Journal of Engineering and Applied Sciences 13 (Special Issue 9): 7074-7080, 2018
ISSN: 1816-949X

A Technical Study on Feature Ranking Techniques and Classification Algorithms

Wareesa Sharif, Noor Azah Samsudin, Mustafa Mat Deris and Shamsul Kamal Ahmad Khalid
Faculty of Computer Science and Inormation Technology,
Universiti Tun Hussein Onn Malaysia (UTHM), 86400 Parit Raja, Johor, Malaysia

Abstract: Since, electronic documents are dramatically increasing therefore document classification becomes a very important task to organise inormation automatically. Text documents are a high dimensional data that create difficulties in classification task. Consequently, various feature ranking techniques are used to reduce the dimensionality of the text data. Features can be selected by using document frequency and term frequency techniques. In term frequency, few researchers have worked on term frequency while keeping the property of document frequency. Little attention has been paid to compare these techniques. In this study, we describe issues of feature ranking techniques and classification document labelling problem. This study also present and discussed experimental result of feature ranking technique with presence or absence of term and term frequency (term count) in document classification problem. The result shows that redesigning of term into term count with document frequency could lead to better classification accuracy than that of term frequency and document frequency separately.

Key words: Feature ranking, filter, wrapper, embedded, machine learning algorithms, dimensional

INTRODUCTION

Printed and online data is the text document which is mostly available free of cost for search and read. Text can be in many different formats such as word, phrase, term, pattern, sentence, paragraph and document. In moderate size text datasets, the number of words can be easily grown in tens of thousands. In other words, documents are represented high dimensional data (Tian *et al*., 2016; Yang *et al*., 2012). This huge amount of data is difficult to manage and extract for mining purposes (Yang *et al*., 2012). Li and Chen (2012) described that the higher dimensionality of feature space impose weighty overhead to mining techniques, e.g., classification because some features can be redundant or irrelevant and misguide classification process (Javed *et al*., 2012). Feature selection plays an important role to reduce the dimensionality of text data. Text datasets have high dimensional data (Forman, 2003) such as reuters 21578 (Javed *et al*., 2015), 20 newsgroup (Feng *et al*., 2015) and Ohsumed (Siddiqui, 2016). Feature selection is a pre-processing step to find minimum subset of features from the large amount of data that use the relevant features from dataset to make adequate classification (Bachrach *et al*., 2004). There are three types of feature selection techniques: filters, wrapper and embedded (Uysal, 2016). During the construction of feature set, embedded and wrapper need frequent classifiers interaction in flow but filter does not. Interaction with classifier may increase computational time. Due to interaction with classifier reason, filter based methods are more preferred to use as compare to wrapper and embedded (Uysal, 2016). Feature selection is divided into feature subset selection and feature ranking. This division is based on how the features are combined for classification evaluation (Gnana *et al*., 2016). In feature ranking method, individual features produce good result by using filtering techniques such as inormation gain (Forman, 2003), Chi-square (Manning *et al*., 2008) and distinguishing feature selector (Uysal and Gunal, 2012). These filter techniques make the dataset small in size, less computation, reduce over fitting and increase generalization (Meenakshi, 2013). Karabulut *et al*. (2012) proved that feature selection techniques affect the classification (Ting *et al*., 2011) performance of algorithms like Naive Bayes, SVM and decision tree. So, feature selection and classifier inluence each other. Forman (Forman, 2003) accomplished in feature selection experimental study, feature vectors are used as Boolean (0, 1). But these feature selection techniques cannot count the word frequency (repetition of word). Most of the feature selection work on document frequency (Lee and Lee, 2006; Azam and Yao, 2012). Document frequency considers the presence/absence of term in positive/negative classes. The number of times a term appears not included in document frequency. So, Li and

Corresponding Author: Wareesa Sharif; Faculty of Computer Science and Information Technology,
Universiti Tun Hussein Onn Malaysia (UTHM), 86400 Parit Raja, Johor, Malaysia

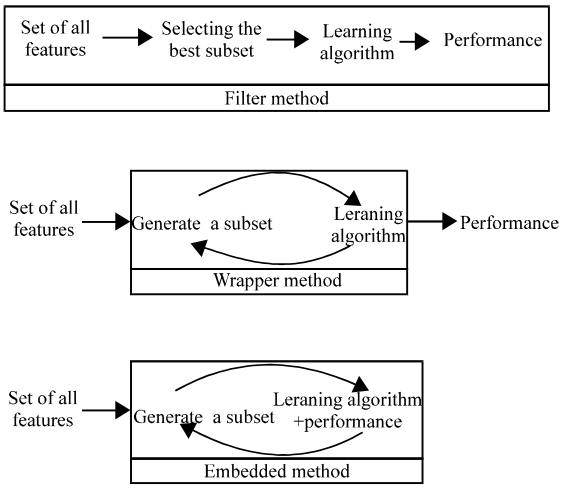

Fig. 1: Types of feature selection techniques

Chen, 2012; Uysal and Gunal, 2012) introduced the "high term frequency and "weighted document frequency", called as "HTF-WDF) use the actual value of the word. For text data, many feature ranking techniques are in literature but it is still an ongoing research area and need new techniques to find the distinctive features to improve the classifier accuracy. Features can be selected by document frequency and term frequency techniques. Based on the literature, few researchers have worked on term with term count but based on literature, very little work done on the comparison of these techniques. In this study, we described issues of feature ranking techniques experimental result of feature ranking technique with presence or absence of term, frequency of terms and its term count in document classification problem. The result shows in table V that use of new redesigned term count by Rehman *et al*., (2015) gave better result.

Literature review: A document is represented by multi-dimensional feature vectors in which each dimension corresponds to a weighted value such as e.g., Chi-square, IG (Forman, 2003; Manning *et al*., 2008). In a text dataset, moderate number of text collection produce result in hundreds and thousands number of features. Therefore, the most important problem is to reduce the high dimensionality feature space in document. So, feature selection is very important in text classification to improve the accuracy, precision, recall and f-Measure. There are three types of feature selection: filters, wrapper and embedded (Uysal, 2016) as shown in Fig. 1.

According to a chosen weighting technique, filter techniques evaluate every term independently. Filter rank the features after evaluation and takes the subset with the highest weights (Forman, 2003). Wrapper algorithms depend on the chosen classifier. They evaluate subsets of the initial term set and best performance subset is selected. It is time consuming process. Heuristic algorithms are used for selection in wrapper (Babu and Vijayan, 2016). Embedded algorithms tried to choose features during training process like artificial neural networks (Joachims, 1998).

MATERIALS AND METHODS

Document frequency (binary) based feature ranking methods: In text classification, many feature ranking methods such as Chi-square (Forman, 2003) inormation gain (Manning *et al*., 2008) and odd ratio (Mladenic and Grobelnik, 1999) can work with binary inormation (presence/absence) of term.

Chi-square: Chi-square is the popular approach for feature selection and is used to study independence of two events like X,Y as:

$$P(XY) = P(X)P(Y) \tag{1}$$

The event X and Y are assume independent and belong to term and class, respectively in text classification (Manning *et al*., 2008). Information Gain (IG) is used to measure the presence and absence of term which contribute for correct classification decision (Forman, 2003):

$$IG(t) = -\sum_{i=1}^{M} P(C_i \log P(C_i) P(t) \sum_{i=1}^{M} P(C_i|t) \log P(C_i|t) + P(\bar{t}) \sum_{i=1}^{M} (C_i|t) \log P(C_i|t) \tag{2}$$

Where:

M = The number of classes
$P(C_i)$ = Probability of class C_i, P(t)
$P(\bar{t})$ = The conditional probabilities of class given presence, absence of term t

IG method select terms distributed in many categories but these terms have less discriminating power in text classification tasks.

Odd ratio is a way to find the membership and non-membership of a class with its nominator and denominator, respectively. By dividing of nominator with denominator each other can normalized the nominator and denominator (Mladenic and Grobelnik, 1999). Denominator should be minimized and nominator should be maximized to get the highest score. It is a one sided metric:

$$OR(t, C_i) = \log \frac{P(t|C_i)\left[1-P(t|\bar{C}_i)\right]}{\left[1-P(t|C_i)\right]P(t|\bar{C}_i)} \tag{3}$$

Where:

$P(t|C_i)$ = Probability of term which tell the presence of class C_i

$P(t|\overline{C}_i)$ = Conditional probability of the term t given classes except C_i

This method is applied to avoid from division by zero error and prevent the nominator become zero.

Term frequency based filter methods: Most of the feature ranking methods are document frequency based like chi-square inormation gain, odd ratio etc. They just show the presence and absence of term in document. These methods do not use the actual value of a term but term frequency based filter use the actual value of term.

Distinguishing feature selector (DFS): Yusal (Uysal and Gunal, 2012) introduced a probabilistic feature ranking method in which it assigns a high rank to the frequently occurring term in one class irrespective of the class size. To rank the term an initial framework of DFS is presented as foloow:

$$DES(t) = \sum_{i=1}^{M} \frac{P(C_i|t)}{P(\overline{t}|C_i)+(t|\overline{C}_i)+1} \quad (4)$$

M denotes number of classes and $P(C_i|t)$ is conditional probability of C_i given presence of term. $P(\overline{t}|C_i)$ is also conditional probability and tell the absence of term t for class C_i, respectively.

µ-document based approach: Baccianella *et al*., 2013) claim that existing feature ranking techniques do not deem the term frequency (term count) to compute the rank of term. By using term frequencies, they logically break the document into "Micro-documents" in which every micro-document contains one word. The result has improved by using the micro-document for ordinal text classification.

New Feature Selection (NFS): Chang *et al.* (2015) described that imbalanced data in document classification occurred because positive class is smaller than negative class. Imbalanced data can cause poor result in minority class. To avoid from poor result, improve the minority class without disturbing overall classification performance. In this study, NFS method selects class inormation words. NFS method use data re-sampling for imbalanced problem. So, data re-sampling technology (SMOTE) improve the minority class issues .The result for macro-F1 is 0.7792 with more than 1000 features.

High Term Frequency, Weighted Term Frequency: (HTF-WDF) is the simplest method but it has some drawbacks. Term is important and appears few times in single category. Due to low requency, it will be eject through feature selection. Some terms appeared frequently in some categories similarly. These terms are difficult to find which category they actually belong to. So, HTF-WDF is measured terms with occurrences and high occurrences. This technique gave result recall 97.41% [3]:

$$WDF_i = DF_i\left(1+1/\max_{1\le j\le n_c} DF'_j\right) \quad (5)$$

Total number of categories is defined by n_c Terms are placed by WDF and term with low WDF discarded.

Relative Discriminative Criterion (RDC): Rehman *et al.* (2015)**:** Rehman redesigned the document frequency of term into each term count. They give the high score to frequently occurring terms in one class but absent in other classes. They also used the frequency graphs to avoid the same score of different terms. They did experiment and produced Micro-F1 59% on reuter 21578:

$$RDC = \frac{(|tpr_{tc}-fpr_{tc}|)}{\min(tpr_{tc},fpr_{tc})*tc} \quad (6)$$

Normalized Relative Discriminative Criterion (NRDC): Wang *et al.* (2015) introduce NRDC feature ranking technique and claimed that terms in long documents are big but in short documents are small. Wang *et al.* (2015) tried to normalise the term count and removes the biasness in the long documents. These techniques are presented in Table 1.

Table 1: Features ranking techniques with classifiers

Research	Techniques	Classifier	Results
Forman (2003)	IG	SVM	F-measure 0.25 Precision 0.41% Recall 0.34
Mladenic (1999)	(OR)	MNB	Fmeasure 0.48% Precision 0.44% Recall 0.80%
Uysal (2012)	DFS	SVM DT	Micro F1, 85.79 Macro F1, 64.93.4 MicroF183.28, Macro-F1 61.25 Micro-F1 85.92, Macro-F1 64.32
Li (2012)	HTF-WDF	SVM	Recall 97.41%
Baccianella (2013)	µdocument	SVM	Rang 2.48% (SVOR on TripAdvisor-15763) to 6.29%
Chang (2015) and Wang *et al.* (2015)	NFS	SVM	Macro-F1 77 %
Rehman (2015)	RDC	MNB SVM	Micro-F1 59% Micro-F1 63%
Wang (2015) Paul (2014)	NRDC	K-ELM	Accuracy 70%

Classification techniques: Due to the increasing number of documents from different sources, text classification

becomes important task. Classification is used to assign data into predefined categories. There are many techniques available to classify documents such as Naive Bayes, SVM, k-NN etc.

Naive Bayes (NB): Naive Bayes is mostly used due to its ease of use and simplicity for text classification (Kulkarni *et al*., 2012). In terms of accuracy and computational efficiency, Naive Bayes is considered best classifier (Paul, 2014) among other classifiers like decision tree, neural network and svm. Two types of NB are Bernoulli model and multinomial model. Multinomial modelis best for large datasets but there are two problems, rough parameter estimation and handling rare categories which contain few training documents (Korde and Mahender, 2012):

$$p(C_j|d) = \frac{p(d|c_j).p(c_j)}{p(d)} \quad (7)$$

$p(c_j)$ is the prior probability of class c_j, $p(d)$ is the inormation of observations which is the knowledge from the text itself to be classified and $p(d|c_j)$ is the distribution probability of document d in classes. Suppose components d_i is independent with each other, conditional probability cannot be computed directly. So:

$$p(d|C_j) = \Pi_i p(d_i|C_j) \quad (8)$$

Ting *et al*. (2011) described that Naive Bayes is effective in text classification. MNB is also used with map reduce framework to remove redundant and irrelevant feature. They achieved good result as 88% f-measure (Wang *et al*., 2015).

Support Vector Machine (SVM): SVM is a supervised learning approach that is used to classify linear and non-linear data (Cortes and Vapnik, 1995). SVM is used to define the optimal decision boundary to classify different classes:

- It is best for binary classification but can be used for multi classification problems
- If suitable pre-processing is used for numerical vectors and sparse distribution, SVM can produce good result

SVM provide consistently better result than Naive Bayes but it is very time consuming and take more computation time, especially in training model (Mladenic and Grobelnik, 1999).

Decision Tree (DT): In text classification, DT is a tree with node that is labelled by term branches. It can also be specified that each internal node represents a test and each branch represent an outcome of the test. Each leaf node holds a class label. It is a top-down method (Nidhi and Gupta, 2011). DT works like 'divide and conquer' approach for classification. It is simple to understand and easy to interpret. Modifications and addition of new possible scenario can also be easily done but if a level of a tree increases the complexity of calculations also increases.

Measuring criteria: In text classification, measure exactness of classifier is:

$$\text{Precision} = \frac{TP}{TP+FP} \quad (9)$$

Where:
TP = Ture positive rate
FP = False positive

$$\text{Recall} = \frac{TP}{TP+FN} \quad (10)$$

Where:
TP = True positive rate
FN = False negative

$$\text{Accuracy} = \frac{TP+TN}{TP+TM+FP+FN} \quad (11)$$

RESULTS AND DISCUSSION

Experimental setup, result and analysis: Feature ranking becomes critical part because text data has large number of attributes. So, it is necessary to reduce the irrelevant features. Different feature ranking techniques are investigated and some of them are tested with supervised learning techniques like multinomial Naive Bayes, SVM and decision tree on reuter dataset. Dataset has 65 different categories in which 8293 documents present. 18933 terms are present. Top 10 classes are selected for experiments. Implementation of feature selection algorithm is done on Java platform and for implementation of machine learning algorithm WEKA tool is used. 10 fold cross validation is used for experimental analysis. Dataset is divided into 10 parts randomly. Process repeated 10 times for training and testing phases and shows the average result in table V.

It is proved in Table 2, that without feature ranking, multinomial Naive Bayes gave 40%, J48 gave 20% and SVM gave 25% accuracy with reuter 21578. So, result is

Table 2: Results of experiment on reuters 21578

Feature ranking/Classifier	Accuracy (%)
Without feature rnking	
Multinomial Naive Bayes	40
Decision tree (J48)	20
Support vector machine	25
Information gain	
Multinomial Naive Bayes	45
Decision tree (J48)	20
Support vector machine	26
Relative discriminative criterion	
Multinomial Naive Bayes	46.66
Decision tree (J48)	53
Support vector machine	29
Normalized relative	
Discriminative criterion	
Multinomial Naive Bayes	66.66
Decision tree (J48)	70
Support vector machine	40

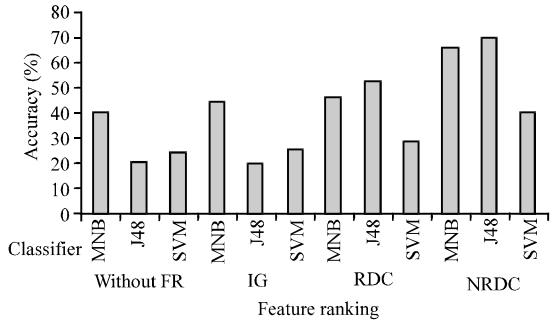

Fig. 2: Result of Reuter21578

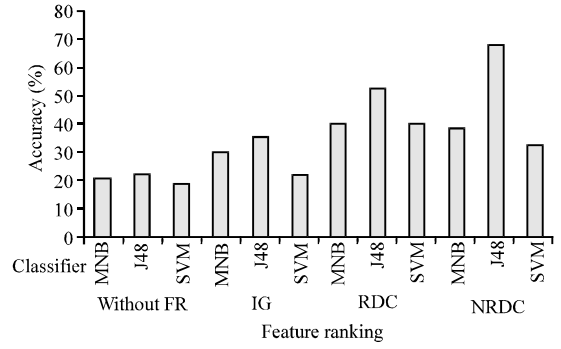

Fig. 3: Result of 20 newsgroups

not good without feature ranking technique because classifier used irrelevant features that are burden for classifier. IG gave lower result than RDC or NRDC. Term frequency and document frequency are used combine in NRDC and RDC. With MNB, NRDC gave 66.66% and SVM gave 40% accuracy. Results are shown in Table 2 and Fig. 2. Researchers (Yang *et al*., 2012; Forman, 2003) consider the presence/absence of term but they ignore the actual value of the term that is important to know the actual size of dataset. We can also know by term count,

Table 3: Results of experiment on 20 newsgroups

Feature ranking/Classifier	Accuracy (%)
Without feature ranking	
Multinomial Naive Bayes	20
Decision tree (J48)	22
Support vector machine	18
Information gain	
Multinomial Naive Bayes	30
Decision tree (J48)	35
Support vector machine	22
Relative discriminative criterion	
Multinomial Naive Bayes	40
Decision tree (J48)	52
Support vector machine	40
Normalized relative discriminative criterion	
Multinomial Naive Bayes	38
Decision tree (J48)	68
Support vector machine	32

that dataset is skew or not. For 20 newsgroup dataset, RDC perform better result than IG, NRDC. NRDC perform better and gave 68% result with DT. With SVM, RDC perform better than other feature ranking techniques present in Table 3 and Fig. 3. Rehman *et al*. (2015), Kulkarni *et al*. (2012) gave value to the frequently occurring terms in one class. They did not used for highly skew dataset. Wang *et al*. (2015), Paul (2014) proposed NRDC for long and short documents to normalize the term count but increase the complexity of the algorithm. Experimental result can be seen in Fig. 2. By using reuter 21578 dataset, RDC and NRDC provide better result than IG with MNB, SVM and DT because RDC, NRDC used actual value of term.

CONCLUSION

Various feature ranking techniques have been used in document classification to improve the performance of classifier. RDC proved that the impact of term and its term count with graphs enhance the performance of filters techniques. Normalized relative discriminative criterion is used to handle the skew problems in text data. Nine feature ranking techniques are reviewed. Experiment is conducted on IG, RDC and NRDC with SVM, decision tree (J48), multinomial Naive Bayes in document classification. There is need to explore more feature ranking techniques and suitable classifier to label the documents.

ACKNOWLEDGEMENTS

This research is supported by Fundamental Research Grant Scheme 1609, Research Acculturation Grant Scheme R045 and U497 at Universiti Tun Hussein Onn Malaysia and in part by a grant from Research Gates IT Solution Sdn. Bhd.

REFERENCES

Azam, N. and J. Yao, 2012. Comparison of term frequency and document frequency based feature selection metrics in text categorization. Expert Syst. Appl., 39: 4760-4768.

Babu, R.L. and S. Vijayan, 2016. Wrapper based feature selection in semantic medical information retrieval. J. Med. Imaging Health Inf., 6: 802-805.

Baccianella, S., A. Esuli and F. Sebastiani, 2013. Using micro-documents for feature selection: The case of ordinal text classification. Expert Syst. Appl., 4: 4687-4696.

Bachrach, R.G., A. Navot and N. Tishby, 2004. Margin based feature selection-theory and algorithms. Proceedings of the 21st ACM International Conference on Machine Learning, July 04-08, 2004, ACM, Banff, Alberta, ISBN:1-58113-838-5, pp: 1-43.

Chang, F., J. Guo, W. Xu and K. Yao, 2015. A feature selection method to handle imbalanced data in text classification. J. Digital Inf. Manage., 13: 169-175.

Cortes, C. and V. Vapnik, 1995. Support-vector networks. Mach. Learn., 20: 273-297.

Feng, G., J. Guo, B.Y. Jing and T. Sun, 2015. Feature subset selection using Naive Bayes for text classification. Pattern Recognit. Lett., 65: 109-115.

Forman, G., 2003. An extensive empirical study of feature selection metrics for text classification. J. Machine Learn. Res., 3: 1289-1305.

Gnana, D.A.A., S. Appavu and E.J. Leavline, 2016. Literature review on feature selection methods for high-dimensional data. Methods, 136: 1-10.

Javed, K., H.A. Babri and M. Saeed, 2012. Feature selection based on class-dependent densities for high-dimensional binary data. Knowl. Data Eng. IEEE. Trans., 24: 465-477.

Javed, K., S. Maruf and H.A. Babri, 2015. A two-stage Markov blanket based feature selection algorithm for text classification. Neurocomputing, 157: 91-104.

Joachims, T., 1998. Text categorization with support vector machines: Learning with many relevant features. Proceedings of the 10th European Conference on Machine Learning, Chemnitz, Germany, April 21-23, 1998, Springer, Berlin, Heidelberg, pp: 137-142.

Karabulut, E.M., S.A. Ozel and T. Ibrikci, 2012. A comparative study on the effect of feature selection on classification accuracy. Procedia Technol., 1: 323-327.

Korde, V. and C.N. Mahender, 2012. Text classification and classifiers: A survey. Intl. J. Artif. Intell. Appl., 3: 85-99.

Kulkarni, A.R., V. Tokekar and P. Kulkarni, 2012. Identifying context of text documents using Naïve Bayes classification and Apriori association rule mining. Proceedings of the 6th IEEE International Conference on Software Engineering (CONSEG), September 5-7, 2012, IEEE, Indore, India, ISBN: 978-1-4673-2174-7, pp: 1-4.

Lee, C. and G.G. Lee, 2006. Information gain and divergence-based feature selection for machine learning-based text categorization. Inform. Process. Manage., 42: 155-165.

Li, Y. and C. Chen, 2012. Research on the feature selection techniques used in text classification. Proceeding of the 2012 9th International Conference on Fuzzy Systems and Knowledge Discovery, May 29-31, 2012, IEEE, Xi'an, China, ISBN: 978-1-4673-0024-7, pp: 725-729.

Manning, C.D., P. Raghavan and H. Schutze, 2008. An Introduction to Information Retrieval. Cambridge University Press, Cambridge, UK., ISBN-13: 9780521865715, pp: 482.

Meenakshi, R.V.S., 2013. Structured data extraction from the deep web. IOSR. J. Comput. Eng., 8: 93-96.

Mladenic, D. and M. Grobelnik, 1999. Feature selection for unbalanced class distribution and naive bayes. Proceedings of the 16th International Conference on Machine Learning, June 27-30, 1999, Morgan Kaufmann, San Francisco, CA., USA., pp: 258-267.

Nidhi and V. Gupta, 2011. Recent trends in text classification techniques. Int. J. Comput. Applic., 35: 45-51.

Paul, A., 2014. Effect of imbalanced data on document classification algorithms. Ph.D Thesis, Auckland University of Technology, Auckland, New Zealand.

Rehman, A., K. Javed, H.A. Babri and M. Saeed, 2015. Relative discrimination criterion-a novel feature ranking method for text data. Expert Syst. Appl., 42: 3670-3681.

Siddiqui, M.A., 2016. An empirical evaluation of text classification and feature selection methods. Artif. Intell. Res., 5: 70-81.

Tian, C., Y. Wang, X. Lin, J. Lin and J. Hong, 2016. Research on high-dimensional data reduction. Intl. J. Database Theor. Appl., 9: 87-96.

Ting, S.L., W.H. Ip and A.H. Tsang, 2011. Is naive bayes a good classifier for document classification?. Int. J. Software Eng. Its Appl., 5: 37-46.

Uysal, A.K. and S. Gunal, 2012. A novel probabilistic feature selection method for text classification. Knowl. Based Syst., 36: 226-235.

Uysal, A.K., 2016. An improved global feature selection scheme for text classification. Expert Syst. Appl., 43: 82-92.

Wang, F., Y. Zhang, H. Xiao, L. Kuang and Y. Lai, 2015. Enhancing stock price prediction with a hybrid approach based extreme learning machine. Proceedings of the IEEE International Conference on Data Mining Workshop (ICDMW), November 14-17, 2015, IEEE, Atlantic City, New Jersey, ISBN: 978-1-4673-8494-0, pp: 1568-1575.

Yang, J., Y. Liu, X. Zhu, Z. Liu and X. Zhang, 2012. A new feature selection based on comprehensive measurement both in inter-category and intra-category for text categorization. Inf. Process. Manage., 48: 741-754.