Webology, Volume 12, Number 2, December, 2015

Techniques for text classification: Literature review and current trends

Rajni Jindal

Department of Computer Science & Engineering, Delhi Technological University, Delhi, India. E-mail: rajnijindal (at) dce.ac.in

Ruchika Malhotra

Department of Computer Science & Engineering, Delhi Technological University, Delhi, India. E-mail: ruchikamalhotra2004 (at) yahoo.com

Abha Jain

Department of Computer Science & Engineering, Delhi Technological University, Delhi, India. E-mail: me_abha (at) yahoo.com

Received October 10, 2015; Accepted December 26, 2015

Abstract

Automated classification of text into predefined categories has always been considered as a vital method to manage and process a vast amount of documents in digital forms that are widespread and continuously increasing. This kind of web information, popularly known as the digital/electronic information is in the form of documents, conference material, publications, journals, editorials, web pages, e-mail etc. People largely access information from these online sources rather than being limited to archaic paper sources like books, magazines, newspapers etc. But the main problem is that this enormous information lacks organization which makes it difficult to manage. Text classification is recognized as one of the key techniques used for organizing such kind of digital data. In this paper we have studied the existing work in the area of text classification which will allow us to have a fair evaluation of the progress made in this field till date. We have investigated the papers to the best of our knowledge and have tried to summarize all existing information in a comprehensive and succinct manner. The studies have been summarized in a tabular form according to the publication year considering numerous key

perspectives. The main emphasis is laid on various steps involved in text classification process viz. document representation methods, feature selection methods, data mining methods and the evaluation technique used by each study to carry out the results on a particular dataset.

Keywords

Machine learning; Text classification; Feature selection; Bag-of-words; Vector space model

Introduction

Text classification is the task of classifying a document under a predefined category. More formally, if di is a document of the entire set of documents D and {c1,c2, c3,...,cn} is the set of all the categories, then text classification assigns one category cj to a document di (Ikonomakis et al., 2005). The documents depending upon their characteristics can be labeled for one class or for more than one class. If a document is assigned to only one class, it is called “single-label” and if the document is assigned to more than one class, it is called “multi-label” (Wang & Chiang, 2011). A ‘‘single-label’’ text classification problem can be further categorized into a ‘‘binary class’’ problem if only one of the two classes is assigned to the document and this ‘‘single-label” text classification problem becomes a ‘‘multi-class’’ problem if only N mutually exclusive classes are assigned to the document. Text classification consists of document representation, feature selection or feature transformation, application of data mining algorithm and finally an evaluation of the applied data mining algorithm.

Now-a-days the amount of information available on the web is tremendous and increasing at an exponential rate. Automatic text classification has always been an important application and research topic since the inception of digital documents to manage the enormous amount of data available on the web (Ikonomakis et al., 2005). It is based on machine learning techniques that automatically build a classifier by learning the characteristics of the categories from a set of pre-classified documents (Sebastiani, 2002). It plays an important role in information extraction and summarization, text retrieval, and question- answering. Typically, most of the data for classification is of heterogeneous nature collected from the web, through newsgroups, bulletin boards, and broadcast or printed news scientific articles, news reports, movie reviews, and advertisements. They are multi-source, and consequently have different formats, different preferred vocabularies and often significantly different writing styles even for documents within one genre. Therefore, automatic text classification is highly essential.

This paper provides an extensive study of the work which has been done till date in the area of text classification highlighting the challenges which occur in classifying an unstructured web

content into a structured format. In other words, the paper aims to focus on elaborating the dynamic and diversified nature of techniques available for classifying a given text into its pre-defined categories and how these techniques have evolved over the past. This in turn will offer new opportunities to the software practitioners and engineers working in this area. They can have an in-depth knowledge about the progress made in the area of text classification beginning from how the term 'text-classification' has coined, followed by a summarization of the work done by authors in this area and finally presenting open problems and issues for the researchers intended to work in this area. After a thorough analysis, it was concluded that text classification is a potential area of research and a lot of work can still be done towards improvising the existing techniques and methodologies which have been used for classifying the unstructured text. The paper presents a systematic review of previous text classification studies with a specific focus on data mining methods, feature selection methods, the dataset and the evaluation technique used. This review uses 132 text classification papers which will allow researchers to have a fair evaluation of all the past studies and suggest possible new directions of research in different areas concerned with text classification. The paper is organized as follows. Section 2 describes the review process, in which we have defined our inclusion criteria and explained the selection procedure. In this section we have also posed 7 research questions which will help us to collect the necessary information. In section 3 we have classified the papers according to different categories and have reported our review along with the important findings. Following this, we have section 4 wherein we have reported the results using different graphical methods. Finally, the review is concluded in section 5, in which we have also suggested some future directions.

Review Process

In this section, procedure used for selecting the relevant studies is discussed followed by an inclusion/exclusion criterion. Then the research questions are highlighted which this review is intended to answer.

Formulation of Research Questions

The most important objective of any review is to include maximum number of studies that are filtered according to the defined inclusion criteria. Thus, selection of the relevant studies or to have a suitable relevant subset of the papers is very essential (Malhotra & Jain, 2011). Following two steps are undertaken to make the selection:

Step 1: This is the initial step in which we have searched various research related digital portals such as ACM, IEEE, Springer, Elsevier, etc. Papers have been searched in various journals and conference proceedings for appropriate selection (Sjoberg et al., 2005). There are a number of important journals in which search has been done like Information processing and management, Pattern Analysis and Applications, Information Retrieval, Knowledge Information System,

Pattern Recognition Letters, Journal of Intelligent Information Systems, Expert Systems with Applications, Applied Soft Computing , Knowledge-Based Systems, Wuhan University Journal of Natural Sciences , Information and Knowledge Systems, Neural Computing and Applications, Machine Learning, Soft Computing, Decision Support Systems, Journal of Computer Science& Technology, IJDAR, Information Sciences, Journal of Zhejiag University Science etc. All the previous papers till date concerned with text classification have been collected and studied to carry out an efficient review. This search was done by identifying the papers whose title or abstract contains some of the relevant keywords such as text classification, text classification, etc. Then, all the papers were scanned through and abstracts were read to identify the relevant papers. This helped us to remove the irrelevant papers and obtain a smaller relevant subset.

Step 2: In this step, the subset of papers obtained in the first step was assessed for its actual relevance. Final inclusion/exclusion decisions were made after retrieving the full texts. At this step, we made the final decision or final subset as to which all papers should be included in this review. The introduction and conclusion section of the papers selected in the initial stage were read and hence a final decision was made. It is useful to maintain a list of excluded studies as they are very useful in identifying the reason for exclusion. At the end of this step, we found 132 relevant studies related to our area of text classification.

Inclusion/ Exclusion criterion

Systematic reviews require explicit inclusion and exclusion criteria to assess each potential primary study. The selection of primary studies is governed by inclusion and exclusion criteria which should be based on the research questions (Catal, 2011).We included the papers in our review if the paper describes research on text classification. This review does not describe all the text classification models and the techniques used to develop them in detail for practitioners. Our aim is to classify the papers with respect to their years, datasets, different feature selection techniques, data mining algorithms and an evaluation measure. We included the papers published in various journals and conference proceedings of digital portals which are of high repute like ACM, IEEE, Springer, Elsevier. We have excluded the papers which did not include experimental results. We did not exclude the papers wherein a new data mining algorithm was not proposed, but instead a new feature selection technique or some new evaluation measure was proposed. In other words, we included all the papers which were related to the field of text classification in some or the other way. Our exclusion did not take into account the publication year of paper or methods which have been used.

Formulation of Research Questions

Formulation of research questions (RQs) is very important to carry out the research. The purpose of research questions is to let the readers know what the review is intended to answer. These RQs

were selected in such a way so as to ensure that there is a total coverage of text classification area. In this review paper, we have addressed the following issues related to the area of text classification:

RQ1. Which is the most popular journal in this area?
RQ2. Which year shows the maximum publications?
RQ3. Which data mining methods are widely used?
RQ4. Has the usage of modern machine learning methods increased over traditional statistical methods?
RQ5. Which feature selection methods are commonly used?
RQ6. Which dataset is commonly used?
RQ7. Which is commonly used document representation method?

Classification of Papers

A number of different approaches have been studied to aid the text classification process. The aim of a classifier is to use a set of pre-classified documents to classify those that have not yet been seen. Figure 1 gives the graphical representation of a text classification process. The five major branches include document representation, feature selection, constructing a Vector Space Model (VSM), application of a data mining method and finally an evaluation of the text classifier.

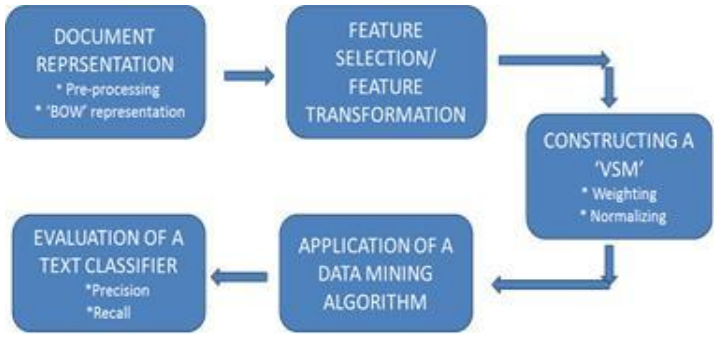

Figure 1. Text Classification Process

Document Representation

Document representation is the task of representing a given document in a form which is suitable for data mining system i.e. in the form of instances with a fixed number of attributes. There are several ways in which the conversion of documents from plain text to instances with a fixed number of attributes in a training set can be carried out. Bag-Of-Words (BOW) is the most commonly used word-based representation method. With this representation a document is considered to be simply a collection of words which occur in it at least once. With this approach, it is possible to have tens of thousands of words occurring in a fairly small set of documents.

Many of them are not important for the learning task and their usage can substantially degrade performance. It is imperative to reduce the size of the feature space. One widely used approach is to use a list of common words that are likely to be useless for classification, known as stop words, and remove all occurrences of these words before creating BOW representation. Another very important way to reduce the number of words is to use stemming which removes words with the same stem and keeps the stem as the feature. For example, the words “train”, “training”, “trainer” and “trains” can be replaced with “train”.

Feature Selection or Feature Transformation

Even after removing stop words from a document and replacing each remaining word by its stem, the number of words in a BOW representation is still very large. Therefore, feature selection method is applied to further reduce the dimensionality of the feature set by removing the irrelevant words. It has a number of advantages like smaller dataset size, considerable shrinking of the search space and lesser computational requirements. The goal is the reduction of the curse of dimensionality to yield improved classification accuracy and reduce over fitting. Methods for feature subset selection for text document classification task use an evaluation function that is applied to a single word. Scoring of individual words (Best Individual Features) can be performed using some of the measures, for instance, Document Frequency (DF), Term Frequency (TF), Mutual Information (MI), Information Gain (IG), Odds Ratio (OR), CHI-square statistic (CHI) and Term Strength (TS). All of these feature-scoring methods rank the features by their independently determined scores, and then select the top scoring features. Another technique to reduce the size of the feature space is referred to as feature transformation. It is also known as feature extraction. This approach does not weight terms in order to discard the lower weighted like feature selection, but compacts the vocabulary based on feature concurrencies. Principal Component Analysis (PCA) is a popularly used method for feature transformation. Some of the well-known feature selection metrics have been summarized in Table 1.

Constructing a Vector Space Model

Once a series of preprocessing tasks have been done (removal of stop words, stemming) and relevant features have been extracted using a particular feature selection method, we will have the total number of features as N which can be represented in some arbitrary order as t1, t2, ..., tN. The ith document is then represented as an ordered set of N values, called an N-dimensional vector which is written as (Xi1, Xi2, ..., XiN) where Xijis a weight measuring the importance of the jth term tjin the ith document. The complete set of vectors for all documents under consideration is called a VSM. There are various methods which can be used for weighting the terms. The most popular method used for calculating the weights is called TFIDF, which stands for Term Frequency Inverse Document Frequency. This combines term frequency with a measure of the rarity of a term in the complete set of documents and has been reported to be the most

efficient of all the methods. Now, before we use the set of N-dimensional vectors, we will first need to normalize the values of the weights. It has been observed that 'normalizing' the feature vectors before submitting them to the learning algorithm is the most necessary and important condition. Comparative analysis of different document representation methods has been provided in Table 2.

Application of a data mining algorithm

After feature selection and transformation, the documents can easily be represented in a form that can be used by a data mining method. A data mining method can either be based on statistical approaches known as the statistical method or can be a machine learning method based on various supervised and un-supervised techniques of machine learning. There are many text classifiers using machine learning techniques like decision trees (DT), naive-bayes (NB), rule induction, neural networks (NN), K- nearest neighbors (KNN), and support vector machines (SVM). They differ in their architecture and the approach adopted. Some of the well- known data mining methods has been summarized in Table 3.

Evaluation of a text classifier

An evaluation measure is used to measure the performance of a text classifier. For each category Ckwe can construct a confusion matrix as shown in the Figure 2 where 'a' denotes the number of true positive classifications, 'b' denotes the number of false positive classifications, c denotes the number of false negative classifications and d denotes the number of true negative classifications. For a perfect classifier b and c would both be zero.

		Predicted Class	
		C_k	Not C_k
Actual Class	C_k	a	c
	Not C_k	b	d

Figure 2: Confusion matrix for Category C_k

The value (a+d)/(a+b+c+d) gives the predictive accuracy. However, the standard performance measures for text classification are recall and precision. Recall is defined as a/(a + c), i.e. the proportion of documents in category C_k that are correctly predicted. Precision is defined as $a/(a + b)$, i.e. the proportion of documents that are predicted as being in category C_k that are actually in that category. Each level of recall is associated with a level of precision. In general, higher the recall, lower the precision, and vice versa (Yang &Pedersen, 1997). The point at which recall equals precision is the break-even point (BEP), which is often used as a single summarizing measure for comparing results. There are instances where a real BEP does not exist. It is

common practice to combine Recall and Precision into a single measure of performance called the F1 Score, which is defined by the formula F1 = 2×Precision×Recall/(Precision+Recall). This the product of precision and recall divided by their average which serves as yet another useful measure used for evaluating the effectiveness of classifiers. These scores are computed for the binary decisions on each individual category first and then averaged. When dealing with multiple classes there are two possible ways of averaging these measures, namely, macro-average and micro-average (Antonie & Zaiane, 2002). In the macro-averaging, one confusion matrix per class is used; the performance measures are computed on each of them and then averaged. In micro-averaging only one contingency table is used for all the classes, an average of all the classes is computed for each cell and the performance measures are obtained therein. The macro-average measure weights all the classes, regardless of how many documents belong to it. The micro-average measure weights all the documents, thus favoring performance on common classes.

Table 1. Comparative analysis of different feature selection methods

S. No.	Paper	Technique	Conclusion/ Advantage
1	Tasci and Gungor (2008)	IG , DF, Accuracy2, AKS	Local policy on IG, DF and Accuracy2 outperformed when the number of keywords is low and global policy outperformed as the number of keywords increases, AKS selected different number of keywords for different classes and improved the performance in skew datasets.
2	Tasci and Gungor (2009)	LDA (Latent Dirichlet Allocation)	Models and discovers the underlying topic structures of textual data, IG performed best at all keyword numbers while the LDA-based metrics performed similar to CHI and DF
3	Wang et al. (2012)	LDA,IG	Combines statistical and semantic information by building SFT, thus improving the accuracy of short text classification
4	Yang and Pedersen (1997)	DF,IG,MI, CHI, TS	IG & CHI are most effective in aggressive term removal, DF has 90% term removal capability and TS has 50-60% capability, MI has inferior performance due to a bias favouring rare terms and a strong sensitivity to probability estimation errors, DF, IG & CHI scores of a term are strongly correlated, thereby meaning that DF thresholding is not an adhoc approach but reliable measure
5	Zhen et al. (2011)	Kullback-Leibler (KL) divergence based global feature evaluation criterion	Measure differences of distributions between two categories and overcomes following disadvantages of CHI:- CHI computes local scores of the term over each category and then takes maximum or average value of these scores as the global term-goodness criterion. Now there is no explicit explanation on how to choose maximum or average, Secondly, CHI cannot reflect the degree of scatter of a term
6	Bakus and Kamel(2006)	Variant of MI (MIFS-C)	Finds optimal value of redundancy parameter, outperformed IG, CHI, OR, CFS (Co-relation based feature selection) and Markov blanket
7	Azam and Yao (2012)	TF,DF	Superior for smaller feature sets, have larger scatter of features among the classes, accumulate information in data at a faster rate.
8	Yang et al. (2012)	CMFS	Measured the significance of a term in both inter-category and intra-category with NB and SVM as the classifiers, superior to DIA, IG, CHI, DF, OCFS when NB was used and superior to DIA, IG, DF, OCFS when SVM was used
9	Liu & Hu (2007)	ARM	Viewed a sentence rather than a document as a transaction
10	Qiu et al. (2008)	DF,TF, TF-IDF,CHI	A two-stage feature selection algorithm consisting of local feature set constructed using DF, TF, TFIDF and global feature set using CHI
11	Meng and Lin (2010)	DF, MI, CHI, LSI	Reduced number of dimensions drastically, introduced the semantic model to overcome the problems existing in the VSM
12	Meng et al. (2011)	FCD, LSI	Reduced number of dimensions drastically, introduced the semantic model to overcome the problems existing in the VSM
13	Zifeng et al. (2007)	CLDA	Selects features using LDA but does not transform high-dimensional feature space into low-dimensional feature space, better than IG and CHI
14	Torkkola (2003)	LDA	Reduced the dimensionality without sacrificing accuracy, 5718 number of features reduced to 12
15	Fragoudis et al. (2005)	Best Terms (BT)	Fast performance, increases classification accuracy of NB and SVM, complexity of BT is linear with respect to number of training set documents and is independent from both the vocabulary size and number of categories

16	Pinheiro et al . (2012)	ALOFT	Ensures that every document in the training set is represented by at least one feature, performs better than the classical Variable Ranking
17	Liu et al. (2012)	Improved AM	Removes those ambiguous features which are not removed by AM
18	Nuntiyagul(2005)	PKIP	Used for item banks, short textual data
19	Ko et al. (2004)	Novel algorithm	Measures the importance of sentences using text summarization techniques, shows difference between important and unimportant sentences, considers features from more important sentences
20	Wilbur and Kim(2009)	NBMBM	Offers no significant advantage over plain MBM, word burstiness is so strong that additional occurrences of a word adds no useful information
21	Chen et al. (2007)	Entropy Label Assignment (ELA), IG, CHI, OCFS	Transforms multi-label data to single-label data before applying feature selection algorithms to solve multi-label feature selection problem, integration of four transformation approaches viz. All Label Assignment (ALA), No Label Assignment (NLA), Largest Label Assignment (LLA) and Smallest Label Assignment (SLA)

Table 2. Comparative analysis of different document representation methods

S. No	Paper	Technique	Conclusion/Advantage
Purpose: Work based on stemming and weighting methods			
1	Song et al. (2005)	Text representation schemes viz. stop words removal, word stemming, indexing, weighting and normalization	Schemes are corpus-dependent, for Reuters indexing and normalizing are important, for 20-NewsGroup weighting and normalizing are important, among the five factors, 'normalization' is the most important, removal of stop words from vocabulary is not harmful, word stemming is harmful on Reuters and helpful on 20 NewsGroup
2	Harrag et al. (2011)	Stemming methods : Light, Root-Based & Dictionary-Lookup Stemming	Used for Arabic text classification, dictionary-lookup stemming is superior for ANN and light-stemming is superior for SVM
3	Leopold (2002)	TFIDF weighting scheme	Has larger impact on SVM performance rather than kernel function alone, no pre-processing and feature selection is needed for SVM
4	Lan et.al (2005)	'tf.rf' (based on discriminating power)	Term weighting scheme has a larger impact on the performance of SVM rather than the kernel function
5	Wu et al. (2012)	Term weighting scheme (based on word clustering)	More accurate than the original weighting methods, reduces dimensionality
6	Altınçay (2012)	Different weighting schemes	Ordering of terms according to their discriminative abilities is dependent on the weighting scheme
Purpose: To handle class imbalance problem			
7	Lu et al. (2009)	TF•Rd redundancy based term weighting scheme	Based on posterior probability distribution, promotes precision-recall, reduces sensitiveness to number of features
8	Chen et al.(2011)	Semantic re-sampling methods based on probabilistic topic models DECOM & DECODER	Uses global semantic information, DECOM deals with class imbalance by generating new samples of rare classes, DECODER smoothens the data by regenerating all samples in each class for data sets with noisy samples & rare classes
9	Sun et al. (2009)	Different re-sampling and term weighting methods using SVM classifiers	SVM learns the best decision surface in most test cases, for classification tasks involving high imbalance ratios it is therefore more critical to find an appropriate threshold than applying any of the re-sampling or weighting strategies
Purpose: To modify the conventional 'BOW'/ 'VSM' representation			
10	Deng (2009)	Singular Value Decomposition (SVD)	Reduced dimensionality to a great extent , discovered important semantic relationships between terms
11	Wang (2009)	Thesaurus of concepts built from Wikipedia	Included semantic relations (synonymy, hyponymy, and associative relations) thus expanding BOW representation
12	Hassan et.al (2011)	Wikitology	Enhanced text categorization by adding background knowledge to documents, better than other knowledge bases like Word Net, Open Project Directory (OPD), Wikipedia
13	Ozgur (2012)	An algorithm based on the extension of BOW	Extracted fewer but informative features using the concepts of lexical dependencies and pruning
14	Yun et.al (2012)	A two-level representation model (2RM)	Represents syntactic information at first level and semantic information at second level, better than VSM
15	Pu et al . (2007)	Local Word Bags (LWB)	Represented a document as a set of tf-idf vectors, considers detailed local text information ignored by BOW model

16	Jo (2009)	Neural Network Classifier (NTC)	Encoded documents into string vectors instead of numerical vectors, removed problems of huge dimensionality and sparse distribution
17	Kehagias (2003)	Word and sense based classifiers	The use of senses for text representation does not result in any significant categorization improvement
18	Zhang et al.(2008)	Concept representation and Sub-topic representation.	Represents the documents using extracted multi-words, have larger impact on performance of SVM rather than kernel, subtopic representation outperformed concept representation, linear kernel outperformed non-linear kernel

Table 3. Comparative analysis of different data mining methods

S. No	Author	Technique	Conclusion/Advantage
1	Lim et al.(2006)	PSVM	Allowed for automatic tuning of the penalty coefficient parameter C and kernel parameter via MCMC method
2	Kumar & Gopal (2010a)	PSVM, Fuzzy PSVM	Maintains constant training time irrespective of the penalty parameter *C* and categories, Fuzzy PSVM showed improved generalization over PSVM
3	Kumar &Gopal (2010 b)	OAA-SVM, OAO-SVM	OAA performed better than OAO for uni-label text classification, OAA is suitable for text corpuses with small number of categories whereas OAO is better on text corpora with large number of categories
4	Lee et al. (2012 a)	Euclidean-SVM	Has low impact on the implementation of kernel function and soft margin parameter C, thus retaining the classification accuracy of SVM classifier, Euclidean distance function replaces the optimal separating hyper-plane as the classification making function of the SVM, consumes a longer time and has lower classification accuracy than conventional SVM as Euclidean distance calculation which inherits the characteristic of nearest neighbor approach suffers from the curse of dimensionality
5	Dai et al. (2008)	CHI based Algorithm	Solves the problem of fine-text-categorization characterized with many redundant features, Outperformed SVM and C4.5 algorithms
Purpose: To solve the multi-label text classification problem			
6	Wang & Chiang (2011)	Multi-label classifier	Sample set from high dimensional space was mapped into a lower dimensional, documents were categorized into multiple classes, probability that a document belongs to a class was predicted
7	Wang & Chiang (2007) & (2009)	OAA-FSVM, OAO-FSVM	Create multi-margin hyperplanes used to distinguish positive class from negative class and then the weight of each data set can be set according to its class, thus solving the Fuzzy data problem, out-performed OAA-SVM and OAO-SVM methods in multi-class text categorization problem
8	Namburuet al. (2005)	PLS	Better than SVM for multiclass categorization, SVM is more suitable for binary classification as for multiclass categorization SVM requires a voting scheme based on the results of pair-wise classification
9	Zelaia et al. (2011)	KNN algorithm	Based on Bayesian voting and SVD, documents represented by 15,000 features in the BOW form and by 11,000 in the Bag-of-Lemmas were simplified to 300 features, consequently saving space and time
10	Schapire et al.(2000)	BoosTexter system	Embodies four versions of boosting, combines many simple and moderately inaccurate categorization rules into a single, highly accurate categorization rule
11	Esuli et al.(2008)	TREEBOOST.MH	It is exponentially cheaper to train and to test than ADABOOST.MH
12	Chen et al.(2004)	Boosting algorithm	Achieves better performance on multi-label Chinese text categorization tasks than other methods viz. NB and Rocchio algorithm.
Purpose: Work based on linear classification methods			
13	Zhang & Oles (2001)	LLSF, LR, NB,SVM	Share similarity by finding hyper-planes that separate a class of document vectors from its compliment, NB is worse, LLSF performed very close to the state-of-art, LR performed as well as SVM
14	Basu et al. (2002)	ANN and SVM	SVM preferable for short text documents, less complex than ANN because parameter that constructs the hyper-plane is very small, ANN performs large matrix calculations on matrices
15	Wang et al.(2006)	Optimal SVM	It outperformed many other conventional algorithms
16	Li et al. (2011)	VPRSVM–RKNN	Combines strengths of both SVM and KNN, VPRSVM filters noisy data which reduces impact on RKNN classifier
17	Mitra et al.(2007)	LS-SVM	Based on LSI coefficient with Gaussian radial basis function (GRBF), LS-SVM outperforms KNN, NB, SVM and NN based system.
Purpose: To solve PU-oriented text classification problem			
18	Peng et al.(2008)	Algorithm based on 1-DNF	Improved 1-DNF obtained more negative data with a lower error rate than 1-DNF, PSOC (Particle Swarm Optimization Classifier) performed better than weighted voting method

19	Shi et al. (2011)	Semi-supervised algorithm	Used positive and unlabeled data based on tolerance rough set and ensemble learning, tolerance rough set theory extracted a set of negative examples. SVM, Rocchio and NB algorithms were used as base classifiers to construct an ensemble classifier, Outperformed algorithms like SEM (Spy EM) and PEBL (Positive Example Based Learning)
20	Pan et al. (2012)	DCEPU	Used concept drift by constructing a validation set and dynamic weighting scheme to assign weight to each base classifier in the ensemble, weighting scheme considers not only the local weight of each base classifier, but also a global weight of each classifier
21	Cabrera et al. (2009)	Semi-supervised algorithm	Removes the problem of supervised learning technique i.e. need of a great number of training instances to construct an accurate classifier, does automatic extraction of unlabeled examples from the Web
22	Lee and Kageura (2007)	A virtual document technique	Enlarged positive training documents, made virtual documents by combining relevant document pairs for a topic in the training set, not only preserved topic but even improved topical representation by using relevant terms that were not given importance in real documents
Purpose: Work based on Multi-lingual text classification			
23	Lee et al. (2006)	LSI (unsupervised), SVM (supervised)	Both the methods are complimentary, a hybrid system to overcome the disadvantages of both approaches is required to give better results
24	Lee & Yang (2009)	LSI and SVM (unsupervised)	SOM-based (self-organizing maps) supervised technique is used, a hybrid system is required to overcome the disadvantages of both the approaches
Purpose: Work based on KNN algorithm			
25	Wan et al.(2012)	SVM-NN approach	Incorporates SVM to training stage of KNN classification, has low impact on the implementation of parameter K thus retaining classification accuracy of KNN, suffers from high time consumption
26	Dong et al.(2012)	kNN algorithm	Based on eager learning, overcomes lazy learning of traditional kNN algorithm, decreases high computational expense
27	Wang & Wang (2007)	TFKNN based on SSR tree	Searches exact k nearest neighbors quickly, ranks all child nodes according to distances between their central points and the central point of their parent reducing searching scope and similarity computing
28	Soucy et al.(2001)	KNN	Reaches impressive results using very few features
29	Guo et al.(2006)	*k*NN model	*k*NN model outperforms the kNN and Rocchio classifiers
30	Wu et al. (2008)	*k*-NN and M3-*k*-NN	Majority voting method performed best when M3-*k*-NN is used while linear voting method performed when *k*-NN is used, Gaussian voting method performed best for both *k*-NN & M3-*k*-NN, M3-*k*-NN used less *k* value than *k*-NN and spent less time to complete prediction than *k*-NN
31	Lu & Bai (2010)	Refined KNN	Weight measurement is based on variance, needs more running time than traditional KNN but is far better than traditional KNN
32	Zhan and Chen (2010)	GC,CNN, SNN,RNN, ENN	GC had highest average generalization accuracy when compared with CNN, SNN, RNN, ENN, especially in the presence of uniform class noise
33	Jiang et al.(2012)	Improved KNN	Based on clustering algorithm, reduced text similarity computation, outperformed KNN, NB and SVM classifiers
34	Haifeng (2010)	Improved KNN	Based on skew sort condition
35	Yang (1999)	DT, NB, NN, kNN Rocchio, LLSF	kNN, LLSF and NN had the best performance, All other learning algorithms performed well except for a NB approach, BPNN has limitations such as slow training speed and can be easily trapped into a local minimum
Purpose: Work based on ANN algorithm			
36	Li et al. (2009)	Revised BP algorithm	Based on automatically constructed thesaurus, removed disadvantages of conventional BPNN
37	Li et al. (2012)	MRBP, LPEBP	MRBP and LPEBP alleviated the problems of standard BPNN, Semantic relations of terms were considered using a CBT and WN
38	Wang et al.(2009)	MBPNN, LSA	Alleviated the problem of traditional BPNN, LSA removed the problem of VSM by including the semantic relations between the terms
39	Zheng et al.(2012)	Framework (LSA+ RELM)	The weights and a bias-variance trade-off was achieved by adding a regularization term into feed-forward NN, Learnt faster than conventional algorithms such as feed-forward NN or SVM
40	Harrag et al.(2010)	SVD-based MLP/RBF	MLP classifier outperformed the RBF classifier, SVD-supported NN classifier was better than the basic NN for Arabic text categorization.
41	Ruiz and Srinivasan (2002)	Feed-forward NN	Hierarchical structure performed better than equivalent flat model, Used divide and conquer principle, comparable performance with respect to the optimized Rocchio algorithm
Purpose: Work based on Centroid-based classifier			
42	Nguyen et al. (2012)	CFC, CFC-KL, CFC-JS (Jensen–Shannon)	CFC leads to poor performance on class-imbalanced data, CFC prunes terms that appeared across all classes discarding non-exclusive but useful terms, CFC–KL was generalized to handle multi-class data by replacing KL measure with multi-class JS divergence (CFC–JS), KL and JS weighted classifier outperformed baseline CFC and unweighted SVM

43	Tan et al. (2011)	Model Adjustment (MA) algorithm	Deals with model misfit problem of centroid classifier, uses training-set errors and training-set margins in contrast to methods like Weight Adjustment, Voting, Refinement, Drag-Pushing and therefore cannot guarantee generalization capability of base classifiers for unseen examples, converges to optimal solution for a linearly separable problem

Lo (2008) proposed a mechanism to facilitate website management, named as 'WebQC' which used P-control chart to control web service quality. It gave a warning signal if the complaining rate is higher than the upper control limit. In the paper by Couto et al.(2006), a comparative study of digital library citations and web links in the context of TC was presented. It was concluded that measures based on co-citation are the best performers for the web directories and bibliographic coupling measures are appropriate for digital library containing scientific papers. The work by Saldarriaga et al. (2010) categorized online handwritten documents based on their textual contents using KNN and SVM algorithms. The effect of word recognition errors on the categorization performances was analyzed, by comparing the performances of a categorization system with the texts obtained through online handwriting recognition and the same texts available as ground truth. Paquet et al.(2012) proposed an approach to categorize handwritten document which is based on the detection of some discriminative keywords prior to the use of tf-idf representation. Results show that the discriminative keyword extraction system leads to better recall/precision tradeoffs than the full recognition strategy. In the paper by Farhoodi & Yari (2010), two efficient machine learning algorithms were examined for Persian text document. Experiments showed that the performance of KNN is better than SVM for Persian text classification. Lia and Mu (2010) proposed an incremental learning algorithm on large-scale corpus for Chinese text classification. In this study, an approach based on SVMs for web text mining of large-scale systems on GBODSS was developed to support enterprise decision making. Experimental results showed that this approach has good classification accuracy by incremental learning and it was seen that speed up of computation time was almost super linear. In the paper by Zakzouk and Mathkour (2012), three binary text classifiers viz. SVM based on evolutionary algorithm, C4.5 and NB were built to test the cricket class of SGSC. It was observed that Naïve-Bayesian leads the pack with best effectiveness ratios overall. Wermter (2000) showed that neural network can be used for tasks like text routing. This was illustrated using different architectures and different corpora. In the paper by Liang et al.(2006), a new dictionary-based text classification approach is proposed to classify the chemical web pages efficiently. After automatic segmentation on the documents to find dictionary terms for document expansion, the approach adopts latent semantic indexing (LSI) to produce the final document vectors, and the relevant categories are finally assigned to the test document by using k-NN algorithm.

Results

74 journal papers and 58 conference proceedings have been evaluated in this review systematically. Each subsection of this section will address the respective research question listed in the above section.

Relevant text classification journals (RQ1)

We used papers on text classification in 24 journals and these journals with three or more papers are displayed in Table 4, together with the corresponding number, proportion, and cumulative proportions of papers (Catal, 2011). Proportions and cumulative proportions have been calculated by considering only the number of journal papers in review. 7 journals shown in Table 4 include 68% of all journal papers in review.

Table 4. Most important text classification journals

Rank	Journal Name	# Papers	Prop (%)	Cumulative Prop (%)
1	Expert Systems with Applications	16	22	22
2	Information Processing &Mgmt	8	11	33
3	Information Retrieval	7	9	42
4	Knowledge Information System	7	9	51
5	Pattern Recognition Letters	5	7	58
6	Pattern Anal Application	4	5	63
7	Journal of Intelligent Information Systems	3	4	67

Year showing the maximum publications (RQ2)

Figure 3 is a curve which plots publication year on the x-axis and the number of papers published in that year on the y-axis for papers in review. We have reviewed in total 132 papers on text classification. Out of these, 44% of papers are conference proceedings and 56% of papers are journal papers. As can be seen from the curve, 22 papers were published in the year 2012 which represents the maximum publication year in this area followed by the year 2011 which shows 17 publications. We can also observe that majority of the papers have been published after year 2004. Figure 4 is a clear indication of our observation. In this figure papers have been classified into two groups: papers published before 2004 and papers published after year 2004. In total, 22 papers have been published before year 2004 and 110 papers have been published after year 2004. It is clearly seen that the popularity of text classification area increased drastically after year 2004 and thus researchers should only examine papers published after year 2004 to reach the most important papers.

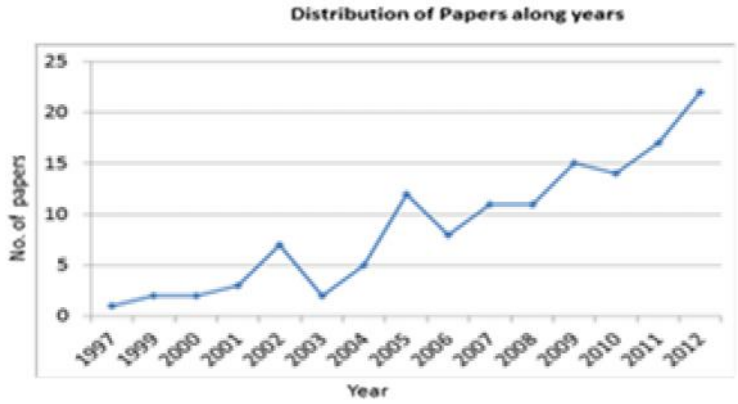

Figure 3. Number of papers per year in review

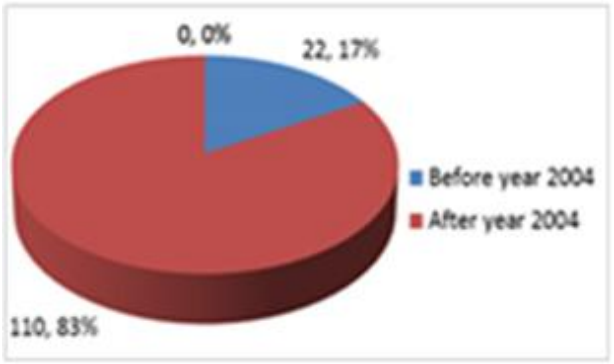

Figure 4. Distribution of papers after year 2004 (Number of papers /percentage of total papers)

Widely used data mining methods (RQ3)

From an extensive literature survey done in the area of text classification, it was observed that the frequently used data mining methods are SVM, KNN, NB, ANN, Rocchio algorithm and Association rule mining (ARM). These methods machine learning algorithm along with the number of papers using these methods is shown in Table 5. As it is clear from the table, SVM is the most popular used by the researchers in their work. Many of the authors have worked on SVM algorithm and proposed its advanced version to better enhance the applicability of this algorithm, thereby improving the performance of text classification. KNN algorithm is the second popular method used by the researchers as it is used in 31 papers. Similar to SVM, the authors of these papers have also proposed different variants of KNN and then compared the performance of their proposed KNN algorithm with the different machine learning algorithms to show that new KNN algorithm performs better than conventional algorithms. Finally, we have NB algorithm which is used in 23 papers and thus falling under rank 3. It can also be clearly seen that 65% of the papers are using SVM and KNN algorithm for categorizing the text and only 35 % of the papers are using other methods apart from KNN and SVM algorithms. This distribution is shown in the Figure 5. This clearly indicates that SVM and KNN algorithms are amongst the

most popular machine learning algorithms used by the researchers. Out of a total of 132 papers, 88 papers used SVM and KNN algorithms and 44 papers used other data mining methods.

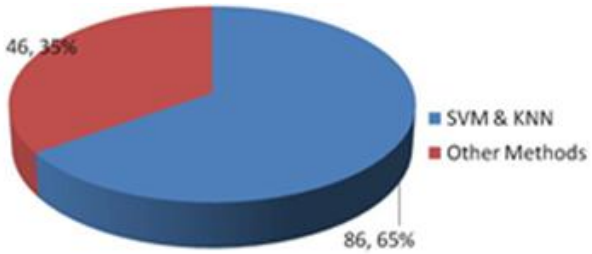

Figure 5. Distribution of machine learning methods (Number of papers/ percentage of total papers)

Table 5. Most important data mining methods used

Rank	Data Mining Methods	# Papers
1	SVM	55
2	KNN	31
3	NB	23
4	ANN	10
5	Rocchio Algorithm	9
6	Association Rule Mining	4

Distribution of data mining methods (RQ4)

Distribution of data mining methods which have been used in the papers is shown in Figure 6. Methods have been divided into four groups: statistical methods, machine learning based methods, statistical methods + machine learning based methods and the unknown category. If machine learning based methods are used together with statistical methods in the same model, method of that paper is marked as 'statistical methods + machine learning based methods'. Some of the authors (Tao et.al, 2005; Altinacy&Erenel, 2010; Luo et.al, 2011) have not used any of the data mining method in their paper as they have proposed a new term-weighting method and have compared its performance with the existing term- weighting schemes. Method of that paper is marked as 'unknown methods'. It has been observed that 86% of papers used machine learning based methods and only 6% of papers used statistical methods like Hidden Markov Model (Frasconi et al., 2002), Logistic Regression (Zhang and Oles 2001; Yen et al. 2011), Partial Least Squares (Namburu et al., 2005), Linear Least Square Fit (Yang &Pedersen, 1997; Yang, 1999; Zhang &Oles, 2001). Because statistical methods are considered black-box solutions and these models are highly dependent on data, it is promising to see that more researchers are exploring the potential of machine learning methods to predict text classification modules. As shown in figure 6, out of a total of 132 papers, 8 papers used statistical methods, 113 papers used machine learning methods, and 8 papers applied statistical methods together with machine learning methods.

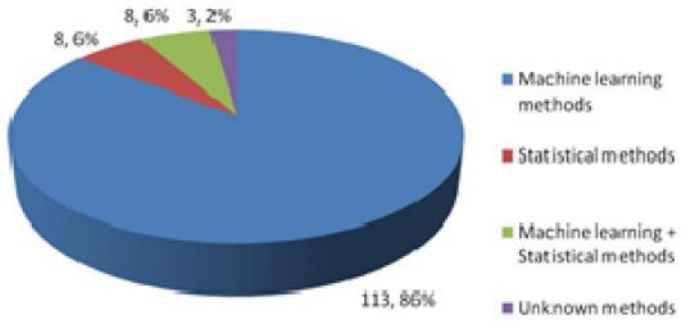

Figure 6. Distribution of data mining methods (Number of papers/ percentage of total papers)

Widely used feature selection methods (RQ5)

Methods which are widely used by the researchers are displayed in Table 6 along with the number of papers using these methods. We can conclude from the table that CHI is the most widely used method followed by IG as the second most popular method. It can be seen from the table that CHI, IG and MI are amongst the most popular feature selection methods used by the researchers in their work. Many of the authors have worked on these methods and proposed their advanced version to better enhance the applicability of the method, thereby improving the performance of text classification. Also many of the authors have proposed their own feature selection method considering these popularly used methods as the base of their theory.

Table 6. Most important feature selection methods used

Rank	Feature Selection Methods	# Papers
1	Chi-squared test(CHI)	21
2	Information Gain (IG)	20
3	Mutual Information(MI)	16
4	Latent Semantic Indexing (LSI), Singular Value Decomposition (SVD)	10
5	Document Frequency (DF)	8
6	Term Strength (TS)	3
7	Odds Ratio(OR)	3
8	Linear Discriminant Analysis (LDA)	3

Widely used datasets (RQ6)

There are a number of datasets which are used by the researchers to conduct the experiment in order to evaluate the performance of the data mining method applied. It has been observed that researchers have mainly used the datasets from the famous machine learning repository called UCI (University of California Irvine) which consists of several public datasets. Table 7 shows the three popularly used dataset. It is clear from the table that the dataset namely Reuters-21578 is the most widely used dataset which is collection of documents that appear on Reuters financial

newswire service. 20NG is the second most popular dataset which is a collection of 20,000 newsgroup documents, partitioned across 20 different newsgroups. Finally, we have Web KB dataset which consists of WWW pages collected from computer science department of various universities. There are various other kinds of datasets used by the researchers like the datasets consisting of medical data, E-mail data, mathematics data etc.

Apart from this, few authors have also analyzed the software project reports available in different open source software repositories for predicting various aspects of software engineering. Menzies and Marcus (2008) have analyzed the defect reports available in the PITS database of NASA and presented an automated method named SEVERIS which is used to assign the severity level to each of the defect found during testing. Assigning the correct severity levels to defect reports is very important as it directly impacts resource allocation and planning of subsequent defect fixing activities. Runeson et al. (2007) and Wang et al. (2008) have also analyzed the defect reports and developed a tool that would be used to detect duplicate reports using Natural Language Processing (NLP). Cubranic and Murphy (2004) analyzed an incoming bug report and proposed an automated method that would assist in bug triage to predict the developer that would work on the bug based on the bug description. Canfora and Cerulo (2005) discussed how software repositories can help developers in managing a new change request, either a bug or an enhancement feature. Lucca et al. (2002) analyzed the maintenance requests coming from the customers in the form of a ticket (containing the description of the request) and developed a router that would work around the clock and would keep dispatching the maintenance requests to an appropriate maintenance team. Huang et al. (2006) analyzed the Non-Functional Requirements (NFRs) as specified by the stakeholders during the requirements gathering process and developed an automated technique that is used to classify them on the basis of its type, thus leading to the detection of NFRs early in the development life cycle.

Table 7. Most important dataset used

Rank	Dataset Used	# Papers
1	Reuters-21578	71
2	20-Newsgroup	36
3	Web KB	15

Distribution of document representation methods (RQ7)

Distribution of document representation methods which have been used in papers is shown in Figure 7. Papers have been classified into two groups: papers using Vector Space Model (VSM) as its document representation method and papers using some other method (apart from VSM).We have done this classification because it was observed from the literature survey that majority of the papers used VSM and only a few papers proposed a new method for document representation. As it is clear from the figure, 83% of the papers used VSM for representing the document and only 17 % of the papers used other methods.

This clearly shows that VSM is the most common document representation method used by the researchers. Out of a total of 132 papers, 109 papers used VSM and only 23 papers used a modified version of VSM for representing the document as Bag-Of –Words. For instance, the paper by Frasconi et al. (2002) used the BOW representation resulting from a multi-nominal word event model using Hidden Markov Model (hmm) for classification. Kehagias et al. (2003) used a word & sense- based method for representation. The work by Kim and Kim (2004) used the concept of passage based document wherein the document is split into passages & categorization is done for each passage & finally document categories are merged with passage categories. An and Chen (2005) represented the document in a subspace of the dimensionality using an algorithm based on concept learning. Doan (2005) proposed a document representation method based on fuzzy set theory. Pu et al.(2007)introduced the concept of Local-Word-Bag and Zhang et al.(2008) proposed multi-word document representation. Srinivas et al.(2008) proposed a MFCC algorithm (Multi-type Features Co-selection for Clusters) for representing text document as a projection on clusters formed from the input dataset. Few authors also used the concept of Rough Set Theory for representing the document (Zhou and Zhang 2008; Shi 2011). Many authors extended the traditional VSM representation by using Wikipedia as a thesaurus to consider semantic relationships between key terms (Li et al., 2009; Wang et al., 2009; Li et al., 2012; Yun et al., 2012). Jo (2009) encoded the documents into string vector (rather than numeric vectors) to avoid the problems of huge dimensionality and sparse distribution which are inherent in encoding documents into numerical vectors. Lee et al. (2012b) used the Bayesian vectorization technique and Wu and Yang (2012) used term clustering algorithm for representation.

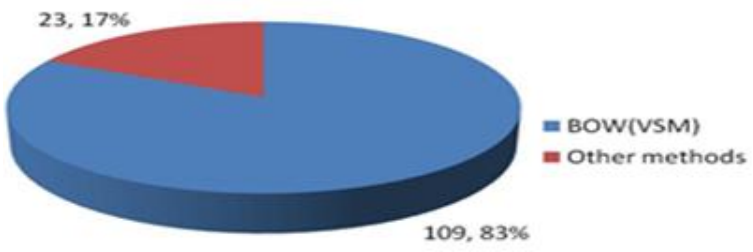

Figure 7. Distribution of document representation methods (# papers/ percentage of total papers)

Conclusion

To retrieve specific information from web is like finding a proverbial needle in the haystack. In this work, the needle is that single piece of information a user needs and the haystack is the large data warehouse built up on the web over a long period of time. Text classification is emerging as one of the most prominent technique to handle this problem. In this paper, we have reviewed the text classification papers since 1997 to 2012 published in conference proceedings and journals of high repute to evaluate the progress made in the area of text classification so far. This review would help future research based on the past studies. We have evaluated the papers with a specific focus on types of data mining methods, feature selection methods, the dataset and the

evaluation techniques used by each study to carry out the results. Following trends were observed in this work:

- Large number of datasets was used to evaluate the results to provide more accurate and generalized results. It was also observed that more number of public datasets was used for text classification because repeatable, refutable and verifiable models can only be built with public datasets. From the review, we can conclude that majority of the researchers have made use of the datasets available in UCI repository which has a collection of wide range of public datasets. As many as 122 papers out of a total of 132 papers have made use of the three most popular datasets available in UCI repository viz. Reuters-21578, 20-Newsgroup and Web KB. Some authors have also made use of private datasets which are not freely available and therefore it is not possible to compare results of the studies using private datasets with results of our own models. Thus, we should make use of public datasets available in UCI repository.
- The review clearly indicates that the most common method for representing a document in text classification is the Vector-Space-Model which represents each document as a vector consisting of an array of words. It is seen from the review that 83% of the papers are using VSM technique for representing the document and only 17% of the papers are using other methods. Once the document is represented as Bag-Of-Words, we reduce of the dimensionality of the dataset by removing features that are considered irrelevant for the classification. There are a number of methods proposed in the literature for feature selection, but Chi-squared statistic and Information Gain are considered as the most widely used methods.
- As specified in this review, machine learning models have better features than statistical methods. Therefore, we should increase the percentage usage of the models based on machine learning techniques. It has been observed that 86% of papers used machine learning based methods and only 6% of papers used statistical methods. Among the various machine learning algorithms studied in the literature, it has been observed that SVM and KNN algorithms are the most widely used machine learning algorithms in the area of text classification. These algorithms have been used by 65% of the papers. While some of the authors have also made use of the statistical methods, but their use has been very limited over the past few years because these methods are black-box solutions and are highly dependent on data. It is promising to see that there is a drastic shift from traditional statistical methods to modern machine learning methods.

From the literature survey done so far, it was observed that very little work has been done in analyzing the software project reports which play a very important role in improving software quality. Software repositories consist of different kinds of project reports which when analyzed using text classification techniques can help in assisting project managers and developers in their SDLC activities. Data contained in software repositories have generated new opportunities in various directions such as change propagation, fault analysis, software complexity, software reuse and social networks.

For instance, defect descriptions of given software can be analyzed in order to predict the severity of defects by developing a new and automated method which can assist the test engineer in assigning severity levels to defect reports. Building an agent using text mining techniques would lead to saving of resources like time, manpower and money as text mining and machine learning methods are low cost, automatic and rapid. Similarly, maintenance requests (i.e. tickets) for a large, distributed telecommunication system can also be analyzed in order to route them to specialized maintenance teams by developing a router that would work around the clock and would keep dispatching the maintenance requests coming from the customers in the form of a ticket (containing the description of the request) to an appropriate maintenance team. The system would be able to balance the workload between different maintenance teams and there would be lowest misclassification error as routing is done without human intervention. Also, Non-Functional Requirements (NFRs) as specified by the stakeholders during the requirements gathering process can be analyzed and then classified on the basis of its type, thus leading to the detection of NFRs early in the development life cycle.

References

Altinçay, H., & Erenel, Z. (2010). Analytical evaluation of term weighting schemes for text categorization. *Pattern Recognition Letters*, 31, 1310–1323.

Altinçay, H., & Erenel, Z. (2012). Using the absolute difference of term occurrence probabilities in binary text categorization. *ApplIntell*, 36, 148–160.

Amine, B.M., & Mimoun, M. (2007).WordNet based cross-language text categorization. *IEEE/ACS International Conference on Computer Systems and Applications*, AICCSA.

An, J., & Chen, Y.P.P. (2005).Keyword Extraction for Text Categorization. Proceedings of the IEEE International Conference on Active Media Technology AMT.

Antonie, M.L., & Zai'ane, O.R. (2002). Text document categorization by term association. *Proceedings of the IEEE International Conference on Data Mining*, ICDM.

Azam, N., & Yao, J.T. (2012). Comparison of term frequency and document frequency based feature selection metrics in text categorization. *Expert Systems with Applications*, 39, 4760–4768.

Bakus, J., & Kamel, M.S. (2006). Higher order feature selection for text classification. *Knowledge Information System*, 9(4), 468-491.

Basu, A., Watters, C., & Shepherd, M. (2002). Support vector machines for text categorization. *Proceedings of the 36th Hawaii IEEE International Conference on System*, HICSS'03.

Cabrera, R.G., Gomez, M.M.Y., Rosso, P., & Pineda, L.V. (2009). Using the Web as corpus for self-training text categorization. *Information Retrieval*, 12, 400–415.

Canfora, G., & Cerulo, L. (2005). How software repositories can help in resolving a new change request. *Workshop on Empirical Studies in Reverse Engineering*.

Catal, C. (2011). Software fault prediction: A literature review and current trends. *Expert Systems with Applications*, 38, 4626-4636.

Chang, Y.C., Chen, S.M., & Liau, C.J. (2008). Multi-label text categorization based on a new linear classifier learning method and a category-sensitive refinement method. *Expert Systems with Applications*, 34, 1948–1953.

Chen, E., Lin, Y., Xiong, H., Luo, Q., & Ma, H. (2011). Exploiting probabilistic topic models to improve text categorization under class imbalance. *Information Processing and Management*, 47, 202-214.

Chen, J., Zhou, X., & Wu, Z. (2004). A multi-label Chinese text categorization system based on boosting algorithm. *Proceedings of the Fourth IEEE International Conference on Computer and Information Technology*.

Chen, L., Guo, G., & Wang, K. (2011). Class-dependent projection based method for text categorization. *Pattern Recognition Letters*, 32, 1493-1501.

Chen, W., Yan, J., Zhang, B., Chen, Z., & Yang, Q. (2007). Document transformation for multi-label feature selection in text categorization. *Seventh IEEE International Conference on Data Mining*.

Chen, X.Y., Chen, Y., Wang, L., & Hu, Y.F. (2004). Text categorization based on frequent patterns with term frequency. *Proceedings of the Third International Conference on Machine Learning and Cybernetics*, Shanghai, 26-29 August.

Couto, T., Cristo, M., Goncalves, M.A., Calado, P., Ziviani, N., Moura, E., & Neto, B.R. (2006). A comparative study of citations and links in document classification. *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries JCDL*, 75-84.

Cubranic, D., & Murphy, G.C. (2004). Automatic bug triage using text categorization. *Proceedings of the Sixteenth International Conference on Software Engineering and Knowledge Engineering*.

Dai, L., Hu, J., & Liu, W.C. (2008). Using modified chi square and rough set for text categorization with many redundant features. *IEEE International Symposium on Computational Intelligence and Design*.

Deng, Y.X.W.W. (2009). A New Text Categorization Method Based on SVD and Cascade Correlation Algorithm. *IEEE International Conference on Artificial Intelligence and Computational Intelligence*.

Doan, S. (2005). A fuzzy-based approach for text representation in text categorization. *IEEE International Conference on Fuzzy Systems*.

Dong, T., Cheng, W., & Shang, W. (2012). The research of kNN text categorization algorithm based on eager learning. *IEEE International Conference on Industrial Control and Electronics Engineering*.

Esuli, A., Fagni, T., & Sebastiani, F. (2008). Boosting multi-label hierarchical text categorization. *Information Retrieval*, 11, 287-313.

Farhoodi, M., & Yari, A. (2010). Applying machine learning algorithms for automatic Persian text classification. *Advanced Information Management and Service (IMS), 6th IEEE International Conference*, 318 – 323.

Ferrer, J.A., & Juan, A. (2010). Constrained domain maximum likelihood estimation for naïve Bayes text classification. *Pattern Analysis and Applications*, 13(2), 189–196.

Fragoudis, D., Meretakis, D., & Likothanassis, S. (2005). Best terms: an efficient feature-selection algorithm for text categorization. *Knowledge and Information Systems*, 8, 16–33.

Frasconi, P., Soda. G., & Vullo, A. (2002). Hidden markov models for text categorization in multi-page documents. *Journal of Intelligent Information Systems*, 18(2/3), 195-217.

Gao, G., & Guan, S. (2012). Text categorization based on improved Rocchio algorithm. *IEEE International Conference on Systems and Informatics ICSAI*.

Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. (2006). Using kNN model for automatic text categorization. *Soft Computing*, 10(5), 423-43.

Hadni, M., Lachkar, A., & Ouatik, S.A. (2012). A new and efficient stemming technique for arabic text categorization. *IEEE International Conference on Multimedia Computing and Systems (ICMCS)*.

Haifeng, L., Shousheng, L., & Zhan, S. (2010). An Improved kNN Text Categorization on Skew Sort Condition. *IEEE International Conference on Computer Application and System Modeling ICCASM*.

Harrag, F., El-Qawasmah, E., & AL-Salman, A.M.S. (2011). Stemming as a Feature Reduction Technique for Arabic Text Categorization. *10th IEEE International Symposium on Programming and Systems (ISPS)*.

Harrag, F., Al-Salman, A.M.S., & Mohammed, M.B. (2010). A comparative study of neural networks architectures on arabic text categorization using feature extraction. *IEEE International Conference on Machine and Web Intelligence (ICMWI)*.

Hassan, S., Rafi, M., & Shaikh, M.S. (2011). Comparing SVM and naive bayes classifiers for text categorization with Wikitology as knowledge enrichment. *14th IEEE International Multitopic Conference*.

Huang, J.C., Settimi, R., Zou, X., & Solc, P. (2006). The detection and classification of non-functional requirements with application to early aspects. *14th IEEE International Conference on Requirements Engineering*, 39-48.

Ikonomakis, M., Kotsiantis, S., & Tampakas, V. (2005). Text classification using machine learning techniques. *WSEAS Transactions on Computers*, 8(4), 966-974.

Jiang, S., Pang, G., Wu, M., & Kuang, L. (2012). An improved K-nearest-neighbor algorithm for text categorization. *Expert Systems with Applications*, 39, 1503–1509.

Jo, T. (2009). Automatic Text Categorization using NTC. *First IEEE International Conference on Networked Digital Technologies NDT*.

Kehagias, A., Petridis, V., Kaburlasos, V. G., & Fragkou. P. (2003). A comparison of word- and sense-based text categorization using several classification algorithms. *Journal of Intelligent Information Systems*, 21(3), 227–247.

kim, J., & kim, M.H. (2004). An evaluation of passage-based text categorization. *Journal of Intelligent Information Systems*, 23(1), 47–65.

Ko, Y., Park, J., & Seo, J. (2004). Improving text categorization using the importance of sentences. *Information Processing and Management*, 40, 65–79.

Kumar, M.A., & Gopal, M. (2010). An investigation on linear SVM and its variants for text categorization. *Second IEEE International Conference on Machine Learning and Computing*.

Kumar, M.A., Gopal, M. (2010). A comparison study on multiple binary-class SVM methods for unilabel text categorization. *Pattern Recognition Letters*, 31, 1437–1444.

Lakshmi, K., & Mukherjee, S. (2005). Profile extraction from mean profile for automatic text categorization. *Proceedings of the IEEE International Conference on Computational Intelligence for Modeling, Control and Automation, and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'05)*.

Lam. W., Ruiz, M., & Srinivasan, P. (1999). Automatic text categorization and its application to text retrieval. *IEEE Transactions on Knowledge and Data Engineering*, 11(6), November/December.

Lan, M., Sung, S.Y., Low, H.B., & Tan, C.L. (2005). A Comparative Study on Term weighting Schemes for Text Categorization. *Proceedings of International Joint Conference on Neural Networks*, Montreal, Canada, July 31 - August 4.

Lee, C.H., & Yang, H.C. (2009). Construction of supervised and unsupervised learning systems for multilingual text categorization. *Expert Systems with Applications*, 36, 2400–2410.

Lee, C.H., Yang, H.C., Chen, T.C., & Ma, S.M. (2006). A comparative study on supervised and unsupervised learning approaches for multilingual text categorization. *Proceedings of the First IEEE International Conference on Innovative Computing, Information and Control ICICIC*.

Lee, K.S., & Kageura, K. (2007). Virtual relevant documents in text categorization with support vector machines. *Information Processing and Management*, 43, 902–913.

Lee, L.H., Wan, C.H., Rajkumar, R., & Isa, D. (2012). An enhanced Support Vector Machine classification framework by using Euclidean distance function for text document categorization. *ApplIntell*, 37, 80–99.

Lee, L.H., Isa, D., Choo, W.O., & Chue, W.Y. (2012). High relevance keyword extraction facility for bayesian text classification on different domains of varying characteristic. *Expert Systems with Applications*, 39, 1147–1155.

Leopold, E., & Kindermann, J. (2002). Text categorization with support vector machines. how to represent texts in input space? *Machine Learning*, 46, 423–444.

Li, C.H., Song, W., & Park, S.C. (2009). An automatically constructed thesaurus for neural network based document categorization. *Expert Systems with Applications*, 36, 10969–10975.

Li, C.H., Yang, J.C., & Park, S.C. (2012). Text categorization algorithms using semantic approaches, corpus-based thesaurus and WordNet. *Expert Systems with Applications*, 39, 765–772.

Li, H., & Yamanishi, K. (2002). Text classification using ESC-based stochastic decision lists. *Information Processing and Management*, 38, 343-362.

Li, W., Miao, D., & Wang, W. (2011). Two-level hierarchical combination method for text classification. *Expert Systems with Applications*, 38, 2030–2039.

Li, Y., Hsu, D.F., & Chung, S.M. (2009). Combining multiple feature selection methods for text categorization by using rank-score characteristics. *21st IEEE International Conference on Tools with Artificial Intelligence*.

Lia, Z., & Mu, J. (2010). Web text categorization for large-scale corpus. *IEEE International Conference on Computer Application and System Modeling ICCASM*.

Liang, C.Y., Guo, L., Xia, Z.J., Nie, F.G., Li, X.X., Su, L., & Yang, Z.Y. (2006). Dictionary-based text categorization of chemical web pages. *Information Processing and Management*, 42, 1017–1029.

Lim, B.P.C., Tsui, M.H., Charastrakul, V., & Shi, D. (2006). Web search with text categorization using probabilistic framework of SVM. *IEEE International Conference on Systems, Man, and Cybernetics*, October 8-11, Taipei, Taiwan.

Liu, L., & Liang, Q. (2011). A high-performing comprehensive learning algorithm for text classification without pre-labeled training set. *Knowledge Information System*, 29, 727–738.

Liu, S.Z., & Hu, H.P. (2007). Text classification using sentential frequent item sets. *Journal of Computer Science & Technology*, 22(2), 334.

Liu, T., & Guo, J. (2005). Automatic text categorization based on angle distribution. *Proceedings of the Fourth International Conference on Machine Learning and Cybernetics*, Guangzhou, 18-21 August.

Liu, Z., & Yang, J. (2012). An improved ambiguity measure feature selection for text categorization. *4th IEEE International Conference on Intelligent Human-Machine Systems and Cybernetics*.

Lo, S. (2008). Web service quality control based on text mining using support vector machine. *Expert Systems with Applications*, 34, 603–610.

Lu, F., & Bai, Q. (2010). A Refined Weighted K-Nearest Neighbors Algorithm for Text Categorization. *IEEE International Conference on Intelligent Systems and Knowledge Engineering (ISKE)*.

Lu, Z.Y., Lin, Y.M., Zhao, S., Chen, J.N, & Zhu, W.D. (2009). A redundancy based term weighting approach for text categorization. *World Congress on Software Engineering, IEEE*.

Lucca, G.A.D., Penta, M.D., & Gradara, S. (2002). An approach to classify software maintenance requests. *International Conference on Software Maintenance (ICSM)*.

Luo, Q., Chen, E., & Xiong, H. (2011). A semantic term weighting scheme for text categorization. *Expert Systems with Applications*, 38, 12708–12716.

Malhotra, R., & Jain, A. (2011). Software fault prediction for object oriented systems: A literature review. *ACM SIGSOFT Software Engineering Notes*, 36(5).

Malik, H.H., Fradkin, D., & Moerchen, F. (2011). Single pass text classification by direct feature weighting. *Knowledge Information System*, 2879–98.

Meng, J., & Lin, H. (2010). A two-stage feature selection method for text categorization. *Seventh IEEE International Conference on Fuzzy Systems and Knowledge Discovery FSKD*.

Meng, J., Lin, H., & Yu, Y. (2011). A two-stage feature selection method for text categorization. *Computers and Mathematics with Applications*, 62, 2793–2800.

Menzies, T., & Marcus, A. (2008). Automated severity assessment of software defect reports. *IEEE International Conference on Software Maintenance (ICSM)*.

Miao, Y.Q., & Kamel, M. (2011). Pairwise optimized Rocchio algorithm for text categorization. *Pattern Recognition Letters*, 32, 375–382.

Mitra, V., Wang, C.J., & Banerjee, S. (2007). Text classification: A least square support vector machine approach. *Applied Soft Computing*, 7, 908–914.

namburu, S.M., Tu, H., Luo, J., & Pattipati, K.R. (2005). Experiments on supervised learning algorithms for text categorization. *IEEE Aerospace Conference, 1-8.*

Nguyen, T.T., Chang, K., & Hui, S.C. (2012). Supervised term weighting centroid-based classifiers for text categorization. *Knowledge Information System*.

Nuntiyagul, A., Naruedomkul, K., & Cercone, N. (2005). Recovering "lack of words" in text categorization for item bank. *Proceedings of the 29th Annual IEEE International Computer Software and Applications Conference (COMPSAC'05)*.

Nuntiyagul, A., Naruedomkul, K., Cercone, N., & Wongsawang, D. (2005). PKIP: Feature selection in text categorization for item banks. *Proceedings of the 17th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'05)*.

Ozgur, L., & Gungor, T. (2012). Optimization of dependency and pruning usage in text classification. *Pattern Analysis and Applications*, 15(1), 45–58.

Pan, S., Zhang, Y., & Li, X. (2012). Dynamic classifier ensemble for positive unlabeled text stream classification. *Knowledge Information System*, 33, 267–287.

Pan, X., & Zhang, S. (2011). Semi-supervised fuzzy learning in text categorization. *Eighth IEEE International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*.

Paquet, T., Heutte, L., Koch, G., & Chatelain, C. (2012). A categorization system for handwritten documents. *IJDAR*, 15, 315–330.

Peng, T., Zuo, W., & He, F. (2008). SVM based adaptive learning method for text classification from positive and unlabeled documents. *Knowledge Information System*, 16, 281–301.

Pinheiro, R.H.W., Cavalcanti, G.D.C., Correa, R.F., & Ren, T.I. (2012). A global-ranking local feature selection method for text categorization. *Expert Systems with Applications*, 39 12851–12857.

Pu, W., Liu, N., Yan, S, Yan, J., Xie, K., & Chen, Z. (2007). Local word bag model for text categorization. *Seventh IEEE International Conference on Data Mining*.

Qiu, L.Q., Zhao, R.Y., Zhou, G., & Yi, S.W. (2008). An extensive empirical study of feature selection for text categorization. *Seventh IEEE/ACIS International Conference on Computer and Information Science*.

Rigutini, L., Maggini, M., & Liu, B. (2005). An EM based training algorithm for cross-language text categorization. *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI'05)*.

Ruiz, M.E., & Srinivasan, P. (2002). Hierarchical text categorization using neural networks. *Information Retrieval*, 5, 87–118.

Runeson, P., Alexandersson, M., & Nyholm, O. (2007). Detection of duplicate defect reports using natural language processing. *29th IEEE International Conference on Software Engineering (ICSE)*, 499 – 508.

Saldarriaga, S.P., Gaudin, C.V., & Morin, E. (2010). Impact of online handwriting recognition performance on text categorization. *IJDAR*, 13, 159–171.

Schapire, R.E., & Singer, Y. (2000). BoosTexter: A boosting-based system for text categorization. *Machine Learning*, 39, 135–168.

Shanahan, J. (2001). Modeling with words: an approach to text categorization. *IEEE International Fuzzy Systems Conference*.

Shi, L., Ma, X., Xi, L., Duan, Q., & Zhao, J. (2011). Rough set and ensemble learning based semi-supervised algorithm for text classification. *Expert Systems with Applications*, 38, 6300–6306.

Sjøberg, D.I.K., Hannay, J.E., Hansen, O., Kampenes, V.B., Karahasanovic, A., Liborg, N.K., & Rekdal, A.C. (2005). A survey of controlled experiments in software engineering. *IEEE Transactions on Software Engineering*, 31(9).

Song, F., Liu, S., & Yang, J. (2005). A comparative study on text representation schemes in text Categorization. *Pattern Analysis and Applications*, 8, 199–209.

Soucy, P., & Mineau, G.W. (2001).A simple kNN algorithm for text categorization. *Proceedings of the IEEE International Conference on Data Mining, ICDM*.

Srinivas, M., Spreethi, K.P., Prasad, E.V. DR., & Kumari, S.A. (2008). MFCC and ARM algorithms for text categorization. *Proceedings of the IEEE International Conference on Computing, Communication and Networking ICCCN*.

Sun, A., Lim, E.P., & Liu, Y. (2009). On strategies for imbalanced text classification using SVM: A comparative study. *Decision Support Systems*, 48, 191–201.

Suzuki, M., Yamagishi, N., Tsai, Y.C., Ishidaand, T., & Goto, M. (2010). English and Taiwanese text categorization using n-gram based on vector space model. *ISITA*, Taichung, Taiwan, October 17-20.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, Vol. 34.

Tan, S., Wang, Y., & Wu, G. (2011). Adapting centroid classifier for document categorization. *Expert Systems with Applications*, 38(8), 10264–10273.

Tan, C.M., Wang, Y.F., & Lee, C.D. (2002). The use of bigrams to enhance text categorization. *Information Processing and Management*, 38, 529–546.

Tao, Z.Y., Ling, G., & Chang, W.Y. (2005). An Improved TF-IDF approach for text classification. *Journal of Zhejiang University Science*, 6A(1), 49-55.

Taşci, S., & Güngör, T. (2008). An evaluation of existing and new feature selection metrics in text categorization. *23rd IEEE International Symposium on Computer and Information Sciences ISCIS*.

Taşci, S., & Güngör, T. (2009). LDA-based keyword selection in text categorization. *24th IEEE International Symposium on Computer and Information Sciences ISCIS*.

Torkkola, K. (2003). Discriminative features for text document classification. *Pattern Analysis and Applications*, 6, 301-308.

Wan, C.H., Lee, L.H., Rajkumar , R., & Isa, D. (2012). A hybrid text classification approach with low dependency on parameter by integrating K-nearest neighbor and support vector machine. *Expert Systems with Applications*, 39, 11880–11888.

Wang , Z., Sun, X., & Zhang, D. (2006). An optimal text categorization algorithm based on SVM. *Proceedings of the IEEE International Conference on Communications, Circuits and Systems*.

Wang, B.K., Huang, Y.F., Yang, W.X., & Li, X. (2012). Short text classification based on strong feature thesaurus. *Journal of Zhejiang University Science C (Computers & Electronics)*, 13(9), 649-659.

Wang, P., Hu, J., Zeng, H.J., & Chen, Z. (2009). Using Wikipedia knowledge to improve text classification. *Knowledge Information System*, 19, 265–281.

Wang, T.Y., & Chiang, H.M. (2007). Fuzzy support vector machine for multi-class text categorization. *Information Processing and Management*, 43, 914–929.

Wang, T.Y., & Chiang, H.M. (2009). One-against-one fuzzy support vector machine classifier: An approach to text categorization. *Expert Systems with Applications*, 36, 10030–10034.

Wang, T.Y., & Chiang, H.M. (2011). Solving multi- label text categorization problem using support vector machine approach with membership function. *Neurocomputing*, 74, 3682–3689.

Wang, W., & Yu, B. (2009). Text categorization based on combination of modified back propagation neural network and latent semantic analysis. *Neural Computing and Applications*, 18, 875–881.

Wang, X., Zhang, L., Xie, T., Anvik, J., & Sun, J. (2008). An approach to detecting duplicate bug reports using natural language and execution information. *Proceedings of the 30th international conference on Software engineering*, pp. 461-470.

Wang, Y., & Wang, Z.O. (2007). A fast kNN algorithm for text categorization. *Proceedings of the Sixth IEEE International Conference on Machine Learning and Cybernetics*, Hong Kong, 19-22 August.

Wermter, S. (2000). Neural network agents for learning semantic text classification. *Information Retrieval*, 3, 87–103.

Wilbur, W.J., & Kim, W. (2009). The ineffectiveness of within-document term frequency in text classification. *Information Retrieval*, 12, 509–525.

Wu, K., Lu, B.L., Utiyama, M., & Isahara, H. (2008). An empirical comparison of min–max-modular k-NN with different voting methods to large-scale text categorization. *Soft Computing*, 12(7), 647-655.

Wu, Y.C., & Yang, J.C. (2012). A weighted cluster-based Chinese text categorization approach: incorporating with word clusters. *IIAI International Conference on Advanced Applied Informatics IEEE*.

Xu, J.S., & Wang, Z.O. (2004). A new method of text categorization based on pa and Kohonen network. *Proceedings of the Third International Conference on Machine Learning and Cybernetics*, Shanghai, 26-29 August.

Xu, J.S. (2007). TCBPLK: A new method of text categorization. *Proceedings of the Sixth IEEE International Conference on Machine Learning and Cybernetics*, Hong Kong, 19-22 August.

Yan, B., & Qian, D. (2007). Building a simple and effective text categorization system using relative importance in category. *Third IEEE International Conference on Natural Computation*.

Yang, J., Liu, Y., Zhu, X., Liu, Z., & Zhang, X. (2012).A new feature selection based on comprehensive measurement both in inter-category and intra-category for text categorization. *Information Processing and Management*, 48, 741–754.

Yang, Y. (1999). An evaluation of statistical approaches to text categorization. *Information Retrieval*, 1(1-2), 69–90.

Yang, Y., & Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. *Proceedings of the Fourteenth International Conference on Machine Learning ICM*, 412-420.

Yen, S.J., Lee, Y.S., Ying, J.C., & Wu, Y.C. (2011). A logistic regression-based smoothing method for Chinese text categorization. 2011. *Expert Systems with Applications*, 38, 11581–11590.

Youn, E., & Jeong, M.K. (2009). Class dependent feature scaling method using naive Bayes classifier for text data mining. *Pattern Recognition Letters*, 30, 477–485.

Yun, J., Jing, L., Yu, J., & Huang, H.(2012). A multi-layer text classification framework based on two-level representation model. *Expert Systems with Applications*, 39, 2035–2046.

Zakzouk, T.S., & Mathkour, H.I. (2012). Comparing text classifiers for sports news. *Procedia Technology*, 1, 474 – 480.

Zelaia, A., Alegria, I., Arregi, O., & Sierra, B. (2011). A multiclass/multilabel document categorization system: Combining multiple classifiers in a reduced dimension. *Applied Soft Computing*, 11, 4981–4990.

Zhan, Y., & Chen, H.(2010). Reducing samples learning for text categorization. *3rd IEEE International Conference on Information Management, Innovation Management and Industrial Engineering*.

Zhang, M.L., Peña, J.M., & Robles, V. (2009). Feature selection for multi-label naive bayes classification. *Information Sciences*, 179, 3218–3229.

Zhang, T., & Oles, F.J. (2001). Text categorization based on regularized linear classification methods. *Information Retrieval*, 4, 5–31.

Zhang, W., Yoshida, T., & Tang, X. (2008). Text classification based on multi-word with support vector machine. *Knowledge-Based Systems*, 21, 879–886.

Zhen, Z., Zeng, X., Wang, H., & Han, L. (2011). A global evaluation criterion for feature selection in text categorization using Kullback-Leibler Divergence. *IEEE International Conference of Soft Computing and Pattern Recognition (SoCPaR)*.

Zheng, W., Qian, Y., & Lu, H. (2012). Text categorization based on regularization extreme learning machine. *Neural Computing and Applications*, 22(3), 447-456.

Zhou, X., & Zhang, H. (2008). An algorithm of text categorization based on similar rough set and fuzzy cognitive map. *Fifth IEEE International Conference on Fuzzy Systems and Knowledge Discovery*.

Zifeng, C., Baowen, X., Weifeng, Z., & Junling, X. (2006). A new approach of feature selection for text categorization. *Wuhan University Journal of Natural Sciences*, 11(5), 1335-1339.

Zifeng, C., Baowen, X., Weifeng, Z., Dawei, J., & Junling, X. (2007). CLDA: Feature selection for text categorization based on constrained LDA. *IEEE International Conference on Semantic Computing*.

Zuo, M., Zeng, G., Tu, X., & Zuo, M. (2010). Study on an Improved Naive Bayesian Classifier used in the Chinese text categorization. Second IEEE International Conference on Modeling, Simulation and Visualization Methods.

Appendix

AKS: Adaptive Keyword Selection, CMFS: Comprehensive Measurement Feature Selection, ARM: Association Rule Mining, LSI: Latent Semantic Indexing, FCD: Feature Contribution Degree, CLDA: Constrained Linear Discriminant Analysis, ALOFT: At Least One FeaTure, PKIP: Patterned Keywords In Phrase, NBMBM: NB multivariate Bernoulli model, PSVM: Probabilistic Framework for SVM, MCMC: Markov Chain Monte Carlo, PLS: Partial Least Square, LS-SVM: Least Square SVM, DCEPU: Dynamic Classifier Ensemble method for Positive and Unlabeled Text stream, TFKNN: Tree-Fast KNN, M3-*k*-NN: Min-Max-Modular KNN, GC: Generalization Capability algorithm, CNN: Condensed Nearest Neighbor, SNN: Selective Nearest Neighbor, RNN: Reduced Nearest Neighbor, ENN: Edited Nearest Neighbor, BPNN: Backward Propagation NN, MRBP: Morbidity neurons Rectified BPNN, LPEBP: Learning Phase Evaluation BPNN, CBT: Corpus Based Thesaurus, WN: WordNet thesaurus, RELM: Regularization Extreme Learning Machine, MLP/RBF NN: Multilayer Perceptron/Radial Basis Function NN, CFC: Class Feature Centroid.

Bibliographic information of this paper for citing:

Jindal, Rajni, Malhotra, Ruchika, & Jain, Abha (2015). "Techniques for text classification: Literature review and current trends." *Webology*, 12(2), Article 139. Available at: http://www.webology.org/2015/v12n2/a139.pdf