International Journal of Computer Applications (0975 - 8887)
Volume 164 - No.8, April 2017

A Comprehensive Survey on various Feature Selection Methods to Categorize Text Documents

B. S. Harish
Department of Information Science & Engineering
Sri Jayachamarajendra College of Engineering,
Mysuru, Karnataka, India

M. B. Revanasiddappa
Department of Information Science & Engineering
Sri Jayachamarajendra College of Engineering,
Mysuru, Karnataka, India

ABSTRACT

Feature selection is one of the well known solution to high dimensionality problem of text categorization. In text categorization, selection of good features (terms) plays a very important role. Feature selection is a strategy that can be used to improve categorization accuracy, effectiveness and computational efficiency. This paper presents an empirical study of most widely used feature selection methods viz. Term Frequency-Inverse Document Frequency (*tf·idf*), Information Gain (IG), Mutual Information(MI), CHI-Square (χ^2), Ambiguity Measure (AM), Term Strength (TS), Term Frequency-Relevance Frequency (*tf·rf*) and Symbolic Feature Selection (SFS) with five different classifiers (Nave Bayes, K-Nearest Neighbor, Centroid Based Classifier, Support Vector Machine and Symbolic Classifier). Experimentations are carried out on standard bench mark datasets like Reuters-21578, 20-Newsgroups and 4 University dataset.

Keywords

High Dimensionality, Feature Selection, Classifiers, Text Categorization

1. INTRODUCTION

In 1980s, the task of text categorization was based on knowledge engineering (KE). A set of rules was defined manually to encode the expert knowledge to categorize the text documents under the given categories [16]. Since, there is a requirement of human intervention in knowledge engineering; later day's researchers proposed many machine learning techniques to automatically manage the text documents. The advantages of machine learning based approaches are that the accuracy is comparable to that of human experts and no intervention from either knowledge engineers or domain experts needed for the construction of a document management tool. Many text mining methods like document retrieval, clustering, categorization, routing and filtering are often used for effective management of text documents [32]. Out of several tasks, text categorization is the one which is commonly used in text information systems. Therefore, current requirement is devising effective and efficient models for text representation and categorization of text documents for real time applications.

In a text document, each term is considered as a feature. The large number of feature set can typically be reduced through a variety of feature selection techniques. Correctly identifying the relevant features in a text documents is a vital importance for text categorization. The main purpose of feature selection is to reduce the high dimensionality of the feature space by selecting the most relevant and discriminating features for the categorization task [38]. Feature Selection (FS) method keeps the terms with highest score according to the predetermined measure of the importance of the term. It has been widely observed that feature selection can be a powerful tool for simplifying or speeding up computations, and when employed appropriately it can lead to less loss in categorization quality. The fundamental goal of feature selection method is to improve the categorization effectiveness and computational efficiency. In spite of numerous approaches in the literature, feature selection is still an ongoing research topic. The researchers are still looking for new techniques to select distinctive features so that the categorization accuracy can be improved and the processing time can be reduced as well. There are good numbers of filter based techniques and wrapper based techniques for the selection of distinctive features in text categorization. Automatic feature selection (FS) methods include removal of trivial and non informative terms, according to corpus statistics [15].

In this paper, following are the feature selection methods evaluated: Document Frequency (DF) [14], Information Gain (IG) [35][11], CHI-Square (χ^2) [14], Mutual Information (MI) [19][18], Term Strength (TS) [14], Ambiguity Measures (AM) [30], Term Frequency-Inverse Document Frequency (*tf·idf*) [16][34], Term Frequency-Relevance Frequency (*tf·rf*) [24][23] and Symbolic Feature Selection (SFS) [10]. The DF is the number of documents in which the term occurs. The DF is computed for each term in the training corpus and those terms whose document frequency is less than a predetermined threshold are removed from the feature space. Hoque et al., [26] proposed greedy feature selection method with the help of mutual information. It is combination of feature?feature mutual information and feature class mutual information to determine optimal subset of features. This method increases the relevancy between features and decreases redundancy. Further, performance of selected feature subset is evaluated using different classifiers on 12 real-life datasets. Bakus and Kamel., [17] present the MIFS-CI (MIFS with common information) variant of the mutual information feature selection algorithm. It finds the optimal value of the redundancy parameter, which is a key parameter in the MIFS-type algorithms. Zhiying and Yang [21] proposed an improved ambiguity measure feature selection method. It improves the performance of the ambiguity measure by using Nave Bayes (NB) and Support Vector Machine (SVM) classifiers. The AM feature selec-

tion method is used as a pre-processing step for the SVM classifier [29], and showed that AM reduces the training time of the SVM classifier.

Lan et al., [25] gives a detailed study of widely used supervised term weighting methods on standard datasets with combination of K-Nearest Neighbor (KNN) and SVM techniques. Lan et al., [24] proposed a new supervised term weighting method, i.e. term frequency-relevance frequency (*tf·rf*) based on the analysis of discriminating power. The *tf·rf* is used as a term weighting method in text categorization, and this method is robust in nature and works consistently [25]. Scott., [31] investigated a number of feature engineering methods for text categorization in the context of symbolic rule-based learning algorithm. The focus is on changing the standard bag of words representation of a text by incorporating some shallow linguistic processing techniques. Harish et al., [10] proposed a Symbolic Feature Selection (SFS) method, which represent documents based on clustering of term frequency vectors and to create multiple clusters to preserve the intraclass variations for each class of documents. SFS keeps the best features for effective text representation and reduces the time taken to classify a given document. This method basically uses the symbolic similarity and dissimilarity measures [22].

Bidi and Elberrichi [27] proposed a feature selection method with the help of genetic algorithm. This method carry out two goals: first goal is to find discriminating feature subset, which improves the classifier performance. Another goal is to determine feature subset with small dimensionality of feature space. Basically, most of the feature selection methods are based on Balanced Accuracy Measure (ACC2). The ACC2 assigns equal ranks to terms, which have equal differences. The main problem of ACC2 is, it ignores their relative document frequencies. To resolve this problem, Rehman et al., [4] presents a novel feature ranking metric, named as normalized difference measure (NDM). Uysal and Gunal [6] proposed a Distinguishing Feature Selection method, which gives importance to term that appears only one time in a class and it is discriminating to other classes. The improved Distinguishing Feature Selection method is described in [5] for text categorization.

All in all, there are many types of feature selection methods exist in the literature. However, in this research work we have restricted ourselves to evaluate the various text classifiers using most widely used feature selection methods. Further, we evaluated the feature selection (FS) methods on benchmark datasets and presented the results obtained using various classifiers.

The rest of the paper is organized as follows: The Feature Selection Methods (FSM) are presented in section II. The Classifiers and datasets used for experimentation are presented in section III. Section IV presents the experimental setup and quantitative comparative study. The paper is concluded in section V.

2. FEATURE SELECTION METHODS

Feature selection methods are used to remove trivial terms and reduce high dimension of feature set to optimize the categorization efficiency and effectiveness. In the following section, we briefly describe the following feature selection methods: Term Frequency-Inverse Document Frequency (*tf·idf*), Mutual Information (MI), Information Gain (IG), CHI-Square (χ^2), Term Frequency-Relevance Frequency (*tf·rf*), Term Strength (TS), Ambiguity Measure (AM) and Symbolic Feature Selection (SFS).

2.1 Term Frequency-Inverse Document Frequency (*tf·idf*)

The term frequency-inverse document frequency (*tf·idf*) is commonly used technique for term weighting in the field of text classification [34]. It determines the relative frequency of terms in a specific document through an inverse proportion of the term over the entire document corpus [20]. The *tf·idf* weight is composed by two conditions: the first condition computes the normalized term frequency tf, and the second condition is the inverse document frequency (*idf*). The term frequency (*tf*) measures the number of times a term occurs in a document and it is used to calculate the describing ability of the term.

$$tf(t) = \frac{(Number\ of\ times\ term\ t\ appears\ in\ a\ document)}{(Total\ number\ of\ terms\ in\ the\ document)} \tag{1}$$

The inverse document frequency (*idf*) is used to calculate the distinguishing ability of the term and also it measures the importance of the term. A higher idf of a term indicates that the term appears in relatively few documents.

$$idf(t) = log\left(\frac{N}{n_i}\right) \tag{2}$$

Where, N is the total number of documents and n_i is the number of documents containing term i. The $tf{\cdot}idf$ of a term t is defined as:

$$tf{\cdot}idf(t) = tf(t) \times \log\left(\frac{N}{n_i}\right) \tag{3}$$

2.2 Mutual Information (MI)

The Mutual Information (MI) [39] is a criterion commonly used in statistical language modeling of word association [19]. MI measures how much information presence or absence and term contribution to make the correct categorization decision on a category.

Given a category c and a term t, let A denote the number of times c and t co-occur, B denotes the number of times t occurs without c, C denotes the number of times c occur without t, and N denotes the total number of documents in c.

$$I(t,c) = log\left(\frac{P(t,c)}{P(t) * P(c)}\right) \tag{4}$$

and is estimated using

$$I(t,c) \approx \log\left(\frac{A \times N}{(A+C) \times (A+B)}\right) \tag{5}$$

If t and c are independent, $I(t,c)$ has a natural value zero. To measure the goodness of a term in a global feature selection, combine the category specific scores of a term into two alternative way.

$$I_{avg}(t) = \sum_{i=1}^{m} P_r(c_i) I(t,c_i) \tag{6}$$

$$I_{max}(t) = \max_{i=1}^{m}\{I(t,c_i)\} \tag{7}$$

The weakness of MI is that the score is strongly influenced by the marginal probabilities of terms, because rare terms will have a higher score than common terms.

2.3 Information Gain (IG)

Information Gain (IG) measures the number of bits of information obtained for category prediction and by knowing presence or absence of a term in a document [14]. The idea behind IG is to select features that reveal the most related information about the classes. IG reaches its maximum value if a term is an ideal indicator for class association, i.e., if the term is present in a document if and only if the document belongs to the respective class. The IG method fails to identify discriminatory features, particularly when they are distributed over multiple classes [28][32]. The information gain of term t is:

$$IG(t,c) = \sum_{c\epsilon\{c_i,\bar{c}_i\}} \sum_{t\epsilon\{t_k,\bar{t}_k\}} P(t,c) * log\left(\frac{P(t,c)}{P(t) * P(c)}\right) \quad (8)$$

Where, $P(t,c)$ is the joint probability of category c, and occurrence of the term t. $P(t)$ is the probability of term, $P(c)$ is the probability of category. The term $\bar{t_k}$ means that the term not present, the category $\bar{c_i}$ means that the category not present.

2.4 CHI-Square (χ^2)

Chi-Square (χ^2) is a statistical feature selection method [14]. χ^2 is used to measure the association between a term and category in text categorization. It also used to test whether the occurrence of a specific term and the occurrence of a specific category are independent. Thus we estimate the quantity for each term and we rank them by their score. If a term is close to more categories, then the score of that term is higher. High scores on χ^2 indicate that the null hypothesis of independence should be rejected and thus that the occurrence of the term and category are dependent. If they are dependent, then we select the feature for the text categorization. The χ^2 measure of a term t for a category c is defined as:

$$\chi^2(t,c) = \frac{N \times (R_a R_d - R_c R_b)^2}{(R_a + R_c) \times (R_b + R_d) \times (R_a + R_b) \times (R_c + R_d)} \quad (9)$$

Where, N is the total number of training samples, R_a is the number of times t and c co-occur, R_b is the number of times t occurs without c, R_c is the number of times c occurs without t, R_d is the number of times neither c nor t occurs. The score of a term is calculated for individual category. This score can be globalized over all category in two ways. The first way is to compute the weighted average score for all category while the second way is to choose the maximum score among all category.

$$\chi^2_{avg}(t) = \sum_{i=1}^{m} P_r(c_i)\chi^2(t,c_i) \quad (10)$$

$$\chi^2_{max}(t) = \max_{i=1}^{m}\{\chi^2(t,c_i)\} \quad (11)$$

2.5 Term Strength (TS)

Term Strength (TS) is proposed and evaluated by Wilbur and Sirotkin [37] for vocabulary reduction in text retrieval, and later applied to text classification [40]. It defines the weight of a term as the probability of finding it in some document d_i given that it has also appeared in the document d_j, similar to d_i. It uses a training set of documents to derive documents pairs whose similarity is above threshold. It is computed based on the conditional probability that a term occurs in the second half of a pair of related documents given that it occurs in the first half [14]. The term strength weight for t was then calculated as the conditional probability of t appearing in d_i given that t appears also in d_j.

$$TS(t) = p(t \in d_j | t \in d_i) \quad (12)$$

Where, d_i and d_j denote an arbitrary pair of distinct but related documents.

2.6 Ambiguity Measure (AM)

Ambiguity Measure (AM) selects the most unambiguous features from the feature set and also to predict the category of a document. Unambiguous features are those features whose presence in a document indicates a strong degree of confidence that a document belongs to only one specific category [29]. AM for each term probability falls into a particular category, the maximum AM score for term t with respect to all categories is assigned as the AM score of term t. If the term is unambiguous, the AM score is near to 1. Conversely, if AM score near to 0, the term is considered more ambiguous and may point to more than one category.

$$AM(t,c) = \left(\frac{tf(t,c)}{tf(t)}\right) \quad (13)$$

$$AM(t) = max(AM(t,c)) \quad (14)$$

Where, $tf(t,c)$ is the term frequency of a term t in category c and $tf(t)$ is the term frequency of a term t in the entire collection.

2.7 Term Frequency-Relevance Frequency (*tf·rf*)

The *tf·rf* is a supervised term weighting method. It improves the term discriminating power for text classification [24]. The fundamental idea of this method is to focus on a high frequency term which is in positive category rather than in the negative category. And main contribution is selecting the positive features from the negative features [25]. This method is robust in nature and works consistently in the best either cross classifier or cross corpus.

$$rf = log\left(1 + \frac{n_i}{\bar{n_i}}\right) \quad (15)$$

Where n_i is the number of documents to which a term is assigned, $\bar{n_i}$ is the number of documents which contain the term and belong to the negative categories. Further rf (relevance frequency) combined with tf (term frequency) by a multiplication operation:

$$tf{\cdot}rf = tf \times \log\left(1 + \frac{n_i}{\bar{n_i}}\right) \quad (16)$$

Term frequency represents a close relationship between the term and the content of documents which contain that term. It is observed that if high frequency terms are spread widely over a large number of documents, we may not retrieve the relevant documents from the whole collection.

2.8 Symbolic Feature Selection (SFS)

The Symbolic Feature Selection (SFS) is an unconventional method and this method works in a proximity space rather than working in the original feature space [10]. SFS makes use of both, similarity and dissimilarity measures to classify the text documents. SFS gives the best features for text document categorization; it mainly reduces the human effort and time to categorize a given document. Features are of interval valued type [22], the degree of similarity between class representative vectors is estimated based on

degrees of overlapping of features. It can be noticed that, these relative overlapping of interval type features are not equal and hence the degree of similarity between two symbolic vectors may not necessarily be symmetric.
Let R_a and R_b be the representative symbolic feature vectors of the classes and respectively. The similarity of C_i to C_j with respect to the feature is given by:

$$S^l_{a\to b} = \left(\frac{|I_{al} \cap I_{bl}|}{|I_{bl}|}\right) \tag{17}$$

Where, $I_{al} = [f^-_{al}, f^+_{al}] \; \forall \; l = 1, 2, ..., m$ are the type features of the representative vector R_a(class C_i) and $I_{bl} = [f^-_{bl}, f^+_{bl}] \; \forall \; l = 1, 2, ..., m$ are the interval type features of the representative vector R_b(class C_j). using the above similarity measure formula and compute the similarity matrix of size $k \times k$, each element of which is a multivalued type data. A matrix M of size $k^2 \times m$ is then constructed for the computed similarity matrix of size $k \times k$ whose elements are multivalued of dimension m by listing out each multivalued type element one by one in the form of rows.
Now, calculate the sum of the total correlations of each column of M with all other columns. Among all the features, we select some of them which have their respective total correlation greater than the average correlation and subsequently retain only those for representation. These selected features are responsible for high cohesion of the classes of documents.

$$TCorr_l = \sum_{q=1}^{m} Corr\left(l^{th}column, q^{th}column\right) \tag{18}$$

$$AvgTCorr = \frac{\sum_{q=1}^{m} TCorr_l}{m} \tag{19}$$

Where $TCorr_l$ be the total correlation of the l^{th} column with all other column of matrix M and $AvgTCorr$ be the average of all total correlation obtained due to all columns. If $TCorr_l > AvgTCorr$ then the feature corresponding to the l^{th} column is selected to be the feature which is capable of increasing the cohesion among the documents belonging to a class. The selected features are stored in the knowledge base for classification purpose.

3. CLASSIFIERS AND DATASETS

In this paper, we used five different existing classifiers with text feature selection methods. Naive Bayes (NB) [7][9][8], k-Nearest Neighbor (k-NN) [13][33], Centroid based Classifier [9][12][36], Support Vector Machine (SVM) [26][9][36], Symbolic Classifier [22]. For all our experimentation, we used standard benchmark datasets i.e. 20 Newsgroups [1], 4 University dataset [2] and Reuters-21578 [3].

4. PERFORMANCE EVALUATION

4.1 Experimental Setup

In our experiments, we used eight feature selection method: Document Frequency (DF), Information Gain (IG), CHI-Square( ), Mutual Information (MI), Term Strength (TS), Ambiguity Measures (AM), Term Frequency-Inverse Document Frequency (*tf·idf*), Term Frequency-Relevance Frequency (*tf·rf*) and Symbolic Feature Selection (SFS). These feature selection methods are used to categorize text documents using five different classifiers (Nave Bayes, K-Nearest Neighbor, Centroid Based Classifier, Support Vector Machine and Symbolic Classifier). The effectiveness of feature selection methods are demonstrated by using three standard datasets viz. 20 Newsgroups, 4 Universities and Reuters 21578. The dataset is separated into training and testing set respectively. We used 60 text documents of each class of a dataset to create training set and remaining 40 text documents for testing purpose. Experimentations are repeated 5 times by choosing the training samples randomly. We use F-Measure to evaluate the performance of each classifier.

4.2 Quantitative Comparative Study and Discussion

The visibility of features is chosen by feature selection methods, which is one of the good indicators for effectiveness of the method. The feature selection method assigns high score to distinctive features. These distinctive features are given to categorization model to categorize given text documents. The categorization performance increases through these selected distinctive features. Out of eight different feature selection methods, the results obtained from symbolic feature selection (SFS) found to be better. In SFS, the selected distinctive features are in interval (multivalued) form, which reduces the dimensionality of feature space and increases the accuracy of classifiers. In order to validate the effectiveness of all these feature selection methods, three standard datasets were utilized to observe performance. The experimental results are presented in Table 1, 2 & 3.
In Table 1 F-measure results are summarized for different feature selection methods (FSM) on 20 Newsgroups dataset. SFS method showed good performance with symbolic classifier compare to other classifiers. However, SFS shows consistent performance with other classifier (NB, KNN, CBC & SVM). On the other hand, SVM classifier achieved better results with χ^2,TS, AM and *tf·rf* feature selection methods. Table 2 presents 4-Unversities dataset result. The SFS method achieved very good result with Symbolic classifier i.e 93.4%. On the other hand SVM performed better with MI, IG, χ^2, TS & *tf·rf*. The Table 3 presents the results on Reuters-21578 dataset. Symbolic classifier outperforms with SFS method by achieving the result of 94.0%. From the above observations, the symbolic feature selection method with symbolic classifier showed very good result on Reuters-21578 dataset. Thus, the symbolic feature selection method with symbolic classifier outperforms all remaining feature selection methods.

5. CONCLUSION

In text categorization, high dimensionality of feature space is a major issue. This issue is resolved by using various feature selection approaches, which increases the efficiency of text categorization. In this paper, we reported an empirical evaluation on the most widely used text feature selection methods with five different existing classifiers (Naive Bayes, k-NN, Centroid based Classifier, SVM, Symbolic Classifier) to categorize text documents. In the experiments, Symbolic Feature Selection (SFS) method performed consistently well on all the other methods using three widely used standard benchmark datasets with symbolic classifier. The evaluation demonstrated that the SFS method has tremendous influence on improving the categorization accuracy. In future, It is also intended to work on the computational complexity of various feature selection methods (FSM) using different classifiers.

Table 1. F-Measure on 20 Newsgroups dataset.

Dataset Name	F-Measure (FM)					
	FSM	**FM using NB**	**FM using KNN**	**FM using CBC**	**FM using SVM**	**FM using SC**
20 Newsgroups Dataset	**tf·idf**	76.8	80.1	83.4	84.3	86.4
	MI	75.4	79.8	81.1	81.2	87.6
	IG	73.2	79.1	82.1	81.4	85.2
	χ^2	75.8	81.4	83.6	83.2	79.8
	TS	76.1	82.2	84.2	84.3	83.4
	AM	79.5	83.4	86.1	87.8	84.6
	tf·rf	81.4	83.9	85.4	88.4	86.2
	SFS	83.2	85.1	88.2	89.9	91.2

Table 2. F-Measure on 4-Universities dataset.

Dataset Name	F-Measure (FM)					
	FSM	**FM using NB**	**FM using KNN**	**FM using CBC**	**FM using SVM**	**FM using SC**
4 University Dataset	**tf·idf**	67.3	69.1	68.8	85.4	88.6
	MI	70.1	72.5	69.1	85.1	84.5
	IG	69.4	73.6	69.4	86.2	83.2
	χ^2	69.4	74.8	68.4	85.1	84.2
	TS	71.1	71.1	69.0	88.4	82.4
	AM	72.5	73.9	70.4	89.1	87.9
	tf·rf	71.9	73.5	71.6	88.4	89.8
	SFS	73.2	75.5	79.4	90.1	93.4

Table 3. F-Measure on Reuters-21578 dataset.

Dataset Name	F-Measure (FM)					
	FSM	**FM using NB**	**FM using KNN**	**FM using CBC**	**FM using SVM**	**FM using SC**
Reuters-21578 Dataset	**tf·idf**	69.0	71.0	69.5	86.0	89.0
	MI	71.5	72.0	70.1	86.1	86.7
	IG	70.5	73.5	71.0	86.0	85.4
	χ^2	69.0	75.0	69.0	84.5	85.2
	TS	72.5	73.2	70.0	89.0	84.6
	AM	72.5	74.2	71.0	89.5	88.2
	tf·rf	73.5	74.0	72.5	89.1	90.1
	SFS	74.1	76.1	78.5	90.1	94.0

FSM – Feature Selection Method
NB – Naive Bayes
KNN – K-Nearest Neighbor
CBC – Centroid Based Classifier
SVM – Support Vector Machine
SC – Symbolic Classifier

6. REFERENCES

[1] 20newsgroups. *http://people.csail.mit.edu/jrennie/20Newsgroups/*.

[2] 4 universities. *http://www.cs.cmu.edu/afs/cs/project/theo-20/www/data/*.

[3] Reuters-21578. *http://www.daviddlewis.com/resources/ test-collections/reuters21578/*.

[4] Rehman Abdur, Kashif Javed, and Haroon A Babri. Feature selection based on a normalized difference measure for text classification. *Information Processing & Management*, 53(2):473–489, 2017.

[5] Uysal Alper Kursat. An improved global feature selection scheme for text classification. *Expert systems with Applications*, 43:82–92, 2016.

[6] Uysal Alper Kursat and Gunal Serkan. A novel probabilistic feature selection method for text classification. *Knowledge-Based Systems*, 36:226–235, 2012.

[7] Hotho Andreas, Andreas Nürnberger, and Gerhard Paaß. A brief survey of text mining. In *Ldv Forum*, volume 20, pages 19–62, 2005.

[8] Khan Aurangzeb, Baharum Baharudin, Lam Hong Lee, and Khairullah Khan. A review of machine learning algorithms for text-documents classification. *Journal of advances in information technology*, 1(1):4–20, 2010.

[9] Harish B S, D S Guru, and S Manjunath. Representation and classification of text documents: A brief review. *IJCA, Special Issue on RTIPPR (2)*, pages 110–119, 2010.

[10] Harish B S, D S Guru, S Manjunath, and Bapu B Kiranagi. A symbolic approach for text classification based on dissimilarity measure. In *Proceedings of the First International Conference on Intelligent Interactive Technologies and Multimedia*, pages 104–108. ACM, 2010.

[11] Lee Changki and Lee Gary-Geunbae. Information gain and divergence-based feature selection for machine learning-based text categorization. *Information processing & management*, 42(1):155–165, 2006.

[12] Han Eui-Hong Sam and Karypis George. Centroid-based document classification: Analysis and experimental results. In *European conference on principles of data mining and knowledge discovery*, pages 424–431. Springer, 2000.

[13] Han Eui-Hong Sam, George Karypis, and Vipin Kumar. Text categorization using weight adjusted k-nearest neighbor classification. In *Pacific-asia conference on knowledge discovery and data mining*, pages 53–65. Springer, 2001.

[14] Song Fengxi, Shuhai Liu, and Jingyu Yang. A comparative study on text representation schemes in text categorization. *Pattern analysis and applications*, 8(1-2):199–209, 2005.

[15] Forman George. An extensive empirical study of feature selection metrics for text classification. *Journal of machine learning research*, 3(Mar):1289–1305, 2003.

[16] Salton Gerard and Buckley Christopher. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988.

[17] Bakus Jan and Kamel Mohamed S. Higher order feature selection for text classification. *Knowledge and Information Systems*, 9(4):468–491, 2006.

[18] Novovičová Jana, Antonín Malík, and Pavel Pudil. Feature selection using improved mutual information for text classification. In *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*, pages 1010–1017. Springer, 2004.

[19] Church Kenneth-Ward and Hanks Patrick. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29, 1990.

[20] Jing Li-Ping, Hou-Kuan Huang, and Hong-Bo Shi. Improved feature selection approach tfidf in text mining. In *Machine Learning and Cybernetics, 2002. Proceedings. 2002 International Conference on*, volume 2, pages 944–946. IEEE, 2002.

[21] Zhiying Liu and Yang Jieming. An improved ambiguity measure feature selection for text categorization. In *Intelligent Human-Machine Systems and Cybernetics (IHMSC), 2012 4th International Conference on*, volume 1, pages 220–223. IEEE, 2012.

[22] Revanasiddappa M B, B S Harish, and S Manjunath. Document classification using symbolic classifiers. In *Contemporary Computing and Informatics (IC3I), 2014 International Conference on*, pages 299–303. IEEE, 2014.

[23] Lan Man, Sam-Yuan Sung, Hwee-Boon Low, and Chew-Lim Tan. A comparative study on term weighting schemes for text categorization. In *Neural Networks, 2005. IJCNN'05. Proceedings. 2005 IEEE International Joint Conference on*, volume 1, pages 546–551. IEEE, 2005.

[24] Lan Man, Chew Lim Tan, and Hwee-Boon Low. Proposing a new term weighting scheme for text categorization. In *AAAI*, volume 6, pages 763–768, 2006.

[25] Lan Man, Chew Lim Tan, Jian Su, and Yue Lu. Supervised and traditional term weighting methods for automatic text categorization. *IEEE transactions on pattern analysis and machine intelligence*, 31(4):721–735, 2009.

[26] Hoque Nazrul, DK Bhattacharyya, and Jugal K Kalita. Mifs-nd: a mutual information-based feature selection method. *Expert Systems with Applications*, 41(14):6371–6385, 2014.

[27] Bidi Noria and Elberrichi Zakaria. Feature selection for text classification using genetic algorithms. In *Modelling, Identification and Control (ICMIC), 2016 8th International Conference on*, pages 806–810. IEEE, 2016.

[28] Mukras Rahman, Nirmalie Wiratunga, Robert Lothian, Sutanu Chakraborti, and David Harper. Information gain feature selection for ordinal text classification using probability redistribution. In *Proceedings of the Textlink workshop at IJCAI*, volume 7, page 16, 2007.

[29] Mengle Saket SR and Goharian Nazli. Using ambiguity measure feature selection algorithm for support vector machine classifier. In *Proceedings of the 2008 ACM symposium on Applied computing*, pages 916–920. ACM, 2008.

[30] Mengle Saket SR and Goharian Nazli. Ambiguity measure feature-selection algorithm. *Journal of the American Society for Information Science and Technology*, 60(5):1037–1050, 2009.

[31] Scott Sam. *Feature engineering for a symbolic approach to text classification.* University of Ottawa (Canada), 1998.

[32] Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47, 2002.

[33] Jiang Shengyi, Guansong Pang, Meiling Wu, and Limin Kuang. An improved k-nearest-neighbor algorithm for text categorization. *Expert Systems with Applications*, 39(1):1503–1509, 2012.

[34] Qu Shouning, Sujuan Wang, and Yan Zou. Improvement of text feature selection method based on tfidf. In *Future Information Technology and Management Engineering, 2008. FITME'08. International Seminar on*, pages 79–81. IEEE, 2008.

[35] Mitchell T. *Machine learning.* McGraw-Hill, Inc. New York, NY, USA, 1997.

[36] Korde Vandana and Mahender C Namrata. Text classification and classifiers: A survey. *International Journal of Artificial Intelligence & Applications*, 3(2):85, 2012.

[37] Wilbur W John and Sirotkin Karl. The automatic identification of stop words. *Journal of information science*, 18(1):45–55, 1992.

[38] Shang Wenqian, Houkuan Huang, Haibin Zhu, Yongmin Lin, Youli Qu, and Zhihai Wang. A novel feature selection algorithm for text categorization. *Expert Systems with Applications*, 33(1):1–5, 2007.

[39] Xu Yang, Gareth JF Jones, JinTao Li, Bin Wang, and Chun-Ming Sun. A study on mutual information-based feature selection for text categorization. *Journal of Computational Information Systems*, 3(3):1007–1012, 2007.

[40] Yang Yiming and Wilbur John. Using corpus statistics to remove redundant words in text categorization. *JASIS*, 47(5):357–369, 1996.