I.J. Intelligent Systems and Applications, **2015, 06, 33-40**
Published Online May 2015 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijisa.2015.06.03

Evaluation of Feature Selection Approaches for Urdu Text Categorization

Tehseen Zia
Department of Computer Science & IT, University of Sargodha, Sargodha, 4100, Sargodha
Email: tehseen_zia@yahoo.com

Qaiser Abbas
Department of Computer Science & IT, University of Sargodha, Sargodha, 4100, Sargodha
Email: qaiser.abbas@uos.edu.pk

Muhammad Pervez Akhtar
Department of Computer Science & IT, University of Sargodha, Sargodha, 4100, Sargodha
E-mail: pervezbcs@gmail.com

Abstract—Efficient feature selection is an important phase of designing an effective text categorization system. Various feature selection methods have been proposed for selecting dissimilar feature sets. It is often essential to evaluate that which method is more effective for a given task and what size of feature set is an effective model selection choice. Aim of this paper is to answer these questions for designing Urdu text categorization system. Five widely used feature selection methods were examined using six well-known classification algorithms: naive Bays (NB), k-nearest neighbor (KNN), support vector machines (SVM) with linear, polynomial and radial basis kernels and decision tree (i.e. J48). The study was conducted over two test collections: EMILLE collection and a naive collection. We have observed that three feature selection methods i.e. information gain, Chi statistics, and symmetrical uncertain, have performed uniformly in most of the cases if not all. Moreover, we have found that no single feature selection method is best for all classifiers. While gain ratio out-performed others for naive Bays and J48, information gain has shown top performance for KNN and SVM with polynomial and radial basis kernels. Overall, linear SVM with any of feature selection methods including information gain, Chi statistics or symmetric uncertain methods is turned-out to be first choice across other combinations of classifiers and feature selection methods on moderate size naive collection. On the other hand, naive Bays with any of feature selection method have shown its advantage for a small sized EMILLE corpus.

Index Terms— Text Categorization, Feature Selection, Urdu, Performance Evaluation, Test Collection

I. Introduction

The domain of text categorization deals with the question that how predefined labels can be assigned to documents. It is an important task with the necessity to automatically organize [16, 21], summarize [2, 33], filter [12] huge amount of textual data. To perform text categorization, one methodology is to use machine learning approach where text classifier is automatically learned through an inductive process that uses pre-labeled text documents as an input [17]. This methodology is very widely used since last two decade [25] where different machine learning techniques have been applied including k-nearest neighbor classification [29], decision trees classification [6], support vector machines [9], artificial neural networks [20] and Bays probabilistic approach [10], etc.

Despite remarkable work on text categorization of English documents, the work on Urdu language text is still in infancy [32] though online Urdu text data is rapidly growing and necessitating the need to develop methods to organize and handle data. An important reason for this lack of interest is unavailability of publically accessible Urdu data collections [23, 30]. The term test collection refer to a collection of documents to which human indexer has assigned categories from a pre-defined set and it allows the researchers to test ideas and compare results without hiring human indexers. For example, Reuters Corpus Volume 1 (RCV1) is a test collection of over 800,000 documents containing newswire stories. The documents are categorized and manually coded using three category sets: topic codes, industrial codes and region code, where each category consists of further sub-categories [14]. According to Lewis [14], test collection is the key resource for research over text categorization.

A major difficulty while employing machine learning approach for text categorization is to handle high dimensionality of feature space. That is, a feature in the textual data represents a unique word and even a moderate size text collection can contain tens or hundreds of thousands of features. This size of feature space adversely affect the classifiers in two ways: firstly, it is prohibitively high for some classifiers. That is, it makes some classifiers computationally intractable like Bays' belief models while makes some others computationally inefficient like e.g. artificial neural networks cannot handle such a size of feature space [29]. Secondly, it badly affects the performance of classifiers due to phenomena well known as over-fitting [25] in which classifiers are trained over the contingent characteristics of data despite just constitutive characteristics. Overfitted classifiers perform well over the training examples but poorly over the testing data. Empirically, it is shown that over-fitting can be avoided if classifiers are trained over number of training examples that are propositional to number of features [6]. In other words, over-fitting can be avoided even only with the availability of small training data by reducing feature space. That's why; reducing the size of feature space is an essential pre-processing task before applying machine learning techniques. However, while reducing feature space; there are two crucial requirements that must be considered: firstly the performance of classification technique must not be degraded substantially. Secondly, the task should be performed automatically since manual procedure can be very laborious.

II. RELATED WORK

To automatically reduce the feature space, several methods commonly referred as feature selection methods have been proposed. The underlying task of these methods is either to select a subset of highly effective features given a set of features or construction of new (high-level) features by combining original features into high-level orthogonal features. The methods that are used to select effective features are mainly relied on evaluation methods to determine the effectiveness of features. Lewis & Ringuette [14] proposed information gain (IG) as feature evaluation measure to reduce feature space for naive Bayes and decision tree based document classification. Wiener [26] used Chi-square (CHI) and mutual information (MI) feature evaluation measures while applying neural networks for document classification [19] have used a wrapper approach for classification algorithms where initially a subset of features is chosen to train a classifier and find its effectiveness. A new feature is then added into features set if by adding this feature into the training set improves over previous performance of classifier. In order to comparatively evaluate the effective of different feature selection methods, Yang [14] has conducted an empirical study over feature selection methods including IG, CHI, MI, document frequency (DF) and term strength (TS). The study was conducted over two test collections: Reuters-22173 and OHSUMED. It is found that IG with removal of up to 98% non-informative features outperformed the rest. A similar comparative study was conducted by Rogati& Yang [24] using same methods (where different variations of each method are also considered) and over two different test collections: Reuters-21578 and Reuter Corpus Version 1 (RCV1). The results have shown that Chi-square statistics consistently outperform other criteria (including IG). Interestingly, both studies have been conducted on similar feature selection methods but on different test collections and yield different results. This phenomenon (i.e. a different test collection yield different results) often exist with such empirical studies. That is, sometimes some test collection happens to be more suited to the underlying assumptions of some methods and sometime not. That's why; with new test collections, benchmark results are reproduced [14]. Lewis has explained this phenomenon as: just like ML classifiers can over-fit if its parameters are tuned over the accidental characteristics of data, research community can over-fit by improving classifiers that have already performed well over existing datasets. Therefore, by recertifying the feature selection methods and classifiers over new test collections periodically, progress can be made.

The objective of the paper is to comparatively evaluate feature selection methods and produce benchmark results for feature selection in Urdu text categorization. To evaluate the performance of feature selection methods, six widely studied, top-performing and scalable classifiers have been selected: naive Bayes (NB), -nearest neighbors (KNN), decision tree (DT) and support vector machines (SVMs) with linear, polynomial and radial basis kernels. We want to answer following questions with empirical results:

- Which feature selection methods are both computationally scalable and top-performing across classifiers?
- Which combination of classifier and feature selection method perform best across classifiers and feature selection methods.

In the prior work on Urdu text categorization, the impact of dimensionality reduction using stop words removal and stemming is empirically analyzed [31]. It is shown that stop word removal has positive impact whereas stemming has caused negative impact in mostly cases. However, the impact of feature selection methods for Urdu text categorization has not been considered as a subject for empirical evaluation.

III. FEATURE SELECTION METHODS

In the empirical study, we have used four feature selection methods: information gain, gain ratio, Chi statistics, symmetric uncertain. In each method, features are evaluated based on its evaluation criteria and ranked based on feature weights. A short introduction of methods is given below:

A. *Information Gain*

Information gain (IG) is a quantitative measure often used in ML to find the worthiness of feature [17]. IG is often defined with the help of entropy. Given a dataset S contain some positive and negative examples related to some binary classification problem. Then entropy H of S is measured as:

$$\mathrm{H(S)} \equiv -\mathrm{p} \oplus \log_2 \mathrm{p} \oplus -\mathrm{p} \ominus \log_2 \mathrm{p} \ominus \qquad (1)$$

Where $\mathrm{p} \oplus$ denotes ratio of positive examples and $\mathrm{p} \ominus$ is the ratio of negative examples. The IG of a feature is then measured as expected reduction in entropy if dataset is partitioned according to the feature. Formally, IG of a feature t relative to a dataset S denoted as $\mathrm{IG(S, t)}$ can be defined as:

$$\mathrm{IG}(t) \equiv \mathrm{H}(S) - \sum_{v \in \mathrm{val}(t)} \frac{|S_v|}{|S|} \mathrm{H}(S) \qquad (2)$$

Where $\mathrm{val}(t)$ represents the set of all values of feature t and S_v is subset of C in which feature t has value v.

B. Gain Ratio

A characteristic of information gain is that it favors features that have many values over features that have few values. To deal with this issue, an extra term with information gain measure is introduced to account that how a feature splits the data. The resultant feature evaluation measure is called gain ratio (GR). Given a dataset S, the gain ratio score of a feature t can be estimated as:

$$\mathrm{GR}(t) = \frac{\mathrm{IG}(t)}{\mathrm{SI}(t)} \qquad (3)$$

Where $\mathrm{IG}(t)$is information gain of feature t as defined in Equation 1. $\mathrm{SI}(t)$ is known as split information of feature t with respect to dataset S:

$$\mathrm{SI}(t) = - \sum_{v \in \mathrm{val}(t)} \frac{|S_v|}{|S|} \log_2 \frac{|S_v|}{|S|} \qquad (4)$$

Where S_v is subset of S in which feature t has value v.

C. χ^2 Statistics (Chi)

Chi is a popular statistical method for measuring the independence between occurrences of two events. When applied to feature selection for measuring the score of a feature, the events are occurrence of a feature t and occurrence of category c. With the aid of a two-way contingency table between t and c, we can measure the effectiveness score of a feature as:

$$\mathrm{Chi}(t, c) = \frac{N \times (AD - CB)^2}{(A + C) \times (B + D) \times (A + B) \times (C + D)} \qquad (5)$$

Where A stands for number of times t co-occur with c, B represents number of times t occur without c, C symbolizes number times c occur without t and D represents number of times when neither t nor c are occurs. In case of independency between $t$and c, Chi measure gives zero value. When performing feature selection, Chi measure is used in a way that for each category the score of each feature is measured initially. Then, the category-specific score of each feature is combined so that each feature has only single score. This combination can be performed in one of the following ways:

$$\chi^2_{\mathrm{avg}}(t) = \sum_{i=1}^{m} p_r(c_i) \chi^2(t, c_i) \qquad (6)$$

$$\chi^2_{\max}(t) = \max_{i=1}^{m} \{\chi^2(t, c_i)\} \qquad (7)$$

D.Symmetrical Uncertain Feature Evaluation

An inadequacy of Chi method is that the method will not detect features that are redundant due to their correlation with other features. One way to select a subset of features that are individually correlated well with class (as in Chi method) but have little correlation with each other. Correlation between two features t_i and t_j can be measured using symmetric uncertainty [7]:

$$U(t_i, t_j) = 2 \frac{H(t_i) + H(t_j) - H(t_i, t_j)}{H(t_i) + H(t_j)} \qquad (8)$$

Where Hentropy function as is described in Equation 1. $H(t_i, t_j)$is joint entropy of t_i and t_j and is calculated from joint probabilities of all combinations of t_i and t_j. The value of symmetric uncertainty (SC) lies between 0and 1. Based on SC measure, effectiveness of set of features is determined as:

$$\frac{\sum_j U(t_j, Cl)}{\sqrt{\sum_i \sum_j U(t_i, t_j)}} \qquad (9)$$

Where Clis the class attribute and indices i and j range over all attributes in the set. When different subsets have same value, one with smallest number of features is selected.

IV. Classification Algorithms

For assessing the performance of a feature selection methods six classification algorithms has been tested including k-nearest neighbours (KNN), naive Bayes (NB), decision tree (DT) and support vector machines (SVM) with linear, polynomial and radial basis kernels.

A. k-Nearest Neighbour (kNN)

kNN Algorithm belongs to the family of instance based learning where rather learning an explicit target function, training examples is simply stored in the database [17]. While classifying a document in predetermined category, its proximity with the existing instances (i.e. documents with already known categories) is measured. Training examples are ranked according to their proximity score and top k ranked examples are then selected to participate in decision making. The decision is then made based on the voting of selected documents; each training example vote for its own category and the category who won in voting is assigned to the queried document. Despite this simplest approach, a more sophisticated way to make decision based on training data is known as distance weighted kNN. In this approach, each training example participates in decision making through voting. However, the votes are weighted according to the proximity of the training examples with the queried document. Finally, the category with maximum aggregated weight is assigned as category of the queried document.

B. Naive Bays Classifier

In NB classifier, text categorization is viewed as estimating posterior probabilities of categories given documents, i.e. $P(c_i|\overrightarrow{d_j})$; the probability that j^{th} document (as represented with a weight vector $\overrightarrow{d_j} =< q_{1j}, q_{2j}, \ldots q_{|T|j} >$ where q_{kj} is the weight of k^{th}

feature in jth document) belongs to class c_i . These posterior probabilities are estimated by using Bayes theorem as:

$$P(c_i|\overrightarrow{d_j}) = \frac{P(d_j|c_i)P(c_i)}{P(\overrightarrow{d_j})} \quad (10\,)$$

where$P(c_i)$ is the prior probability that represents the probability of selecting a random document belongs to classc_i, $P(\overrightarrow{d_j})$ is the probability that a randomly chosen document has weight vector $\overrightarrow{d_j}$ and $P(d_j|c_i)$ is the probability that the document d_j belongs to class c_i. To make $P(d_j|c_i)$computation tractable, it is assumed that coordinates of the document vector are conditionally independent of each other. By following the assumption, the term $P(d_j|c_i)$ can be estimated as:

$$P(\overrightarrow{d_j}|c_i) = \prod_{k=1}^{|T|} P(w_{kj}|c_i) \quad (11)$$

C.Support Vector Machines (SVM)

SVM is based on the idea well known as structural risk minimization principle [3] where objective is to find a hypothesis that can guarantee lowest true error (error of a hypothesis for classifying an unseen instance drawn from the same distribution as training data). The direct estimation of true error is not possible unless learner knows true target concept. However, according to structural risk minimization principle, the true error can be bounded using training error and complexity of hypothesis space. The complexity of hypothesis is represented by using well known Vapnik-Chervonenkic dimension or VC dimension) [17]. The task underlying SVM is to minimize true error of resultant hypothesis by efficiently controlling the VC dimension of hypothesis space [3].

Some sophisticated mathematical transformations are also applied (known as kernel trick) to data prior to learning hypothesis when instances are not linearly separable. The objective of transformation is to transform data instances to higher dimensions features space in order to make them linearly separable. Depending on the choice of kernel function, SVM have three popular variants: SVM with linear or no kernel, polynomial kernel and radial basis kernel.

D.Decision Tree

In this classifier, the hypothesis is represented as a tree: node of the tree corresponds to a feature, an edge of the node corresponds to the feature values and leafs correspond to categories [17]. Various methods have been proposed to automatically learn DT from training data and most of them follow the approach known as top-down greedy search [22]. The well-known examples of this approach for DT learning are ID3 algorithm and C4.5. In the algorithms, a node is selected is a way that candidate attributes (i.e. words) are evaluated using a quantitative measure (known as information gain) and the best among them (with maximum information gain) is selected. The DT is learned branch-by-branch where a branch is continued to be grown until either of two stopping criterion is met: every attribute is chosen along the path or training example associated with the leaf node belonged to same class.

V. Test Collections and Experimental Setup

To conduct this empirical study, we have used two test collections. The first test collection is well known as EMILLE corpus 1 which is distributed by European Language Resource Association. The corpus is prepared during a collaborative venture between Lancaster University, UK and Central Institute of Indian Languages (CIIL). The corpus is monolingual data of 14 Indian languages including Urdu and each language includes three components: monolingual, parallel and annotated versions. We have used free downloadable version of the corpus known as EMILLE corpus (Beta release version)2. For Urdu language only parallel text corpus is available with this release which include few documents belong to four categories: education, health, legal and social (the categories with one document are not considered such as housing).

The second test collection is a self-collected naive collection of 5000 documents distributed over four categories: politics, commerce, sports and entertainment. The collection contains newswire stories (during session November 1st 2011 to January 31st 2013) and is collected from two news channels: British Broadcasting (BBC) and Voice of America (VOA Urdu). The category-wise distribution of document is shown in Table 1.

Table 1. Distribution of documents across categories in naive collection

Politics	Commerce	Sports	Entertainment	Total
1500	1300	1500	1200	5000

To evaluate the performances of classifiers; we have used standard f-measure according to the recommendation of [14]. Accuracy may not be an effective measure since good accuracy may be achieved by always predicting one class (e.g. positive class). In contract f-measure evaluate the category-wise prediction abilities of the classifiers as:

$$F = \frac{2 * p * r}{p + r} \quad (12)$$

Where p and r are respectively the precision and recall of classifiers:

$$r = \frac{a}{a + c} \quad (13)$$

$$p = \frac{a}{a + b} \quad (14)$$

[1] http://www.lancaster.ac.uk/fass/projects/corpus/emille/

[2] http://www.ota.ox.ac.uk/scripts/download.php?approval=9d5c5288a573453a422f

Where the symbol a denotes number of documents a classifier correctly assigned to the category b denotes the number of documents a classifier incorrectly assigned to the category.

VI. Parameter Settings

In this section, it is described the methodology for tuning parameters within the classifiers. The introduction of these classifiers is described in Section 3.2.

A. *SVM*

To perform experimentations with SVM classifiers, a wrapper SVM tool for Weka toolkit known as LIBSVM [5] is used where values of other parameters were not changed. SVM classifier is analyzed for three of its popular variations, i.e. linear SVM, SVM with polynomial kernel and SVM with radial basis kernel. SVM with polynomial kernel is tested with respect to degree of polynomial (i.e. parameter d). Results of the analysis is shown in Fig. 1 where it can be seen that the classifier perform at its best at $d = 1$. Moreover, it can also be noticed from the table that performance of the classifier significantly degrades as value of d increases.

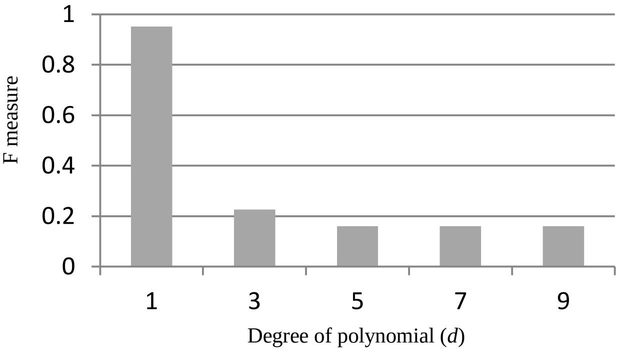

Fig. 1. Performance evaluation of polynomial SVM with respect to degree of polynomial (i.e. d).

SVM with radial basis kernel is analyzed for various values of gamma. It is found that the classifier shows better performance over gamma $\gamma = 0$ as shown in Fig. 2.

B. *kNN*

KNN classifier has one free parameters: k (neighbourhood size). We have found the optimal values of k by using $fivefold\ cross\ validation$. In order to choose the values of free parameters, following values are tried.

k: 1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 25, 27, 29, 31, 33, 35, 37, 39, 41, 43, 45, 47, 49, 51, 53, 55, 57, 59, 61, 63, 65, 67, 69, 71, 73, 75, 77, 79

As shown in Fig.3, the precision is improved until the value $k = 19$ and then it remains almost consistent till $k = 70$. However, recall and f-measure continuously decreased as k increased. Hence, it is a trade off since if you chose value of k to maximize precision, recall and f-measure will not be at their best values. Because, even the maximum values of recall and f-measure (i.e. at $k = 13$) are not promising, the value of $k = 13$ is chosen so as not to compromise the effectiveness of recall and f-measure any further.

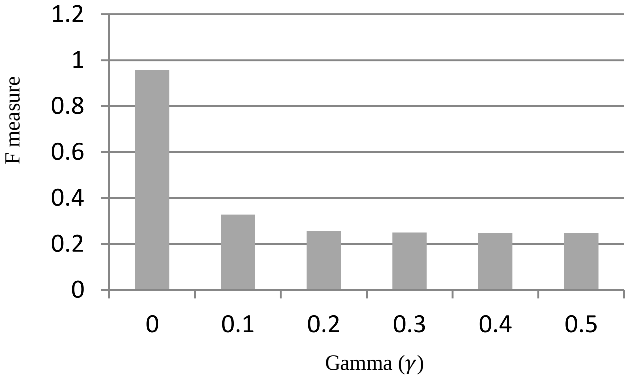

Fig. 2. Performance evaluation of radial basis SVM for different γ values

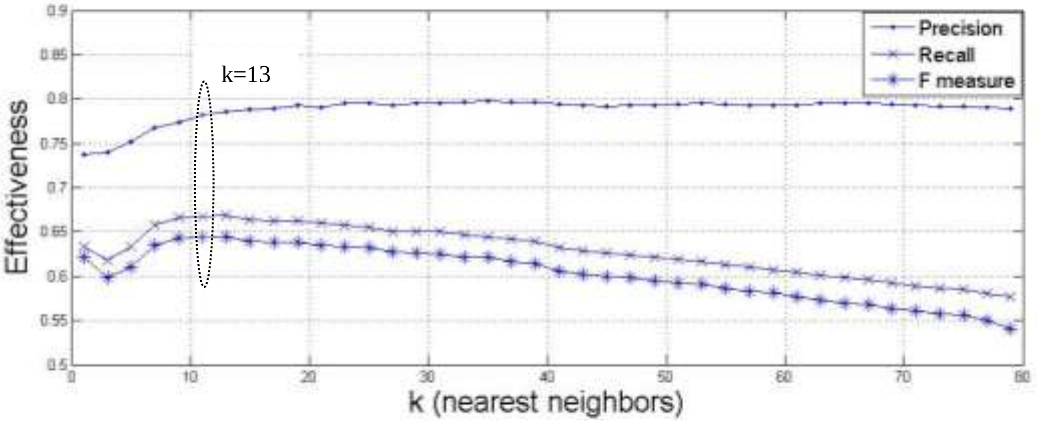

Fig. 3. K (nearest neighbours) vs. effectiveness.

VII. Results

A collective view of results of all four classifiers with top performing feature selection methods is shown in Fig. 4 for naive collection. The results over EMILLE collection are shown in Fig. 5. The performance of classifiers with respect to OR feature selection method is skipped during the illustration of results since OR consistently perform poorly. Other feature selection methods that have performed equivalently are shown with one curve. We have observed that information gain and chi square have an advantage over other methods. We have noticed that these results are consistent with previous study, conducted for English language [24].

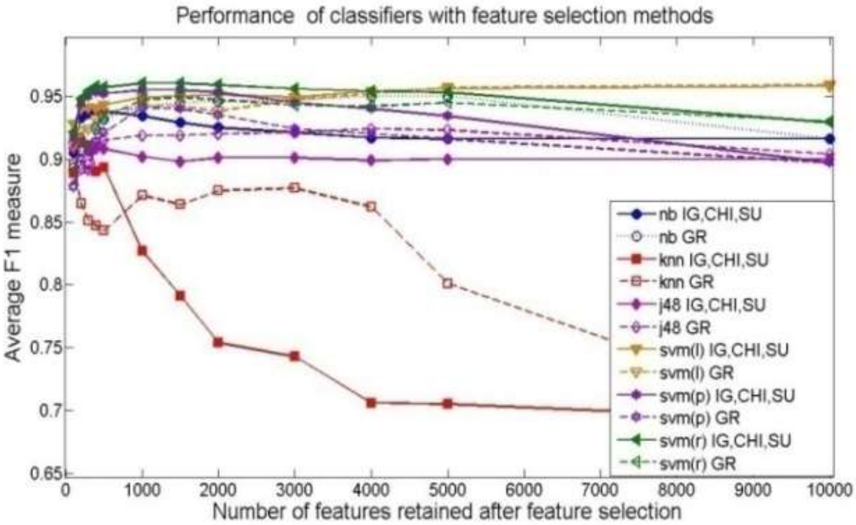

Fig. 4. Macroaveraged F1 measure of six classifiers in combination with five top feature selection methods. Feature selection methods IG, CHI, FA and SC have performed equivalently so they are illustrated with a single curve. For example, nb IG+CHI+FA+SC shows the performance of na ïve Bayes with feature selection methods: IG, CHI, FA and SC. The legends svm (l), svm(p) and svm (r) respectively represents SVM with linear, polynomial and radial basis kernel.

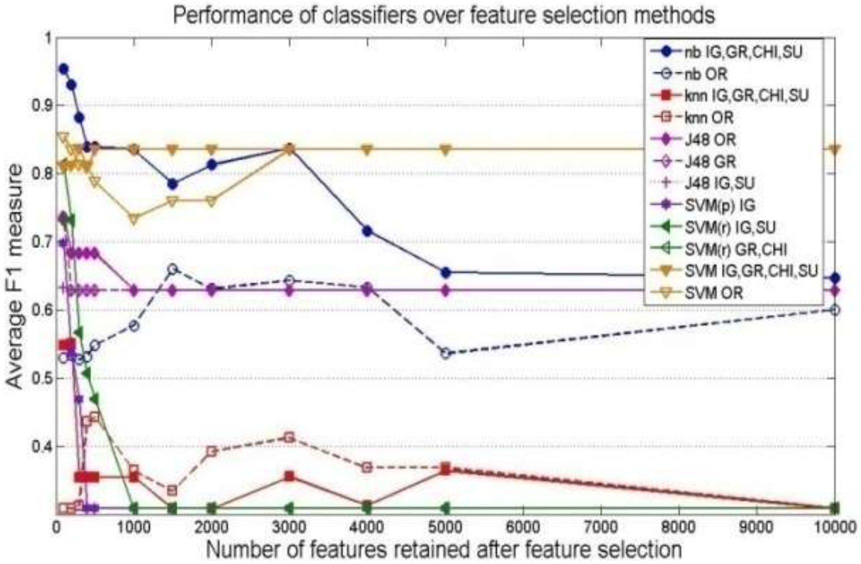

Fig.5. Macroaveraged F1 measure of six classifiers in combination with five top feature selection methods with Emile corpus.

We have found that linear SVM has an advantage over other classifiers irrespectively of feature selection methods. This result endorses the previous finding that SVM is least dependent over feature selection since it has an internal over-fitting avoidance mechanism [9]. We have achieved 96% f-measure by using linear SVM with IG and Chi square as feature selection methods on a na ïve collection. On the other hand, KNN classifier is found to be most vulnerable with respect to number of features and feature selection method. Naive Bayes is observed to be second top performing classifier. However, its performance also depends on the choice of feature selection method and number of features.

Finally, as an outcome of this empirical study, we will recommend linear SVM along with IG or Chi square methods for designing Urdu text classifier. However, this combination is useful when a moderate size (few thousand documents) test collection is available. In case of a small test collection (few hundred documents), we suggest na ïve Bayes classifier along with IG or Chi square method where the size of feature set should not be more than 500 hundred.

VIII. CONCLUSION

We have conducted an empirical study to analyse performance of five feature selection methods (i.e.

information gain, gain ratio, Chi statistics, symmetric uncertain and OneR) using six classifiers (naive Bayes, KNN, support vector machine with linear, polynomial and radial basis kernels and decision tree) on two Urdu test collections: naive collection and EMILLE collection. We have observed that four feature selection methods i.e. information gain, Chi statistics, symmetrical uncertain and filter attribute, have performed uniformly in most of the cases if not all. Moreover, it is observed that no single feature selection method dominate in all classifiers: while gain ratio out-perform others for naive Bayes and J48, IG and companions have shown top performance for KNN and SVM with polynomial and radial basis kernels. Compared with other classifiers, SVM does not get much benefit from feature selection. Linear SVM with any of feature selection methods IG, Chi or SC is outperformed other combinations of classifiers and feature selection methods over a moderate size naive collection. On the other hand for a small sized EMILLE corpus, naive Bayes with any of feature selection method has shown its advantage.

REFERENCES

[1] D.Balie, “Baseline Information Extraction: Multilingual Information Extraction from Text with Machine Learning and Natural Language Techniques”. Technical Report, University of Ottawa, 2005.

[2] D.Blei, “Probabilistic Topic Models”. Communication of The ACM, 55(4), 2012.

[3] C. Burges, “A Tutorial on Support Vector Machines for Pattern Recognition”, Data Mining and Knowledge Discovery, 2, 121-167, 1998.

[4] C.Chung Chang and Chih-Jen Lin, “LIBSVM : A Library for Support Vector Machines”, ACM Transactions on Intelligent Systems and Technology, 2(3), 2011.

[5] W.Cohen, Y.Singer, “Context Sensitive Learning Methods for Text Categorization”, Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, 307-315, 1991.

[6] N.Fuhr, C.Buckley, “A Probabilistic Learning Approach for Document Indexing”. ACM Transaction on Information Systems, 9(3), 223-248, 1991.

[7] M.Hall, E.Frank, G.Holmes, B.Pfahringer, P.Reutemann, L. H.Witten, “The Weka Data Mining Software: An Update”. SIGKKD Exploration, 11(1), 10-18, 2009.

[8] B.Jiang, D.Xiang-Qian, M.Lin-Tao, H.Ying, T.Wang, X.Wei-Wie, The Second International Symposium on Optimization and Systems Biology, 2008.

[9] T.Joachims, “Text Catagorization with Support Vector Machines: Learning with many Relevant Features”, Tenth European Conference on Machine Learning (ECML-98), 137-142, 1998.

[10] T.Joachims, “A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization”, Proceedings of the Fourteenth International Conference on Machine Learning, 143-151, 1997.

[11] K.Kira, L.A.Rendell, In Proceedings of the ninth international workshop on Machine learning, pp: 249-256, 1992.

[12] Shrivastava, J.N., Bindu, M.H., “E-mail Spam Filtering Using Adaptive Genetic Algorithm”, I.J. Intelligent Systems and Applications, 02, 54-60, 2014,

[13] D.D.Lewis, M.Ringuette, “Comparision of Two Learning Algorithms for Text Categorization”, In Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval, 1991.

[14] D.D.Lewis, Y.Yang, T.Ross, F.Li, “RCV1: A New Benchmark Collection for Text Categorization Research”, Journal of Machine Learning Research, 5, 361-397, 2004.

[15] H.Y.Li, A.K.Jain, “Classification of Text Documents”.Computer Journal. 41(8), 537-546, 1998.

[16] McCallum, A.Kachites, "MALLET: A Machine Learning for Language Toolkit." http://mallet.cs.umass.edu, 2002.

[17] T.Mitchell, “Machine Learning”.McGraw-Hill, New York, 1997.

[18] D.Mladenic, M.Grobelnik, “Feature selection for unbalanced class distribution and Naïve Bayes”. Proc. of the 16th Int. Conference on Machine Learning San Francisco: Morgan Kaufmann, pp. 258–267, 1999.

[19] I.Moulinier, G.Raskinis, J.Ganascia, “Text Categorization: A Symbolic Approach”, In Proceedings of the Fifth Annual Symposium on Document Analysis and Information Retrieval, 1996.

[20] H. T.Ng, W. B.Goh, K. L.Low, “Feature Selection, Perceptron Learning, and a Usability Case Study for Text Categorization”, In Proceedings of SIGIR-97, 20th ACM International Conference on Research and Development, 1997.

[21] K.Nigam, K.A.Mccallum S.Thrun, T.Mitchell, “Text Classification from Labeled and Unlabeled Documents using EM”, Machine Learning, 39(2/3), 103-134, 2000.

[22] J. R. Quinlan, “C4.5: Programs for Machine Learning”, Morgan Kaufmann Publishers Inc. San Francisco, CA, USA, 1993.

[23] K.Riaz “Rule-Based Name Entity Recognition in Urdu”, Proceeding to Name Entity Workshop, 126-135, 2010.

[24] M.Rogati, Y.Yang, “High-Performing Feature Selection for Text Classification”, In Proceedings of the eleventh international conference on Information and knowledge management, 2002.

[25] F.Sebastiani, “Machine Learning in Automated Text Categorization”.ACM Computing Surveys, 34(1), 1-47, 2002.

[26] E.Wiener, J.O.Pedersen, A.S.Wiegend, “A Neural Network Approach to Topic Spotting”, In Proceedings of the Fifth Annual Symposium on Document Analysis and Information Retrieval, 1995.

[27] I.H.Witten, G.W.Paynter, E.Frank, C.Gutwin, and C.G. Nevill-Manning, “Kea: Practical Automatic Keyphrase Extraction”, Design and Usability of Digital Libraries: Case Studies in the Asia Pacific, Information Science Publishing, 129-152, 2005.

[28] I.H.Witten, E.Frank, “Data Mining: Practical Machine Learning Tools and Techiques”, Second edition, Morgan Kaufman Publishers, 2005.

[29] Y.Yang, “Expert network: Effective and Efficient Learning from Human Decisions in Text Categorization and Retrieval”, Proceedings of ACMSIGIR, 13-2, 1994.

[30] Qaiser, A., “ A Stochastic Prediction Interface for Urdu”,I.J. Intelligent Systems and Applications, 01, 94-100, 2015.

[31] Ali, R.A., Ijaz, M., “Urdu Text Classification”, Proceedings of the 7th International Conference on Frontiers of Information Technology (FIT09), 2009.

[32] Abbas, Q., "Building a hierarchical annotated corpus of urdu: the URDU. KON-TB treebank." In *Computational Linguistics and Intelligent Text Processing*, pp. 66-79. Springer Berlin Heidelberg, 2012.

[33] Shahzadi, F., Zia, T., “An Empirical Study on Sentiment Polarity Classification of Book Reviews”, VFAST Transaction on Software Engineering, 2013.

Authors’ Profiles

Tehseen Zia is working as faculty member in Department of Computer Science and IT, University of Sargodha since 2005.Currently, he is working as assistant professor. He got PhD scholarship from Higher Education Commission in 2007 and completed his PhD degree from Institute of Computer Technology, Vienna University of Technology in November 2010. His research interest is in machine learning and text mining.

Qaiser Abbas did his doctorate in Computational Linguistics or Natural Language Processing from University of Konstanz, Germany. After working as a research assistant at University of Konstanz, Germany, currently he is working as a lecturer in University of Sargodha, Pakistan at Department of Computer Science. The detailed profile along with the research work can be seen on the following URL http://clsp.org/qabbas

Muhammad Pervez Akhtar dis his MS in text processing from University of Sargodha, Pakistan. His research interest includes text processing, information retrieval and extraction.

How to cite this paper: Tehseen Zia, Qaiser Abbas, Muhammad Pervez Akhtar,"Evaluation of Feature Selection Approaches for Urdu Text Categorization", International Journal of Intelligent Systems and Applications (IJISA), vol.7, no.6, pp.33-40, 2015. DOI: 10.5815/ijisa.2015.06.03