: k Nearest Neighbours model [12][16]

- **N-Gram**: Statistical N-Gram Language Modeling

To systematically evaluate the Vietnamese document classification models, we investigate the comparison of several feature selection methods (MI, IG, CHI, GSS, OR, OCFS), different discounting smoothing techniques (used in N-Gram model), and different learning machine models (SVMs, k-NN, Naive Bayesian classification for documents represented by N-gram). Additionally, the total accuracy is calculated from the average accuracy of all categories for each experiment.

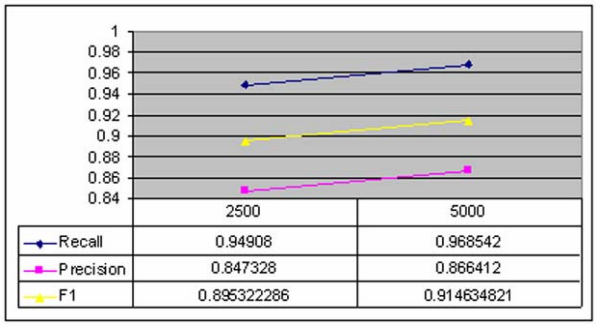


Figure 3. OCFS Feature Selection Evaluation with different number of terms (Corpus Level 1)

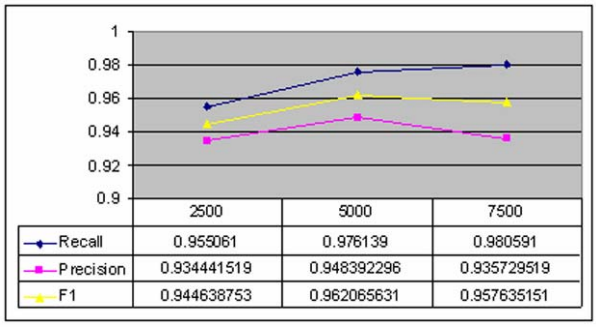


Figure 4. OCFS Feature Selection Evaluation with different number of terms (Corpus Level 2)

The results show that more the number of terms give more accuracy but classification speed is quite slower. So we choose the number of terms with 2,500 for next experiments.

[7] We use LIBSVM library (www.csie.ntu.edu.tw/~cjlin/**libsvm**/) in our experiment.

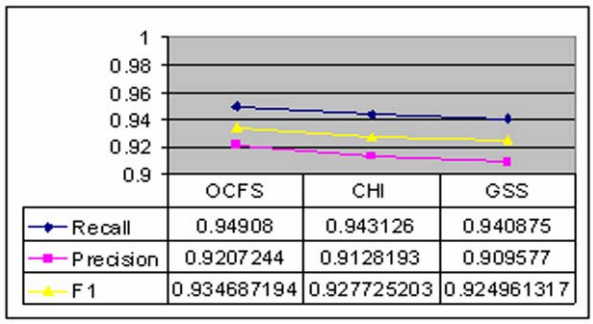


Figure 5. Feature Selection Methods Evaluation (2,500 terms) (Corpus Level 1)

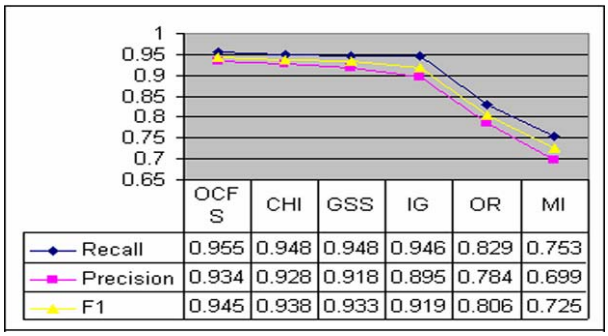


Figure 6. Feature Selection Methods Evaluation (2,500 terms) (Corpus Level 2)

Figure 5 and Figure 6 showed that the OCFS feature selection method gets the best performance on six feature selection methods which use for experiments on Vietnamese text classification.

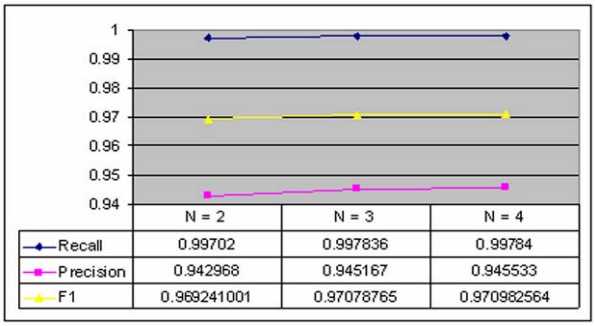


Figure 7. N-Gram evaluation with different N-Order values (Good-Turing smoothing) (Corpus Level 1)

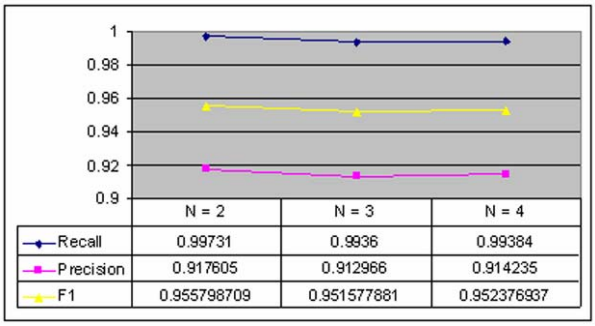
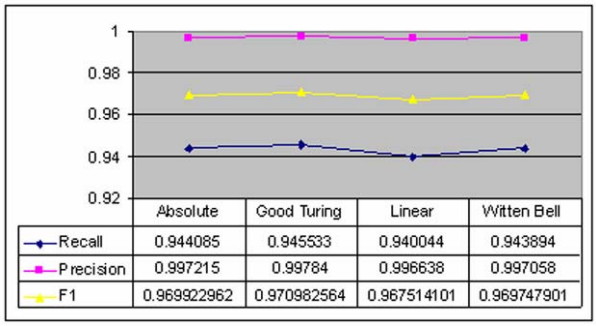


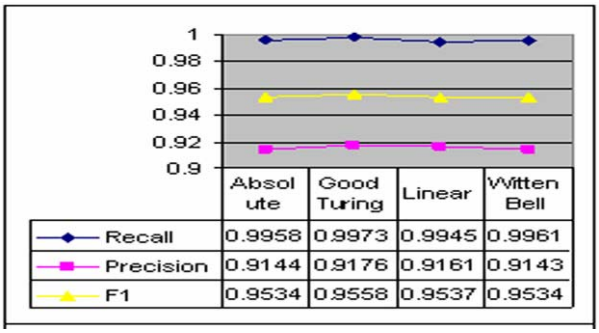
Figure 8. N-Gram evaluation with different N-Order values (Good-Turing smoothing) (Corpus Level 2)

For Corpus Level 1, the number of the training examples is so large (about 50,000 docs) that 4-gram frequency becomes higher. So perplexity of 4-gram is small and the performance is better.



	Absolute	Good Turing	Linear	Witten Bell
Recall	0.944085	0.945533	0.940044	0.943894
Precision	0.997215	0.99784	0.996638	0.997058
F1	0.969922962	0.970982564	0.967514101	0.969747901

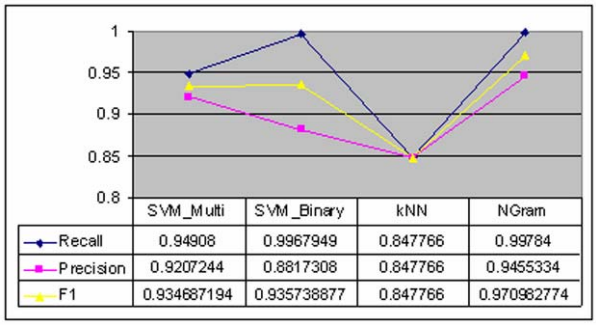
Figure 9. N-Gram evaluation with different discounting smoothing methods (N=4) (Corpus Level 1)



	Absolute	Good Turing	Linear	Witten Bell
Recall	0.9958	0.9973	0.9945	0.9961
Precision	0.9144	0.9176	0.9161	0.9143
F1	0.9534	0.9558	0.9537	0.9534

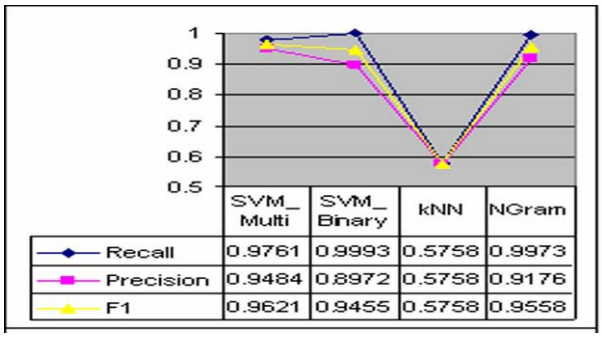
Figure 10. N-Gram evaluation with different discounting smoothing methods (N=2) (Corpus Level 2)

The above result shows that Good-Turing discounting smoothing method is best for Vietnamese document classification with N-Gram model.



	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

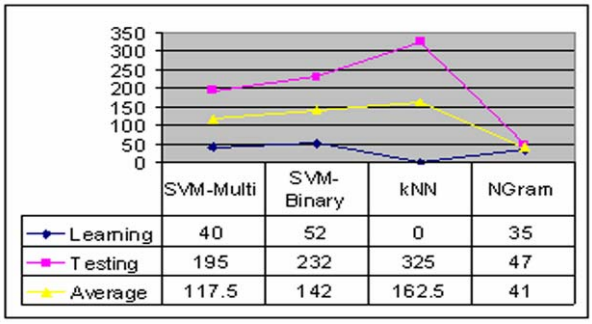
Figure 11. Evaluation with different document classification methods (Corpus Level 1)



	SVM_ Multi	SVM_ Binary	kNN	NGram
Recall	0.9761	0.9993	0.5758	0.9973
Precision	0.9484	0.8972	0.5758	0.9176
F1	0.9621	0.9455	0.5758	0.9558

Figure 12. Evaluation with different document classification methods (Corpus Level 2)

For the Corpus Level 1, the number of training examples is very large and N-Gram method becomes very effective.



	SVM-Multi	SVM-Binary	kNN	NGram
Learning	40	52	0	35
Testing	195	232	325	47
Average	117.5	142	162.5	41

Figure 13. Evaluation with time of learning (14375 docs) and testing (12075 docs) (Corpus Level 2)

In SVM training models, we choose the following parameters: C=1, 10; kernel-function = linear; SVM-type = C-SVM; other default parameters.

In kNNs model, we choose k=30 [12][17].

In N-Gram model, we choose N=2, 3, 4 and other default parameters.

V. CONCLUSION AND FUTURE WORK

With the differences between Vietnamese and English, finding an feasible approach for Vietnamese TC is very interesting. With our experiments, we prove that both SVM with average accuracy 96.21% and N-Gram with average accuracy 95.58% absolutely suitable to use for Vietnamese TC. Especially, the N-Gram model seems to be preferable to SVM for the following reasons: the higher classification speed, avoidance of the word segmentation and explicit feature selection procedure, and giving the equivalent F_1-score result.

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

1) The limitations from tokenizer (word segmentation tool) affects to classification performance (in BOW approach)

2) The documents have the ambiguities between two or many topics because these documents have too many tokens or phrases which both express the content of topic.

In the future, we could combine more semantic and contextual features (e.g. Latent Semantic Indexing – LSI [16]) to improve our system for handling polysemy and synonymy.

ACKNOWLEDGMENT

Thanks go to Mr. Chih-Chung Chang and Mr. Chih-Jen Lin for their Support Vector Machines tool, LIBSVM. We would like to thank the Global Liason Office of National Institute of Informatics in Tokyo for granting us the travel fund to research this problem. Finally, we also sincerely thank colleagues in the VCL Group (Vietnamese Computational Linguistics) for their invaluable and insightful comments.

REFERENCES

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

[2] Hayes, P. J., Andersen, P. M., Nirenburg, I. B., and Schmandt, L. M. 1990. Tcs: a shell for content-based text categorization. In *Proceedings of CAIA-90, 6th IEEE Conference on Artificial Intelligence Applications* (Santa Barbara, US, 1990), pp. 320–326.

[3] Ciya Liao, Shamim Alpha, Paul Dixon. Oracle Corporation. 2003. Feature preparation in Text Categorization, *AusDM03 Conference.*

[4] Fuchen Peng, Dale Schuurmans, Shaojun Wang. (2004). Augmenting Naïve Bayes Classifiers with Statistical Language Models, *Information Retrieval*, 7, p317-345.

[5] 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.

[6] Ng, H. T., Goh, W. B., and Low, K. L. 1997. Feature selection, perceptron learning, and a usability case study for text categorization. In *Proceedings of SIGIR-97, 20th ACM International Conference on Research and Development in Information Retrieval* (Philadelphia, US, 1997), pp. 67–73.

[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] Lewis D.D. Feature Selection and Feature Extraction for Text Categorization. *In Proceedings of the Speech and Natural Language Workshop, (1992)*

[9] Liu, H. and Motoda, H. Feature Extraction, Construction and Selection: A Data Mining Perspective. *Kluwer Academic*, Norwel, MA, USA, *1998*

[10] Jun Yan, Ning Liu, Benyu Zhang, Shuicheng Yan, Zheng Chen, Qiansheng Cheng, Weiguo Fan. (2005). OCFS: Optimal Orthogonal Centroid Feature Selection for Text Categorization (2005). *ACM* 2005.

[11] Maria Fernanda Caropreso, Stan Matwin, Fabrizio Sebastiani (2001). A Learner-Independent Evaluation of the Usefulness of Statistical Phrases for Automated Text Categorization, *Text Databases and Document Management: Theory and Practice, Idea Group Publishing, Hershey, US*, pp. 78--102.

[12] Yang, Y. and Pedersen, J.O., A comparative Study on Feature Selection in Text Categorization. *In Proceedings of the 14th International Conference on Machine Learning (ICML)*, (1997), 412-420.

[13] 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.

[14] Dinh Dien, Vu Thuy (2006), "A maximum entropy approach for Vietnamese word segmentation". *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, pp 247 – 252.

[15] 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.

[16] Tao Liu, Zheng Chen, Benyu Zhang, Wei-ying Ma, Gongyi Wu (2004). Improving Text Classification using Local Latent Semantic Indexing, Data Mining, 2004. *ICDM 2004. Proceedings, Fourth IEEE International Conference*.

[17] Yang, Y. M., & Chute, C. G. (1994). An Example-Based Mapping Method for Text Categorization and Retrieval. *ACM Transactions on Information Systems*, 12 (3), 252-277.