An Evaluation of Existing and New Feature Selection Metrics 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 widely used for organizing and manipulating the documents in the electronic medium. Since the data in text categorization field are high-dimensional, feature selection is crucial to make the task more efficient and precise. In this paper, we make an extensive evaluation of the feature selection metrics used in text categorization by using local and global policies. For the experiments, we use three datasets which vary in size, complexity and skewness. We use SVM as the classifier and tf-idf weighting for term weighting. We observed that almost in all metrics, local policy outperforms when the number of keywords is low and global policy outperforms as the number of keywords increases.**

In addition to the evaluation of the existing feature selection metrics, we propose new metrics, which have shown high success rates especially in datasets with a low number of keywords. Moreover, we propose a keyword selection policy called *Adaptive Keyword Selection* (AKS). It is based on selecting different number of keywords for different classes and it improved the performance significantly in skew datasets.

I. INTRODUCTION

In recent years, the amount of available documents in the electronic medium such as electronic books, digital libraries and email messages increased rapidly. Therefore, the task of organizing and manipulating these resources has gained more importance and has become more difficult. For this task, many machine learning and information retrieval methods have been proposed and promising results were obtained by some of these methods.

Text categorization is the task of automatically assigning documents to some predefined categories. For text categorization, supervised machine learning techniques such as bayesian methods, decision trees, neural networks and support vector machines (SVM) are widely used.

In text classification, one typically uses a 'bag of words' model where each position in the input feature vector corresponds to a given word or phrase. Since generally there are thousands of words in a document corpus, the data is of very high dimensionality. This high dimensionality is an important challenge for the learner. Therefore, generally feature selection is applied to this high dimensional data for eliminating some of the dimensions without decreasing the categorization accuracy. Moreover, feature selection can help to prevent overfitting which is seen in very high dimensional data.

This paper presents the study of four popular feature selection metrics with both local and global policies. In local policy, each category has a different set of keywords while in global policy the reduced feature set is the same for all categories. Local policy helps us to find the most important terms for each class, while global policy favors the prevailing classes and gives penalty to classes with small number of training documents.

We have used three different datasets in our experiments that vary in size, complexity and class skewness (class imbalance). Thus, they reveal the behaviour of the metrics with different kinds of datasets. Especially, since high skewness in the document collection is a major difficulty for text categorization, we conducted experiments with skew datasets such as Wap and Reuters. Results have shown that feature selection is especially important in such datasets and local policy performs better than the global policy. In homogenous datasets, both policies have similar performances and global policy can achieve higher accuracies for a large number of keywords.

We also introduce some new feature selection metrics that are at least as good as the well-known metrics in all datasets. In some datasets such as Wap and Hitech, we have seen that they are better than the existing metrics. In addition, these new metrics have shown high performances at a small number of keywords such as 100 keywords. This makes them invaluable especially when the practitioner is constrained to use a small number of keywords. Another new method called *Adaptive*

Keyword Selection (AKS) which selects different number of keywords for classes that have different sizes has shown significant improvements on datasets that have a limited number of training instances.

For the experiments we use SVM as the learning method, since it was shown by different studies that it is one of the best classifiers for text categorization. We use SVM-Light package with default parameters and a linear kernel [4].

II. RELATED WORK

Text categorization is the task of finding the categories of some unlabeled documents by using a labeled training set of documents. Therefore, most of the machine learning algorithms such as SVMs, neural networks, Naive Bayes and k-nearest neighbor can be used for this task. There are several studies in the literature where these learning algorithms have been compared, e.g. [5,7]. It was found that SVM is generally the top performer in text classification [2,4,9].

Regardless of the learning algorithm, text classification is a quite hard problem since the dimensionality of the data is very high. Due to this reason, feature selection is a fundamental issue in text classification problems. There are numerous studies on feature selection which evaluate and compare most of the popular feature selection metrics [3,8]. In the experiments conducted in these studies, there are many variations of the parameters, such as dataset selection, policy used, classifier algorithm and so on.

In the study of Yang and Pedersen, five of the popular feature selection metrics are evaluated on the Reuters and Ohsumed datasets [8]. In this study, they use kNN and LLSF as the classifiers instead of SVM. In a later study, they also consider SVM and compare it with other classifiers[14]. However, both of these studies are based on feature selection with global policy and local policy is not considered.

Forman [3] considers local policy and gives a comprehensive evaluation of many feature selection metrics. SVM is used as the classifier and different types of datasets including skew datasets as well as homogenous ones are considered. However, despite the diversity in the datasets and the metrics, this study also lacks the comparison of local and global policies. Since the experimantal settings in these two studies are different, it is not possible to utilize the results for a comparison of local and global policies.

A study that includes the comparison of local and global policies is given by Debole and Sebastiani [1]. In this study, they focus on term weighting using the feature selection scores and thus they do not give a detailed comparison. In addition, they only use the Reuters dataset and it is hard to generalize the results to other datasets with different class sizes and skewness.

Ozgur and Gungor [6] analyzes two keyword selection policies named as class- and corpus-based keyword selection by using SVM on datasets of different skewness and sizes. However, they only use the metric 'tf-idf keyword selection' and do not consider the popular feature selection metrics.

In addition, there are studies in which new keyword selection metrics are proposed. One such example is the study of Forman[3] where a method called Bi-normal Seperation (BNS), which is especially successful in high-skew datasets, is proposed. Another example is Gain Ratio (GR), which is acquired by normalizing IG score of a term by its entropy.

III. FEATURE SELECTION METRICS

In this study, four popular feature selection metrics are analyzed. For obtaining the global versions of the local metrics, we use the globalization technique in which the global score of a term is calculated from its local scores:

$$f_{\max}(t_k) = \max_{i=1}^{m} f(t_k, c_i) \tag{1}$$

where t_k is a term, c_i is a class, m is the number of classes and the function $f(\ )$ denotes the score of a term.

A. Existing Metrics

1) Information Gain (IG): Information Gain 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:

$$\begin{aligned} IG(t) = & -\sum_{i=1}^{m} P_r(c_i) \log P_r(c_i) \\ & + P_r(t) \sum_{i=1}^{m} P_r(c_i|t) \log P_r(c_i|t) \\ & + P_r(\bar{t}) \sum_{i=1}^{m} P_r(c_i|\bar{t}) \log P_r(c_i|\bar{t}) \end{aligned} \tag{2}$$

where $P_r(c_i)$ is the probability of a document to have class label c_i, $P_r(t)$ is the probability of a term t to appear in a document, $P_r(c_i|t)$ is the probability of a document to have class label c_i given that term t appears in the document and $P_r(c_i|\bar{t})$ is the probability of a document to have class label c_i given that term t does not appear in the document.

2) Chi-square Statistics (CHI): Chi-square test is applied to test the independence of two random variables. In the domain of text categorization, the two random variables are the occurrence of the term t and the occurrence of the class c:

$$\chi^2(t,c) = \frac{n(AD-CB)^2}{(A+C)(B+D)(A+B)(C+D)} \quad (3)$$

where A and C denote the number of documents in class c in which term t, respectively, appears and does not appear; B and D denote the number of documents in other classes in which term t, respectively, appears and does not appear. n is the total number of documents.

3) Document Frequency Thresholding (DF): This method is based on the assumption that infrequent terms are not reliable and effective in the category prediction. Document frequency refers to the number of documents in which a term appears and the method favors the terms whose document frequencies are the highest:

$$DF(t) = n_t \quad (4)$$

where n_t is the number of documents in which term t appears.

4) Accuracy2 (Acc2): This metric is based on the difference of the distributions of a term in the documents belonging to a class and the documents not belonging to that class. It was first studied by Forman [3]:

$$Acc2(t,c) = \frac{A}{n_1} - \frac{B}{n_2} \quad (5)$$

where A and B are as defined in CHI; n_1 and n_2 correspond to the number of documents, respectively, belonging to class c and not belonging to class c.

B. Proposed Metrics

1) M_1 Method: This metric is a different version of Acc2. In Acc2, only the number of documents in which the term occurs is taken into account without considering the number of actual occurrences of the term in the documents. In this method, we multiply two scores: the score calculated by using the number of documents in which the term occurs and the score calculated by using the actual occurrences of the term in the documents:

$$M_1(t,c) = \left[\frac{A}{n_1} - \frac{B}{n_2}\right] \cdot \left[\frac{C}{t_1} - \frac{D}{t_2}\right] \quad (6)$$

where A, B, n_1 and n_2 are as defined in Acc2; C and D are the number of the occurrences of the term t in the documents, respectively, belonging to class c and not belonging to class c; t_1 and t_2 correspond to the number of terms, respectively, in class c and in other classes.

2) M_2 Method: This method is similar to the Acc2 method, but we measure the correlation between a term and a class in a different way. Here we take the documents in the whole corpus in which the term appears as a group and we find the proportion of the documents with class label c in this group:

$$M_2(t,c) = DF(t)\left[\frac{A}{d_1} - \frac{C}{d_2}\right] \quad (7)$$

where A and C are as defined in CHI; d_1 and d_2 correspond to the number of documents in which, respectively, the term occurs and the term does not occur.

3) M_3 Method: This method is simply the multiplication of the M_1 score of a term with the document frequency of the term. The M_1 method alone does not consider the document frequency; it may give equal weights to frequent and rare terms:

$$M_3(t,c) = DF(t) \cdot M_1(t,c) \quad (8)$$

4) M_4 Method: In the experiments, we have observed that despite the fact that the M_1 Method gives very good results for a low number of keywords, it is not as good as global methods when the number of keywords increases. For handling the deficiency of the M_1 Method at a high number of keywords, we select the first n keywords by the M_1 Method, where n is the number of documents in that class. Then we select the remaining keywords from the list of keywords found by the global IG metric.

C. Adaptive Keyword Selection (AKS)

In this method, we apply the idea that different classes in a dataset may require different number of keywords for the best accuracy. For determining the number of keywords for each class, we divided the classes into groups with respect to the number of documents they contain. Then we carried out several tests to determine the best number of keywords for each group. It may not be the optimal solution but a simple one that can be improved in later studies.

Below is the keyword number selection procedure for a class with respect to its training document size, where n represents the number of documents in the training set of the class:

$$Use \begin{cases} 100 \text{ keywords} & n > 0 \text{ and } n \le 15 \\ 20 \text{ keywords} & n > 15 \text{ and } n \le 30 \\ 100 \text{ keywords} & n > 30 \text{ and } n \le 100 \\ 500 \text{ keywords} & n > 100 \text{ and } n \le 200 \\ 1000 \text{ keywords} & n > 200 \end{cases} \text{, if} \quad (9)$$

Basically, it selects more keywords as the document number in a class increases. The only exception is for classes that have less than 15 examples. The reason may be that for a class that has such a low number of documents, a few reliable keywords describing the class cannot be determined. Therefore, we have to use more keywords for a better classification. As can be seen in the next section, AKS strategy increased the results of most keyword selection metrics in skew datasets.

III. EXPERIMENTAL SETTINGS

In this study, we used SVM as the learning method, which is reported as a top classifier in text categorization consistently in previous studies. We used the SVM-Light implementation with default parameter settings and a linear kernel[10].

We performed experiments on three datasets with different characteristics: Wap dataset is a skew dataset with 20 classes and very few training instances (1047 documents). Hitech dataset is homogenous and it is easier compared to the Wap dataset, because it has only six classes all of which contain sufficient training instances. Finally, Reuters-21578 dataset, a standard in text categorization, is used. It has 90 classes and 9603 training instances after 'ModApte' splitting is applied.

In all experiments, we have removed the stop words according to the stop words list of the SMART system[11]. In addition, non-alphabetic characters are discarded, all letters are converted to lowercase and stemming is applied by means of the Porter's stemmer[12]. For term weighting, we have used tf-idf weighting with length normalization. We have measured the results in terms of Micro- and Macro-averaged F1-measures at different keyword selection points. The former reflects the overall accuracy 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.

IV. RESULTS AND DISCUSSION

In this study, we carried out several experiments with local and global policies using keyword numbers ranging from 30 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 remain constant or start to decline.

Tables 1-3 show the Micro- and Macro-averaged F-measures in the datasets, using the old feature selection metrics as well as the proposed ones. The local and global policies of the previous metrics are denoted by *(l)* and *(g)*, respectively, in the tables.

A. General Observations

When we consider the results in the tables, we see that in skew datasets (Wap and Reuters), the difference among the Micro- and Macro-averaged F-measures is higher compared to the homogenous dataset Hitech. This situation can be explained by the fact that Macro-averaged F-measure gives equal weights to all classes while Micro-averaged F-measure gives equal weight to each document. Therefore, rare classes which are difficult to classify decrease the Macro-averaged F-measure.

Another observation about skew datasets is that Macro-averaged F-measure reaches its maximum value at about 100 keywords with local policy. Probably this situation is related with the number of training documents. Rare classes which have only a few documents in the training corpus are best classified by using a small number of keywords that are selected locally. It is not possible to find many reasonable keywords for a class when we do not have enough training data for it. Since Macro-averaged F-measure is highly affected by the success of the classifier in rare classes, the success increases if we achieve to classify rare classes more successfully.

When we compare the existing metrics, we observe that most of the time IG gives the best results independent of the dataset. The CHI method can also be regarded as successful and it achieves accuracies comparable to that of the IG method. On the other hand, despite its success at a high number of keywords, DF is not comparable to IG and CHI when the number of keywords is low. This may be due to the fact that when there are many keywords, most important features (i.e. those that are the best discriminators of classes) are almost always selected. However, since DF has a very simple logic, when we use a very small number of keywords such as 30 or 50, feature selection by the DF method may ignore some of the major keywords, which in turn deteriorates the results significantly.

We also see that while global policy is better than local policy at a large number of keywords, it is generally beaten by local policy when the number of keywords is lower than 1000. This indicates that when we select a small number of keywords, global policy cannot identify the keywords that can represent all the classes well.

B. Analysis of the Proposed Methods

The success of local policy at a low number of keywords motivated us to concentrate on local feature selection methods. All of the methods that are proposed in this paper are local methods. In fact, methods M_2, M_3 and M_4 are modified versions of the M_1 Method. But we include all of them since each one shows different behavior in different environments. For instance, the M_3 Method is the best method in Wap dataset while the M_4 Method is the best one for Hitech dataset.

In Wap dataset, the M_1 Method reaches 76.7% Micro- and 59.4% Macro-averaged F-measures, while IG can reach at most

75.2% Micro- and 54.8% Macro-averaged F-measures with local policy. In addition, the IG method reaches its highest values at different number of keywords while 100 keywords is an optimal choice for the M_1 Method for both measures. The success of the proposed methods in Wap dataset shows that they are very good at finding the best features even when a class does not have too many training documents.

Micro-F	30	50	100	500	1000	2000	All
IG(l)	0.610	0.617	0.638	0.654	0.644	0.638	0.649
IG(g)	0.523	0.559	0.621	0.645	0.649	0.666	0.649
CHI(l)	0.590	0.620	0.631	0.636	0.619	0.632	0.649
CHI(g)	0.559	0.597	0.621	0.633	0.651	0.667	0.649
Acc2(l)	0.612	0.636	0.649	0.651	0.659	0.646	0.649
Acc2(g)	0.581	0.575	0.606	0.657	0.642	0.661	0.649
DF(l)	0.550	0.578	0.624	0.622	0.644	0.661	0.649
DF(g)	0.546	0.538	0.583	0.609	0.624	0.629	0.649
M_4	**0.625**	0.637	0.658	**0.656**	**0.666**	**0.673**	0.649
M_3	0.610	0.637	0.638	0.652	0.652	0.653	0.649
M_2	0.617	**0.638**	0.645	0.645	0.655	0.645	0.649
M_1	0.623	0.630	**0.657**	0.648	0.655	0.629	0.649
Macro-F	**30**	**50**	**100**	**500**	**1000**	**2000**	**All**
IG(l)	0.529	0.539	0.577	0.591	0.573	0.557	0.558
IG(g)	0.433	0.461	0.538	0.572	0.597	0.602	0.558
CHI(l)	0.495	0.536	0.572	0.551	0.545	0.567	0.558
CHI(g)	0.437	0.509	0.528	0.570	0.610	0.605	0.558
Acc2(l)	0.522	0.550	0.571	**0.596**	0.600	0.593	0.558
Acc2(g)	0.507	0.496	0.521	0.567	0.582	0.603	0.558
DF(l)	0.485	0.507	0.549	0.549	0.592	0.603	0.558
DF(g)	0.389	0.383	0.461	0.510	0.524	0.532	0.558
M_4	**0.553**	**0.578**	0.582	0.590	**0.615**	**0.615**	0.558
M_3	0.533	0.563	0.559	0.585	0.581	0.586	0.558
M_2	0.527	0.546	0.568	0.594	0.597	0.588	0.558
M_1	0.542	0.556	**0.594**	0.578	0.600	0.561	0.558

Table 1. Micro- and Macro-averaged F-measures for Hitech Dataset

Micro-F	30	50	100	500	1000	2000	All	AKS
IG(l)	0.735	0.750	0.742	0.744	0.742	0.749	0.752	0.769
IG(g)	0.526	0.577	0.644	0.753	0.755	0.755	0.752	-
CHI(l)	0.714	0.732	0.732	0.736	0.742	**0.758**	0.752	0.751
CHI(g)	0.523	0.540	0.607	0.712	0.730	0.749	0.752	-
Acc2(l)	0.728	0.757	0.770	0.755	0.752	0.752	0.752	**0.795**
Acc2(g)	0.476	0.529	0.629	0.730	0.743	**0.758**	0.752	-
DF(l)	0.567	0.704	0.751	0.747	0.760	0.747	0.752	0.777
DF(g)	0.341	0.395	0.543	0.723	0.756	0.758	0.752	-
M_4	**0.739**	**0.769**	0.765	**0.761**	0.755	0.754	0.752	0.790
M_3	0.701	0.735	**0.776**	**0.761**	**0.758**	0.750	0.752	0.793
M_2	0.732	0.738	0.757	0.759	0.756	0.753	0.752	0.774
M_1	0.738	0.762	0.767	0.750	0.747	0.748	0.752	0.790
Macro-F	**30**	**50**	**100**	**500**	**1000**	**2000**	**All**	**All**
IG(l)	0.531	0.548	0.517	0.508	0.460	0.482	0.450	0.545
IG(g)	0.185	0.284	0.375	0.501	0.473	0.467	0.450	-
CHI(l)	0.511	0.520	0.509	0.491	0.475	0.491	0.450	0.538
CHI(g)	0.239	0.256	0.336	0.451	0.486	0.478	0.450	-
Acc2(l)	0.554	0.564	0.551	0.516	0.488	0.485	0.450	0.574
Acc2(g)	0.235	0.278	0.411	0.489	0.480	**0.492**	0.450	-
DF(l)	0.353	0.513	0.550	0.483	0.524	0.481	0.450	0.587
DF(g)	0.053	0.095	0.237	0.430	0.474	0.462	0.450	-
M_4	0.554	**0.572**	0.558	0.497	0.471	0.467	0.450	0.585
M_3	0.497	0.546	0.577	**0.522**	**0.495**	0.481	0.450	**0.590**
M_2	**0.562**	0.539	0.558	0.519	0.492	0.486	0.450	0.539
M_1	0.559	0.551	**0.594**	0.520	0.488	0.482	0.450	0.630

Table 2. Micro- and Macro-averaged F-measures for Wap Dataset

Micro-F	30	50	100	500	1000	2000	All	AKS
IG(l)	0.820	0.838	0.842	0.850	0.856	0.856	0.855	0.858
IG(g)	0.661	0.705	0.765	0.849	0.857	0.861	0.855	-
CHI(l)	**0.823**	**0.840**	0.842	0.845	0.852	0.854	0.855	0.853
CHI(g)	0.367	0.531	0.626	0.798	0.844	**0.862**	0.855	-
Acc2(l)	0.811	0.835	0.846	0.860	**0.862**	0.859	0.855	0.863
Acc2(g)	0.388	0.513	0.622	0.814	0.832	0.860	0.855	-
DF(l)	0.802	0.820	0.841	0.854	0.859	0.859	0.855	0.862
DF(g)	0.542	0.624	0.679	0.802	0.839	0.857	0.855	-
M_4	0.815	0.823	0.852	0.861	0.857	0.861	0.855	0.861
M_3	0.803	0.819	0.846	**0.863**	0.861	0.860	0.855	0.862
M_2	0.815	0.828	0.847	0.861	0.861	0.860	0.855	0.864
M_1	0.817	0.835	**0.854**	0.858	0.861	**0.862**	0.855	**0.866**
Macro-F	**30**	**50**	**100**	**500**	**1000**	**2000**	**All**	**AKS**
IG(l)	0.530	0.512	0.517	0.495	0.493	0.490	0.438	0.527
IG(g)	0.099	0.140	0.195	0.392	0.457	0.476	0.438	-
CHI(l)	0.491	0.493	0.500	0.493	0.493	0.491	0.438	0.497
CHI(g)	0.107	0.163	0.242	0.439	0.476	0.482	0.438	-
Acc2(l)	0.525	0.524	0.527	**0.513**	**0.500**	0.489	0.438	0.531
Acc2(g)	0.113	0.145	0.215	0.484	0.488	0.490	0.438	-
DF(l)	0.497	0.515	**0.539**	0.511	0.500	0.493	0.438	**0.538**
DF(g)	0.034	0.058	0.090	0.243	0.364	0.438	0.438	-
M_4	0.485	0.477	0.491	0.491	0.472	0.478	0.438	0.499
M_3	0.459	0.495	0.506	0.506	0.498	0.489	0.438	0.499
M_2	**0.531**	0.519	0.529	**0.513**	0.499	0.489	0.438	0.531
M_1	0.512	**0.531**	0.529	0.505	0.496	**0.494**	0.438	0.535

Table 3. Micro- and Macro-averaged F-measures for Reuters Dataset

In Hitech dataset, the M_4 Method reaches 67.3% Micro- and 61.5% Macro-averaged F-measures while CHI, the best method in this dataset, can achieve at most 66.7% Micro- and 61.0% Macro-averaged F-measures. Again, the gap enlarges when the number of keywords is decreased.

In Reuters dataset, the proposed methods are again more successful than the previous methods. However, they do not increase the success rates significantly in this dataset. Nevertheless, these new methods can still be regarded as successful, since we have observed that they are at least as good as the existing methods regardless of the dataset.

Another remarkable property of the proposed methods is that they reach their maximum values or at least give satisfactory results about 100 keywords. This may be explained by the fact that all of them are based on local policy. Nevertheless, in Hitech dataset we see that the M_4 Method preserves its success rate when the number of keywords is increased from 100 to 2000, although the other three methods are not successful when the keyword number is high. This situation is expected since the M_4 Method uses some of the keywords found by the IG(g) method. Therefore, it is affected by the success of IG(g) at a large number of keywords.

The last columns of Tables 2 and 3 display the results of Adaptive Keyword Selection on Wap and Reuters datasets. We have not carried out experiments on Hitech dataset, since it is a reasonable strategy only if the dataset is skew. When we look at the tables, we see that AKS improves the results of almost all keyword selection metrics in Wap dataset, while it improves the results slightly in Reuters dataset. In Wap dataset, the M_1 Method with AKS improves the Micro- and Macro-averaged F-

measures up to 79.0% and 63.0%, respectively, which were under 76.0% and 55.0% with IG or CHI. This indicates that if we find the optimal number of keywords for each class, AKS can be very valuable for skew datasets that have a small number of training instances.

In addition, it has a high value in both Micro- and Macro-averaged F-measures. This is particularly important since no other method has proved the best in both of the F-measures at the same time. For instance, if we consider IG (l) at 2000 keywords for Reuters dataset, Micro-averaged F-measure is quite high (85.6%) but Macro-averaged F-measure is only 49.0%. On the other hand, if we select 100 keywords, Macro-averaged F-measure increases to 51.7% but Micro-averaged F-measure decreases to 84.2%. When we use AKS strategy, IG (l) Micro- and Macro-averaged F-measures are both at their highest values (85.8% and 52.7%, respectively). This situation is a consequence of its success in classifying both rare and common classes correctly.

V. CONCLUSIONS AND FUTURE WORK

In this study, we have made an extensive study of the feature selection metrics in text categorization with SVM as the classifier. We have compared some of the well-known feature selection metrics such as IG, CHI and DF-Thresholding by varying the number of selected features from 30 to 2000 and also compared the local and global policies on each metric. In the experiments, we have used three datasets with different skewness, size and complexity.

We have also introduced some new feature selection metrics that are at least as good as the well-known metrics in all datasets. In some datasets such as Wap and Hitech, we have seen that they are better than the existing metrics. In addition, these new metrics have shown high performances especially at a small number of keywords such as 100 keywords. This makes them invaluable when the practitioner is constrained to use a small number of keywords.

Another contribution of this study is a new feature selection policy called Adaptive Keyword Selection which selects different number of keywords for classes that have different sizes. It has shown significant improvements especially with datasets that have a limited number of training instances.

Future work includes the experiments of the proposed feature selection metrics with other term weighting approaches such as Supervised Term Weighting [1, 13] and learning algorithms apart from Support Vector Machines. In addition, Adaptive Keyword Selection can be extended to make it capable of adjusting the number of features automatically according to the properties of the dataset used.

REFERENCES

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

[2] Dumais, S., Platt, J., Heckerman, D., Sahami, M.: Inductive Learning Algorithms and Representations for Text Categorization. In Proceedings of the 17th International Conference on Information and Knowledge Management, pg. 148-155, Maryland, 1998.

[3] Forman, G.: An Extensive Empirical Study of Feature Selection Metrics for Text Classification. Journal of Machine Learning Research 3 (2003) 1289–1305

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

[5] Özgür, A.: Supervised and Unsupervised Machine Learning Techniques for Text Document Categorization. Master's Thesis (2004), Bogazici University, Turkey

[6] Özgür, A., Güngör, T.: 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.

[7] Yang, Y., Liu, X.: A Re-examination of Text Categorization Methods. In Proceedings of the Twenty-Second International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), pages 42-49, 1999.

[8] Yang, Y., Pedersen, J.O.: A Comparative Study on Feature Selection in Text Categorization. In: Proceedings of the 14th International Conference on Machine Learning (1997) 412–420

[9] Susan Dumais, John Platt, David Heckerman and Mehran Sahami. Inductive Learning Algorithms and Representations for Text Categorization. In Proceedings of the 17th International Conference on Information and Knowledge Management, pages 148-155, Maryland, 1998.

[10] http://svmlight.joachims.org/

[11] ftp://ftp.cs.cornell.edu/pub/smart/, 2004.

[12] http://www.tartarus.org/ martin/PorterStemmer/, 2004.

[13] Pascal Soucy, Guy W. Mineau: Beyond TFIDF Weighting for Text Categorization in the Vector Space Model. IJCAI 2005: 1130-113.