COMPARATIVE STUDY OF FEATURE SELECTION APPROACHES FOR URDU TEXT CATEGORIZATION

Tehseen Zia[1], Muhammad Pervez Akhter[2]and Qaiser Abbas[3]

[1, 2, 3]Department of Computer Science & IT
University of Sargodha, 40100, Sargodha, Pakistan

Email: [1]tehseen.zia@uos.edu.pk, [2]pervezbcs@gmail.com,[3]qaiser.abbas@uos.edu.pk

ABSTRACT

This paper presentsacomparative study of feature selection methods for Urdu text categorization. Fivewell-knownfeature selection methods were analyzedby means ofsixrecognized classification algorithms: support vector machines (with linear, polynomial and radial basis kernels), naive Bayes, k-nearest neighbour (KNN), and decision tree (i.e. J48). Experimentations are performed on two test collections includinga standard EMILLE collection and a naive collection. We have found that information gain, Chi statistics, and symmetrical uncertainfeature selection methods have uniformly performed in mostly cases. We also found that no solo feature selection technique is best for every classifier.That is,naive Bayes and J48 have advantage with gain ratio than other feature selection methods. Similarly, support vector machines (SVM) and KNN classifiers have shown top performance with information gain.Generally,linear SVM with any of feature selection methods out-performed other classifiers on moderate-size naive collection.Conversely, naive Bayes with any of feature selection technique has an advantage over other classifiers for a small-size EMILLE corpus.

Keywords: Text Categorization, Feature Selection, Urdu, Performance Evaluation, Test Collection

1.0 INTRODUCTION

Urdu is an Indo-European language and lies in the category of morphologically rich languages [1, 7]. According to an estimate [27], there are more than 450 million speakers of Urdu around the globe. Urdu is a national language in Pakistan and an official language in India. It is also spoken in other countries where migrants from Pakistan and India are residing in majority like Gulf, UK, Canada, USA, etc. Its script is inherited from Arabic and it contains a vast vocabulary of Arabic, Persian, and Turkish. Although Urdu shares grammar with Hindi, there is a substantial difference in the vocabulary and script writing style between these two languages. Hindi is written in Davanagri script, a variant of Sanskrit, while Urdu follows Arabic script. That's why; though linguistically both Hindi and Urdu are considered as the same languages [2], the resources (e.g. test collections) used for Hindi are not useful for Urdu language. A dynamic characteristic of Urdu is that it is flexible for accepting the lexical features and vocabulary from other languages. Because of this feature, it is very common to experience foreign words during processing of Urdu text. Like other morphologically rich languages, it is necessary to use Unicode encoding scheme for Urdu text processing.

Content base analysis of natural language text and assignment of predefined markers to the text documents lie under the umbrella of text categorization. Traditional knowledge engineering (KE), in which a set of rules is defined explicitly, requires expert knowledge [28]. Similarly, in machine learning (ML), a general learner e.g., an inductive process, builds a text classifier automatically based on a set of pre-marked or pre-labelled text documents [18, 19, 28]. In this construction of text classifier, neither the involvement of domain expert or knowledge engineer is required nor is the hectic encoding of explicit rules needed. Therefore, ML is a useful text categorization methodology as compared to its rival KE. However, it is not only the case. The availability of immense volume of online text data and digital libraries for handling and organizing text data and text categorization plays an important role in these tasks. Such online resources have been successfully utilized in cataloguing new articles [14, 17], in sorting and filtering emails [17] and in learning the reading interests of users [24].

As the work regarding English is concerned, categorization of text documents has been conducted massively; however, no substantial work on Urdu text categorization exists. The major reason of this non-existence is the unavailability of Urdu test collections [27]. According to Lewis [19], test collection is the key resource for the text categorization. The term test collection here refers to a collection of documents. The human indexer then indexes this collection of documents by assigning categories from a pre-defined set. This can be done without hiring human indexers and it permits the researchers to test ideas and compare the results. Reuters Corpus Volume 1 (RCV1) is an example of this test collection. It consists of more than 800,000 newswire stories. This corpus is manually encoded with three category sets, which include topic, industrial and region codes. Moreover, each category is further categorized into sub-categories [19].

High feature space dimensionality remainsa major hurdlefor modelling effective text classifiers because it makes classifiers computationally intractable and inefficient [33]. The phenomena behind the inefficiency of classifiers are well-known as over-fitting [29]. In this phenomenon,the classifiers perform better over training instances but badly over testing instances. Studies have shown that over-fitting can be preventedby collectingtraining instances propositional to number of features [12]. These results can lead to infer thatby reducing number of features, over-fitting can be avoided. That's why; feature spacereduction is acrucial task of modelling text classifier.

Several feature selection methods have been proposed and can be categorized into two classes depending on either a subset of features is selected or new features are constructed by combining original features. The methods that are used to select features are mainly relied on feature evaluation metrics.Lewis &Ringuette[18]applied information gain(IG)for feature selection before modelling naive Bayes and decision tree based document classifiers. Wiener [32] employedChisquare and mutual information (MI)for selecting feature whiledesigning neural networks based document classifier.Moohebat et al. [23] have proposed a wrapper based feature selection methods for text classifiers.In this method, a classifier is trained with an initial subset of features to find its efficacy. A feature is included into the subset whenitsinclusion improves the performance of classifier. Yang [33]has performed acomparative study of feature selection methods including IG, MI, document frequency (DF), Chi, and term strength (TS). It is reported that IG has out-performed other by reducing 98% features. In another empirical study,Rogati& Yang [28]has reported thatChi-square outperform other methods (including IG). Intriguingly, both empirical studies have been performedby using same feature selection techniqueshowever test collections are different which yield different results. This phenomenon (i.e. a different test collection yield different results) often exist with such empirical studies. That is, sometimessome 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 (e.g. [19]). Lewis [19] has explainedthis phenomenonas: just like ML classifiers can overfit 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 goal of this study is to empiricallyassess feature selection approachesfor Urdu text categorization and generate benchmark results.We have analyzedfive feature selection methods includingIG, Chi, gain ratio (GR), symmetric uncertain and oneR. To evaluate effectiveness the methods, we have employed six classifiers including support vector machines (with linear, polynomial and radial basis kernels), naive Bayes, k-nearest neighbour (KNN), and decision tree (i.e. J48).This empirical study is focused on answering following questions:

- Whichfeature selection methods are best across classifiers?
- Which feature selection method is more suitable for which classifier?
- How much features are sufficient for each classifier to make good predictions?

The organization of the paper is as follows: feature selection methods are introduced in Section 2, a brief introduction of classifiers are given in Section 3,experimental setupisgiven in Section 4, results are presented in Section 5. Finally, the conclusion of the study is given in Section 6.

2.0 FEATURE SELECTION METHODS

To evaluate performance of feature selection methods, we have employedfivewidely used feature selection methods including information gain, Chi statistics, gain ratio, symmetric uncertain, relief feature evaluation and

OneR. We have presented a short introduction of the methods in this section before presenting experimental results.

2.1 Information gain (IG)

Information gain is a quantitative measurefor finding the worthiness of feature for classification task[22, pp:55-57]. IG is defined with the aid of entropy. Entropy can be characterized as a quantification of (im)purity of a dataset collection. For instance, assume a dataset(S)containing positive and negative instancesrelated toa binary classification problem. Entropy (H) of S is then can be measured as:

$$H(S) \equiv -p \oplus log_2 p \oplus -p \ominus log_2 p \ominus \quad (\mathbf{1})$$

where$p \oplus$and $p \ominus$ isrespectively the ratio of positive and negative instances. Themeasureof expected reduction in entropy bypartitioning the dataset with respect to the feature is known as IG of the feature. IG of afeature$f$ can be defined as:

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

where$val(f)$symbolizes the set of all values of feature$t$ and S_v is subset of C in which feature$t$ has value v.

2.2 Gain Ratio

Onetrait of IG is that it favours features with many values over features with few values. However,the issues can be rectified by using an extra term with IG measure to account that how a feature splits the data. The resultant measure is called gain ratio (GR) and defined as:

$$GR(t) = \frac{IG(f)}{SI(f)} (\mathbf{3})$$

Where $IG(f)$ representsIG of feature f as defined in Equation 1. $SI(f)$is called split information of feature $f$with respect to dataset S:

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

WhereS_v is subset of S in which feature$f$ has value v.

2.3 χ^2 Statistics (Chi)

Chi is a well-knownstatistical measure for quantifying independence of two events. Asa feature selection measure, it is used to quantify independence between a feature f and category$c$ as.

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

Where$A$isa countof co-occurrence of $f$and c, $B$is a countof occurrence of f without c, C is a count of occurrence of c without f and $D$is a count when neither f nor c are occurred. In feature selection process, Chi measure is employedto score each feature for each class. To obtain a single score of a feature, all category-specific scores of the feature are combined. Two standard ways to perform this combination are:

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

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

2.4 Symmetrical Uncertain

Aninsufficiency of χ^2 is to consider redundant (i.e. correlated) features. Symmetric uncertainty is a feature selection method where the selected features are not correlated with each other. Correlation between two features t_i and t_j are measured as[32, pp: 323-324]:

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

Where $H(t_i)$ is entropy of feature t_i. $H(t_i, t_j)$ismutual entropy of features t_i and t_j. Symmetric uncertainty scoreof set of features is then determined as:

$$\frac{\sum_j U(t_j, C)}{\sqrt{\sum_i \sum_j U(t_i, t_j)}} \quad (\mathbf{9})$$

Where Cdenotes class attribute, indices i and j range over all attributes in the set.

2.5 OneR Feature Selection

OneR feature selection is mainlybased on accuracy measure as implementedinOneR classifier [31]. Standard cross-validation test can be applied for feature evaluation.The method is flexible toallow different search and evaluation techniques to be used.

3.0 CLASSIFICATION TECHNIQUES AND METHODS

For measuring performance of a feature selection approaches six well-studied classifiers 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. Because each classifier made certain assumption about data, another objective to use six classifiers is to comparatively analyse their performance. An introduction of the algorithms is given below.

3.1 Modelling support vector machines

Principles of computational learning theories had become ideal for support vector machines (SVM) and particularly, the principle of structural risk minimization [9]. This principle finds a hypothesis, which can guarantee the lowest true error. This error is the error of a hypothesis to classify an unseen instance drawn from the same distribution as with the training data. The true error can be estimated directly (unless learner knows true target concept). On the other hand, the concept of training error and the complexity of hypothesis space can be applied as a binding condition to estimate the true error (well known as Vapnik-Chervonenkic dimension or VC dimension) [22, pp. 214-220].Similarly, SVM lessens the true error of consequential hypothesis. It is also the distinguishing characteristics of SVM. It lessens the true error by controlling the VC dimension of hypothesis space, efficiently [9].

Binary classification problem is a best example to explain SVM. This problem can be viewed and understood geometrically. Various separating hyper-planes can be preferred as a decision boundary as can be seen in Fig. 1. However, SVM picks the one that has the maximum distance (known as margin) with respect to instances laying on the boundary (known as support vectors).The problem of finding the maximum margin is mathematically outlined as follows.

$$minimize_{w,b} < w.w >$$

$$subjecttoy_i(< w.x_i > +b) \geq 1 \qquad i = 1, \ldots l$$

where l represents the number of training examples, x_i is the input vector, y_i is the desired output. The problem is reformulated for computational convenience and is presented in Equation 3.

$$L(w,b,\alpha) = \frac{1}{2} < w.w > - \sum_{i=1}^{l} \alpha_i [y_i (< w_i . x_i > + b) - 1] \, (\mathbf{3})$$

In Equation 3, $\alpha_i \geq 0$ is the Lagrange multiplier. TheLagrange formulation is often named as primal formulation. Differentiating Equation 1 with respect to w and b, and substituting their values in Equation3, the problem can be formulated into another form known as dual form in Equation 4.

$$L(w,b,\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j < x_i . x_j > (\mathbf{4})$$

Some sophisticated mathematical transformation such as kernel trick is applied to data prior to learning decision boundary when the instances are not linearly separable. The basic idea behind the transformation is to map the data instance x_i to higher dimension feature space X. The instance $x_i$in new feature space is referred as$\Phi(x_i)$. In Equation 4, the dot product between data instances is an essential computation for the dual formulation. In higher dimensional space, the dot product can be computed as $\Phi(x_i).\Phi(x_j)$, which is one way to do this but using kernel functions $k(x_i, x_j)$, the dot product can easily be computed as given in Equation 5.

$$\Phi(x_i).\Phi(x_j) = K(x_i, x_j) (\mathbf{5})$$

SVM can learn polynomial and radial base classifiers depending on the choice of kernel function as follows in Equation 6 and 7, respectively.

$$K_{poly}(x_i, x_j) = (x_1 + x_2 + 1)^2 (\mathbf{6})$$

$$K_{rbf}(x_i, x_j) = exp(-\gamma(x_1 + x_2)^2) (\mathbf{7})$$

In text categorization problem, SVM method was primarily introduced by [15]. According to his findings, SVM offers following advantages for text categorization. First, SVM has the mechanism of over-fitting protection, which does not depend upon the number of features. It is least affected from the higher dimensionality of features in text categorization. Second, because the over-fitting mechanism of SVM is independent from size of features, feature selection is often not needed.

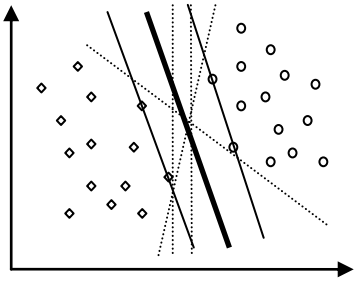

Fig. 1.A prototypical problem for learning support vector machine. The example is a binary classification problem where circles are the instances of negative class and diamonds are the instances of positive class. The bold line hyper-plane is best among the others shown in dotted lines.

3.2 Modelling Naive Bayes classifier

Among the best know classifiers for natural language text categorization, Naive Bayes (NB) classifier has a great importance [20, 22; pp: 177-183, 25]. In this method, text categorization is observed as estimating the probabilities $P(c_i|\overrightarrow{d_j})$. In other words, the probability for the j^{th} documentthat belongs to class c_i is calculated. The detailed form of the probability for the j^{th} document with a weight vector is $\overrightarrow{d_j} =< q_{1j}, q_{2j}, \ldots q_{|T|j} >$, where q_{kj} is the weight of k^{th} feature in j^{th} document). To compute the probabilities, Bayes theorem can be utilized as given in Equation 8.

$$P(c_i|\vec{d_j}) = \frac{P(d_j|c_i)P(c_i)}{P(\vec{d_j})} \textbf{(8)}$$

$P(c_i)$is the probability of a randomly selected document, which belongs to class c_i. $P(\vec{d_j})$is the probability of a randomly chosen document with weight vector $\vec{d_j}$and $P(d_j|c_i)$ is the probability of document $d_j$belongs to given class c_i. During the computation of the term $P(d_j|c_i)$, it was assumed that the coordinates of the document vector are conditionally independent of each other. By following the assumption, the term $P(d_j|c_i)$ was managed and estimated according to Equation 9.

$$P(\vec{d_j}|c_i) = \prod_{k=1}^{|T|} P(w_{kj}|c_i) \textbf{(9)}$$

The effectiveness of naive Bayes classifier for text categorization has been demonstrated empirically number of times. For example, in a benchmark study where the task is to classify USENET news articles, 89% accuracy was achieved [16]. Similar results were reported in [18] and [20]over NEWSWEEDERand Yahoo new articles test collections. In [24], a naive Bayes approach is proposed for learning text classifiers from small test collections.

3.3 Modelling decision trees

Hypothesis is presented in the form of a tree in decision tree (DT) classifier. Each node of the tree represents a feature and edges represent tests on the weights that the feature has on test documents. Similarly, the leaf nodes of tree correspond to categories [22; pp. 52-78].In text categorization the nodes of the DT correspond to words because words are used as features. Edges correspond to tests on weights of words in a test document. Now, passing the document through different tests on words performs the classification. This classification starts from the word at the root node and ends after the leaf is encountered, where the label of the node is assigned to the document.

In order to learn DT, various methods have been proposed in literature, which have the same fundamental approach i.e. top-down greedy search [26]. The well-known examples of this are ID3 algorithm and C4.5. In these algorithms, the selection of a node is a way that candidate attributes (i.e. words) are evaluated using a quantitative measure known as information gain and finally, 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 the training example associated with the leaf node belongs to the same class.

In text classification, DT is used in various practices such as a main classification method in [12], a baseline classifier or a member of classifiers committee in [29]. The advantage of using DTs over most of the other machine learning methods is that humans can interpret it easily. Other methods like Naive Bayes and artificial neural network are quantitative in nature and cannot be interpreted easily.

3.4 Modelling K-nearest neighbour

k-nearest neighbour (kNN) is a paradigm of instance-based learning [22, pp.230-235]. Unlike learning an explicit target function as in the case of decision tree or neural networks, training examples are simply stored in the database. When the classification of a query (i.e. a document with unknown category) is needed, its relationship with the existing instances (i.e. documents with already known categories) is inspected to find most similar instance(s) with the query. The retrieved instance(s) are then used to classify the query. This is done in such a way that the query is assigned with the class on which majority of the retrieved instances are agreed. Besides this approach, distance weighted *kNN* is another method to make decision on the basis of retrieved instances. In this method, weights are assigned to the categories of retrieved instances predicated on their distances from the query and category with maximum aggregated weight is assigned as category of the query. As each category is weighted for each query, threshold-predicated technique can additionally be acclimated to make decision; category with weight slaking the threshold function is assigned as category of the query. However, it requires determining experimentally the threshold function from training data (i.e. genuinely from

validation dataset separate from training dataset). Despite threshold function, the method withal requires experimentally determining the number of top instances (i.e. value of k) to be considered for decision making.

Since in kNN (or distance weighted kNN) method, generalization beyond the training instances is deferred until each incipient query is not arrived, such methods are withal kenned as indolent or lazy learners.As with each query, each instance is revisited, hence, the drawback of this method is computationally expensive. Despite that, the method is widely studied and quite effective for text categorization [29]. This is the reason; we have chosen it to study its performance with other classifiers.

4.0 TEST COLLECTION AND EXPERIMENTAL SETUP

To conduct this empirical study, we have used two test collections. The first test collection is well known as EMILLE corpus1 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 includesthree 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 documents are incipient stories that were amassed from two Urdu news websites: British Broadcasting Company (BBC Urdu) and Voice of America (VOA Urdu), during the session November 1, 2011 to January 31, 2013.The documents were formatted according to the XML coding standards. An exemplary encoded document is shown in Fig. 2. Documents were categorized into four categories. These include politics, entertainment, sports and business. A category-wise distribution of the documents 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

Unfortunately, the category-wise distribution of the documents is not uniform like Entertainment category contains fewer documents as compared to other categories. The documents contained 64563 unique words. A list of 450 functional words was defined in order to perform feature selection. The list contained words such as case makers (e.g., 'kE' کے, 'sE' سے, 'nE'نے, etc., in [3]), conjunctions (e.g. 'albatah' البتہ, 'aor' اور, 'voh' وہ, 'jo' جو, 'magar' مگر, 'cUNkEh' چونکہ, agarcEh' اگرچہ, 'balkEh' بلکہ, etc., in [4]), manner adverbs (e.g. 'EsEj' جیسے, طرح 'Hara2t-sij', 'EsEa' ایسے, 'AyOg' گویا, 'tarU2is-s' صورت, and others in [5]), etc., in [6]. After removing functional or stop words, 64113 words were left. To further reduce the size of words, we applied feature selection using information gain as described in Section 2.1. This information gain was used as a measure to evaluate the effectiveness of each word to individually categorize the documents. Based on the measures words were then ranked, we then chose 30,000 top ranked words for further experimentation.

We used three standard evaluation measures known as precision, recall and f-measure to evaluate the performance of classifiers. Because accuracy may not be an effective measure here for example, in binary classification problem good accuracy may be achieved always by predicting negative class in data in presence of few positive cases. On the other hand, precision, recall and f-measure have the ability to evaluate a category-wise prediction of the classifiers. These measures with respect to positive class can be defined as given in Equations 10, 11 and 12.

[1] http://www.lancaster.ac.uk/fass/projects/corpus/emille/
[2] http://www.ota.ox.ac.uk/scripts/download.php?approval=9d5c5288a573453a422f

$$recall = \frac{\#\ of\ correct\ positive\ predictions}{\#\ of\ positive\ examples} (\mathbf{10})$$

$$precision = \frac{\#\ of\ correct\ positive\ predictions}{\#\ of\ positive\ predictions} (\mathbf{11})$$

$$f - measure = \frac{2 * precision * recall}{precision + recall} (\mathbf{12})$$

Where the symbol a denotes number of documents a classifier correctly assigned to the category (also known as true positives), b denotes the number of documents a classifier incorrectly assigned to the category (also known as false positives) and c denotes the number of documents that actually belongs to a category but classifier does not assign them to the category (also known as false negatives). To measure performance of a classifier across set of categories, we have used macroaverage: un-weighted mean of F score of all categories.

We used 5-folded cross validation in order to validate the results because number of documents in Entertainment category was limited, and it might not be useful to perform more fine-grained cross validation e.g. 10-folded cross validation.

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE Corpus SYSTEM "src/test.dtd" >
<Corpus>
<Header>
<fileDesc>
titleStmt>      Urdu Corpus          </titleStmt>
<editionStmt>   Version 1.0          </editionStmt>
<publicationStmt>
<distributor>   UOS       </distributor>
<telephone></telephone>
<eAddress>      tehseen.zia@uos.edu.pk          </eAddress>
</publicationStmt>
</fileDesc>
</Header>
<Doc>
<text>
<body>
<title>          آئی پی ایل: میکسول سب سے مہنگے کھلاڑی          </title>
<p>
آسٹریلیا کے گلین میکسویل کو ابھرتے ہوئے ہونہار کھلاڑی کی شکل میں دیکھا جا رہا ہے۔ بھارت کے جنوبی شہر چنئی میں اتوار کو ہونے والی انڈین
پریمیئر لیگ (آئی پی ایل) کی دو ہزار
تیرہ کی نیلامی میں آسٹریلیا کے گلین میکسول سب سے مہنگے کھلاڑی کے طور پر سامنے آئے۔ میکسول نیلامی میں واحد کھلاڑی رہے جن کی خدمات
ایک ملین ڈالر میں حاصل کی گئیں۔ دو ہزار تیرہ کے
سیزن کے لیے ان کی خدمات ممبئی انڈینز نے حاصل کی ہیں۔ ممبئی انڈینز کی مالکن نیتا امبانی نے کہا ہے کہ ‘میکسول ابھرتے ہوئے کھلاڑی ہیں۔ وہ عمدہ
بلے بازی کرتے ہیں، بولنگ کر سکتے ہیں اور اچھی
فیلڈنگ کرتے ہیں۔ ہم لوگوں نے جن چند ناموں پر غور کیا تھا ان میں وہ بھی تھے۔ میکسول کے بعد سب سے زیادہ قیمت میں سری لنکا کے بولر اجنتھا
مینڈس فروخت ہوئے جنہیں پونے واریئرز نے سات لاکھ
پچیس ہزار ڈالر میں خریدا۔ زیادہ تر لوگوں کو امید تھی کہ آسٹریلیا کے کپتان اور ان دنوں زبردست فارم میں نظر آنے والے مائیکل کلارک کی بولی سب
سے زیادہ لگے گی لیکن ایسا نہ ہو سکا اور ان کی
نیلامی ان کی بنیادی قیمت چار لاکھ ڈالر پر ہی ہوئی۔ پانچ گھنٹے تک جاری رہنے والی اس نیلامی میں کل سینتیس کرکٹر خریدے گئے۔ نیلامی میں کل ایک
سو ایک کھلاڑی شامل تھے اور اس بار بھی نیلامی سے
پاکستانی کھلاڑیوں کو باہر رکھا گیا۔
</p>
</body>
</text>
</Doc>
</Corpus>
```

Fig.2. An example test collection document in naive collection

The experimentation of this work was performed in WEKA[3]tool. In essence, WEKA is open source software that provides a unified workbench, which includes state of the art ML techniques such as data analysis and visualization, pre-processing, classification, clustering, etc[14].There are several reasons for which we chose WEKA to perform the experimentation. Firstly, it has state of the art built-in functionalities for tokenization, stop words removal, attribution selection, feature weighting, classification and classifier evaluation. Secondly, it is a popular tool used for text and natural language processing tasks such as language identification, named entity recognition, sentence boundary detection [8], word sense disambiguation and key phrase detection [30]. Despite that, famous natural language processing workbenches (e.g. GATE) also use the interface of WEKA [14]. Moreover, it can easily be transformed to handle Unicode format.

5.0 PARAMETER SETTING FOR CLASSIFIERS

In this section, it is described the methodology for setting parameters within the classifiers: SVM and kNN. The decision tree and naive Bayes classifiers have no such parameter that requires to be tuned. The parameter setting is performed over naivecollection with five folded cross validation test.

SVM: The experimentation with SVM classifiers is performed using a wrapper SVM tool for Weka toolkit: LIBSVM3.17[10]. We have analysed three popular types of SVM classifier: linear SVM, SVM with polynomial kernel and SVM with radial basis kernel. The critical parameter setting for SVM with polynomial kernel is valueof degree of polynomial (i.e. parameter d). We have analysed the parameter with various values of dand results is shown in Table 2 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.

Table 2. Performance evaluation of polynomial SVM with respect to degree of polynomial (i.e. d)

d	Precision	Recall	F Measure
1	**0.953**	**0.952**	**0.952**
3	0.661	0.361	0.227
5	0.107	0.327	0.161
7	0.107	0.327	0.161
9	0.107	0.327	0.161

Table 3.Performance evaluation of radial basis SVM with respect to gamma (i.e.γ)

γ	Precision	Recall	F Measure
0	**0.957**	**0.957**	**0.957**
0.1	0.707	0.415	0.328
0.2	0.741	0.374	0.256
0.3	0.758	0.371	0.25
0.4	0.758	0.371	0.249
0.5	0.757	0.37	0.247

[3]Acronym of "Waikato Environment for Knowledge Analysis"

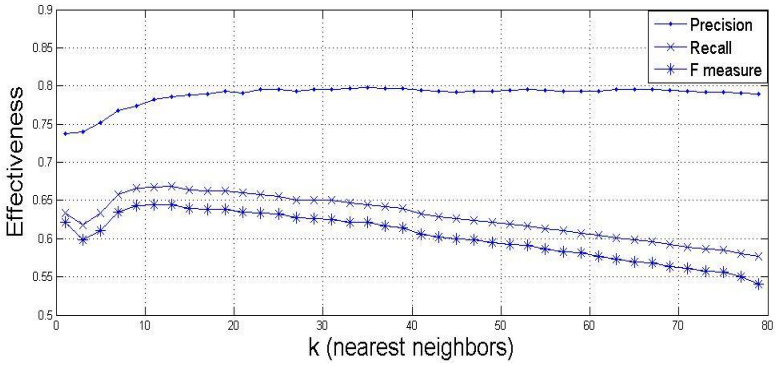

Fig. 3.K (nearest neighbours) vs. effectiveness.

The critical parameter setting for SVM with radial basis kernel is the value of gamma (γ) whichwe haveanalysed for its various values. It is found that the classifier shows better performance over gamma $\gamma = 0$ as shown in Table 5.

kNN: The classifier has two free parameters: k (neighbourhood size) and attribute set size. Since the analysis of attribute set size is main topic of the paper, it is analysed in detail below in Section 7. In this section, we have presented the result of analysis of parameter k which is tested for following values:

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

The result of the analysis is shown in Fig. 2. We have selected value of k to be 13 for rest of the experimentations.

6.0 RESULTS

In this section we have presented effectiveness of classification algorithms trained on top-ranked features sets of varied size.Features are ranked based on their effectiveness which is measured by using feature evaluation criterion underlying feature selection methods.After ranking features, sets of top-ranked features are used for training the classifiers. The performance of classification methods are then measured using f-measure.

Table 4.Macroaveraged F1 measure of different feature selection methods in combination with naive Bayes for self-collected test collection. Top-ranked feature set sizes are shown along horizontal axis and effectiveness (F1 measure) of the classifier is shown vertically.

No. of top ranked features	IG	GR	CHI	SU	OR	RA
100	0.905	0.893	0.918	0.906	0.879	0.203
200	0.933	0.907	0.936	0.932	0.891	0.234
300	0.936	0.906	0.939	0.937	0.904	0.242
400	0.939	0.924	0.942	0.941	0.909	0.251
500	0.938	0.932	0.943	0.942	0.908	0.308
1000	0.934	0.942	0.935	0.937	0.904	0.381
1500	0.929	0.945	0.929	0.93	0.899	0.42
2000	0.925	0.946	0.925	0.924	0.897	0.455
3000	0.921	0.947	0.92	0.921	0.898	0.508
4000	0.917	**0.95**	0.918	0.92	0.898	0.557
5000	0.916	0.95	0.916	0.916	0.897	0.596
10000	0.916	0.916	0.916	0.916	0.898	0.671
15000	0.916	0.916	0.916	0.916	0.898	0.7
20000	0.916	0.916	0.916	0.916	0.899	0.885

The experimental results as shown in Table4 (for naive collection)and Table5 (for EMILLE collection)have endorsed the known fact that naive Bayes classifier takes advantage of appropriate feature selection[18, 19]. Gain ratio (GR)has shown consistently high performance in both test collections.

Table 5.Macroaveraged F1 measure of different feature selection methods in combination with naive BayesforEMILLEtest collection.

No. of top ranked features	IG	GR	CHI	SU	OR	RA
100	0.86	0.93	**0.954**	0.93	0.907	0.53
200	0.84	0.93	0.93	0.905	0.887	0.534
300	0.882	0.882	0.882	0.882	0.907	0.527
400	0.839	0.839	0.839	0.839	0.907	0.531
500	0.839	0.839	0.839	0.839	0.907	0.548
1000	0.835	0.835	0.835	0.835	0.883	0.577
1500	0.785	0.785	0.785	0.785	0.883	0.66
2000	0.813	0.813	0.813	0.813	0.862	0.631
3000	0.837	0.837	0.837	0.837	0.885	0.643
4000	0.716	0.716	0.716	0.716	0.812	0.633
5000	0.655	0.655	0.655	0.655	0.785	0.536
10000	0.647	0.647	0.647	0.647	0.645	0.6

Similar to naive Bayes, KNN is also known to be highly susceptible to number of irrelevant features [22, pp:231-236; 33, pp:323] and our results as shown in Table 6 (for naive collection) and Table 7 (for EMILLE collection) have further endorsed the fact. IG has consistently shown high performance in both test collections. However, due to the small size of EMILLE test collection, the results of KNN over this test collection are not encouraging.

Table6.Macroaveraged F1 measure of different feature selection methods in combination with KNN for naive test collection**.**

No. of top ranked features	IG	GR	CHI	SU	OR	RA
100	0.889	0.896	0.885	0.885	0.882	0.218
200	**0.913**	0.865	0.909	0.907	0.881	0.244
300	0.911	0.851	0.908	0.915	0.854	0.232
400	0.89	0.847	0.887	0.894	0.825	0.235
500	0.893	0.843	0.874	0.885	0.786	0.276
1000	0.827	0.871	0.824	0.833	0.682	0.285
1500	0.791	0.864	0.784	0.794	0.575	0.256
2000	0.754	0.875	0.743	0.752	0.494	0.232
3000	0.743	0.877	0.729	0.748	0.422	0.21
4000	0.706	0.862	0.705	0.718	0.395	0.196
5000	0.705	0.801	0.7	0.706	0.367	0.262
10000	0.692	0.695	0.693	0.692	0.324	0.268
15000	0.689	0.689	0.689	0.689	0.304	0.197
20000	0.679	0.679	0.679	0.679	0.295	0.357

The results of feature selection methods over J48 classifier are shown in Table 8 and Table 9. GR has shown top resultsin both collections. During experimentations, we have observed that IG, CHI and SU performed uniformly and have similar effects over the performance of classifiers. All of them have the potential to reduce about 99% or more of total features with promising performance of classifiers (as measured by macroaveraged F1).

Table 7.Macroaveraged F1 measure of different feature selection methods in combination with KNN for EMILLE corpus. Feature selection methods IG, GR, FA, SC are shown with curve since they have same results.

No. of top ranked features	IG	GR	CHI	SU	OR	RA
100	0.548	0.308	0.518	**0.55**	0.354	0.308
200	**0.55**	0.518	0.444	0.55	0.308	0.308
300	0.354	0.354	0.354	0.354	0.308	0.313
400	0.354	0.354	0.354	0.354	0.308	0.436
500	0.354	0.354	0.354	0.354	0.308	0.442
1000	0.354	0.354	0.354	0.354	0.308	0.364
1500	0.308	0.308	0.308	0.308	0.308	0.334
2000	0.308	0.308	0.308	0.308	0.308	0.392
3000	0.355	0.355	0.355	0.355	0.308	0.412
4000	0.313	0.313	0.313	0.313	0.308	0.368
5000	0.364	0.364	0.364	0.364	0.308	0.368
10000	0.308		0.308	0.308	0.308	0.308

Table 8.Macroaveraged F1 measure of different feature selection methods in combination with J48 for naive collection.

No. of top ranked features	IG	GR	CHI	SU	OR	RA
100	0.909	0.879	0.905	0.908	0.888	0.189
200	0.917	0.892	0.915	0.915	0.894	0.208
300	0.907	0.896	0.911	0.912	0.899	0.208
400	0.909	0.911	0.91	0.911	0.898	0.215
500	0.908	0.915	0.909	0.909	0.897	0.254
1000	0.903	0.919	0.904	0.901	0.897	0.306
1500	0.899	0.919	0.897	0.902	0.896	0.337
2000	0.902	0.92	0.901	0.901	0.896	0.337
3000	0.902	0.921	0.901	0.9	0.896	0.372
4000	0.9	**0.924**	0.9	0.9	0.896	0.408
5000	0.901	0.923	0.9	0.9	0.896	0.494
10000	0.901	0.904	0.9	0.9	0.896	0.61
15000	0.901	0.904	0.9	0.9	0.896	0.659

Table 9.Macroaveraged F1 measure of different feature selection methods in combination with J48 forEMILLE corpus.

No. of top ranked features	IG	GR	CHI	SU	OR	RA
100	0.633	**0.73**	0.67	0.633	**0.73**	0.561
200	0.629	0.629	0.629	0.629	0.683	0.604
300	0.629	0.629	0.629	0.629	0.683	0.6
400	0.629	0.629	0.629	0.629	0.683	0.571
500	0.629	0.629	0.629	0.629	0.683	0.571
1000	0.629	0.629	0.629	0.629	0.629	0.607
1500	0.629	0.629	0.629	0.629	0.629	0.573
2000	0.629	0.629	0.629	0.629	0.629	0.573
3000	0.629	0.629	0.629	0.629	0.629	0.573
4000	0.629	0.629	0.629	0.629	0.629	0.623
5000	0.629	0.629	0.629	0.629	0.629	0.615
10000	0.629		0.629	0.629	0.629	0.588

In comparison to naive Bayes and KNN classifiers, SVM does not get much benefit from feature selection.Our resultsas shown in Table10and Table 11havefurther endorsed the finding of [17]: one reason why SVM works well for text categorization is its overfitting protection mechanism. Therefore, considering sufficiently large number of features (i.e. approximately 10,000), the selection of appropriate feature selection methods (except RA) does not seem to be a question of choice.

Table10.Macroaveraged F1 measure of different feature selection methods in combination with SVM for self collected test collection.

No. of top ranked features	IG	GR	CHI	SU	OR	RA
100	0.928	0.911	0.927	0.926	0.917	0.232
200	0.94	0.922	0.94	0.94	0.923	0.275
300	0.941	0.924	0.936	0.937	0.925	0.301
400	0.94	0.932	0.941	0.941	0.927	0.323
500	0.943	0.937	0.94	0.942	0.931	0.365
1000	0.949	0.942	0.948	0.946	0.941	0.411
1500	0.95	0.942	0.953	0.952	0.943	0.431
2000	0.951	0.938	0.951	0.95	0.942	0.46
3000	0.949	0.947	0.954	0.954	0.944	0.51
4000	0.954	0.952	0.956	0.953	0.943	0.59
5000	0.956	0.957	0.958	0.956	0.943	0.648
10000	0.958	**0.959**	0.958	0.958	0.946	0.767
15000	0.958	0.958	0.958	0.958	0.945	0.818
20000	0.957	0.957	0.956	0.956	0.944	

Table 11.Macroaveraged F1 measure of different feature selection methods in combination with SVM forEMILLE corpus.

No. of top ranked features	IG	GR	CHI	SU	OR	RA
100	**0.838**	0.811	0.885	0.838	0.854	0.453
200	0.812	0.812	0.837	0.812	0.836	0.4
300	0.837	0.837	0.837	0.837	0.813	0.403
400	0.809	0.809	0.809	0.809	0.812	0.402
500	0.836	0.836	0.836	0.836	0.788	0.418
1000	0.836	0.836	0.757	0.757	0.734	0.467
1500	0.836	0.836	0.681	0.681	0.76	0.505
2000	0.836	0.836	0.752	0.752	0.76	0.535
3000	0.836	0.836	0.752	0.752	0.835	0.608
4000	0.836	0.836	0.654	0.654	0.835	0.634
5000	0.836	0.836	0.627	0.627	0.835	0.612
10000	0.836	0.836	0.436	0.436	0.835	0.436

Results of feature selection methods for SVM with polynomial kernel are shown in Table 12 and Table 13. This variant of SVM turned out to be very vulnerable with respect to number of features. Performance of the classifier gradually decreases with number of features: the decrease is more smoothed over naive collection due to its moderate size and sharp over EMILLE collection due to its small size. IG feature selection method has shown consistently top performance in both collections.

Table 12.Macroaveraged F1 measure of different feature selection methods in combination with SVM with polynomial kernel for self collected collection.

No. of top ranked features	IG	GR	CHI	SU	OR	RA
100	0.917	0.878	0.921	0.916	0.908	0.205
200	0.944	0.9	0.944	0.942	0.93	0.197
300	0.95	0.891	0.947	0.95	0.941	0.184
400	0.953	0.914	0.951	0.952	0.946	0.333
500	0.952	0.921	0.953	0.954	0.945	0.172
1000	0.955	0.941	0.954	0.955	0.948	0.17
1500	**0.955**	0.94	**0.956**	0.955	0.943	0.167
2000	0.953	0.935	0.953	0.953	0.94	0.163
3000	0.945	0.924	0.944	0.945	0.933	0.166
4000	0.94	0.92	0.941	0.94	0.927	0.171
5000	0.934	0.916	0.934	0.933	0.918	0.176
10000	0.897	0.897	0.897	0.897	0.874	0.183
15000	0.844	0.844	0.844	0.844	0.808	0.185

Table 13.Macroaveraged F1 measure of different feature selection methods in combination with SVM with polynomial kernel forEMILLEcollection.

No. of top ranked features	IG	GR	CHI	SU	OR	RA
100	**0.698**	0.566	0.68	0.669	0.566	0.308
200	0.538	0.506	0.506	0.538	0.566	0.308
300	0.468	0.468	0.468	0.468	0.588	0.308
400	0.308	0.308	0.308	0.308	0.518	0.308
500	0.308	0.308	0.308	0.308	0.518	0.308
1000	0.308	0.308	0.308	0.308	0.308	0.308
1500	0.308	0.308	0.308	0.308	0.308	0.308
2000	0.308	0.308	0.308	0.308	0.308	0.308
3000	0.308	0.308	0.308	0.308	0.308	0.308
4000	0.308	0.308	0.308	0.308	0.308	0.308
5000	0.308	0.308	0.308	0.308	0.308	0.308
10000	0.308	0.308	0.308	0.308	0.308	0.308

Results of feature selection methods over SVM with radial basis kernel are shown in Table 14 and Table 15. IG feature selection method again (like SVM with polynomial kernel) has shown consistently top performance in both collections.

Table 14.Macroaveraged F1 measure of different feature selection methods in combination with SVM with radial basis kernel for self collected collection.

No. of top ranked features	IG	GR	CHI	SU	OR
100	0.921	0.891	0.926	0.919	0.916
200	0.948	0.913	0.945	0.947	0.936
300	0.954	0.909	0.949	0.954	0.942
400	0.958	0.926	0.955	0.956	0.949
500	0.957	0.933	0.957	0.959	0.951
1000	**0.96**	0.947	0.958	0.959	0.951
1500	0.96	0.949	0.961	0.961	0.951
2000	0.959	0.947	0.959	0.959	0.951
3000	0.956	0.943	0.955	0.956	0.947
4000	0.954	0.942	0.954	0.954	0.94
5000	0.953	0.945	0.952	0.952	0.937
10000	0.929	0.93	0.93	0.929	0.908
15000	0.906	0.906	0.906	0.906	0.879

Table 15.Macroaveraged F1 measure of different feature selection methods in combination with SVM with radial basis kernel for EMILLE collection.

No. of top ranked features	IG	GR	CHI	SU	OR	RA
100	**0.813**	0.731	0.731	0.813	0.783	0.359
200	0.731	0.731	0.731	0.731	0.73	0.359
300	0.566	0.566	0.566	0.566	0.647	0.391
400	0.506	0.506	0.506	0.506	0.649	0.308
500	0.468	0.468	0.468	0.468	0.649	0.308
1000	0.308	0.308	0.308	0.308	0.556	0.308
1500	0.308	0.308	0.308	0.308	0.308	0.308
2000	0.308	0.308	0.308	0.308	0.308	0.308
3000	0.308	0.308	0.308	0.308	0.308	0.308
4000	0.308	0.308	0.308	0.308	0.308	0.308
5000	0.308	0.308	0.308	0.308	0.308	0.308
10000	0.308	0.308	0.308	0.308	0.308	0.308

Finally we have shown a collective viewof results of all four classifiers with top performing feature selection methods in Fig.4 for naive collection and in Fig.5 for EMILLE collection. The performance of classifiers with

respect to OR feature selection method is skipped during the illustration of results since OR consistently perform badly. Other feature selection methods that have performed equivalently are shown with one curve.

It can be seen from figure that linear SVM with any of five feature selection methods has outperformed other classifiers for moderate size naive collection. Moreover, it has no dependency over feature selection method. On the other hand, KNN classifier gives worse performance than others and greatly depends on the 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. That is, at some point, it shows comparable performance with SVM when GR is used as feature selection method.

On the other hand, naive Bayes classifier with any feature selection method has shown its supremacy for small size EMILLE collection. Linear SVM with any feature selection methods (except OR) is found to be second top performing classifier with the advantage that selection of appropriate feature set size is not as critical parameter to choose as in naive Bayes.

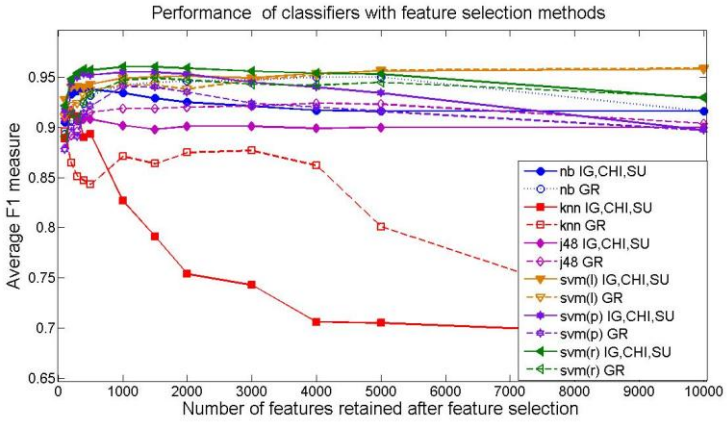

Fig.4.Macroaveraged F1 measure of six classifiers in combination with five top feature selection methods. Feature selection methods IG, CHI, FA and SC have performed equivalently so they are illustrated with a single curve. For example, nb IG+CHI+FA+SC shows the performance of naive Bayes with feature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

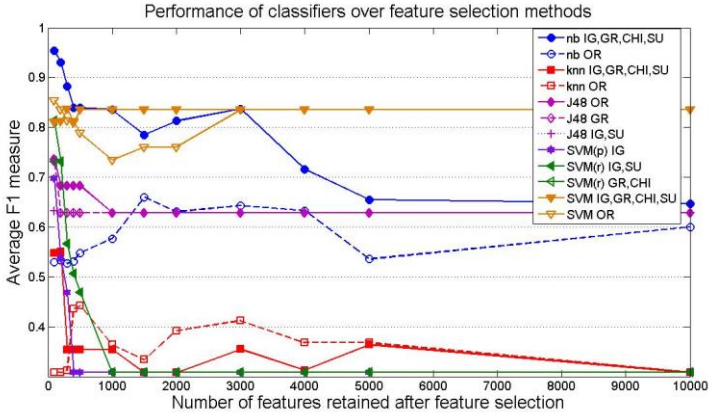

Fig. 5.Macroaveraged F1 measure of six classifiers in combination with five top feature selection methods with emile corpus.

7.0 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.

8.0 FUTURE WORK

This work can further be extended in three directions. The first direction is to analyse the impact of these feature selection methods on other text classifiers such as Rocchio and recently proposed Fuzzy Soft Set based classifier [13]. The second direction is to include other recently proposed feature selection methods such as distinguishing feature selector (a new probabilistic feature selection method) in analysis. An interesting analysis in this regard could be to empirically evaluate the effectiveness of topic models (probabilistic graphical models primarily used for document indexing) as feature selection methods [34]. Third direction is to develop a huge test collection like RCV1 [18] and reproduce the results.

REFERENCES

[1] Abbas, Q. "Exploiting Language Variants Via Grammar Parsing Having Morphologically Rich Information". In Proceedings of the EMNLP'2014 Language Technology for Closely Related Languages and Language Variants, Association of Computational Linguistics, 2014, P 35-45, Qatar.

[2] Abbas, Q., Zia, T. and Khan, A.N. "Syntactic and Semantic Analysis of Urdu Modal Verbs using XLE Parser". International Journal of Computer Applications, Vol. 107(10), pp. 39-46, 2014,USA

[3] Abbas Q. and Raza G. "A Computational Classification of Urdu Dynamic Copula Verb". International Journal of Computer Applications (IJCA), Vol. 85(10), P 1-12. 2014, ISSN 0975 - 8887. Published by Foundation of Computer Science, New York, USA.

[4] Abbas, Q. "Building a Hierarchical Annotated Corpus of Urdu: The URDU.KON-TB Treebank". Lecture Notes in Computer Science (LNCS). Vol. 7181(1), P 66-79, 2012,ISSN 0302-9743, Springer-Verlag Berlin/Heidelberg.

[5] Abbas Q. "Semi-Semantic Part of Speech Annotation and Evaluation". In Proceedings of ACL 8th Linguistic Annotation Workshop (COLING), Association of Computational Linguistics, 2014, P 75-81, Ireland.

[6] Abbas, Q. "A Stochastic Prediction Interface for Urdu", International Journal of Intelligent Systems and Applications (IJISA), Vol.7, No.1, PP.94-100, 2015. DOI: 10.5815/ijisa.2015.01.09

[7] Abbas, Q. Khan, N.A. "Lexical functional grammar for Urdu modal verbs". In Proceedings of 5th IEEE International Conference on Engineering and Technology(ICET), 2009.

[8] Balie D. Nadeau.,"Baseline Information Extraction: Multilingual Information Extraction from Textwith Machine Learning and Natural Language Techniques". Technical Report, University of Ottawa, 2005.

[9] Burges, C., J., C., "A Tutorial on Support Vector Machines for Pattern Recognition". *Data Mining and Knowledge Discovery*, Vol.2, 1998, pp.121-167.

[10] Chih-Chung Chang and Chih-Jen Lin.,"LIBSVM : A Library for Support Vector Machines". *ACM Transactions on Intelligent Systems and Technology*, Vol. 2, No. 3, 2011.

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

[12] Fuhr, N., Buckley, C.,"A Probabilistic Learning Approach for Document Indexing". *ACM Transaction on Information Systems*,Vol. 9, No.3, 1991, pp.223-248.

[13] Handaga,B., Deris, M., "Text Categorization Based on Fuzzy Soft Set Theory". Lecture Notes in Computer Science Volume 7336, 2012, pp. 340-352.

[14] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, L. H., "The Weka Data Mining Software: An Update", *SIGKKD Exploration*, Vol.11, No.1, 2009, pp.10-18.

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

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

[17] Lewis, D, D. Knowles, K, A.,"Threading Electronic Mail: A Preliminary Study". *Information Processing and Management*, Vol. 33, No.2, 1997, pp.209-217.

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

[19] Lewis, D, D., Yang, Y, G. Ross, T. Li, F., "RCV1: A New Benchmark Collection for Text Categorization Research". *Journal of Machine Learning Research*, Vol.5, 2004, pp.361-397.

[20] Li, Y. H., Jain, A. K.,"Classification of Text Documents",*Computer Journal*,Vol. 41, No.8, 1998, pp. 537-546.

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

[22] Mitchell, T.,*Machine Learning*, McGraw-Hill, New York, 1997.

[23] Moohebat, M., Raj, R.G. , Kareem, S.B.A., Thorleuchter, D., "Identifying ISI-indexed articles by their lexical usage: A text analysis approach", *Journal of the Association for Information Science and Technology*, Vol. 66, No. 3, pp. 501–511. doi: 10.1002/asi.23194.

[24] Nigam, K. Mccallum K, A. Thrun, S. Mitchell, T., "Text Classification from Labeled and Unlabeled Documents using EM",*Machine Learning*, Vol.39 (2/3), 2002, pp.103-134.

[25] Pazzani, M, J., Muramatsu, J., &Billsus, D., "Syskill and Webert: Identifying Interesting Websites",*Proceedings of Thirteenth Conference on Artificial Intelligence*, 1996, pp. 54-59.

[26] Raj, R.G., Abdul-Kareem, S., "Information Dissemination And Storage For Tele-Text Based Conversational Systems' Learning",*Malaysian Journal of Computer Science*, Vol. 22(2):2009. Pp. 138-159.

[27] Riaz, K.,"Rule-Based Name Entity Recognition in Urdu",*Proceeding to Name Entity Workshop*, 2010, pp. 126-135.

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

[29] Sebastiani, F.,"Machine Learning in Automated Text Categorization",*ACM Computing Surveys*, Vol. 34, No.1, 2002, pp.1-47.

[30] Yeow, W.L., Mahmud, R., Raj, R.G., "An application of case-based reasoning with machine learning for forensic autopsy", *Expert Systems with Applications*, Vol 41, No. 7, 2014, pp. 3497-3505, ISSN 0957- 4174, http://dx.doi.org/10.1016/j.eswa.2013.10.054. (http://www.sciencedirect.com/science/article/pii/S0957417413008713).

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

[32] Witten I.H., Frank, E.,*Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufman Publishers, Second edition, 2005.

[33] Yang, Y. "Expert network: Effective and Efficient Learning from Human Decisions in Text Categorization and Retrieval". *Proceedings of ACMSIGIR*, 1994, pp. 13-22.

[34] Zhu, Y., Li, L., Luo, L., "Learning to Classify Short Text with Topic Model and External Knowledge", Knowledge Science, Engineering and Management, Lecture Notes in Computer Science, Vol. 8041, 2013, pp 493-503.