I J C T A, 10(8), 2017, pp. 763-773 ISSN: 0974-5572
© International Science Press

Filter Feature selection Approaches for Automated Text Categorization

Smita Shedbale*, Kailash Shaw and Pradeep Kumar Mallick*****

ABSTRACT

Automated categorization of texts into predefined categories has become a booming interest over last few decades, due to the increasing availability of documents in digital form as well as emerging need to organize them for classification and prediction purpose. To deal with a big challenge in text categorization which is learning from high dimensional data, feature selection becomes very important. By selecting only subsets of features that are relevant and important for processing, feature selection speeds up learning process. There are various feature selection approaches that are addressed well in literature. This paper surveys existing literature about feature selection approaches to text categorization.

Keywords: Text Categorization, Feature selection, Filter, Wrapper, Information Gain

1. INTRODUCTION

As the representation of document is widely increasing in digital form, it is necessary to categorize them into predermined set of categories in automatic way; this process is called as automated text categorization. Each document can be in multiple, exactly one, or no category at all. There are various machine learning algorithms have been developed to address this challenge by composing it as a classification problem [1][2]. Generally, automatic text classification process automatically constructs a classifier by learning, from a set of prelabeled documents.

There are number of ways to represent the document to calculate the form of term and the calculation of weight of term. But most widely used document representation for text categorization is by using one of the basic model called as "bag-of-words", where occurrence of each word is considered as a feature for training a classifier. This method of document representation is called as a Vector Space Model [3], where each feature in a feature space corresponds to term or a phrase in a vocabulary collected from a particular data set. The value of each feature represents the importance of the term in the document, according to a specific feature measurement.

A major challenge in text categorization is the learning from high dimensional data. Firstly, a document may consist of large number of words that is hundreds, thousands of words. Applying original document directly which contains this large number of words, may result into a high computational load for the learning process. And second problem is that, many of the words from original document may be irrelevant to the topic or redundant, so use of those words will reduce the performance of classifier. To avoid the issue of the learning from high dimensional data and to make the learning process faster, it is important to reduce original feature space to few important words/feature preserving semantic meaning of original document.

* M.E. Student, Dept. of Computer Engineering, D.Y. Patil College of Engineering, Akurdi, Savitribai Phule Pune University, India, *Email: Simran1008@gmail.com*

** Assistant Professor, Dept. of Computer Engineering, D.Y. Patil College of Engineering, Akurdi, Savitribai Phule Pune University, India, *Email: kailash.shaw@gmail.com*

*** Department of Computer Science & Engineering, Vignana Bharathi Institute of Technology, Hyderabad, India, *Email: pradeepmallick84@gmail.com*

The process of reducing the original size of feature space is called as dimensionality reduction. A most common dimensionality reduction approach used in text categorization is feature selection. Feature selection aims at selecting only a subset of relevant features from the set of original large set of features. Reducing irrelevant and redundant features, feature selection approach improves the overall performance of classification.

Over a last decade, a number of feature selection approaches have been proposed, which can be generally categorized into the following two types: the filter approach and the wrapper approach [4]. The filter approach selects feature subsets based on the general characteristics of the data without involving the learning algorithms that will use the selected features. A score which indicates the "importance" of the term is assigned to each individual feature based on an independent evaluation criterion, such as distance measure, dependency measure, entropy measure, and consistency measure. Hence, the filter approach only selects a number of high ranked features and ignores the rest. Alternatively, the wrapper approach intensely searches for best features with an evaluation criterion based on the learning algorithm. Although it has been shown that the wrapper approach usually performs better than the filter approach, it has much more computational cost than the filter approach, which sometimes makes it difficult to use in real world.

Typically, the filter approach is predominant and most widely used in text classification because of its simplicity and efficiency. However, the filter approach evaluates the importance of a feature by only exploiting essential characteristics of the training data without considering the learning algorithm for discrimination, which may lead to an unexpected classification performance. Given a particular learning algorithm, it is hard to guess which filter feature selection approach will give best discrimination [5].

The rest of the paper is organized as follows. Section II provides filter feature selection approaches for automated text categorization. Section III provides materials and analyses methods used in the literature. Section IV provides result analysis of the methods used in various papers in the literature. Section V provides details of performance evaluation metrics which are widely used for text categorization.

2. FILTER FEATURE SELECTION APPROACHES

Feature selection which is also called term selection, is a widely used approach for reducing dimensionality in text categorization. Suppose original vocabulary of document contains M number of terms, the number of terms to be selected is considered as t which is a specified integer value; the feature selection approach tries to select t out of M terms from the original document. Yang and Pedersen have shown that implementation of feature selection approach can remove 98 percent unique terms without adversely affecting the classification performance too much, and thus feature selection approach can reduce the computational burden for classification [6].

Wrapper feature selection approach used by Kohavi and John is a feature selection method which a feature is either added or removed at each step towards the optimal feature subset selection [7]. Once a new feature set is generated, the classifier is retrained with new generated features and it is next tested on a validation data set. This approach greedily searches the feature space and is always finds a better feature subset in order to get an improved classification performance. But, the high computational cost and complexity makes this approach impractical to use in text categorization applications.

The another possibility of choice is the filter approach for feature selection, in which each feature in the original feature space is assigned with a score according to its importance and only the high ranked features are selected to form subset of relevant features [6][8]. This gives an advantage of this approach in the context of its easy implementation with low computational cost. Several state-of-the-art measures that are widely used in text categorization are described as below.

Document frequency which is one of the basic feature selection approaches is defined as the number of documents in which a particular term present. It removes terms which are found to be useless for classification from the original large feature space. This approach is very effective and also inspired researcher mainly in

the domain of text mining in their experiments to remove all the terms that occur no more than a specified times (a usually range specified is from 1 to 3) in the training set as a preprocessing stage. By doing so, tens of hundreds rare features can be removed before the step of feature selection. TF-IDF measure considers both term frequency and inverse document frequency to calculate the importance or weight of features [9].

Term frequency tf(*t, d*) represents the number of times that term *t* occurs in document *d.* Document frequency df_t, defined to be the number of documents in the collection that contain a term t. The inverse document frequency is a measure of how much information the word provides, that is, whether the term is common or rare across all documents. It is the logarithmically scaled inverse fraction of the documents that contain the word, obtained by dividing the total number of documents by the number of documents containing the term, and then taking the logarithm of that quotient. Inverse document frequency of a term t is defined as,

$$Idf_t = \log (N/df_t)$$

where N denotes total number of documents in a collection. For rare terms, idf value is high and conversely for frequent terms idf value is low. A generalized TF-IDF measure [10] is proposed which considers various different level of hierarchies among words which effectively analyzes tweet user behaviors. Generalized TF-IDF measure is the measure which involves TF and generalized form of IDF.

Many other filter approaches are based on the information theory measures, including mutual information, information gain, relevancy score, chi-square statistic, odds ratio, expected cross entropy for text (CET), GSS coefficient, power, relevancy score and many more. Some of these measures are described below.

Mutual information (MI) [5] measures the mutual dependency of two variables. For a term t_k and a category c_i, the MI measure between t_k and c_i is defined to be

$$MI(t_k, c_i) = \log \frac{p(t_k, c_i)}{p(t_k)\,p(c_i)}$$

where $p(t_k, c_i)$ denotes the probability that the term t_k appears in a document and this document belongs to the category c_i, $p(t_k)$ is the probability that the term t_k appears in a document, and $p(c_i)$ is the probability that a document belongs to the category ci. One can see that $MI(t_k, c_i)$ is zero if t_k and c_i is independent, i.e., the term t_k is useless for discriminating the documents belonging to the category c_i.

Information Gain (IG) [6] measures the decrease in entropy when the feature is present vs. feature is absent.

$$IG(t_k, c_i) = p(t_k, c_i)\log\frac{p(t_k, c)}{p(t_k)\,p(c)} + p(\overline{t_k}, c_i)\log\frac{p(\overline{t_k}, c)}{p(\overline{t_k})\,p(c)}$$

where $p(\overline{t_k}, c_i)$ denotes the probability that the term t_k does not present in a document and this document belongs to the category c_i, $p(\overline{t_k})$ is the probability that the term t_k does not present in a document. IG principle usually has good performance than the MI principle and is less influenced by the terms which appear rare.

Expected *Cross Entropy for Text* (CET) [8] is proposed as

$$CET(t_k, c_i) = p(t_k, c_i) log \frac{p(t_k, c_i)}{p(t_k)\,p(c_i)}$$

Chi-square statistic is proposed in [6] to measure the lack of independence between the term t_k and the category c_i, which is modeled by a Chi-square (χ^2) distribution. By considering the negative evidence of term in a document, a general χ^2 statistic measure is defined as

$$CHI\left(t_k, c_i\right) = \frac{\left[\, p\left(t_k, c_i\right) p\left(\overline{t}_k, \overline{c}_i\right) - p\left(\overline{t}_k, \overline{c}_i\right) p\left(\overline{t}_k, \overline{c}_i\right)\right]^2}{p\left(t_k, c_i\right) p\left(t_k, \overline{c}_i\right) p\left(\overline{t}_k, c_i\right) p\left(\overline{t}_k, \overline{c}_i\right)}$$

where the document space is divided into two categories, c_i denotes ith category and its complement $\overline{c}_i$ that is the pool of all the remaining categories, p($\overline{t}_k$, $\overline{c}_i$) denotes the probability that the term t_k does not appear in a document and also this document does not belong to the category c_i, and p (t_k, $\overline{c}_i$) denotes the probability that the term t_k appears in a document but this document does not belong to the category c_i. A modified measure which is called as GSS coefficient using negative evidence is proposed by Galavotti et al. in [11], which is defined as,

$$GSS(t_k, c_i) = p(t_{k,} c_i)\, p(\overline{t}_{k,} \overline{c}_i) - p(t_k, \overline{c}_i)\, p(\overline{t}_k, c_i)$$

It has been shown that this measure outperforms the original chi-square measure on several data sets [11].

Another measure which is based on Information Theory is Kullback–Leibler divergence (KLD) is also called as discrimination information. Kullback and Leibler is a measure of information from statistical aspects of view, involving two probability distributions associated with the same experiment, calling discrimination function [12]. Considering a two-class classification problem each class is represented by a particular distribution, say P_1 for class c1 and for P_2 for class c2. A test procedure for classification can be considered as a binary hypothesis testing such that if a sample is drawn from P1, hypothesis H1 will be accepted (hypothesis H2 will be rejected), and if a sample is drawn from P2, hypothesis H2 will be accepted (H1 will be rejected). P ($x|H_i$) is considered as class conditional probability distribution [5]. According to the Information Theory, KL-divergence KL (P_1, P_2) between two probability distributions (from P_1 to P_2) is defined as

$$KL\left(P_1, P_2\right) = \int_x p\left(x\middle|H_1\right) \log \frac{p\left(x\middle|H_1\right)}{P\left(x\middle|H_2\right)} dx$$

This KL divergence is a non-symmetric information theoretic measure of distance between two probability distributions (from P_1 to P_2). The smaller the relative entropy, the more similar the distribution of the two variables. This non symmetric form of expression is not strictly a distance metric. To obtain a symmetric measure, one can define J-divergence which is based on the Kullback–Leibler divergence, with some useful differences, including that it is symmetric and it is always a finite value, defined as

$$J\,(P_1, P_2) = KL\,(P_1, P_2) + KL\,(P_2, P_1)$$

$$= \int_x \left[\, p\left(x\middle|H_1\right) - p\left(x\middle|H_2\right)\right] \log \frac{p\left(x\middle|H_1\right)}{P\left(x\middle|H_2\right)} dx$$

J divergence is only defined for binary hypothesis. It can be extended for multiple hypothesis testing (i.e., multi-class classification) in which the divergences of each individual distribution with a reference distribution are calculated and summed together. Here, first J-divergence is generalized to the multi-distribution using the scheme of “one-vs-all”[13] which is called as Jeffreys-Multi-Hypothesis Divergence, which is defined as follows:

$$JMH\left(P_1, P_{2\ldots}, P_N\right) = \sum_{i=1}^{N} KL\left(P_i, \overline{P}_i\right)$$

Where, $\overline{P}$ is the combination of all remaining N – 1 distributions.

Almost all of these filter approaches based on the information theory measures use binary variables, e.g. the presence (t_k) or the absence ($\bar{t}_k$) of a term in a document, and a document belonging to a category (c_i) or not ($\overline{c_i}$). Some of the approaches make use of the term occurrence to measure the term importance in the document, and hence richer information is contained. Meanwhile, these existing filter approaches rank the features by only exploring the intrinsic characteristics of data based on the feature relevancy without considering their discriminative information in classifiers. It is also possible that class-specific features can be used for text categorization. Instead of using the combination operation to select a global feature subset for all classes, such approach selects a specific feature subset for each class, namely class specific features. Existing feature importance evaluation criteria can still be applied in this kind of approach [14], [15].

3. MATERIALS AND METHODS

This section gives details of the datasets which are widely used in text categorization and various feature selection approaches for text categorization.

3.1. MATERIALS

This section focuses on data sets which are widely used by many of the research based on text mining such as text categorization and text clustering.

3.1.1. 20 NEWSGROUP

The 20-NEWSGROUPS benchmark consists of about 20,000 documents collected from the postings of 20 different online newsgroups or topics. The 20-Newsgroups were collected and had become one of the standard corpora for text categorization. This dataset is known for its large size and balanced categories. In this benchmark, note that some topics are hierarchically categorized, and few different topics could be very closely related to each other. The “bydate” version of this dataset contains a standard train/test split. Most of the experiments are carried on this benchmark which is created by Deng et al.[16].

3.1.2. The REUTERS

The REUTERS originally contains 21, 578 documents with 135 topics, but some documents belong to multiple topics. Most of the experiments uses the ModApte version of the Reuters by removing those documents with multiple labels. This version consists of 8; 293 documents in 65 topics. Most of the experiments divide REUTERS into three data sets, named REUTERS-10, REUTERS-20 and REUTERS-30, consisting of the documents of the first 10, 20 and 30 topics, respectively [16][17].

3.1.3. Topic Detection and Tracking (TDT2)

The TDT2 benchmark consists of 11, 201 documents taken from two newswires (AP WorldStream and New York Times Newservice), two radio programs (PRI The World and VOA World News) and two television program (CNN Headline News and ABC World News Tonight). Also, those documents that belong two or more topics have been removed. Because of the extremely imbalanced data for some categories, some experiments in literature only uses the first 10 topics.

3.1.4. WebKB

WebKB is a collection of web pages from four different college web sites. The 8282 web pages are nouniformly assigned to 7 categories like student, faculty, staff, department, course, project, and other.

3.1.5. CLASSIC3

CLASSIC3 dataset (comprising of 3 collections of abstracts). CRAN, CISI, MED. Classic3 whose class distribution is nearly homogenous among three classes. Classic-3 data set has 273 features, 3000 instances and 3 classes.

3.1.6. Bloggender

These datasets contain both numerical and categorical values with various dimensionalities and numbers of instances. This dataset has Bloggender-male and Bloggender-female which consists of 3232 number of instances and 101 number of attributes of each.

3.1.7. Oshumed

The first version of Ohsumed dataset is a subset containing most frequent 10 categories called 0h10, and the other set is complete Ohsumed dataset with all 23 categories called oh23. First twenty thousand documents from Ohsumed dataset as used by Joachims (1998).This dataset is considered very hard to classify due to its high sparsity.

3.1.8. BCII

A binary categorization data set Data BCII is obtained from the Protein Interaction Article Sub-task (IAS) of BioCreAtIvE II challenge. The Data BCII is composed of abstracts of 6,172 articles in total, which are taken from a set of MEDLINE articles that are annotated as interaction articles.

3.2. Methods

This section focuses on the Methods used in different research papers in the context of text categorization.

1. “Toward Optimal feature selection in Naïve Bayes for Text Categorization” [5] present a novel and efficient feature selection framework based on the Information Theory, which aims to rank the features with their discriminative capacity for classification. Paper introduces a new divergence measure, called Jeffreys-Multi-Hypothesis (JMH) divergence, to measure multi-distribution divergence for multi-class classification. Based on the JMH-divergence, author developed two efficient feature selection methods, called maximum discrimination (MD) and MD_Chi^2 methods, for text categorization. Experimentations are carried out over 20-Newsgroups, Reuters, and Topic Detection and Tracking (TDT2).
2. “A term weighting scheme based on the measure of relevance and distinction for text categorization” [18] proposed a novel feature-selection algorithm, named *AD*, which comprehensively measures the degree of relevance and distinction of terms occur in document set. Author have used three benchmark data collections 20-Newsgroups, Reuters-21578 and WebKB for evaluation.
3. “A study on term weighting for text categorization: a novel supervised variant of TF. IDF” [19] proposed a supervised variant of the tf-idf scheme, based on computing the usual idf factor without considering documents of the category to be recognized, so that importance of terms frequently appearing only within it is not underestimated. A further proposed variant is additionally based on relevance frequency, considering occurrences of words within the category itself. Approach use in paper used Reuters-21578 corpus and 20 Newsgroups to evaluate its effectiveness.
4. “A discriminative and semantic feature selection method for text categorization” [20] proposed a novel feature selection method that first selects features in documents with discriminative power

and then computes the semantic similarity between features and documents. All the experiments were conducted on two benchmark datasets, Reuter-21578 and 20-Newsgroups.

5. "Chi-square statistics feature selection based on term frequency and distribution for text categorization" [21] propose a modified CHI feature selection approach which is called term frequency and distribution based CHI to overcome these weaknesses. Sample variance is used to calculate the term distribution, and improve the classic CHI with maximum term frequency. Behaviour of proposed method were analysed using different types of data Reuters-21578, 20-Newsgroup WebKB.
6. "New feature selection methods based on context similarity for text categorization." [22] proposed four new context similarity based feature selection methods, GIcs, DFcs, CDMcs and Acc2cs. Before the process of text categorization, each raw document in the article collection is transformed into a big vector according to the bag-of-words document representation. Then the feature filter methods, such as the GI, DF, CDM and Acc2, are utilized to selection most important features from the feature vector space based on document frequency. Two benchmark data sets from different application domain are used for experiments, DataBCII and Reuters-21578.

4. RESULT AND ANALYSIS

Tang, Bo, Steven Kay et al. [5] proposed "Toward optimal feature selection in naive Bayes for text categorization". To compare the performance of these feature selection methods, author evaluated the classification accuracy and the F1 measure metric of naive Bayes and SVM classifiers with different number of features ranging from 10 to 2,000. Authors first tested these feature selection approaches when naive Bayes is used as the classifier. Fig. 1 shows the results on the 20-NEWSGROUPS data set. It can be shown that the performance is improved when more features are selected. The proposed two approaches commonly perform better than others. As shown in the figure, the proposed MD method performs better than the others. The DF method is the worst one for this data set [5].

Fig. 2 shows the result on the data set of ALT-COMP that is a subset of 20-NEWSGROUPS with the categories alt.* and the categories comp.*. For this data set, the MD-χ^2 is the best one among all others.

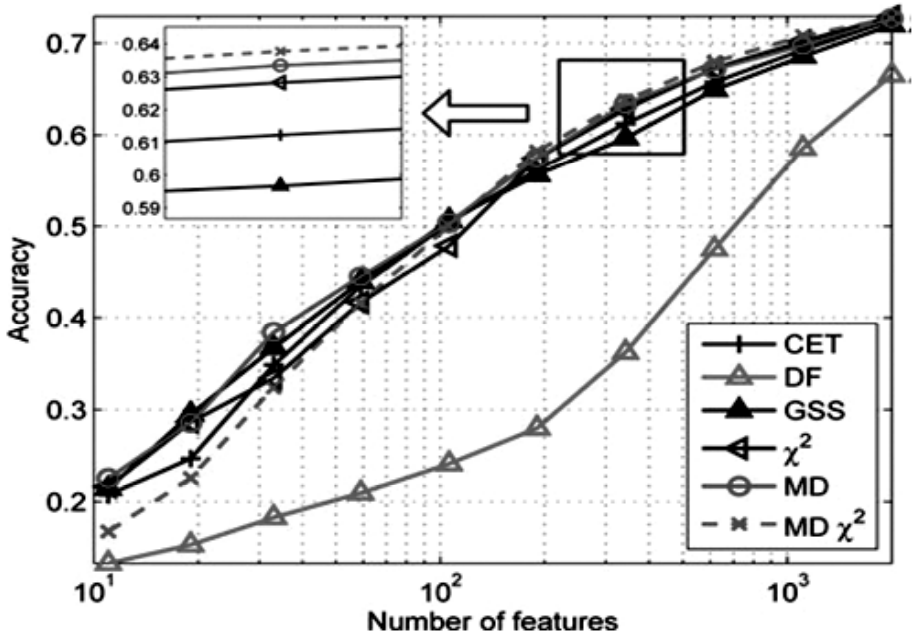
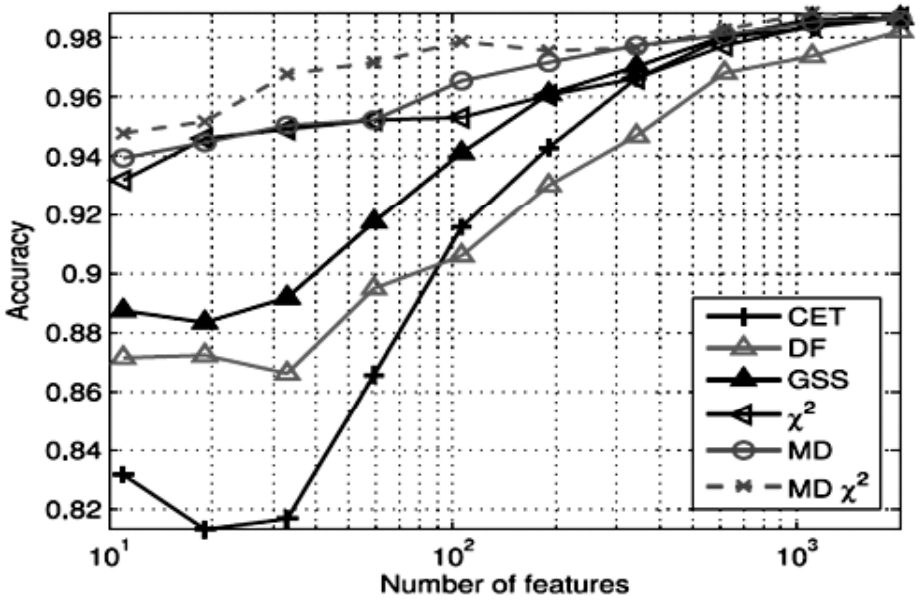


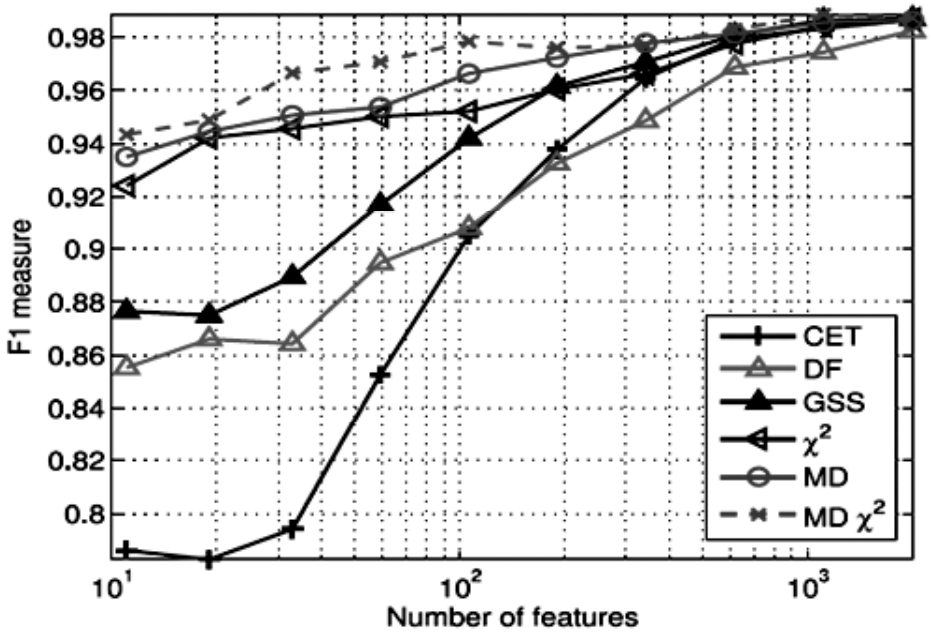
Figure 1: An accuracy comparison of feature selection methods on the 20-NEWSGROUPS with 20 topics for naive Bayes [5]

The comparison of the F1 measure on ALT-COMP is given in Fig. 2b and shows that the proposed MD and its asymptotic χ^2 statistic are the best two approaches.

Yang, Jieming, et al. [18] proposed "A term weighting scheme based on the measure of relevance and distinction for text categorization." The efficiency of the proposed measure AD (Association and distinction) was examined through the experiments of text categorization with NB and SVM classifier. The results, comparing with six classic feature-selection algorithms, (Information Gain (IG), Mutual Information (MI), Odds Ratio (OR), DIA association factor (DIA), Orthogonal Centroid Feature Selection (OCFS) and Ambiguity Measure(AM)), show that the proposed method AD is significantly superior to MI, OR, DIA, OCFS, AM when Naive Bayes is used and significantly outperforms IG, MI, OR, DIA, AM when Support Vector Machines is used.



(a) Accuracy for ALT-COMP [5]



(b) Accuracy for ALT-COMP[5]

Figure 2: Performance comparisons of feature selection methods on the ALT-COMP data set: (a) accuracy and (b) F1 measure, when naive Bayes is used as the classifier. [5]

Domeniconi, Giacomo, et al. [19] proposed "A study on term weighting for text categorization: a novel supervised variant of TF. IDF". For each of the three dataset used in paper author tested the classification varying the number of features selected for each category and tested against the performance of other term weighting methods used in the paper in terms of micro-F1 and macro-F1 on the different datasets. Proposed scheme tf.idfec-based achieved top results in all datasets and with each classifier.

Zong, Wei, et al. [20] proposed "A discriminative and semantic feature selection method for text categorization". Proposed feature selection method from this framework i.e., the discriminative feature selection method (DFS), and the discriminative and semantic feature selection method (DFS + Similarity) is compared with some of the traditional feature selection methods. When compared with the traditional feature selection methods, the proposed method produced better and more stable results than the traditional methods.

Jin, Chuanxin, et al. [21] proposed "Chi-square statistics feature selection based on term frequency and distribution for text categorization". TF often impacts the topics of document collection therefore, author proposed modified approaches based on CHI. In these approaches, it is also consider the term distribution which is another important parameter for CHI. Through extensive experiments on three common text corpora with kNN classifer, it is observed that the Micro-F1 and Macro-F1 performances of proposed approach is better than the classic methods in most instances.

Chen Y, B Han et al. [22] proposed "New feature selection methods based on context similarity for text categorization". Using different data sets from different application domain, the effectiveness of the proposed methods were investigated and compared against well-known frequency based techniques. The proposed methods can achieve better performances on both binary and multi-classification problems. Through experimental analysis, the results shows that the context similarity based methods outperform the corresponding frequency based methods in terms of the micro and macro F1 measures both on binary and multi-classification problems.

5. EVALUATION METRICS

Generally, these following metrics are used to evaluate the classification performance: accuracy, precision, recall, and F1 measure. The accuracy metric which is widely used in machine learning domain indicates the overall classification performance. The precision is the percentage of documents that are correctly classified as positive out of all the documents that are classified as positive, and the recall is the percentage of documents that are correctly classified as positive out of all the documents that are actually positive [5]. The metrics of precision and recall and accuracy are defined as

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

where TP denotes the number of true positive, FP denotes the number of false positive, and FN denotes the number of false negative. These two metrics are inversely proportional to each other. In other words, increasing the precision is at the cost of reducing the recall, and vice versa.

Another metrics which combines the precision and recall called F-measure or F_1 score can be interpreted as a weighted average of the precision and recall which is define as

$$\text{F1 measure} = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

The metrics of precision, recall and F1 measure are originally defined for binary class. For multi-class classification, we follow several other studies [1], [23], [24] in which binary classifiers are built for each individual class and a global F1 measure is obtained by averaging the F1 measure of each class weighted by the class prior.

6. CONCLUSION

The recent developments in feature selection have addressed the problem from the point of view of improving the performance of classification and prediction. They have met the challenge of operating on input spaces of large dimensions consisting of hundreds or several thousand variables. Various feature selection methods are widely used in text categorization with the aim to reduce dimensionality of original feature space and hence improving overall performance of classification. Performance of feature selection also varies based on the type of data it works on. Feature selection mainly categorizes into filter and wrapper approaches. Filter approach ranks feature according to importance and based upon general characteristics without involving classifiers feedback on the selected subset. And hence provides significant performance over wrapper approach. Wrapper approach finds best subset of feature from original large set of feature because it evaluates the selected features with the help of classifiers feedback. It works best for small data set. But as the number of feature increases performance of wrapper becomes worse because of high computational cost. Because of its efficiency and simplicity, filter approach is most widely used approach in automated text categorization. There are various filter feature selection techniques which are based on information theory. Maximum Discrimination which is one of the feature selection technique based on information theory shown promising performance improvement compared to the existing feature selection methods.

ACKNOWLEDGMNT

The author would like to thank the publishers and researchers for making their resources available. We also thank the college authority for providing the required infrastructure and support. Finally we would like to extend our heartfelt gratitude to friends and family members.

REFERENCES

[1] T. Joachims, “Text categorization with support vector machines: Learning with many relevant features,” in Proc. Eur. Conf. Mach. Learn., 1998, pp. 137–142.

[2] W. Lam, M. Ruiz, and P. Srinivasan, “Automatic text categorization and its application to text retrieval,” IEEE Trans. Knowl. Data Eng., vol. 11, no. 6, pp. 865–879, Nov./Dec. 1999

[3] Harish, Bhat S., Devanur S. Guru, and Shantharamu Manjunath. “Representation and classification of text documents: A brief review.” *IJCA, Special Issue on RTIPPR (2)* (2010): 110-119.

[4] H. Liu and L. Yu, “Toward integrating feature selection algorithms for classification and clustering,” IEEE Trans. Knowl. Data Eng., vol. 17, no. 4, pp. 491–502, Apr. 2005.

[5] Tang, Bo, Steven Kay, and Haibo He. “Toward optimal feature selection in naive Bayes for text categorization.” (2016).

[6] Y. Yang and J. O. Pedersen, “A comparative study on feature selection in text categorization,” in Proc. Int. Conf. Mach. Learn., vol. 97, 1997, pp. 412–420.

[7] R. Kohavi and G. H. John, “Wrappers for feature subset selection,” Artif. Intell., vol. 97, no. 1, pp. 273–324, 1997.

[8] D. Mladenic and M. Grobelnik, “Feature selection for unbalanced class distribution and naive Bayes,” in Proc. Int. Conf. Mach. Learn., 1999, vol. 99, pp. 258–267.

[9] G. Salton, A. Wong, and C.-S. Yang, “A vector space model for automatic indexing,” Commun ACM, vol. 18, no. 11, pp. 613–620, 1975.

[10] F. Bouillot, P. N. Hai, N. B_echet, S. Bringay, D. Ienco, S. Matwin, P. Poncelet, M. Roche, and M. Teisseire, “How to

extract relevant knowledge from tweets?" in Proc. Int. Workshop Inform. Search, Integr. Personalization, 2013, pp. 111–120.

[11] L. Galavotti, F. Sebastiani, and M. Simi, "Experiments on the use of feature selection and negative evidence in automated text categorization," in Proc. 4th Eur. Conf. Res. Adv. Technol. Digit. Libraries, 2000, pp. 59–68.

[12] Bigi, Brigitte. "Using Kullback-Leibler distance for text categorization." *European Conference on Information Retrieval*. Springer Berlin Heidelberg, 2003.

[13] R. Rifkin and A. Klautau, "In defense of one-vs-all classification," J. Mach. Learn. Res., vol. 5, pp. 101–141, 2004.

[14] B. Tang, H. He, P. M. Baggenstoss, and S. Kay, "A Bayesian Classification approach using class-specific features for text categorization," IEEE Trans. Knowl. Data Eng., vol. 28, no. 6, pp. 1602–1606, Jun. 2016.

[15] B. Tang, S. Kay, H. He, and P. M. Baggenstoss, "EEF: Exponentially embedded families with class-specific features for classification," IEEE Signal Process. Lett., in press, 2016.

[16] D. Cai, X. He, and J. Han, "Document clustering using locality preserving indexing," IEEE Trans. Knowl. Data Eng., vol. 17, no. 12, pp. 1624–1637, Dec. 2005.

[17] D. Cai, Q. Mei, J. Han, and C. Zhai, "Modeling hidden topics on document manifold," in Proc. ACM Conf. Inform. Knowl. Manage., 2008, pp. 911–920.

[18] Yang, Jieming, et al. "A term weighting scheme based on the measure of relevance and distinction for text categorization." Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), 2015 16th IEEE/ACIS International Conference on. IEEE, 2015.

[19] Domeniconi, Giacomo, et al. "A study on term weighting for text categorization: a novel supervised variant of TF. IDF." Proceedings of the 4th international conference on data management technologies and applications (DATA). Candidate to the best conference paper award. 2015.

[20] Zong, Wei, et al. "A discriminative and semantic feature selection method for text categorization." International Journal of Production Economics 165 (2015): 215-222.

[21] Jin, Chuanxin, et al. "Chi-square statistics feature selection based on term frequency and distribution for text categorization." IETE Journal of Research 61.4 (2015): 351-362.

[22] Chen, Yifei, Bingqing Han, and Ping Hou. "New feature selection methods based on context similarity for text categorization." *Fuzzy Systems and Knowledge Discovery (FSKD), 2014 11th International Conference on*. IEEE, 2014.

[23] R. Liere and P. Tadepalli, "Active learning with committees for text categorization," in Proc. 14th Conf. Amer. Assoc. Artif. Intell., 1997, pp. 591–596.

[24] E. F. Combarro, E. Montanes, I. Diaz, J. Ranilla, and R. Mones, "Introducing a family of linear measures for feature selection in text categorization," IEEE Trans. Knowl. Data Eng., vol. 17, no. 9, pp. 1223–1232, Sep. 2005.