Empirical Study on Filter based Feature Selection Methods for Text Classification

Subhajit Dey Sarkar
M.Tech Final Year
Department of Computer Science and Engineering
Institute of Engineering and Management
West Bengal, India

Saptarsi Goswami
Asst. Professor
Department of Computer Science and Engineering
Institute of Engineering and Management
West Bengal, India

ABSTRACT
Text classification has become much more relevant with the increased volume of unstructured data from various sources. Several techniques have been developed for text classification. High dimensionality of feature space is one of the established problems in text classification. Feature selection is one of the techniques to reduce dimensionality. Feature selection helps in increasing classifier performance, reduce over filtering to speed up the classification model construction and testing and make models more interpretable. This paper presents an empirical study comparing performance of few feature selection techniques (Chi-squared, Information Gain, Mutual Information and Symmetrical Uncertainty) employed with different classifiers like naive bayes, SVM, decision tree and k-NN. Motivation of the paper is to present results of feature selection methods on various classifiers on text datasets. The study further allows comparing the relative performance of the classifiers and the methods.

Keywords
Feature Selection, Filter Method, High Dimensionality, Text Classification, Text Categorization.

1. INTRODUCTION
Text mining is the process of retrieving useful information from textual data. Textual data are now-a-days present in various forms like email, tweets, product reviews, chats etc. One of the major tasks of text mining is text classification which has variety of applications like detecting spam emails, topic categorization of news articles, sentiment analysis to name a few [1], [2], [3]. The demand of text classification for the management of text document has increased rapidly with the increase of textual data [4][5]. One of the major problems of text classification is the high dimensionality of feature space [6], [7]. This problem arises where a document is constructed as a "bag of words" (BoW) where every single word is used as feature which occurs in a document [8], [9]. Feature selection for text classification helps in reduction of dimensionality of feature space. The paper shows preference on feature selection over methods like feature transformation (Principle Component Analysis) because it retains the semantics of the data. Feature selection is applied to identify informative features which helps in the reduction of high dimensional of feature space [3], [10], [11]. The use of feature selection has the below three advantages (I) improving classification performance (II) to speed up the classification model construction & testing (III) Interpretability of the model will increase.

This paper represents an empirical study of four feature selection metrics namely :- (I) Chi-squared (II) Information gain (III) Mutual Information and (IV) Symmetrical Uncertainty. These are employed with four classifiers namely naive bayes, support vector machines (SVM), decision tree and k- nearest neighbors (k-NN) respectively.

The remainder of this paper is organized as follows. Section 2 reviews few related works. Section 3 involves discussion about the various feature selection methods. Section 4 describes the experiment setup. Section 5 provides a detail analysis on the experiment result. Section 6 concludes this paper.

2. RELATED WORK
Various studies have addressed text classification using different techniques to classify text documents, and different metrics to evaluate the accuracies of these techniques [10], [12], [13], [14]. At present most of the studies are based on feature selection, and are mainly focused on its different methods performance in statistical learning on text classification [11], [15], [16], [17]. The methods like Chi-squared, Information Gain (IG), Mutual Information (MI), Term Strength (TS) and Document frequency thresholding (DF) of feature selection are being used to determine the performance of information retrieval from the dataset [18]. Another method of feature selection, Gini Index is also used in information retrieval process [2], [15]. Previous feature selection studies show the chi-squared, information gain and mutual information more effective than the rest [19]. An empirical study on feature selection methods observes the inverse relation between the accuracy and feature reduction [13].Also study on feature selection based on odd ratio shows beneficial effects on naïve bayes while normal prediction from SVM shows better result [20]. This comparison shows the need of feature selection in text classification to obtain a better accuracy rate. Feature selection metrics used in our paper are Chi-squared, Information Gain, Mutual Information and Symmetrical Uncertainty. The use of symmetrical uncertainty metric for text classification is not used in previous studies.

3. FEATURE SELECTION METHODS

In machine learning and statistics, feature selection is the process of selecting a subset of relevant features from a large feature set, without compromising on the quality of the output. The paper focused on four different feature selection metrics which are Chi-squared, Information Gain, Mutual Information and Symmetrical Uncertainty.

An idea about the working procedure of this feature selection methods are discussed below:

Chi-squared (X^2 statistic):

Chi-squared is generally used to measure the lack of independence between t and c (where t is for term and c is for class or category) and compared to the X^2 distribution with one degree of freedom. The expression for X^2static is defined as:

$$X^2_{(t,c)} = \frac{D\times(PE-MQ)^2}{(P+M)\times(Q+N)\times(P+Q)\times(M+N)}$$

where D = total number of documents

P = the number of documents of class c containing term t

Q = the number of documents containing t occurs without c.

M= the number of documents class c occurs without t.

N=the number of documents of other class without t.

Information gain (IG):

Information gain also known as gain ratio is generally used for measuring the reduction in entropy required for category prediction by knowing the presence or the absence of a term or feature in the document. It is frequently used as a term goodness criterion in machine learning.

The following expression describes information gain:

$$\begin{aligned} IG &= -\sum_i \Pr(c_i)\, logPr(c_i) \\ &+ \Pr(t)\sum_i \Pr(c_i|t)\, logPr(c_i|t) \\ &+ \Pr(t)\sum_i \Pr(c_i|t)\, logPr(c_i|t) \end{aligned}$$

Mutual Information (MI):

Mutual information is a criterion commonly used in statistical language modeling of word association and related applications [21]. Mutual information between a term t and class c is defined by

$$MI(t,c) = log\frac{\Pr(t,c)}{\Pr(t)\Pr(c)}$$

To measure the global goodness of a term in feature selection, we combine the category specific scores of a term into two alternate ways

$$MI_{max}(t) = max_i MI(t,c_i)$$

$$MI_{max}(t) = \sum_i \Pr(c_i)\, MI(t,c_i)$$

Symmetrical uncertainty (SU):

The expression for symmetrical uncertainty is define as

$$SU(A,B) = 2\frac{IG(A|B)}{H(A)+H(B)}$$

where IG(A|B) is the information gain of the feature A which is an independent attribute and B belongs to an class attribute. Both H(A) and H(B) are entropy of feature A and B respectively [21]

4. EXPERIMENTAL SETUP

This section describes the detailed information about the experimental dataset and the various steps involved in analyzing the performance of the different classifier with and without using feature selection metrics. This section contains two part I) dataset information and II) methodology.

4.1 Data Set Information

Table-1 summarizes the characteristics of the dataset used in our experiment. A detail information of the datasets as follows:

Table-1 Characteristics of the datasets

Datasets	CNAE-9	SMS spam collection	Preview of hotel data
Number of documents	1080	5572	50
Number of terms	856	6632	3360
Numbers of categories	9	2	2

4.2 Methodology

The various steps used for the experiment are described below:

(i) Text document are stripped of space and punctuation.

(ii) Numbers and stop words are removed

(iii) Stemming is applied.

(iv) Term document matrix is prepared.

(v) The term document matrix is then split into two subsets. 70% of the term document matrix is used for training and 30% is held out for testing.

(vi) The performance of different classifier with and without feature selection was compared.

(vii) Classification Accuracy has been used as the basis of comparison.

(viii) The performance result of the construction for prediction of test model are mentioned in table 2

(ix) Feature selection helps in reduction of train data by selecting the features having non-zero value.

(x) The performance of the prediction model using different feature selection metrics employed in text classification is mentioned in table-3.

5. RESULT AND DISCUSSION

1. The accuracy rate obtained by applying the different classification algorithms on the data sets are discussed in table 2.

Table-2 Classification accuracy rate of classification algorithms

Classification algorithms	CNAE-9	SPAMHAM	Hotel dataset
Decision tree	46%	87%	47%
SVM	83%	90%	60%
Naive Bayes	19.8%	13%	40%
k-NN	77%	90%	53%

2. The individual accuracy rates obtained from different feature selection methods on the classifier are discussed in table-3. Different feature selection metrics are applied on the classifiers.

Table-3 Classification accuracy rate of classification algorithms using feature selection

CA	FS methods	CNAE-9	SPAM-HAM	Hotel dataset
Decision Tree	CHI	61%	93%	73%
	IG	63%	93%	73%
	MI	65%	93%	67%
	SU	54%	94%	60%
SVM	CHI	89%	95%	73%
	IG	89%	96%	93%
	MI	87%	96%	80%
	SU	87%	96%	80%
Naïve Bayes	CHI	53%	92%	60%
	IG	60%	92%	67%
	MI	52%	93%	60%
	SU	53%	92%	60%
k-NN	CHI	86%	96%	73%
	IG	85%	96%	80%
	MI	86%	94%	73%
	SU	89%	96%	73%

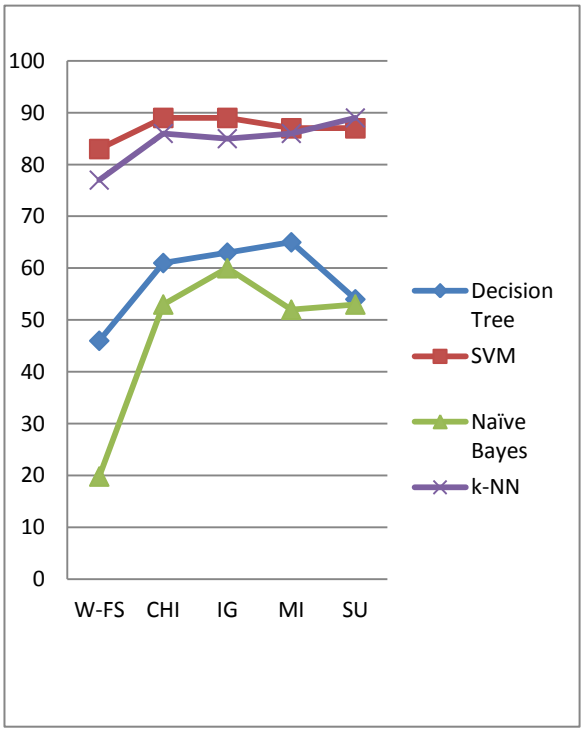
*CA= Classification algorithm,
*FS-=Feature Selection

3. Table-4 provides the average of the accuracy rates for all the dataset involved in the experiment obtained from table-2 and table-3. In table-4, without feature selection column provides the average of the classifiers accuracy rate obtained from table-2 and the rest of the column provides the average of the individual feature selection metrics accuracy rate obtained from table-3.

Table-4 Average performance of datasets for CNAE-9, spam ham and hotel database

CA	Without Feature Selection	Using CHI	Using IG	Using MI	Using SU
Decision Tree	60%	75.6%	76.3%	75%	69.3%
SVM	77.6%	85.6%	92.6%	87.6%	87.6%
Naïve Bayes	24.2%	68.3%	73%	68.3%	68.3%
k-NN	73.3%	85%	87.6%	84.3%	86%

*CA- Classification algorithm



*W-FS= Without Feature Selection

Figure-1 Performance of set accuracy rate involved in CNAE-9 dataset

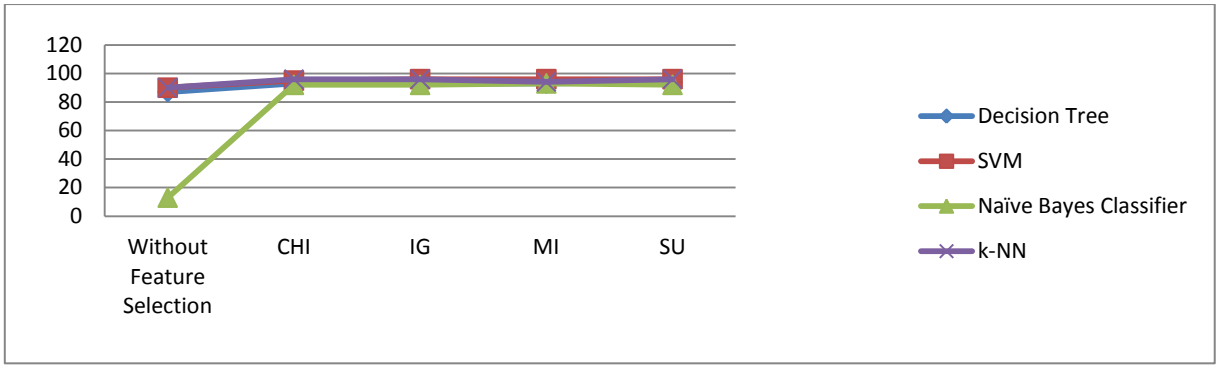


Figure-2 Performance of set accuracy rate involved in SMS spam collection dataset

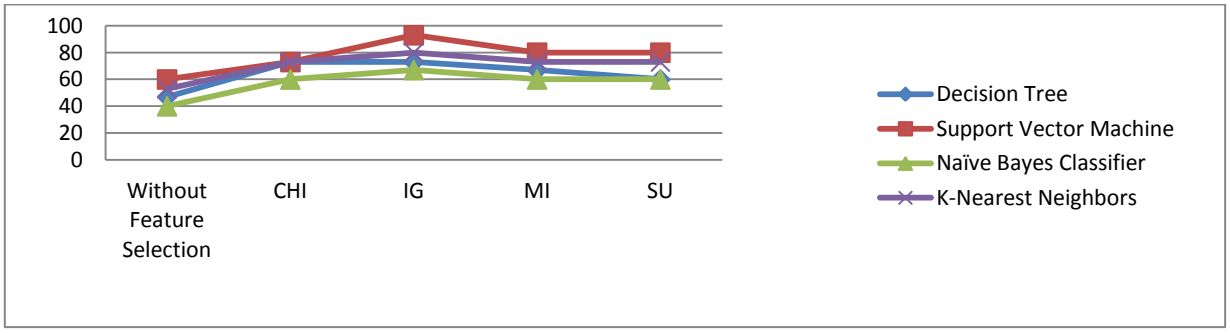


Figure-3 Performance of set accuracy rate involved in preview of hotel dataset

.

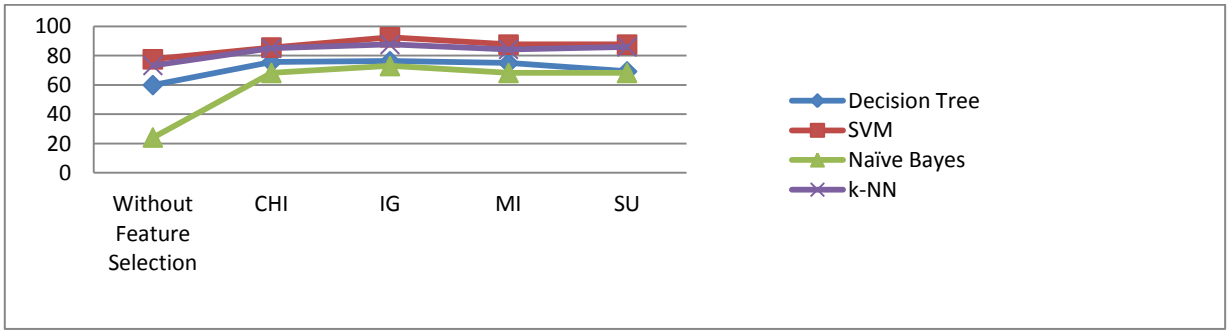


Figure-4 Performance of the average set prediction

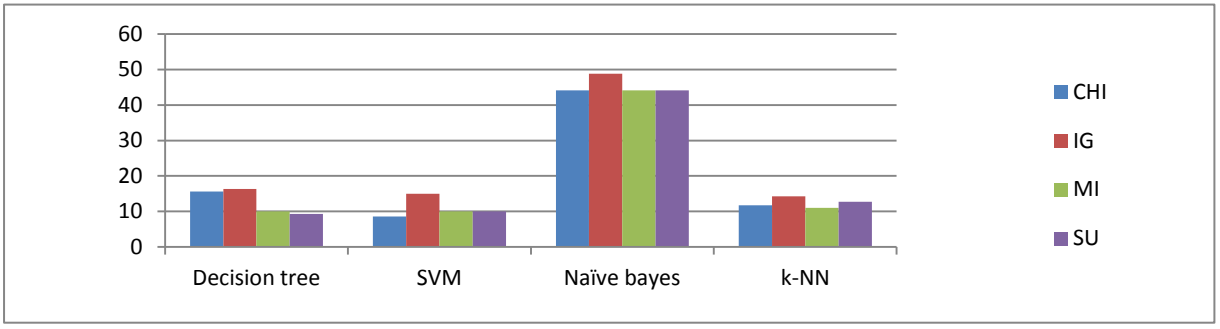


Figure-5 Improvement rate of different classifiers

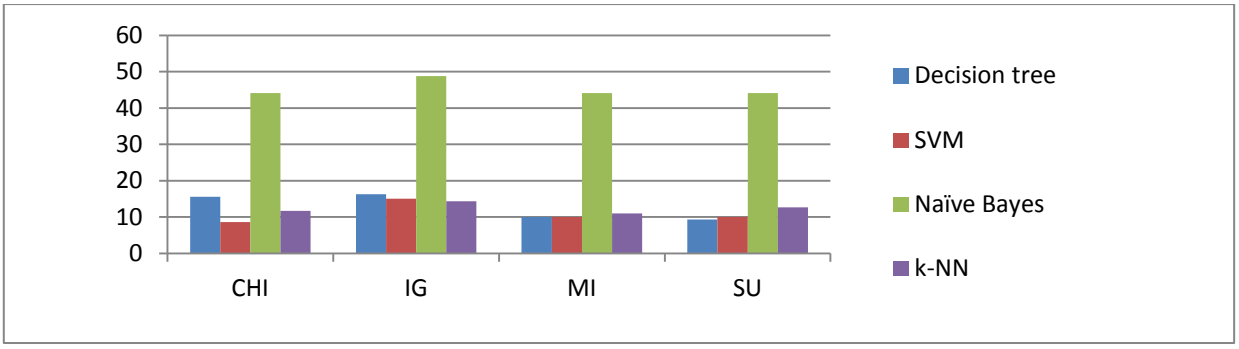


Figure-6 Performance rate in different feature selection metrics

5.1 Analysis of the result

The datasets used for the experiment are described in table-1. The experiment result is obtained in two steps one using classifiers algorithms (table-2) and result obtain applying feature selection methods to the classifiers (table-3). Table-2 describes the analysis of four classifiers naïve bayes, SVM, decision tree and k-NN on the three dataset CNAE-9, SMS spamham collection and preview of hotel dataset. The use of different feature selection metrics Chi-squared, Information Gain, Mutual Information and Symmetrical Uncertainty with the four classifiers are describe in table-3. Table-4 shows the average performance of testing set for CNAE-9, SMS spamham collection and preview of hotel datasets.

Among the four classifiers naïve bayes perform the least with an average testing set prediction rate of 24.2% but from table-3 applying different feature selection methods the testing set accuracy has increased significantly. SVM classifier lies in the top of the table with an average rate of testing set 77.6% but still applying feature selection methods on the classifier has increased the test data accuracy rate. Decision tree classifier performs moderately on both the table (2 and 3). Among the four feature selection methods IG performs better. It was also witnessed that the performance of naïve bayes classifier increased but it performed worst than other classifiers.

Figure 1, 2 & 3 shows the performance of the prediction of set by classifiers with and without using four feature selection metrics i.e. chi-square metric, information gain metric, mutual information metric and symmetrical uncertainty metric and on the three data set used in our experiment i.e. CNAE-9 dataset, SMS spam collection dataset and preview of hotel dataset. The three figures 1, 2 and 3 clearly points out the poor performance of naïve bayes classifier. Naïve bayes stands in bottom of the graph as compared to the performance of the other mentioned classifier on the two experimental dataset. The experiment result shows that SVM and k-NN classifiers prove to be the best in prediction whereas decision tree classifier performs on an average scale. Figure-4 shows the average performance of the testing set for CNAE-9 dataset, SMS spam collection dataset and preview of hotel dataset. Figure-5 shows the improvement rate of the different classification algorithms using feature selection metrics and figure-6 shows the performance rate of different feature selection metrics.

6. CONCLUSION

The experimental result illustrates the effect of application of feature selection methods on text classification. The use of feature selection methods in the experiment seems to enhance the performance of the classifiers. So it is highly recommended for using feature selection in the domain of text classification, which is high dimensional in nature. From the results it can be concluded that among the four feature selection metrics chi-squared, information gain, mutual information and symmetrical uncertainty, the use of information gain metric has the most positive impact over other metrics. SVM outperformed other classifiers in all the occasions. Another finding of the study is, naïve bayes was the worst in terms of accuracy with an average prediction of rate of 24.2% and with feature selection there is a dramatic improvement and the result is then comparable to other classifiers with reduced set of features. There will be a need to work further on theoretical foundations to make it more appropriate for text classification.

7. REFERENCES

[1] Yang, Yiming, and Thorsten Joachims. "Text categorization." *Scholarpedia* 3.5 (2008): 4242.

[2] Sriram, Bharath, et al. "Short text classification in twitter to improve information filtering."*Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2010"

[3] Aggarwal, Charu C., and ChengXiang Zhai. "A survey of text clustering algorithms" *Mining Text Data*. Springer US, 2012. 77-128.

[4] Fabrizio Sebastiani. Text categorization. In Alessandro Zanasi (ed.), *Text Mining and its Applications*, WIT Press, Southampton, UK, 2005, pp. 109-129.

[5] Dasgupta, Anirban, et al. "Feature selection methods for text classification."*Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2007.

[6] Singh, Sanasam Ranbir, Hema A. Murthy, and Timothy A. Gonsalves. "Feature Selection for Text Classification Based on Gini Coefficient of Inequality." *Journal of Machine Learning Research-Proceedings Track* 10 (2010): 76-85.

[7] Joachims, Thorsten. "A statistical learning learning model of text classification for support vector machines." *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2001.

[8] Zhang, Yin, Rong Jin, and Zhi-Hua Zhou. "Understanding bag-of-words model: a statistical framework." *International Journal of Machine Learning and Cybernetics* 1.1-4 (2010): 43-52.

[9] Wallach, Hanna M. "Topic modeling: beyond bag-of-words. "*Proceedings of the 23rd international conference on Machine learning"*. ACM, 2006.

[10] Refaeilzadeh, Payam, Lei Tang, and Huan Liu. "On comparison of feature selection algorithms." *Proceedings of AAAI Workshop on Evaluation Methods for Machine Learning II*. 2007.

[11] Novovicova, Jana, and Antonin Malik. "Information-theoretic feature selection algorithms for text classification." *Neural Networks, 2005. IJCNN'05. Proceedings. 2005 IEEE International Joint Conference on*. Vol. 5. IEEE, 2005.

[12] Singh, Sanasam Ranbir, Hema A. Murthy, and Timothy A. Gonsalves."Feature Selection for Text Classification Based on Gini Coefficient of Inequality." *Journal of Machine Learning Research-Proceedings Track* 10 (2010): 76-85.

[13] Arauzo-Azofra, Antonio, José Luis Aznarte, and José M. Benítez. "Empirical study of feature selection methods based on individual feature evaluation for classification problems." *Expert Systems with Applications* 38.7 (2011): 8170-8177.

[14] Rong-zong, S. U. N. "An Improved KNN Algorithm for Text Classification [J]."*Computer Knowledge and Technology* 1 (2010): 073.

[15] Yang, Yiming, and Jan O. Pedersen. "A comparative study on feature selection in text categorization." *ICML*. Vol. 97. 1997.

[16] Baharudin, Baharum, Lam Hong Lee, and Khairullah Khan. " A review of machine learning algorithms for text-documents classification." *Journal of advances in information technology* 1.1 (2010): 4-20.

[17] Chen, Jingnian, et al. "Feature selection for text classification with Naïve Bayes." *Expert Systems with Applications* 36.3 (2009): 5432-5435.

[18] Li, Shoushan, et al. "A framework of feature selection methods for text categorization." *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*. Association for Computational Linguistics, 2009 *Systems with Applications* 38.7 (2011): 8170-8177.

[19] DING, Xiaoming, and Yan TANG. "Improved Mutual Information Method For Text Feature Selection."

[20] Brank, Janez, et al. "Interaction of feature selection methods and linear classification models." *Workshop on Text Learning held at ICML*. 2002.

[21] Ali, Syed Imran, and Waseem Shahzad. "A feature subset selection method based on symmetric uncertainty and Ant Colony Optimization." *Emerging Technologies (ICET), 2012 International Conference on*. IEEE, 2012