LDA-based Keyword Selection in Text Categorization

Şerafettin Taşcı
Computer Engineering Department
Bogazici University
Bebek, 34342 Istanbul, Turkey
SERAFETTIN.TASCI@BOUN.EDU.TR

Tunga Güngör
Computer Engineering Department
Bogazici University
Bebek, 34342 Istanbul, Turkey
GUNGORT@BOUN.EDU.TR

***Abstract*-Text categorization is the task of automatically assigning unlabeled text documents to some predefined category labels by means of an induction algorithm. Since the data in text categorization are high-dimensional, feature selection is broadly used in text categorization systems for reducing the dimensionality. In the literature, there are some widely known metrics such as information gain and document frequency thresholding. Recently, a generative graphical model called latent dirichlet allocation (LDA) that can be used to model and discover the underlying topic structures of textual data, was proposed. In this paper, we use the hidden topic analysis of LDA for feature selection and compare it with the classical feature selection metrics in text categorization. For the experiments, we use SVM as the classifier and tf*idf weighting for weighting the terms. We observed that almost in all metrics, information gain performs best at all keyword numbers while the LDA-based metrics perform similar to chi-square and document frequency thresholding.**

Keywords: document categorization, feature selection, latent dirichlet allocation

I. INTRODUCTION

Text categorization is a supervised learning task in which documents are assigned to categories based on the training on a labeled document set. It has gained great popularity and importance in recent years since the amount of documents in electronic medium which necessitate organization and arrangement increased considerably. A large amount of statistical techniques and machine learning approaches have been used for this task such as naive Bayes, linear regression, rocchio, neural network, k-nearest neighbor (kNN) and support vector machines (SVM) [12].

In text categorization, generally a document is represented as a set of words without regarding grammar and word order. This representation is called 'bag of words' model. Since a document set may contain thousands of words, a 'bag of words' representation of a document will probably have a very high dimensionality. This situation is a critical challenge for most learning algorithms. Therefore, feature selection is broadly used in text categorization systems for the purpose of reducing the dimensionality. Dimensionality reduction has many benefits such as improving the interpretability of data, reducing the time and storage requirements and speeding up the learning process. Moreover, it may improve the classification accuracy since it can prevent over fitting by eliminating the terms that are useless or misleading for the classifier.

Feature selection on textual data is mostly based on feature ranking in which all features are ranked by a metric that estimates their importance and then the ones with the highest ranks are selected. In the literature, there are many feature selection metrics such as Information Gain, Chi-square statistics and Document Frequency. The first two metrics are supervised (i.e. they require a labeled training set) while the last one, DF, is an unsupervised metric (i.e. it does not require a labeled training set).

Feature selection is at least as important as the choice of the induction algorithm in text categorization. Accordingly, many studies to evaluate the feature selection metrics have been done in recent years. Reference [15] evaluates five of the most popular feature selection metrics on the Reuters and Ohsumed datasets. In this study, kNN and LLSF are used as the classification algorithms instead of SVM. Only global policy is used and the metrics are evaluated in terms of precision. In a later study [14] SVM is also considered and compared with other classifiers. Reference [4] considers local policy and gives a comprehensive evaluation of many well-known feature selection metrics. In this study, SVM is the classifier and many datasets including skew datasets as well as homogenous ones are used. Reference [2] investigates some more advanced feature-selection approaches that use higher order decisions and take the feature-to-feature correlation into account when selecting the feature set such as odds ratio, CFS and Markov blanket. Finally, in [13] some well-known feature selection metrics and the policies were evaluated in datasets with different characteristics with a focus on the comparison of the feature selection policies.

Recently, a generative model called LDA (latent dirichlet allocation) that allows sets of observations to be explained by unobserved groups which explain why some parts of the data are similar was proposed [1]. LDA is a generative graphical model that can be used to model and discover the underlying topic structures of any kind of discrete data where textual data is a typical example. The outputs of the LDA analysis for a

given dataset are a list of hidden topics each consisting of a list of terms ranked by relevance. The underlying idea of LDA-based feature selection framework is that a good term should only be highly ranked in only a few topics to be more discriminative for classification. So for finding the best terms, entropies of the terms on the term-topic matrix are calculated and the terms that have lower entropy values are selected. In [16], LDA-based feature selection was proposed. However, they only used LDA based on Gibbs sampling, ignoring the variational method that is used in the original LDA framework [1]. In addition, they do not compare LDA-based feature selection with the popular feature selection metrics like information gain.

LDA has been shown to be more effective than pLSA, which uses a latent variable model in which documents are represented as mixtures of topics, in text-related problems such as document classification since it follows a full generation process for document collection [1,5]. However, as far as we know, there is no study aimed at the evaluation of feature selection by LDA in text categorization. In this paper, we present a comparison of the classical feature selection metrics with LDA-based feature selection. In Section 2, we describe the existing feature selection methods that are used in this study and the LDA-based feature selection with a short summary of latent dirichlet allocation. Section 3 mentions about our experimental settings; the datasets, evaluation metrics, preprocessing steps and the classifier. In Section 4, we give the results of the experiments and give a comparative discussion of these results. We conclude the study in Section 5.

II. FEATURE SELECTION METRICS

In this section we give information about the known feature selection metrics that are used in this study as well as the LDA-based feature selection.

A. Existing Metrics

In this study, three popular feature selection metrics are used (see Table I). The function $f(t_k, c_i)$ denotes the feature selection score of a term t_k and it is specified locally to a specific category c_i. In order to assess the value of t_k in a global sense, the maximum $f_{max}(t_k) = max_{i=1}^{|C|} f(t_k, c_i)$ of the category-specific values is computed where |c| denotes the number of categories[3].

1) Information gain (IG): This metric measures the reduction in the entropy by knowing the existence or absence of a term in a document. It is a very popular term-goodness criterion that is widely used in the machine learning community [3,15]

2) Chi-square (x^2) Statistics (CHI): In statistics, chi-square test is applied to measure the independence of two random variables. In the domain of text categorization, the two random variables are the occurrence of term t and the occurrence of class c. It is also used extensively in the text categorization research and in most studies it is claimed to perform comparable to information gain [4,15].

3) Document frequency (DF): This metric is a very simple metric which is independent from the class labels. It is based on the assumption that infrequent terms are not reliable and effective in category prediction. It counts the number of documents in which a term appears and selects the terms whose counts are the highest. In spite of its simplicity, it has a performance similar to IG and CHI if the keyword number is not too low [4,13,15].

TABLE I. FEATURE SELECTION METRICS

Name	Formula
Information Gain	$IG(t_k, c_i) = \sum_{c \in \{c_i, \bar{c}_i\}} \sum_{t \in \{t_k, \bar{t}_k\}} P(t,c).log \frac{P(t,c)}{P(t).P(c)}$
Chi-square	$x^2(t_k, c_i) = N \times \frac{[P(t_k,c_i)P(\bar{t}_k,\bar{c}_i) - P(\bar{t}_k,c_i)P(t_k,\bar{c}_i)]^2}{P(t_k)P(\bar{t}_k)P(c_i)P(\bar{c}_i)}$
Document Frequency	$DF(t_k, c_i) = P(t_k, c_i)$

Notation:

$P(t_k, c_i)$: Percentage of documents belonging to class c_i in which term t_k occurs.

$P(\bar{t}_k, \bar{c}_i)$: Percentage of documents not belonging to class c_i in which term t_k does not occur.

$P(\bar{t}_k, c_i)$: Percentage of documents belonging to class c_i in which term t_k does not occur.

$P(t_k, \bar{c}_i)$: Percentage of documents not belonging to class c_i in which term t_k occurs.

N : Total number of documents in the dataset.

B. LDA-based Feature Selection

Latent Dirichlet Allocation (LDA) was proposed by Blei et al. [1] as a method to find the latent structure of the topics in a text corpus. LDA is closely related to the probabilistic latent semantic analysis (pLSA) proposed by Hofmann [7]. In fact, pLSA is a probabilistic formulation of the well-known latent semantic indexing (LSI) technique. The intuition behind LSI is to find the latent structure of the topics or concepts in a text corpus.

The basic generative process of LDA is highly similar to pLSA. In pLSI, the topic mixture is conditioned on each document whereas in LDA, the topic mixture is drawn from a conjugate Dirichlet prior that is same for all documents (see Fig. 1).

In LDA, a document $w_m = \{w_{m,n}\}_{n=1}^{N_m}$ is generated by first picking a distribution over the topics v_m from a Dirichlet distribution (Dir(α)), which determines topic assignment for the words in that document where $w_{m,n}$ is a particular word for the word placeholder [m,n], N_m is the length of document m and v_m is the topic distribution for document m. Then the topic assignment for each word placeholder [m, n] is performed by sampling a particular topic $z_{m,n}$ from multinomial distribution *Mult*(v_m) where $z_{m,n}$ is the topic

index of the nth word in document m. And finally, a particular word $w_{m,n}$ is generated for the word placeholder [m, n] by sampling from multinomial distribution, $Mult(\varphi_{z_{m,n}})$ where φ_k is the word distribution for topic k.

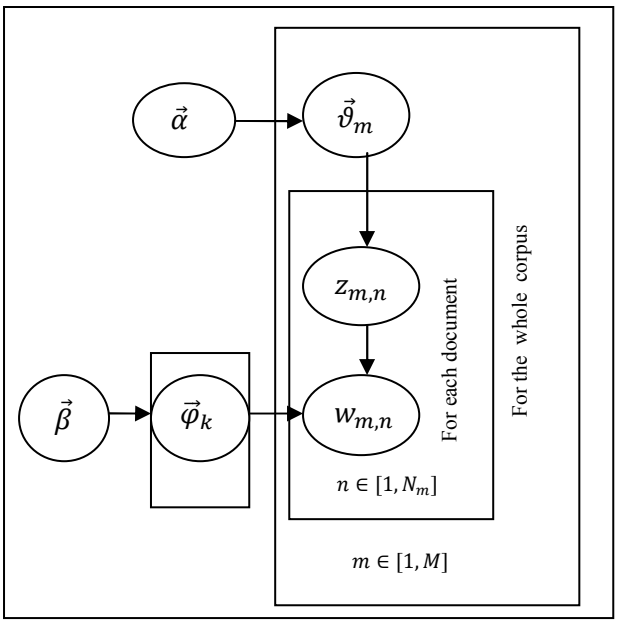

Figure 1. Document generation process in LDA

We can write the likelihood of the full data collection according to the generation process in Fig. 1 as:

$$p(W|\alpha,\beta) = \prod_{m=1}^{M} \iint p(v_m|a)p(\Phi|\beta).\prod_{n=1}^{N_m} p\left(w_{m,n}|v_m,\Phi\right)d\Phi \mathrm{d}\, v_m$$

where Φ is the term-topic matrix and $\alpha\ and\ \beta$ are the Dirichlet parameters.

Since the above function is a hyper geometric function that is infeasible to compute, approximate methods such as Variational Methods [1] and Gibbs Sampling [5] are used. In [1], a variational approximation to the log likelihood is used:

$$log\ p(W|\alpha,\beta) = \log \int_v \sum_z p(w|z,\beta)p(z|v)p(v;\alpha)q(v,z;\gamma,\varphi)dv$$

where a fully factorized variational distribution $q(v,z;\gamma,\varphi) = q(v;\gamma)\prod_n q(z_n;\varphi_n)$ parameterized by φ_n and γ is used. Here, $q(v;\gamma)$ is Dir (γ) and $q(z_n;\varphi_n)$ is Mult (φ_n).

In Griffiths and Stevyers (2004), the topic assignment of a word t depends on the topic assignment of all other word positions using the following multinomial distribution:

$$p(z_i = k|z_i{}',w) = \frac{n_{k,i'}^{(t)} + \beta_t}{\left[\sum_{v=1}^{V} n_k^{(v)} + \beta_v\right] - 1} \cdot \frac{n_{m,i'}^{(k)} + \alpha_k}{\left[\sum_{j=1}^{K} n_m^{(j)} + \alpha_j\right] - 1}$$

where $n_{k,i'}^{(t)}$ is the number of times word t is assigned to topic k except the current one, $\sum_{v=1}^{V} n_k^{(v)}$ is the total number of words assigned to topic k except the current one, $n_{m,i'}^{(k)}$ is the number of words in document m assigned to topic k except the current one and $\sum_{j=1}^{K} n_m^{(j)}$ is the total number of words in document m except the current one. Finally, φ_k is computed as follows:

$$\varphi_{k,t} = \frac{n_k^{(t)} + \beta_t}{\sum_{v=1}^{V} n_k^{(v)} + \beta_v}$$

After the above steps are completed, we calculate the entropies of the words by using term-topic matrices and select the words having smaller entropy values since they should be more discriminative for the text categorization task.

III. EXPERIMENTAL SETTINGS

For this work, we used SVM as the learning method, since in previous studies it was asserted that SVM is almost always a very good classifier in text categorization [4,8]. It is designed for solving binary classification problems by finding a hyper plane in n-dimensional space that separates positive and negative examples with the largest possible margin. By this way, the generalization error on unseen examples is minimized. We used the SVM-Light implementation with default parameter settings and a linear kernel.

We have performed experiments on two datasets (see Table II). Wap dataset is a skew dataset with 20 classes and very few training instances (1047 documents). Reuters-21578 dataset, a standard dataset in text categorization, has 90 classes and 9603 training instances after 'ModApte' splitting is applied.

TABLE II. SUMMARY OF THE USED DATASETS

Dataset	# of Training Documents	# of Test Documents	# of Classes	# of Terms
Wap	1047	513	20	8064
Reuters	9603	3299	90	20308

In all experiments, we have removed the stop words according to the stop words list of the SMART system. In addition, non-alphabetic characters are discarded, all letters are converted to lowercase and stemming is applied by means of the Porter's stemmer. For term weighting, we have used tf*idf weighting with length normalization [17].

We measured the results in terms of Micro-averaged and Macro-averaged F1-measures at different keyword selection points. The former reflects the overall performance better, while the latter is good at measuring the classifier's performance on rare categories since it gives equal weight to all classes regardless of the frequency of the class. We varied the number of keywords from 50 to 2000. We have not carried out experiments with more than 2000 keywords since we have seen in our preliminary experiments that F1 measures

generally reach their maximum values below 2000 keywords and then start to decline.

For LDA based on variational Bayes, we used the C implementation of LDA prepared by Daichi Mochihashi[1]. We preferred it over Blei's implementation since it was claimed to run much faster. For LDA based on Gibbs sampling, we used the Java version[2] of GibbsLDA++[3]. For both methods, the default parameters were used.

IV. RESULTS AND DISCUSSION

In this study, we carried out several experiments using different number of keywords and different feature selection metrics. Here we show the results graphically; the exact results can be found in Appendix A. In the tables in the appendix, the tendency of the changes in the accuracies as a function of the keyword number and the highest accuracy points should be regarded as more important than the absolute accuracy values. In this study, we carried out experiments to answer two primary questions:

- What should be the optimum number of hidden topics in LDA?
- Is LDA-based feature selection as successful as the classical metrics?

The answer to the first question is given in Figure 2. In this figure, we plot the result of the experiment in which we used the Wap dataset with 100 keywords and changed the hidden topic number from 10 to 500. As shown in Table II, the actual number of topics in Wap dataset is 20. We have not used less than 10 topics since it caused early convergence of LDA.

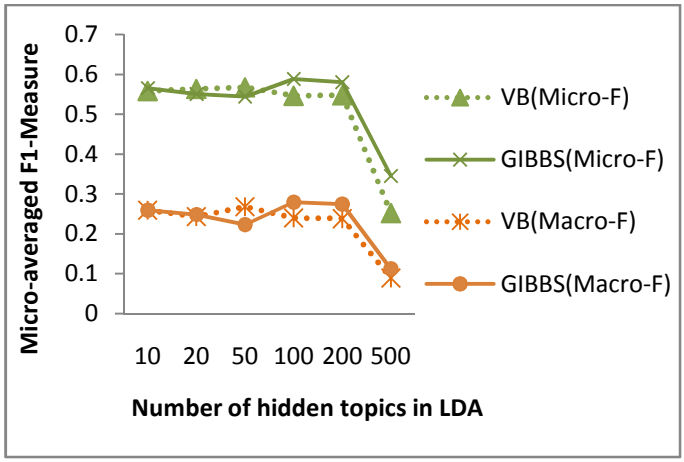

Figure 2. Micro- and Macro-averaged F-measures at 100 keywords for Wap dataset vs. the number of hidden topics in LDA

The first result that we can conclude from Figure 2 is that both variational Bayes (VB) and Gibbs sampling give similar results at low topic numbers while the latter performs better after 50 hidden topics. In addition we see that VB gives the best results at 50 topics for the Wap dataset while Gibbs gives the best results at 100 topics. Due to this reason, in later experiments we have used 50 topics for VB and 100 topics for Gibbs in this dataset. We have not performed this experiment for Reuters dataset since it has a very high time cost. Instead, we used 100 topics for Reuters which was shown to give satisfactory results in [16].

Figures 3 and 4 show the performances of different metrics on the Wap dataset. The figures show the superiority of Information Gain (IG) to other metrics at all keyword numbers. We also see that both VB and Gibbs perform similarly and better than Document Frequency Thresholding (DF). Moreover, their performances exceed Chi-square Statistics (CHI) at high keyword numbers. This is a particularly important observation since, despite their unsupervised nature, they can beat CHI which is a supervised metric. Another point is that they reach the performance of using all words at 2000 keywords. This indicates that we can obtain a 75% reduction in corpus size without any loss in performance by using these metrics.

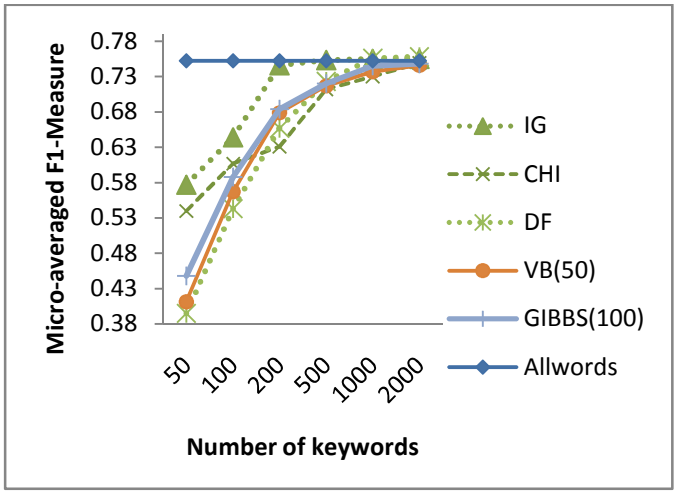

Figure 3. Micro-averaged F-measures for Wap Dataset

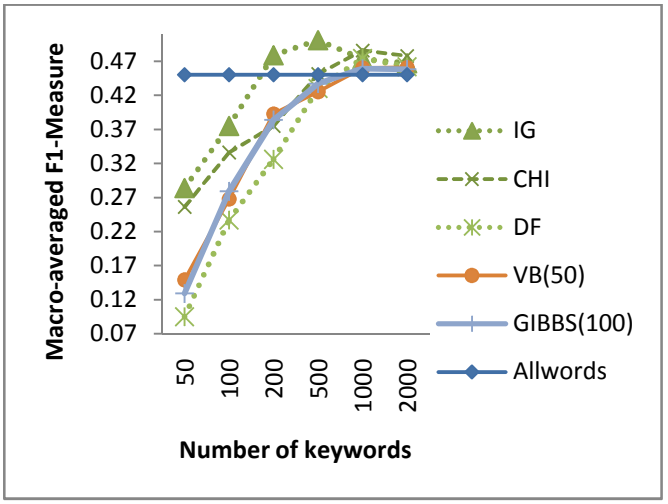

Figure 4. Macro-averaged F-measures for Wap Dataset

[1] http://chasen.org/~daiti-m/dist/lda/lda-0.1.tar.gz
[2] http://www.arbylon.net/projects/LdaGibbsSampler.java
[3] http://gibbslda.sourceforge.net/

In Figures 5 and 6, we see the results for the Reuters dataset. Reuters is given for comparison with previous studies since it is one of the most popular datasets in the text categorization community. The results in Reuters dataset are similar to the results in the Wap dataset except that in our experiments Gibbs is slightly better than VB. This is especially remarkable in the micro-averaged F-measure results. In fact, the performance of Gibbs is acceptable but VB performs even worse than DF. We can probably conclude from this observation that VB is not a suitable method for tasks that contain a very large number of features. In this experiment also, we observe that all feature selection metrics reach the performance of using all words at 2000 keywords. This signals approximately 90% reduction in corpus size without any performance tradeoff.

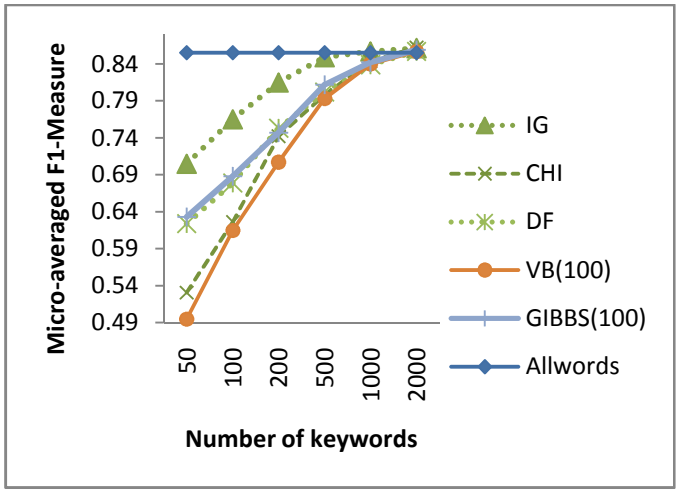

Figure 5. Micro-averaged F-measures for Reuters Dataset

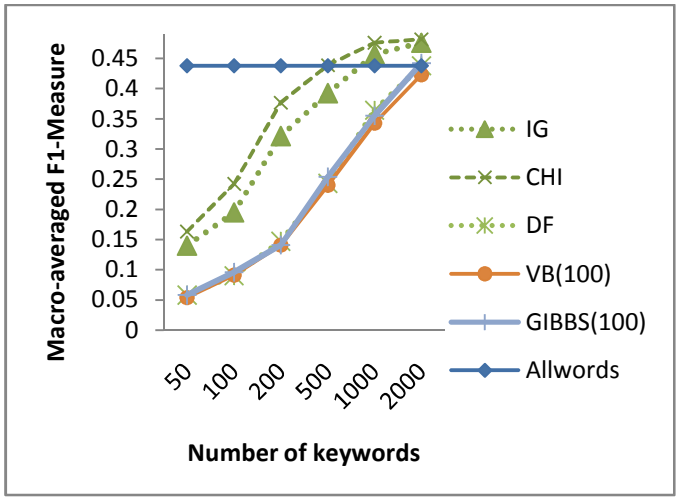

Figure 6. Macro-averaged F-measures for Reuters Dataset

V. CONCLUSIONS AND FUTURE WORK

VI. In this study, we have evaluated the performance of LDA-based feature selection by comparing it with well-known metrics. We used two variations of LDA, namely LDA with Gibbs sampling and LDA with variational Bayes. In these experiments, we observed that they cannot reach the performance of Information Gain in any settings. However, despite their unsupervised nature, they can perform comparably to the well-known supervised metric Chi-square and they are generally better than the simple unsupervised metric DF thresholding. So they can be used reliably when we have a training dataset for which we do not have available the class labels. Furthermore, as the number of keywords increases to about 2000, they can perform as good as using all the terms without any feature selection.

REFERENCES

[1] D. M. Blei, A. Ng, and M. I. Jordan, "Latent Dirichlet Allocation", *JMLR*, Vol.3, 2003, pp.993-1022.

[2] Jan Bakus, Mohamed S. Kamel: Higher order feature selection for text classification. Knowledge Information Systems 9(4): 468-491 (2006).

[3] F. Debole & F. Sebastiani. Supervised Term Weighting for Automated Text Categorization. In: Proceedings of SAC-03, 18th ACM Symposium on Applied Computing. ACM Press (2003) 784–788

[4] George Forman. An extensive empirical study of feature selection metrics for text classification, The Journal of Machine Learning Research, 3, 3/1/2003.

[5] T. L. Griffiths and M. Steyvers, "Finding scientific topics", *The National Academy of Sciences*, Vol.101, 2004, pp.5228-5235.

[6] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. Journal of Machine Learning Research, 3:1157–1182, 2003.

[7] T. Hofmann, "Probabilistic LSA", *Proc. UAI*, 1999.

[8] T. Joachims. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In European Conference on Machine Learning (ECML) (1998)

[9] R. Kohavi and George. H. John. Wrappers for feature subset selection. Artificial Intelligence, 97:273–324, 1997.

[10] Dunja Mladenic and Marko Grobelnik. Feature Selection for Unbalanced Class Distribution and Naïve Bayes. In Proceedings of the Sixteenth International Conference on Machine Learning (ICML), pages 258-267, 1999.

[11] A. Özgür, T. Güngör. Classification of Skewed and Homogeneous Document Corpora with Class-Based and Corpus-Based Keywords, Lecture Notes in Artificial Intelligence, Vol.4314, 2007, p.91-101, Springer-Verlag, Berlin Heidelberg.

[12] F. Sebastiani. Machine learning in automated text categorization. ACM Computing Surveys, 34(1):1-47, 2002.

[13] S. Tasci, T. Güngör, Evaluation of Feature Selection Metrics and Policies in Text Categorization, ISCIS, 2008.

[14] Yiming Yang , Xin Liu, A re-examination of text categorization methods, Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, p.42-49, August 15-19, 1999, Berkeley, California, United States .

[15] Yiming Yang , Jan O. Pedersen, A Comparative Study on Feature Selection in Text Categorization, Proceedings of the 14th International Conference on Machine Learning, p.412-420, July 08-12, 1997.

[16] Zhiwei Zhang, Xuan Hieu Phan, Susumu Horiguchi: An Efficient Feature Selection Using Hidden Topic in Text Categorization. AINA Workshops 2008: 1223-1228.

[17] Gerard Salton, Chris Buckley. Term Weighting Approaches in Automatic Text Retrieval, Cornell University, Ithaca, NY, 1987.

APPENDIX A. RESULTS OF EXPERIMENTS FOR WAP AND REUTERS DATASETS

# of Topics	10	20	50	100	200	500
VB(Micro-F)	0.558	0.563	0.567	0.546	0.547	0.252
GIBBS(Micro-F)	0.565	0.550	0.544	0.588	0.580	0.345
VB(Macro-F)	0.259	0.243	0.268	0.240	0.238	0.089
GIBBS(Macro-F)	0.259	0.248	0.223	0.279	0.274	0.112

Table A1. Micro- and Macro-averaged F-measures at 100 keywords for Wap dataset by varying the number of hidden topics in LDA

Micro-F	**50**	**100**	**200**	**500**	**1000**	**2000**	**All**
IG	0.577	0.644	0.746	0.753	0.755	0.755	0.752
CHI	0.540	0.607	0.631	0.712	0.730	0.749	0.752
DF	0.395	0.543	0.657	0.723	0.756	0.758	0.752
VB(50)	0.411	0.567	0.679	0.717	0.737	0.746	0.752
GIBBS(100)	0.448	0.588	0.684	0.720	0.745	0.748	0.752
Macro-F	**50**	**100**	**200**	**500**	**1000**	**2000**	**All**
IG	0.284	0.375	0.479	0.501	0.473	0.467	0.450
CHI	0.256	0.336	0.375	0.451	0.486	0.478	0.450
DF	0.095	0.237	0.326	0.430	0.474	0.462	0.450
VB(50)	0.149	0.268	0.393	0.425	0.460	0.461	0.450
GIBBS(100)	0.129	0.279	0.384	0.437	0.459	0.458	0.450

Table A2. Micro- and Macro-averaged F-measures for Wap dataset for varying keyword numbers

Micro-F	50	100	200	500	1000	2000	All
IG	0.705	0.765	0.815	0.849	0.857	0.861	0.855
CHI	0.531	0.626	0.742	0.798	0.844	0.862	0.855
DF	0.624	0.679	0.753	0.802	0.839	0.857	0.855
VB(100)	0.495	0.615	0.707	0.793	0.840	0.857	0.855
GIBBS(100)	0.633	0.688	0.747	0.812	0.842	0.859	0.855
Macro-F	**50**	**100**	**200**	**500**	**1000**	**2000**	**All**
IG	0.140	0.195	0.321	0.392	0.457	0.476	0.438
CHI	0.163	0.242	0.377	0.439	0.476	0.482	0.438
DF	0.058	0.090	0.147	0.243	0.364	0.438	0.438
VB(100)	0.054	0.091	0.141	0.240	0.343	0.423	0.438
GIBBS(100)	0.058	0.096	0.141	0.254	0.355	0.442	0.438

Table A3. Micro- and Macro-averaged F-measures for Reuters dataset for varying keyword numbers