International Journal of Mechatronics, Electrical and Computer Technology

Vol. 3(9), Oct, 2013, pp. 314-329, **ISSN**: 2305-0543
Available online at: http://www.aeuso.org

An Efficient Method for Automatic Text Categorization

Mohammad Behrouzian Nejad[1*], Iman Attarzadeh[2] and Mehdi Hosseinzadeh[3]

[1]Department of Computer Engineering, Science and Research Branch, Islamic Azad University, Kerman, Iran

[2]Department of Computer Engineering, Dezful Branch, Islamic Azad University, Dezful, Iran

[3]Department of Computer Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran

*Corresponding Author's E-mail: *m.behrouzian@srbiau.ac.ir*

Abstract

Automatic Text Categorization refers to assigning uncategorized text documents to one or more predefined categories. Texts categorization generally divided into two main sections: feature selection and learning algorithm. For Feature selection and learning algorithms techniques, various methods have been proposed. The purpose of the proposed techniques, increasing the accuracy of classification and achieve optimal performance. In this paper a hybrid method is proposed which uses Filtering feature selection technique to reduce the complexity and works on combining classifiers outputs. The proposed method is homogeneous and uses uniform classifiers with different sampling with replacement from the training set. The results show the superiority of the proposed method compare to Naïve Bayes and j48 classifier and some related works according to the criteria of accuracy, precision, recall, F1 and classification error.

Keywords: Data Mining, Text Mining, Automatic Text Classification, Feature Selection, Learning Algorithm.

1. Introduction

Data mining is one of the most recent advances to serve data management technologies. Text mining is the one of the most important areas of data mining [1,2]. Main operation of text mining is consists of extracting knowledge from text. The

Vol. 3(9), Oct, 2013, pp. 314-329, **ISSN**: 2305-0543
Available online at: http://www.aeuso.org

most important issues facing the text mining is automatic text categorization. Automatic text categorization means of assigned text documents available to pre-define several categories that documents that belong to them [3]. To do this work, must first identify the categories that usually it is done by experts, then documents must be determined for each category. The main purpose finding real category of collected text documents. With the increasing use of the Internet and electronic documents, automatic text categorization has become imperative. Several methods have been proposed for the text categorization.

Among the proposed methods can be cited text categorization based on unorganized data with extracted information [4], text classification based on features [5], Text Categorization using PDDP with Support Vector Machines (SVM) [6], text categorization using K-Nearest Neighbor (KNN) [7], Improved KNN Algorithm by Optimizing Cross-validation [8], Improved KNN based on clustering [9], improved KNN using ant colony [10], Naïve Bayes method [11, 12], text categorization using association rules [13], … that their purpose are improving the accuracy and efficiency of classification.

The purpose of this paper providing a method for automatic text categorization that can classify the text documents with high performance. The results show the superiority of the proposed method compared heterogeneous classifier and single classifiers according to the criteria of accuracy, precision, recall, F1 and classification error. The reminder of this paper is organized as follows. In Section 2, we review the related works. Process of text categorization is discussed in sections 3. Section 4 describes the proposed method. Sections 5 and 6 describe the implementation and evaluation of proposed method. Finally, Sections 7 and 8 contains the conclusion and references of the paper.

Vol. 3(9), Oct, 2013, pp. 314-329, ISSN: 2305-0543
Available online at: http://www.aeuso.org

2. Related Works

KNN algorithm is one of the approaches widely used in text classification because is easy to implement. One of the problems in K nearest neighbor algorithm is determine the proper value of the parameter K, which is necessary for the performance of a classification. The parameter K is usually more than the number of classes and a odd number is considered. According to the KNN is a lazy learning, which means that hold all the training samples to the end of classification. With increasing parameter K, increases the complexity of KNN which is not desirable. In order to solve this problem in [14] a hybrid classification method using two classifiers KNN and SVM are presented. In order to reduce training time, for each category first SVM is used for classification.

Then support vectors of different categories as training data are given with KNN classifier. To calculate the average distance between the test data and each support vector, the Euclidean distance is used. The final decision will be based on batch which its vector have minimum distance supporter with test data. Extracted results show the efficiency of this method. In [15] an improved KNN algorithm for text categorization is proposed. In this method, classification model is built using a clustering algorithm which for clustering uses from minimum distance for divides learning textual examples into hyper Koreas with the same radius.

Then used KNN method for classifying sets of experiments based on the model. This method significantly reduces the computational complexity and dynamically updates the classification model. The results indicate that this method works best from methods such as KNN, NB and SVM. In [16] the combination of two classifiers are used as serial categories, that the first classifier are candidates categories for new document, then second classifier, select final class of new documents between classes. This method with 2850 documents (training and testing) is investigated. The results show that the performance

Vol. 3(9), Oct, 2013, pp. 314-329, ISSN: 2305-0543
Available online at: http://www.aeuso.org

improvement which due to the low number of documents obtained, the reliability is not high enough.

3. Process of Text Categorization

The main steps of the process of text categorization can be classified into three main stages of preprocessing, feature selection and classification stage [17].

3.1. Preprocessing

In the preprocessing stage, usually on the input data operations are separating words, removal of redundant words (stop words), stemming and term (Feature) weighting.

3.2. Feature Selection

This step refers to select a subset of features of the text. In general, feature selection techniques are classified into two general categories: filtering and wrapper methods. Filtering methods are independent of the learning algorithm. These methods Regardless of learning algorithm and using statistical methods to do feature selection and have time complexity lower than the wrapper methods. Wrapper methods uses from learning algorithm as the evaluation function. These methods have higher time complexity and accuracy than filter methods [18]. With the increasing size of the features in text classification, generally these methods could not be used because of the high complexity. Some filtering methods that can be used in many texts classification techniques such as Document Frequency (DF), Information Gain (IG), and Mutual Information (MI) [19]. One of the most versatile and popular methods is information gain method which used in many studies and has good results. Information gain value measures \the number of bits of information obtained for category prediction by knowing presence of absence of a term in a document". Information gain values were calculated as (1) which $P(t, c)$ shows number of

Vol. 3(9), Oct, 2013, pp. 314-329, **ISSN**: 2305-0543
Available online at: http://www.aeuso.org

text documents in category c which have term t and $p(\bar{t}, c)$ shows number of text documents in category c which have nor term t [19]:

$$IG(t_k, c_i) = \sum_{c\in\{c_i,\bar{c}_i\}} \sum_{t\in\{t_k,\bar{t}_k\}} p(t,c) \log_2 \frac{P(t,c)}{P(t)P(c)} \quad (1)$$

3.3. Learning Algorithm

In this step, classifiers from preprocessed text, act to learning [17, 20].

3.3.1. Naïve Bayes: One of the most commonly methods for text classification is Naïve Bayes algorithms. This algorithm is based on conditional probabilities and uses the Bayes theorem. In this approach to classify a new example, the probability of all categories for new example calculated and then category with the highest probability is selected as a category of new example. It is one of the fastest methods.

3.3.2. Decision Tree: decision tree is a tree which internal nodes represent features, outputted edges are feature selection criteria and leaf nodes represent classes. Construct of decision tree has two phases: growth and pruning. In growth phase a decision tree is built from the training data. In The pruning phase, section of the tree is pruned so that the test is not done on the branch. Decision tree classification based on feature selection criteria are divided into two categories CART and C4.5. CART is classification and regression algorithm. In this study, we used from the j48 decision tree that construct the pruning or not pruning tree of C4.5.

3.3.3. Support Vector Machine: This algorithm is one of the most popular classification algorithms which in recent years had a pretty good performance. This algorithm is based on classification of linear data. In the dividing line data, try to choose a line that is more confident border than the other lines. Optimum choice for line data, can done QP methods

Vol. 3(9), Oct, 2013, pp. 314-329, ISSN: 2305-0543
Available online at: http://www.aeuso.org

which related to solve restrictions problems. In [21] linear and nonlinear methods are tested which linear method is little better than nonlinear.

4. Proposed Method

One way to improve classification performance is use combination of classifiers. Use of combination of classifiers by combining multiple alone classifier can improve performance. Conducted Researches shows by use of combination of classifiers, can improve performance [22]. The proposed method combines the outputs of classifiers and because they do not need to know the structure of classifiers and its feature vectors, it is more common [23]. The proposed method is homogeneous and uses uniform classifiers with different sampling with replacement from the training set. For this work we use from bootstrapping method [24-27]. Fig. 1 shows the proposed method.

International Journal of Mechatronics, Electrical and Computer Technology

Vol. 3(9), Oct, 2013, pp. 314-329, **ISSN**: 2305-0543
Available online at: http://www.aeuso.org

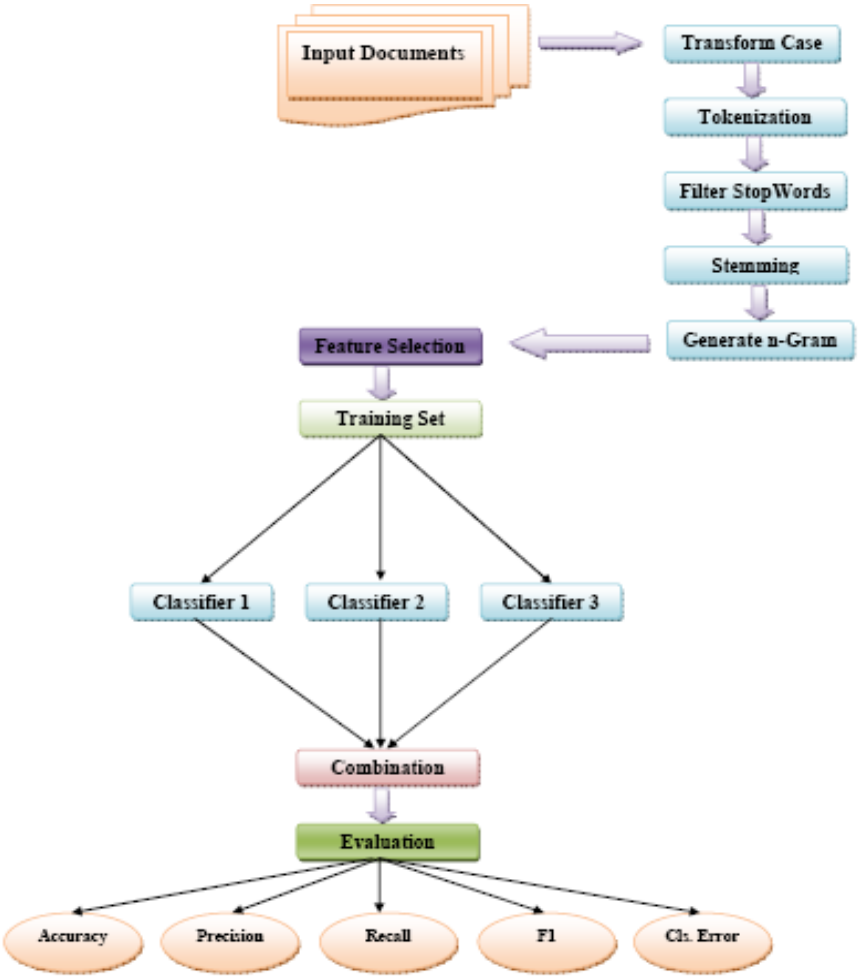

Figure 1: proposed method.

In the proposed method, input documents are in the preprocessing. Feature selection is carried out at next stage. Then we use of different sampling with replacement from the training set, several new training set with the initial size obtained and use by each classifiers. Then classifiers outputs are combine. In combine stage we use majority vote. Last stage is performance evaluation by different criteria.

International Journal of Mechatronics, Electrical and Computer Technology

Vol. 3(9), Oct, 2013, pp. 314-329, ISSN: 2305-0543
Available online at: http://www.aeuso.org

5. Implementation

5.1. Implementation Tools

In this paper for implement proposed method we use from Rapid Miner version 5.2 [28]. This software is an open source data mining tools and written by Java language. Evaluation runs on a system with windows operating system, 2 GHz of CPU and 2 GB of memory.

5.2. Dataset

In this paper we use from reuters-21578 dataset [29]. We use from R (8) subset of this dataset [30]. This data set contains 8 main category and have 7674 text documents. Each document related to a category. Details of R (8) subset of reuters-21578, shows in table 1.

Table 1: Details of R (8) reuters-21578

R (8) reuters-21578			
Categories	Train documents	Test documents	Total documents
Acq	1596	696	2292
Crude	253	121	374
Earn	2840	1083	3923
Grain	41	10	51
Interest	190	81	271
money-fx	206	87	293
Ship	108	36	144
Trade	251	75	326

5.3. Preprocessing

This step in the proposed method contains: Transform Case, Tokenization, Filter Stop Words, Stemming and Generate n-Gram.

Vol. 3(9), Oct, 2013, pp. 314-329, ISSN: 2305-0543
Available online at: http://www.aeuso.org

- Transform Case: in this step, all characters of text are converted to the same form. At this stage, all the characters are converted to lowercase.
- Tokenization: At this stage, the whole text is divided into separate successive words.
- Filter Stop Words: This step will eliminate redundant words like and, the, for and ….
- Stemming: At this stage for eliminate prefixes and suffixes of words, Porter algorithm are used.
- Generate n-Gram: At this stage for indexing and reduce the dimension text n-gram method is used. With using n-gram we can make the text as a series of consecutive words with length n. This model was originally proposed for speech processing issues. But now many different versions of this model have been proposed for text classification problems in natural language processing [31-34]. The experiments performed on different values of n and to avoid increasing the complexity of the n-gram with n = 2 was used.

After these steps, the weighted features are performed. The Tfidf Term weighting method used in this paper [3].

5.4. Feature Selection

In this study, for increasing the efficiency and reducing the complexity, information gain feature selection method used in the feature selection stage.

5.5. Learning Algorithm

The proposed method using libsvm, is implemented. Libsvm which Library Support Vector Machines also say that, based on support vector machines and Java

Vol. 3(9), Oct, 2013, pp. 314-329, ISSN: 2305-0543
Available online at: http://www.aeuso.org

libraries which have been developed by [35]. The proposed methods are compared with naïve bayes and j48.

5.6. Evaluation Criteria

This section will examine the evaluation criteria for the text classification [3]. Different status of categories and documents according to input dataset to classification with TP, FP, TN, FN values for two Negative and Positive categories shows as table 2.

Table 2: Different status of categories and documents

		Predicted	
		Positive	Negative
	Positive	TP	FN
	Negative	FP	TN

Were FP is the number of documents which incorrectly classified under Positive ; TN number of documents which correctly classified under Negative, TP is number of documents which correctly classified under Positive, and FN is number of documents which incorrectly classified under Negative. According to parameters in table 2., different evaluation criteria like Accuracy, Classification Error, Precision, Recall and F1 are presented. How to calculate of these measures are shown (2) to (6) respectively.

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (2)$$

$$ER = \frac{FN+FP}{TP+FP+FN+TN} = 1 - \text{Accuracy} \quad (3)$$

Vol. 3(9), Oct, 2013, pp. 314-329, **ISSN**: 2305-0543
Available online at: http://www.aeuso.org

$$Precision_i = \frac{TP_i}{TP_i + FP_i} \quad (4)$$

$$Recall_i = \frac{TP_i}{TP_i + FN_i} \quad (5)$$

$$F1_i = \frac{2 * Precision_i * Recall_i}{Precision_i + Recall_i} \quad (6)$$

Were i index represent that these parameters should be calculate for each category i.

6. Evaluation of Proposed Method

At this step, the proposed method has been evaluated with naïve bayes and j48 using different criteria. All result are per percent. Table 3. shows the results of proposed method, Table 4. shows the results of naïve bayes classifier and Table 5. shows the results of j48 classifier.

Table 3: the results of proposed method

Category	Precision	Recall
Acq	99.58	99.73
Trade	99.88	100
Ship	100	99.60
Interest	97.35	98.07
Grain	100	98.99
Crude	99.78	99.83
Earn	99.92	99.81
money-fx	97.45	97.31
Avg-precision: 99.25		
Avg-recall: 99.17		
F1: 99.20		
Accuracy: 99.63		
classification_error: 0.37		

International Journal of Mechatronics, Electrical and Computer Technology

Vol. 3(9), Oct, 2013, pp. 314-329, ISSN: 2305-0543
Available online at: http://www.aeuso.org

Table 4: the results of naïve bayes classifier

Category	Precision	Recall
Acq	84.62	90.35
Trade	77.19	70.12
Ship	30	50
Interest	73.96	74.74
Grain	17.72	34.15
Crude	62.59	72.73
Earn	99.31	90.88
money-fx	67.94	68.93
Avg-precision: 64.30		
Avg-recall: 68.90		
F1: 66.52		
Accuracy: 86.33		
classification_error: 13.67		

Table 5: the results of j48 classifier

Category	Precision	Recall
Acq	85.59	88.97
Trade	81.93	81.27
Ship	36.26	30.56
Interest	77.20	78.42
Grain	13.64	7.32
Crude	83.53	84.19
Earn	95.59	94.61
money-fx	67.32	66.99
Avg-precision: 67.88		
Avg-recall: 66.55		
F1: 67.20		
Accuracy: 88.37		
classification_error: 11.63		

Vol. 3(9), Oct, 2013, pp. 314-329, ISSN: 2305-0543
Available online at: http://www.aeuso.org

Table 6. shows the results of proposed method compare to some related works. Result show that our method has better performance than Naïve Bayes and j48 Classifiers and related works according to the criteria of accuracy, precision, recall, F1 and classification error.

Table 6: Comparison of our method

Criteria	Our method	Ref. [9]	Ref. [6]	Ref. [7]
Average Precision	99.25	91.33	-	-
Average Recall	99.17	91.15	-	-
Average F1	99.20	91.23	-	-
Accuracy	99.63	-	98.05	90.04
Cls.Error	0.37	-	-	-

Conclusion

According to this note that one way to improve classification performance is use combination of classifiers; in this paper we proposed a hybrid method which uses filtering method for feature selection technique to reduce the complexity and combine classifiers outputs. The proposed method is homogeneous and uses uniform classifiers with different sampling with replacement from the training set. The results show the superiority of the proposed method compared naïve bayes and j48 classifiers and some references according to the criteria of accuracy, precision, recall, F1 and classification error.

Vol. 3(9), Oct, 2013, pp. 314-329, **ISSN**: 2305-0543
Available online at: http://www.aeuso.org

References

[1] L. H. Witten and E. Frank, Data mining Practical Machine Learning Tools and Techniques, 2nd Edition, MORGAN KAUFMANN PUBLISHERS IS AN IMPRINT OF ELSEVIER, (2005), pp. 4-11.

[2] L. Wang and X. Fu, Dtat Mining with Computetional Intelligence, Advanced Information and Knowledge Processing, Springer-Verlag Berlin Heidelberg, (2005), pp. 1-5.

[3] F. SEBASTIANI, Machine Learning in Automated Text Categorization, ACM Computing Surveys, Vol. 34, 1 (2002), pp. 1–47.

[4] S. Manne, S. S. Fatima, A Novel Approach for Text Categorization of Unorganized data based with Information Extraction, International Journal on Computer Science and Engineering, Vol. 3, 7 (2011), pp. 2846-2854.

[5] K. Nirmala, M. Pushpa, Feature based Text Classification using Application Term Set, International Journal of Computer Applications, Vol. 52, 10 (2012), pp. 1-3.

[6] K.Gayathri, A.Marimuthu, Text Categorization using PDDP with Support Vector Machines, International Journal of Emerging Trends and Technology in Computer Science, Vol. 1, 3 (2012), pp. 224-227,.

[7] E. H. Han, G. Karypis, V. Kumar, Text Categorization Using Weight Adjusted k-Nearest Neighbor Classification, Proceedings of the 5th Pacific-Asia Conference on Knowledge Discovery and Data Mining, (2001), pp. 53-65.

[8] S. S. Dadhania, J. S. Dhobi, Improved kNN Algorithm by Optimizing Cross-validation, International Journal of Engineering Research & Technology, Vol. 1, 3 (2012), pp. 1-6.

[9] Z. Yong, L. Youwen, X. Shixiong, An Improved KNN Text Classification Algorithm Based on Clustering, Journal of Computers, Vol. 4, 3 (2009), pp. 230-237.

[10] Y. Hongwei, Z. Wei, Application of Ant Colony algorithm in KNN text classification, Journal of Changchun University of technology (natural science Edition), Vol. 33, 1 (2010), pp.159-163.

[11] S. Kim, K. Han, H. Rim, S. H. Myaeng, Some effective techniques for naïve bayes text classification, IEEE Transactions on Knowledge and Data Engineering, Vol. 18, 11 (2006), pp. 1457-1466.

[12] M. J. Meena, K. R. Chandran, Naïve bayes text classification with positive features selected by statistical method, Proceedings of the IEEE international conference on Advanced Computing, (2009), pp. 28 – 33.

[13] C. M. Rahman, F. A. Sohel, P. Naushad, S. M. Kamruzzaman, Text Classification using the Concept of Association Rule of Data Mining, Proceedings of the International Conference on Information Technology, Kathmandu, Nepal, (2003), pp. 234-241.

Vol. 3(9), Oct, 2013, pp. 314-329, **ISSN**: 2305-0543
Available online at: http://www.aeuso.org

[14] C. H. Wan, L. H. Lee, R. Rajkumar, D. Lsa, A hybrid text classification approach with low dependency on parameter by integrating K-nearest neighbor and support vector machine, Expert Systems with Applications, vol. 39, (2012), pp. 11880-11888.

[15] L. Baoli, Y. Shiwen, L. Qin, An improved k-nearest neighbor algorithm for text categorization, Proceedings of the 20th International Conference on Computer Processing of Oriental Languages, Shenyang, China, (2003).

[16] Z. Zheng, S. Zhou, A. Zhou, Sequential Classifiers Combination for Text Categorization: An Experimental Study, WAIM 2004, (2004), pp. 509–518.

[17] V. Korde, C. N. Mahender, TEXT CLASSIFICATION AND CLASSIFIERS: A SURVEY, International Journal of Artificial Intelligence & Applications (IJAIA), Vol.3, 2 (2012), pp. 85-99.

[18] R. Jensen, Combining rough and fuzzy sets for feature selection, PhD Thesise, University of Edinburgh, (2005).

[19] K. Dave, Study of feature selection algorithms for text-categorization, University of Nevada, Las Vegas, UNLV Theses/Dissertations/Professional Papers/ Capstones. Paper 1380, (2011).

[20] Mita K. Dalal, Mukesh A. Zaveri, Automatic Text Classification: A Technical Review, International Journal of Computer Applications, Vol. 28, 2 (2011), pp 37-40.

[21] R. Klinkenberg, T. Joachims, Detecting concept drift with support vector machines, Proceedings of the 17th International Conference on Machine Learning, (2000), pp. 487-494.

[22] H. A. Torshizi, H. R. Tahmasebi, Review of Classifier Combination Techniques, Proceedings of the 2nd Iranian Data Mining Conference, AmirKabir University of Technology, Iran, (2008).

[23] S. Tulyakov, S. Jaeger, V. Govindaraju, D. Doermann, Review of Classifier combination Methods, Studies in Computational Inteligence(SCI), Vol. 90, (2008), pp.361-386.

[24] A. Joorabchi, A. E. Mahdi, A New Method for Bootstrapping an Automatic Text Classification System Utilizing Public Library Resources. Proceedings of the 19th Irish Conference on Artificial Intelligence and Cognitive Science (AICS08), (2008), August 27 - 29.

[25] Y. Ko, J. Seo, Learning with unlabeled data for text categorization using bootstrapping and feature projection techniques, Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, (2004).

Vol. 3(9), Oct, 2013, pp. 314-329, **ISSN**: 2305-0543
Available online at: http://www.aeuso.org

[26] A. Mccallum , K. Nigam, Text Classification by Bootstrapping with Keywords, EM and Shrinkage, In ACL99 - Workshop for Unsupervised Learning in Natural Language Processing, (1999).

[27] Y. Ko, J. Seo, Text classification from unlabeled documents with bootstrapping and feature projection techniques, Information Processing and Management, Vol. 45, (2009), pp 70–83.

[28] Software available in: http://rapid-i.com/content/view/181/190/lang,en/, (2013/06/04).

[29] Dataset available in: http://archive.ics.uci.edu/ml/datasets/Reuters-21578+Text+Categorization+ Collection, (2013/06/04).

[30] Dataset available in: http://csmining.org/index.php/r52-and-r8-of-reuters-21578.html, (2013/06/04).

[31] F. Peng, D. Schuurmans, S. Wang, Language and Task Independent Text Categorization with Simple Language Models, Proceedings of the HLT-NAACL, (2003).

[32] J. Fürnkranz, A Study Using n-gram Features for Text Categorization, Austrian Research Institute for Artificial Intelligence, (1998).

[33] Z. WEI, D. MIAO, J. H. CHAUCHAT, R. ZHAO, W. LI, N-grams based feature selection and text representation for Chinese Text Classification, International Journal of Computational Intelligence Systems, Vol.2, 4 (2009), pp 365-374.

[34] P. Náther, N-gram based Text Categorization, Diploma thesis, COMENIUS UNIVERSITY, (2005).

[35] C. Chang, C. J. Lin ,LIBSVM: a library for support vector machines, (2001). Available in: http://www.csie.ntu.edu.tw/~cjlin/libsvm/, (2013/06/04).