A Review of Machine Learning Algorithms for Text-Documents Classification

Article · February 2010

DOI: 10.4304/jait.1.1.4-20 · Source: DOAJ

4 authors, including:

Lam Hong Lee

19 PUBLICATIONS **324** CITATIONS

SEE PROFILE

Khairullah Khan

University of Science & Technology Bannu

26 PUBLICATIONS **153** CITATIONS

SEE PROFILE

Aurangzeb Khan

University of Science & Technology Bannu

34 PUBLICATIONS **192** CITATIONS

SEE PROFILE

A Review of Machine Learning Algorithms for Text-Documents Classification

Aurangzeb Khan, Baharum Baharudin, Lam Hong Lee*, Khairullah khan
Department of Computer and Information Science,
Universiti Teknologi PETRONAS, Tronoh, Malaysia.
*Faculty of Science, Engineering and Technology,
Universiti Tunku Abdul Rahman, Perak Campus, Kampar, Malaysia.
(E-mail: aurangzebb_khan@yahoo.com, baharbh@petronas.com.my,leelh@utar.edu.my,khairullah_k@yahoo.com)

***Abstract*— With the increasing availability of electronic documents and the rapid growth of the World Wide Web, the task of automatic categorization of documents became the key method for organizing the information and knowledge discovery. Proper classification of e-documents, online news, blogs, e-mails and digital libraries need text mining, machine learning and natural language processing techniques to get meaningful knowledge. The aim of this paper is to highlight the important techniques and methodologies that are employed in text documents classification, while at the same time making awareness of some of the interesting challenges that remain to be solved, focused mainly on text representation and machine learning techniques. This paper provides a review of the theory and methods of document classification and text mining, focusing on the existing literature**.

***Index Terms*— Text mining, Web mining, Documents classification, Information retrieval**.

I. INTRODUCTION

The text mining studies are gaining more importance recently because of the availability of the increasing number of the electronic documents from a variety of sources. The resources of unstructured and semi structured information include the word wide web, governmental electronic repositories, news articles, biological databases, chat rooms, digital libraries, online forums, electronic mail and blog repositories. Therefore, proper classification and knowledge discovery from these resources is an important area for research.

Natural Language Processing (NLP), Data Mining, and Machine Learning techniques work together to automatically classify and discover patterns from the electronic documents. The main goal of text mining is to enable users to extract information from textual resources and deals with the operations like, retrieval, classification (supervised, unsupervised and semi supervised) and summarization. However how these documented can be properly annotated, presented and classified. So it consists of several challenges, like proper annotation to the documents, appropriate document representation, dimensionality reduction to handle algorithmic issues [1], and an appropriate classifier function to obtain good generalization and avoid over-fitting. Extraction, Integration and classification of electronic documents from different sources and knowledge discovery from these documents are important for the research communities.

Today the web is the main source for the text documents, the amount of textual data available to us is consistently increasing, and approximately 80% of the information of an organization is stored in unstructured textual format [2], in the form of reports, email, views and news etc. The [3] shows that approximately 90% of the world's data is held in unstructured formats, so Information intensive business processes demand that we transcend from simple document retrieval to knowledge discovery. The need of automatically retrieval of useful knowledge from the huge amount of textual data in order to assist the human analysis is fully apparent [4].

Market trend based on the content of the online news articles, sentiments, and events is an emerging topic for research in data mining and text mining community [5]. For these purpose state-of-the-art approaches to text classifications are presented in [6], in which three problems were discussed: documents representation, classifier construction and classifier evaluation. So constructing a data structure that can represent the documents, and constructing a classifier that can be used to predicate the class label of a document with high accuracy, are the key points in text classification.

One of the purposes of research is to review the available and known work, so an attempt is made to collect what's known about the documents classification and representation. This paper covers the overview of syntactic and semantic matters, domain ontology, tokenization concern and focused on the different machine learning techniques for text classification using the existing literature. The motivated perspective of the related research areas of text mining are:

Information Extraction (IE) methods is aim to extract specific information from text documents. This is the first

Aurangzeb khan and Khairullah Khan are PhD Students, Department of Computer and Information Science at Universiti Teknologi PETRONAS, Tronoh, Malaysia.
Baharum Baharudin is an Assistant Professor at the Department of Computer and Information Science at Universiti Teknologi PETRONAS, Tronoh, Malaysia.
Lam Hong Lee is an Assistant Professor at the Faculty of Science, Engineering and Technology of Universiti Tunku Abdul Rahman, Perak Campus, located in Kampar, Malaysia.
(E-mail:aurangzebb_khan@yahoo.com,baharbh@petronas.com.my leelh@utar.edu.my)., Khairullah_k@yahoo.com)

Manuscript received May 28, 2009; revised September 7, 2009.

doi:10.4304/jait.1.1.4-20

approach assumes that text mining essentially corresponds to information extraction.

Information Retrieval (IR) is the finding of documents which contain answers to questions. In order to achieve this goal statistical measures and methods are used for automatic processing of text data and comparison to the given question. Information retrieval in the broader sense deals with the entire range of information processing, from data retrieval to knowledge retrieval [7].

Natural Language Processing (NLP) is to achieve a better understanding of natural language by use of computers and represent the documents semantically to improve the classification and informational retrieval process. Semantic analysis is the process of linguistically parsing sentences and paragraphs into key concepts, verbs and proper nouns. Using statistics-backed technology, these words are then compared to the taxonomy.

Ontology is the explicit and abstract model representation of already defined finite sets of terms and concepts, involved in knowledge management, knowledge engineering, and intelligent information integration [23].

In this paper we have used system literature review process and followed standard steps for searching, screening, data-extraction, and reporting.

First of all we tried to search for relevant papers, presentations, research reports and policy documents that were broadly concerned with documents classification or text mining. We identified appropriate electronic databases and websites. Potentially relevant papers were identified using the electronic databases and websites, Such as IEEE Explore, Springer Linker, Science Direct, ACM Portal and Googol Search Engine. For best and consistent search a systematic search strategy was adopted. Proper keywords, queries, and phrases were derived from the desired research question. These keywords were arranged into categories and related keywords were arranged. Some facilities of digital libraries like sort by year etc were also used. The search keywords were refined to include only those words which have produced successful results. We used boolean logic for efficient searching, for example (Classification OR text OR recommendations). We also tried combination of words like Text Mining, Trend and Ontology analysis, Documents classification and Subjectivity Analysis etc.

Each search results were checked and assessed on screen to find relevance for inclusion and exclusion with the criteria that we made two categories of papers i.e. in or before 2000 and after 2000. The following studies were included: The result statements written in English, The research is conducted after 1980, Published and/or unpublished research, focused on documents classification, Machine Learning and Natural Language Processing (NLP). The non English writing and study before 1980 were excluded.

To find evidence and check the quality of papers we carried out an in-depth study of the results provided from the research. In our future work we will try to make this step more strong and effective. We have tried to get some reports drawn using tables and graphs on the basis of existing studies.

The rest of the paper is organized as follows. In Section 2 an overview of documents representation approaches, Section 3 presents document classification models, in Section 4 new and hybrid techniques were presented. Section 5 consists of comparative study of different methods and finally in Section 6, some discussions and conclusion were made.

II DOCUMENTS REPRESENTATION

The documents representation is one of the preprocessing technique that is used to reduce the complexity of the documents and make them easier to handle, the document have to be transformed from the full text version to a document vector. Text representation is the important aspect in documents classification, denotes the mapping of a documents into a compact form of its contents. A text document is typically represented as a vector of term weights (word features) from a set of terms (dictionary), where each term occurs at least once in a certain minimum number of document. A major characteristic of the text classification problem is the extremely high dimensionality of text data. The number of potential features often exceeds the number of training documents. A definition of a document is that it is made of a joint membership of terms which have various patterns of occurrence. Text classification is an important component in many informational management tasks, however with the explosive growth of the web data, algorithms that can improve the classification efficiency while maintaining accuracy, are highly desired [8].

Documents pre-processing or dimensionality reduction (DR) allows an efficient data manipulation and representation. Lot of discussions on the pre-processing and DR are there in the current literature and many models and techniques have been proposed. DR is a very important step in text classification, because irrelevant and redundant features often degrade the performance of classification algorithms both in speed and classification accuracy and also its tendency to reduce overfitting.

DR techniques can classified into Feature Extraction (FE) [11] and Feature Selection (FS) approaches, as discussed below.

A. Feature Extraction

The process of pre-processing is to make clear the border of each language structure and to eliminate as much as possible the language dependent factors, tokenization, stop words removal, and stemming [10]. FE is the fist step of pre processing which is used to presents the text documents into clear word format. So removing stop words and stemming words is the pre-processing tasks [12]. The documents in text classification are represented by a great amount of features and most of them could be irrelevant or noisy [9]. DR is the exclusion of a large number of keywords, base preferably on a statistical

process, to create a low dimension vector [13]. DR techniques have inward much attention recently because effective dimension reduction make the learning task more efficient and save more storage space [14]. Commonly the steeps taken please for the feature extractions (Fig.1) are:

Tokenization: A document is treated as a string, and then partitioned into a list of tokens.

Removing stop words: Stop words such as “the”, “a”, “and”… etc are frequently occurring, so the insignificant words need to be removed.

Stemming word: Applying the stemming algorithm that converts different word form into similar canonical form. This step is the process of conflating tokens to their root form, e.g. connection to connect, computing to compute etc.

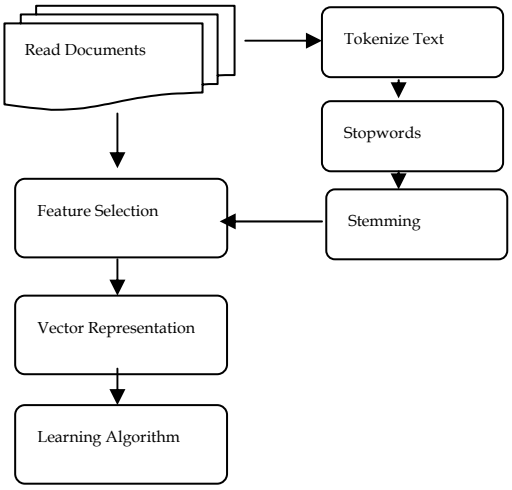

Fig. 1 Document Classification Process

B. Feature Selection

After feature extraction the important step in preprocessing of text classification, is feature selection to construct vector space, which improve the scalability, efficiency and accuracy of a text classifier. In general, a good feature selection method should consider domain and algorithm characteristics [15]. The main idea of FS is to select subset of features from the original documents. FS is performed by keeping the words with highest score according to predetermined measure of the importance of the word [9].The selected features retains original physical meaning and provide a better understanding for the data and learning process [11]. For text classification a major problem is the high dimensionality of the feature space. Almost every text domain has much number of features, most of these features are not relevant and beneficial for text classification task, and even some noise features may sharply reduce the classification accuracy [16]. Hence FS is commonly used in text classification to reduce the dimensionality of feature space and improve the efficiency and accuracy of classifiers.

There are mainly two types of feature selection methods in machine learning; wrappers and filters. Wrappers use the classification accuracy of some learning algorithms as their evaluation function. Since wrappers have to train a classifier for each feature subset to be evaluated, they are usually much more time consuming especially when the number of features is high. So wrappers are generally not suitable for text classification. As opposed to wrappers, filters perform FS independently of the learning algorithm that will use the selected features. In order to evaluate a feature, filters use an evaluation metric that measures the ability of the feature to differentiate each class [17]. In text classification, a text document may partially match many categories. We need to find the best matching category for the text document. The term (word) frequency/inverse document frequency (TF-IDF) approach is commonly used to weight each word in the text document according to how unique it is. In other words, the TF-IDF approach captures the relevancy among words, text documents and particular categories.

Some of the recent literature shows that works are in progress for the efficient feature selection to optimize the classification process. A novel feature selection method is presented in [17], in which the degrees of deviation from poison distribution are utilized to select informative features. Based on ant colony optimization a new feature selection algorithm is presented in [18], to improve the text categorization. Also in [19] the authors introduced a new weighting method based on statistical estimation of the importance of a word categorization problem. The [20] proposed a new feature scaling method, called class–dependent–feature–weighting (CDFW) using naive Bayes (NB) classifier.

Many feature evaluation metrics have been explored, notable among which are information gain (IG), term frequency, Chi-square, expected cross entropy, Odds Ratio, the weight of evidence of text, mutual information, Gini index. Term frequency and document frequency (TF/DF) (Table-1) etc. A good feature selection metric should consider problem domain and algorithm characteristics.

The authors in [21] focused on the document representation techniques and demonstrate that the choice of document representation has a profound impact on the quality of the classifier. They used the centroid-based text classifier, which is a simple and robust text classification scheme, and compare four different types of document representations: N-grams, Single terms, phrases and RDR which is a logic-based documents representation. The N-gram is a string-based representation with no linguistic processing. The Single term approach is based on words with minimum linguistic processing. The phrase approach is based on linguistically formed phrases and single words. The RDR is based on linguistic processing and representing documents as a set of logical predicates. In [22] the authors present significantly more efficient indexing and classification of large document repositories, e.g. to support information retrieval over all enterprise file servers with frequent file updates.

TABLE 1. FEATURE SELECTION TECHNIQUES

Gain Ration	$GR(t_k, c_i) = \frac{\sum_{c \in \{c_i, \bar{c}_i\}} \sum_{t \in \{t_k, \bar{t}_k\}} p(t, c) \log \frac{P(t, c)}{P(t) P(c)}}{-\sum_{c \in \{c_i, \bar{c}_i\}} P(c) \log P(c)}$
Informational Gain(IG)	$IG(w) = -\sum_{j=1}^{K} P(c_j) \log P(c_j) + P(w) \sum_{j=1}^{K} P(c_j \mid w) \log P(c_j \mid w) + P(\bar{w}) \sum_{j=1}^{K} P(c_j \mid \bar{w}) \log P(c_j \mid \bar{w})$ $= H(samples) - H(samples \mid w)$
Chi Square	$\chi^2(f_i, c_j) = \frac{\lvert D \rvert \times (\#(c_j, f_i) \#(\bar{c}_j, \bar{f}_i) - \#(c_j, \bar{f}_i).\#(\bar{c}_j, f_i))^2}{(\#(\#(c_j, f_i) + \#(c_j, \bar{f}_i)) \times (\#(\bar{c}_j, f_i) + \#(\bar{c}_j, \bar{f}_i)) \times ((c_j, f_i) + \#(\bar{c}_j, f_i)) \times (\#(c_j, \bar{f}_i) + \#(\bar{c}_j, \bar{f}_i))}$
Conditional mutual Information	$CMI(C \mid S) = H(C) - H(C \mid S_1, S_2, ..., S_n)$
Document Frequency(DF)	$DF(t_k) = P(t_k)$
Term Frequency(TF)	$tf(f_i, d_j) = \frac{freq_{ij}}{\max_k freq_{Kj}}$
Inverse Document Frequency(IDF)	$\lvert idf \rvert = \log \frac{\lvert D \rvert}{\lvert \#(f1) \rvert}$
Term	$s(t) = P(t \in y \mid t \in x)$
Weighted Ration	$WOddsRation(w) = P(w) \times OddsRatio(w)$
Odd Ration	$OddsRatio(f_i, c_j) = \log \frac{P(f_i \mid c_j)(1 - P(f_i \mid \neg c_j))}{(1 - P(f_i \mid c_j))(P(f_i \mid \neg c_j)}$

C. Semantic and Ontology Base Documents Representation

This section focused no the semantic, ontology techniques, language and the associated issues for documents classification. According to [44] the statistical techniques are not sufficient for the text mining. Better classification will be performed when consider the semantic under consideration. Ontology is a data model that represents a set of concepts within a domain and the relationships between those concepts. It is used to reason about the objects within that domain. Ontology is the explicit and abstract model representation of already defined finite sets of terms and concepts, involved in knowledge management, knowledge engineering, and intelligent information integration [23].The characteristics of objects and entities (individuals, instances) is a real thing and association (relations) with attribute is used for the titles of the two concepts or entities. Ontology is divided into three categories i.e., Natural Language Ontology (NLO), Domain Ontology (DO) and Ontology Instance (OI) [24]. NLO is the relationship between general lexical tokens of statements based on natural language, DO is the knowledge of a particular domain and OI is the automatically generated web page behaves like an object. Web Ontology Language (OWL) is the ontology support language derived from America DAPRA Agent Markup Language (DAML) and based on ontology, inference and European Ontology Interchange Language (OIL)[25]. OWL claims to be an extension in Resource Description Framework (RDF)[26]. In expressing logical statements because it not only describe classes and properties but also provides the concepts of namespace, import, cardinality relationship between the classes and enumerated classes. Ontology has been proposed for handling semantically heterogeneity when extracting informational from various text sources such as internet [27].

Machine learning algorithms automatically builds a classifier by learning the characteristics of the categories from a set of classified documents, and then uses the classifier to classify documents into predefined categories. However, these machine learning methods have some drawbacks: (1) In order to train classifier, human must collect large number of training text terms, the process is very laborious. If the predefined categories changed, these methods must collect a new set of training text terms. (2) Most of these traditional methods haven't considered the semantic relations between words, so it is difficult to improve the accuracy of these classification methods [6]. (3) The issue of translatability, between one natural language into another natural language. These types of issues identify that machine understanding systems are facing problems. Such issues are discussed in the literature, some of these may be addressed if we have machine readable ontology [32], and that's why this is and important potential area for research.

During the text mining process, ontology can be used to provide expert, background knowledge about a domain. Some resent research shows the importance of the domain ontology in the text classification process, the [27] presents automatic classification of incoming news using hierarchical news ontology, based on this classification on one hand, and on the users' profiles on the other hand,

the personalization engine of the system is able to provide a personalized paper to each user on to her mobile reading device. A novel ontology-based automatic classification and ranking method is represented in [34] where Web documents are characterized by a set of weighted terms, categories are represented by ontology. In [35] the authors presented an approach towards mining ontology from natural language, in which they considered a domain-specific dictionary for telecommunications documents.

How to include user context and preferences in the form of an ontology in order to classify unstructured documents into useful categories and the use of a context-based free text interpreter (CFTI) [36], which performs syntactical analysis and lexical semantic processing of sentences, to derive a description of the content of the unstructured documents, with relevance to the context of the user. In [38] the authors presented a novel text categorization method based on ontological knowledge that does not require a training set. Also an Automatic Document Classifier System based on Ontology and the Naïve Bayes Classifier is proposed in [39].

Ontology's have shown their usefulness in application areas such as knowledge management, bioinformatics, e-learning, intelligent information integration [40], information brokering [41] and natural-language processing [42]. Now it is the positional and challenging area for text classification.

Semantic analysis is the process of linguistically parsing sentences and paragraphs into key concepts, verbs and proper nouns. Using statistics-backed technology, these words are then compared to taxonomy (categories) and grouped according to relevance [43]. Better classification will be performed when consider the semantic under consideration, so the semantically representation of text and web document is the key challenge for the documents classification and knowledge management. Recently many researchers addressed such types of issues.

The authors in [45] present the ambiguity issues in natural language text and present anew technique for resoling ambiguity problem in extracting concept/entity from the text which can improve the document classification process. Multilingual text representation and classification is on of the main and challenging issue in text classification.

In [37] the idea of workflow composition is presented, and addressed the important issues of semantic description of such as services for particular text mining task. Moreover, there are other two open problems in text mining: polysemy, synonymy. Polysemy refers to the fact that a word can have multiple meanings. Distinguishing between different meanings of a word (called word sense disambiguation) is not easy, often requiring the context in which the word appears. Synonymy means that different words can have the same or similar meaning. Some of the natural language issues that should be consider during the text mining process shown in overview [46] is listed below in Table-2.

TABLE 2. SEMANTIC ISSUES FOR DOCUMENTS CLASSIFICATION

Sentence Splitting	How we Identifying sentence boundaries in a document.
Tokenization	How the documents are tokenized and tokens are recorded or annotated, by word or phrase. This is important because many down stream components need the tokens to be clearly identified for analysis.
Part-of-Speech (pos) Tagging	What about the part of speech characteristics and the data annotation. How such components are assigning a pos tag to token pos information.
Stop word list	How stop word list will be taken, and which words are to consider as stop word in which domain.
Stemming	If we reduce the words to their stems, how it will affect the meaning of the documents.
Noisy Data	Which steps are required for the document to be clear from noisy data.
Word Sense	How we clarify the meaning of the word in the text, ambiguity problem.
Collocations	What about the compound and technical terms.
Syntax	How should make a syntactic or grammar analysis. What about data dependency, anaphoric problems.
Text Representation	Which will be more important for representation of the documents: Phrases, Word or Concept and Noun or adjective? And for this which techniques will be feasible to use.
Domain and data understanding for Ontology	How to define the area, data availability and its relation for ontology construction.

Semantically representation of documents is the challenging area for research in text mining. By proper implantation of this will be improve the classification and the information retrieval process.

III MACHINE LEARNING TECHNIQUES

The documents can be classified by three ways, unsupervised, supervised and semi supervised methods. Many techniques and algorithms are proposed recently for the clustering and classification of electronic documents. This section focused on the supervised classification techniques, new developments and highlighted some of the opportunities and challenges using the existing literature. The automatic classification of documents into predefined categories has observed as an active attention, as the internet usage rate has quickly enlarged. From last few years , the task of automatic text classification have been extensively studied and rapid progress seems in this area, including the machine learning approaches such as Bayesian classifier, Decision Tree, K-nearest neighbor(KNN), Support Vector Machines(SVMs), Neural Networks, Latent Semantic Analysis, Rocchio's Algorithm, Fuzzy Correlation and Genetic Algorithms etc. Normally supervised learning techniques are used for automatic text classification, where pre-defined category labels are assigned to documents based on the likelihood suggested by a training set of labelled documents. Some of these techniques are described below.

A. *Rocchio's Algorithm*

Rocchio's Algorithm [75] is a vector space method for document routing or filtering in informational retrieval, build prototype vector for each class using a training set of documents, i.e. the average vector over all training document vectors that belong to class ci, and calculate similarity between test document and each of prototype vectors, which assign test document to the class with maximum similarity.

$$C_i = \alpha * centroid_{C_i} - \beta * centroid_{\overline{C_i}} \qquad (1)$$

When given a category, the vector of documents belonging to this category is given a positive weight, and the vectors of remaining documents are given negative weight. The positively and negatively weighted vectors, the prototype vector of this category is obtained.

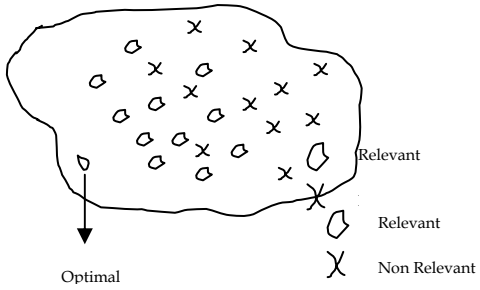

Fig .2 Rocchio Optimal query for separating relevant and non relevant documents

This algorithm [61] is easy to implement, efficient in computation, fast learner and have relevance feedback mechanism but low classification accuracy. Linear combination is too simple for classification and constant α and β are empirical. This is a widely used relevance feedback algorithm that operates in the vector space model [76]. The researchers have used a variation of Rocchio's algorithm in a machine learning context, i.e., for learning a user profile from unstructured text [77] [78], the goal in these applications is to automatically induce a text classifier that can distinguish between classes of documents.

B. *K-nearest neighbor (k-NN)*

The k-nearest neighbor algorithm (k-NN) [66] is used to test the degree of similarity between documents and k training data and to store a certain amount of classification data, thereby determining the category of test documents. This method is an instant-based learning algorithm that categorized objects based on closest feature space in the training set [62]. The training sets are mapped into multi-dimensional feature space. The feature space is partitioned into regions based on the category of the training set. A point in the feature space is assigned to a particular category if it is the most frequent category among the k nearest training data. Usually Euclidean Distance is typically used in computing the distance between the vectors. The key element of this method is the availability of a similarity measure for identifying neighbors of a particular document [62]. The training phase consists only of storing the feature vectors and categories of the training set. In the classification phase, distances from the new vector, representing an input document, to all stored vectors are computed and k closest samples are selected. The annotated category of a document is predicted based on the nearest point which has been assigned to a particular category.

$$\arg\max_i \sum_{j=1}^{k} sim(D_j \mid D) * \delta(C(D_j), i) \qquad (2)$$

Calculate similarity between test document and each neighbour, and assign test document to the class which contains most of the neighbors. Fig.3.

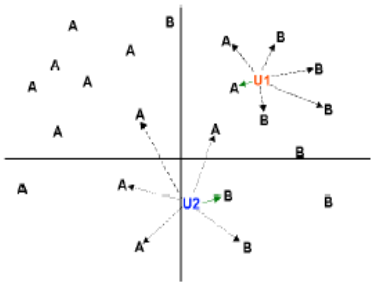

Fig. 3 k-Nearest Neighbor

This method is effective, non parametric and easy to implement. As compare to Rocchio algorithm more local characteristics of documents are considered, however the classification time is long and difficult to find optimal value of k. i.e., to analyze the k-NN and the Rocchio algorithm, some shortcomings of each are identified in [56]. A new algorithm is proposed in [67] which incorporating the relationship of concept-based thesauri into document categorization using a k-NN classifier, while [60] presents the use of phrases as basic features in the email classification problem and performed extensive empirical evaluation using large email collections and tested with three text classification algorithms, namely, a naive Bayes classifier and two k-NN classifiers using TF- IDF weighting and resemblance respectively. The k-nearest neighbor classification method is outstanding with its simplicity and is widely used techniques for text classification. This method performs well even in handling the classification tasks with multi-categorized documents. The major drawback of this method is it uses all features in distance computation, and causes the method computationally intensive, especially when the size of training set grows. Besides, the accuracy of k-nearest neighbor classification is severely degraded by the presence of noisy or irrelevant features.

C. *Decision Tree*

The decision tree rebuilds the manual categorization of training documents by constructing well-defined true/false-queries in the form of a tree structure. In a decision tree structure, leaves represent the corresponding category of documents and branches represent conjunctions of features that lead to those categories. The well-

organized decision tree can easily classify a document by putting it in the root node of the tree and let it run through the query structure until it reaches a certain leaf, which represents the goal for the classification of the document.

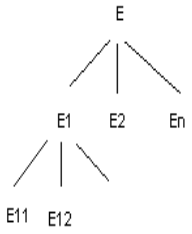

Fig. 4 Decision Tree

The decision tree classification method is outstanding from other decision support tools with several advantages. The main advantage of decision tree is its simplicity in understanding and interpreting, even for non-expert users. Besides, the explanation of a given result can be easily replicated by using simple mathematics algorithms, and provide a consolidated view of the classification logic, which is a useful information of classification.

It can be shown experimentally that text classification tasks frequently involve a large number of relevant features [79]. Therefore, a decision tree's tendency to base classifications on as few tests as possible can lead to poor performance on text classification. However, when there are a small number of structured attributes, the performance, simplicity and understandability of decision trees for content-based models are all advantages. The [80] describe an application of decision trees for personalizing advertisements on web pages.

The major risk of implementing a decision tree is it over fits the training data with the occurrence of an alternative tree that categorizes the training data worse but would categorize the documents to be categorized better [63]. This is due to the classification algorithm of decision tree is made to categorize training data effectively, however neglect the performance of classifying other documents. Besides, huge and excessively complex structure of tree is built from a dataset with very large number of entries.

D. Decision Rules Classification

Decision rules classification method uses the rule-based inference to classify documents to their annotated categories [64] [65]. The algorithms construct a rule set that describe the profile for each category. Rules are typically constructed in the format of "IF condition THEN conclusion", where the condition portion is filled by features of the category, and the conclusion portion is represented with the category's name or another rule to be tested. The rule set for a particular category is then constructed by combining every separate rule from the same category with logical operator, typically use "and" and "or". During the classification tasks, not necessarily every rule in the rule set needs to be satisfied. In the case of handling a dataset with large number of features for each category, heuristics implementation is recommended to reduce the size of rules set without affecting the performance of the classification. The [49] presents a hybrid method of rule-based processing and back-propagation neural networks for spam filtering, Instead of using keywords, this study utilize the spamming behaviours as features for describing emails.

The main advantage of the implementation of decision rules method for classification tasks is the construction of local dictionary for each individual category during the feature extraction phase [64]. Local dictionaries are able to distinguish the meaning of a particular word for different categories. However, the drawback of the decision rule method is the impossibility to assign a document to a category exclusively due to the rules from different rule sets is applicable to each other. Besides, the learning and updating of decision rule methods need extensive involvement of human experts to construct or update the rule sets. Like the decision trees classification method, the decision rules method does not work well when the number of distinguishing features is large.

E. Naïve Bayes Algorithm

Naïve Bayes classifier is a simple probabilistic classifier based on applying Bayes' Theorem with strong independence assumptions. A more descriptive term for the underlying probability model would be independent feature model. These independence assumptions of features make the features order is irrelevant and consequently that the present of one feature does not affect other features in classification tasks [99]. These assumptions make the computation of Bayesian classification approach more efficient, but this assumption severely limits its applicability. Depending on the precise nature of the probability model, the naïve Bayes classifiers can be trained very efficiently by requiring a relatively small amount of training data to estimate the parameters necessary for classification. Because independent variables are assumed, only the variances of the variables for each class need to be determined and not the entire covariance matrix.

Due to its apparently over-simplified assumptions, the naïve Bayes classifiers often work much better in many complex real-world situations than one might expect. The naïve Bayes classifiers has been reported to perform surprisingly well for many real world classification applications under some specific conditions [100] [101] [102] [103] [104].

An advantage of the naïve Bayes classifier is that it requires a small amount of training data to estimate the parameters necessary for classification. Bayesian classification approach arrives at the correct classification as long as the correct category is more probable than the others. Category's probabilities do not have to be estimated very well. In other words, the overall classifier is robust enough to ignore serious deficiencies in its underlying naïve probability model.

The main disadvantage of the naïve Bayes classification approach is its relatively low classification performance compare to other discriminative algorithms, such as the

SVM with its outperformed classification effectiveness. Therefore, many active researches have been carried out to clarify the reasons that the naïve Bayes classifier fails in classification tasks and enhance the traditional approaches by implementing some effective and efficient techniques [100] [102] [103] [104] [105].

$$P(c_i \mid D) = \frac{P(c_i)P(D \mid c_i)}{P(D)} \quad (4)$$

$$P(D \mid c_i) = \prod_{j=1}^{n} P(d_j \mid c_i) \quad (3)$$

Where P(Ci)= $P(C = c_i) = \frac{Ni}{N}$

and P(dj|ci) = $P(d_j \mid c_i) = \frac{1 + N_{ji}}{M + \sum_{k=1}^{M} N_{ki}}$

Naïve Bayes has been one of the popular machine learning methods for many years. Its simplicity makes the framework attractive in various tasks and reasonable performances are obtained in the tasks although this learning is based on an unrealistic independence assumption. For this reason, there also have been many interesting works of investigating naive Bayes. Recently the [83] shows very good results by selecting Naïve Bayes with SVM for text classification also the authors in [84] prove that Naive Bayes with SOM give very good results in clustering the documents. The authors in [85] propose a Poisson Naive Bayes text classification model with weight-enhancing method, and shows that the new model assumes that a document is generated by a multivariate Poisson model. They suggest per-document term frequency normalization to estimate the Poisson parameter, while the traditional multinomial classifier estimates its parameters by considering all the training documents as a unique huge training document. The [86] presented that naive Bayes can perform surprisingly well in the classification tasks where the probability itself calculated by the naive Bayes is not important. The authors in a review [87] described that researcher shows great interest in naïve Bayes classifier for spam filtering. So this technique is most widely used in email, web contents, and spam categorization.

Naive Bayes work well on numeric and textual data, easy to implement and computation comparing with other algorithms, however conditional independence assumption is violated by real-world data and perform very poorly when features are highly correlated and does not consider frequency of word occurrences.

F. Artificial Neural Network

Artificial neural networks are constructed from a large number of elements with an input fan order of magnitudes larger than in computational elements of traditional architectures [106] [107]. These elements, namely artificial neuron are interconnected into group using a mathematical model for information processing based on a connectionist approach to computation. The neural networks make their neuron sensitive to store item. It can be used for distortion tolerant storing of a large number of cases represented by high dimensional vectors.

Different types of neural network approaches have been implemented to document classification tasks. Some of the researches use the single-layer perceptron, which contains only an input layer and an output layer due to its simplicity of implementing [108]. Inputs are fed directly to the outputs via a series of weights. In this way it can be considered the simplest kind of feed-forward network. The multi-layer perceptron which is more sophisticated, which consists of an input layer, one or more hidden layers, and an output layer in its structure, also widely implemented for classification tasks [106].

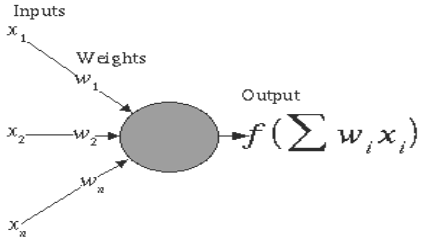

Fig. 5 Artificial Neural Network

The main advantage of the implementation of artificial neural network in classification tasks is the ability in handling documents with high-dimensional features, and documents with noisy and contradictory data. Furthermore, linear speed up in the matching process with respect of the large number of computational elements is provided by a computing architecture which is inherently parallel, where each element can compare its input value against the value of stored cases independently from others [107].

The drawback of the artificial neural networks is their high computing cost which consumes high CPU and physical memory usage. Another disadvantage is that the artificial neural networks are extremely difficult to understand for average users. This may negatively influence the acceptance of these methods.

In recent years, neural network has been applied in document classification systems to improve efficiency. Text categorization models using back-propagation neural network (BPNN) and modified back-propagation neural network (MBPNN) are proposed in [54] for documents classification. An efficient feature selection method is used to reduce the dimensionality as well as improve the performance. New Neural network based document classification method [68], was presented, which is helpful for companies to manage patent documents more effectively.

The ANN can get Inputs xi arrives through pre-synaptic connections, Synaptic efficacy is modelled using real

weights wi and the response of the neuron is a nonlinear function f of its weighted inputs.

The output from neuron j for pattern p is Opj where

$$O_{pj}(net_j) = \frac{1}{1 + e^{-\lambda net_j}} \quad (5)$$

and

$$net_j = bias * W_{bias} + \sum_k O_{pk} W_{jk} \quad (6)$$

Neural network for document classification produce good results in complex domains and suitable for both discrete and continuous data (especially better for the continuous domain). Testing is very fast however training is relatively slow and learned results are difficult for users to interpret than learned rules (comparing with Decision tree), Empirical Risk Minimization (ERM) makes ANN try to minimize training error, may lead to overfitting.

G. Fuzzy correlation

Fuzzy correlation can deal with fuzzy information or incomplete data, and also convert the property value into fuzzy sets for multiple document classification [69].

In [55] the authors explores the challenges of multi-class text categorization using one-against-one fuzzy support vector machine with Reuter's news as the example data, and shows better results using one-against-one fuzzy support vector machine as a new technique when compare with one-against-one support vector machine. [61] presented the improvement of decision rule and design a new algorithm of f-k-NN (fuzzy k-NN) to improve categorization performance when the class distribution is uneven, and show that the new method is more effective. So the researchers shows great interest recently to use the fuzzy rules and sets to improve the classification accuracy, by incorporating the fuzzy correlation or fuzzy logic with the machine learning algorithm and the feature selection methods to improve the classification process.

H. Genetic Algorithm

Genetic algorithm [81] aims to find optimum characteristic parameters using the mechanisms of genetic evolution and survival of the fittest in natural selection. Genetic algorithms make it possible to remove misleading judgments in the algorithms and improve the accuracy of document classification. This is an adaptive probability global optimization algorithm, which simulated in a natural environment of biological and genetic evolution, and is widely used for their simplicity and strength. Now several researchers used this method for the improvement of the text classification process. In authors in [82] introduced the genetic algorithm to text categorization and used to build and optimize the user template, and also introduced simulated annealing to improve the shortcomings of genetic algorithm. In the experimental analysis, they show that the improved method is feasible and effective for text classification.

I. Support Vector Machine (SVM)

Support vector machines (SVMs) are one of the discriminative classification methods which are commonly recognized to be more accurate. The SVM classification method is based on the Structural Risk Minimization principle from computational learning theory [109]. The idea of this principle is to find a hypothesis to guarantee the lowest true error. Besides, the SVM are well-founded that very open to theoretical understanding and analysis [110].

The SVM need both positive and negative training set which are uncommon for other classification methods. These positive and negative training set are needed for the SVM to seek for the decision surface that best separates the positive from the negative data in the n-dimensional space, so called the hyper plane. The document representatives which are closest to the decision surface are called the support vector. The performance of the SVM classification remains unchanged if documents that do not belong to the support vectors are removed from the set of training data [99].

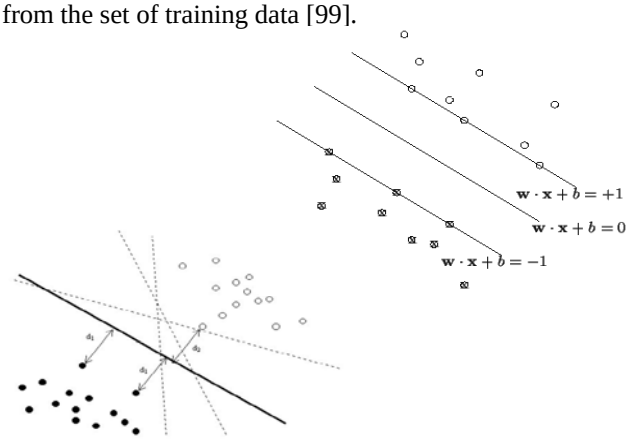

Fig. 6 Illustration of optimal separating hyper plane, hyper planes and support vectors

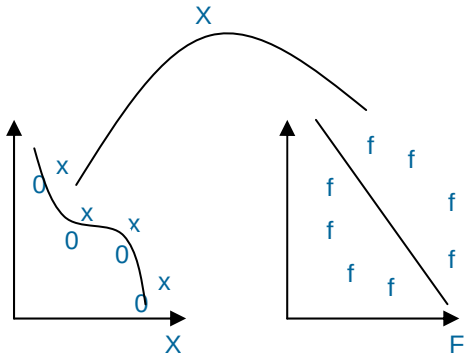

Fig. 7 Mapping non linear input space onto high dimensional space

The SVM classification method is outstanding from the others with its outstanding classification effectiveness [99] [111] [112] [110] [113] [70]. Furthermore, it can handle documents with high-dimensional input space, and culls out most of the irrelevant features. However, the major drawback of the SVM is their relatively complex training and categorizing algorithms and also the high time and memory consumptions during training stage and classifying stage. Besides, confusions occur during the classification tasks due to the documents could be a notated to several categories because of the similarity is typically calculated individually for each category [99].

So SVM is supervised learning method for classification to find out the linear separating hyperplane which maximize the margin, i.e., the optimal separating hyperplane (OSH) and maximizes the margin between the two data sets. To calculate the margin, two parallel hyperplanes are constructed, one on each side of the separating hyperplane, which are "pushed up against" the two data sets. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the neighboring data points of both classes, since in general the larger the margin the lower the generalization error of the classifier.

Maximizing the margin is equivalent to

$$\begin{aligned} \underset{w,b,\zeta_i}{\text{minimize}} \quad & \frac{1}{2} w^T w + C\left(\sum_{i=1}^{N} \zeta_i\right) \\ \text{subject to} \quad & y_i (w^T x_i - b) + \zeta_i - 1 \geq 0, \quad 1 \leq i \leq N \\ & \zeta_i \geq 0, \quad 1 \leq i \leq N \end{aligned} \tag{9}$$

Introducing Lagrange multipliers α, β , the Lagrangian is:

$$\begin{aligned} \ell(w, b, \zeta_i; \alpha, \beta) &= \frac{1}{2} w^T w + C \sum_{i=1}^{N} \zeta_i \\ &\quad - \sum_{i=1}^{N} \alpha_i [y_i (w^T x_i - b) + \zeta_i - 1] - \sum_{i=1}^{N} \mu_i \zeta_i \\ &= \frac{1}{2} w^T w + \sum_{i=1}^{N} (C - \alpha_i - \mu_i) \zeta_i \\ &\quad - \left(\sum_{i=1}^{N} \alpha_i y_i x_i^T \right) w - \left(\sum_{i=1}^{N} \alpha_i y_i \right) b + \sum_{i=1}^{N} \alpha_i \end{aligned} \tag{10}$$

The authors in [50] implemented and measured the performance of the leading supervised and unsupervised approaches for multilingual text categorization; they selected support vector machines (SVM) as representative of supervised techniques as well as latent semantic indexing (LSI) and self-organizing maps (SOM) techniques for unsupervised methods for system implementation. In [52] the authors analyses and compares SVM ensembles with four different ensemble constructing techniques, namely bagging, AdaBoost, Arc-X4 and a modified AdaBoost. Twenty real-world data sets from the UCI repository are used as benchmarks to evaluate and compare the performance of these SVM ensemble classifiers by their classification accuracy.

An optimal SVM algorithm via multiple optimal strategies is developed in [47], such as a novel importance weight definition, the feature selection using the entropy weighting scheme, the optimal parameter settings. The SVM is a best technique for the documents classification [83].

IV HYBRID TECHNIQUES

Many new hybrid methods and techniques are proposed recently in the area of Machine Learning and text mining. The concept of combining classifiers is proposed as a new direction for the improvement of the performance of individual classifiers. Recently many methods have been suggested for the creation of ensemble of classifiers. Mechanisms that are used to build ensemble of classifiers[114] include: i) Using different subset of training data with a single learning method, ii) Using different training parameters with a single training method (e.g. using different initial weights for each neural network in an ensemble) and iii) Using different learning methods.[88].

The benefits of local versus global feature sets and local versus global dictionaries in text categorization have examined in [121]. Local features are class dependent features while global features are class independent features. Local dictionaries are class dependent dictionaries while global dictionaries are class independent dictionaries. The best text categorization is obtained using local features and local dictionaries [121].

A New hybrid text document classification approach is proposed in [83], used naive Bayes method at the front end for raw text data vectorization, in conjunction with a SVM classifier at the back end to classify the documents to the right category. They shows that the proposed hybrid approach of the Naive Bayes vectorizer and SVM classifier has improved classification accuracy compared to the pure naive Bayes classification approach. The [84] presents another hybrid method of naïve Bayes with self organizing map (SOM). Proposed Bayes classifier used at the front end, while SOM performs the indexing steps to retrieve the best match cases.

So In the context of combining multiple classifiers for text categorization, a number of researchers have shown that combining different classifiers can improve classification accuracy [89]. It is observed from the Comparison between the best individual classifier and the combined method, that the performance of the combined method is superior [90] [91] [92].

A hybrid method is proposed in [93] in which the learning phase evaluation back propagation neural network (LPEBP) to improve the traditional BPNN. And adopt singular value decomposition (SVD) technique to reduce the dimension and construct the latent semantics between terms, and show that the LPEBP is much faster than the traditional BPNN, which enhances the performance of the traditional BPNN. The SVD technique cannot only greatly reduce the high dimensionality but also enhance the performance. So SVD is to further improve the document classification systems precisely and efficiently.

The [94] a new hybrid technique for text classification is proposed that requires less training data and less computational time and show that text classification that requires fewer documents for training instead of using words, word relation i.e. association rules from these words is used to derive feature set from pre-classified text documents. The concept of Naïve Bayes classifier is then used on derived features and finally only a single concept of genetic presented has been added for final classification.

In[55] the authors explores the challenges of multi-class text categorization using one-against-one fuzzy support vector machine with reuter's news as the example data, and shows better results using one-against-one fuzzy support vector machine as a new technique when compare with one-against-one support vector machine.

A hybrid algorithm is proposed in [56], based on variable precision rough set to combine the strength of both k-NN and Rocchio techniques to improve the text classification accuracy and overcome the weaknesses of Rocchio algorithm.

The authors in [95] suggest a new hybrid approach to web document classification built upon both, graph and vector representations. K-NN algorithm shows that the proposed graph and vector approaches performing better in terms of classification accuracy along with a significant reduction in classification time.

The [96] proposed two methods to modify the standard BPNN and adopt the semantic feature space (SFS) method to reduce the number of dimensions as well as construct latent semantics between terms, and show that the modified methods enhanced the performance of the standard BPNN and were more efficient than the standard BPNN. The SFS method cannot only greatly reduce the dimensionality, but also enhances performance and can therefore be used to further improve text classification systems precisely and efficiently.

The [97] presents a semi-supervised learning method SSRANK for classification task. It leverages the uses of both labelled data and unlabeled data, utilizes views from both traditional IR and supervised learning to conduct data labelling, and relies on a criterion to control the process of data labelling.

A new algorithm of f-k-NN (fuzzy k-NN) proposed in [98] for the improvement of decision rule and design to improve classification performance when the class distribution is uneven, and show that the new method is more effective. The approach of [48] is a nontrivial extension of document classification methodology from a fixed set of classes to a knowledge hierarchy like Gene Ontology.

In [51], the authors proposed a new approach to automatic discovery of implicit rhetorical information from texts based on evolutionary computation methods. In order to guide the search for rhetorical connections from natural-language texts. And in [53], the authors present a segmentation methodology of handwritten documents in their distinct entities, namely, text lines and words.

In [120] the combination of similarity-based learning algorithms and associated thresholding strategies significantly influences the overall performance of text classification. After investigating two similarity-based classifiers (k-NN and Rocchio) and three common thresholding techniques (RCut, PCut, and SCut), they described a new learning algorithm known as the keyword association network (KAN) and a new thresholding strategy (RinSCut) to improve performance over existing techniques, and shows that the new approaches give better results.

A new machine learning method is proposed for constructing ranking models in document retrieval [57]. The method, aims to use the advantages of both the traditional Information Retrieval (IR) methods and the supervised learning methods for IR proposed recently.

The main concern of authors in [58] is to investigate the effectiveness of using multi-words for text representation on the performances of text classification. Firstly, a practical method is proposed to implement the multi-word extraction from documents based on the syntactical structure. Secondly, two strategies as general concept representation and subtopic representation are presented to represent the documents using the extracted multi-words. The proposed method launches in [59] for text classification tasks with only unlabeled documents and the title word of each category for learning, and then it automatically learns text classifier by using bootstrapping and feature projection techniques.

V COMPARATIVE STUDY

The growing phenomenon of the textual data needs text mining, machine learning and natural language processing techniques and methodologies to organize and extract pattern and knowledge from the documents. This review focused on the existing literature and explored the documents representation and classification techniques. Text representation is a crucial issue. Most of the literature gives the statistical of syntactic solution for the text representation. However the representation model depend on the informational that we require. Concept base or semantically representations of documents require more attention.

The performance of a classification algorithm in data mining is greatly affected by the quality of data source. Irrelevant and redundant features of data not only increase the cost of mining process, but also degrade the quality of the result in some cases [71]. Each algorithm has its own advantages and disadvantages as described in section II and III.

However, in [6] the author compare the different text classification techniques and have to bear in mind that comparisons are reliable only when based on experiments performed by the same author under carefully controlled conditions. They are instead more problematic when they involve different experiments performed by different authors. In this case various "background conditions," often extraneous to the learning algorithm itself may influence the results. These may include, among others, different

choices in pre-processing (stemming, etc.), indexing, dimensionality reduction and classifier parameter values etc.

A performance compression in [115] presented a controlled study on a large number of filter feature selection methods for text classification. Over 100 variants of five major feature selection criteria were examined using four well-known classification algorithms: Naive Bayesian (NB) approach, Rocchio-style classifier, k-NN method and SVM system. Two benchmark collections were chosen as the testbeds: Reuters-21578 and small portion of Reuters Corpus Version 1 (RCV1), making the new results comparable to published results. They presents that feature selection methods based on $\chi 2$ statistics consistently outperformed those based on other criteria (including information gain) for all four classifiers and both data collections, and that a further increase in performance was obtained by combining uncorrelated and high-performing feature selection methods. The results they obtained using only 3% of the available features are among the best reported, including results obtained with the full feature set. The empirical results of their study suggest that using filter methods which include the $\chi 2$ statistic, combining them with DF or IG, and eliminating the rare words. Such methods were consistently better.

In [116] the authors discussed, that some studies compared feature selection techniques or feature space transformation whereas some others compared the performance of different algorithms. Recently the rising interest towards the Support Vector Machine, various studies showed that SVM outperforms then other classification algorithms. So should we just not problem about other classification algorithms and opt always for SVM? They have decided to investigate this issue and compared SVM to k-NN and naive Bayes on binary classification tasks. An important issue is to compare optimized versions of these algorithms; from their results it shows all the classifiers achieved comparable performance on most problems. One surprising result is that SVM was not a clear winner, despite quite good overall performance. If a suitable pre-processing is used with k-NN, this algorithm continues to achieve very good results and scales up well with the number of documents, which is not the case for SVM. As for Naive Bayes, it also achieved good performance.

The [117] deals with the performance of different classification algorithms and the impact of feature selection algorithm on Logistic Regression Classifier, How it controls False Discovery Rate (FDR) and thus improves the efficiency of Logistic Regression classifier. As per the analysis support vector machine has more parameters than logistics regression and decision tree classifier, SVM has the highest classification precision most of the time, however SVM is very time consuming because of more parameters, demands more computation time. Compared to SVM, logistic regression is computationally efficient. Its results usually have static meaning. However it does not perform well when data set exhibits explicit data structures.

In [118] compression on four machine learning algorithms, which are Naive Bayesian (NB), neural network (NN), support vector machine (SVM) and relevance vector machine (RVM), are proposed for spam classification. An empirical evaluation for them on the benchmark spam filtering corpora is presented. The experiments are performed based on different training set size and extracted feature size. Experimental results show that NN classifier is unsuitable for using alone as a spam rejection tool. Generally, the performances of SVM and RVM classifiers are obviously superior to NB classifier. Compared with SVM, RVM is shown to provide the similar classification result with less relevance vectors and much faster testing time despite the slower learning procedure, they show that RVM is more suitable than SVM for spam classification in terms of the applications that require low complexity.

In [119] email data was classified using four different classifiers (Neural Network, SVM classifier, Naïve Bayesian Classifier, and J48 classifier). The experiment was performed based on different data size and different feature size. The final classification result should be '1' if it is finally spam, otherwise, it should be '0'. This paper shows that simple J48 classifier which make a binary tree, could be efficient for the dataset which could be classified as binary tree.

The [120] shows that two main research areas in statistical text categorization are: similarity-based learning algorithms and associated thresholding strategies. The combination of these techniques significantly influences the overall performance of text categorization. After investigating two similarity-based classifiers (k-NN and Rocchio) and three common thresholding techniques (RCut, PCut, and SCut), they described a new learning algorithm known as the keyword association network (KAN) and a new thresholding strategy (RinSCut) to improve performance over existing techniques. Extensive experiments have been conducted on the Reuters-21578 and 20-Newsgroups data sets, and shows that the new approaches give better results.

Comparing with ANN, SVM capture the inherent characteristics of the data better and embedding the Structural Risk Minimization (SRM) principle which minimizes the upper bound on the generalization error (better than the Empirical Risk Minimization principle) also ability to learn can be independent of the dimensionality of the feature space and global minima vs. local minima, However there are some difficulties in parameter tuning and kernel selection.

VI Discussion and Conclusions

This paper provides a review of machine learning approaches and documents representation techniques. An analysis of feature selection methods and classification algorithms were presented. It was verified from the study that information Gain and Chi square statistics are the most commonly used and well performed methods for feature selection, however many other FS methods are

proposed as single or hybrid technique recently, shown good results, and needs more exploration for efficient classification process. Several algorithms or combination of algorithms as hybrid approaches was proposed for the automatic classification of documents, among these algorithms, SVM, NB and kNN classifiers are shown most appropriate in the existing literature.

Most researchers in text classification assume the documents representation as a Bag of Word (BOG), although according to [44] the statistical techniques are not sufficient for the text mining. Text representation is a crucial issue. Most of the literature gives the statistical of syntactic solution for the text representation. However the representation model depend on the informational that we require. Concept base or semantically representation of documents requires more research. Better classification will be performed when consider the semantic under considerations, semantically and ontology base documents representation opportunities were discussed in this paper. With the addition of the ontology and semantic to represent the documents will be more improve accuracy and the classification process. So the identification of features that capture semantic content is one of the important areas for research. The general multiple learning issues in the presence of noise is a tremendously challenging problem that is just now being formulated and will likely require more work in order to successfully develop strategies to find the underlying nature of the manifold.

Several algorithms or combination of algorithms as hybrid approaches were proposed for the automatics classification of documents. Among these algorithms, SVM, NB , kNN and their hybrid system with the combination of different other algorithms and feature selection techniques are shown most appropriate in the existing literature. However the NB is perform well in spam filtering and email categorization, requires a small amount of training data to estimate the parameters necessary for classification. Naive Bayes works well on numeric and textual data, easy to implement comparing with other algorithms, however conditional independence assumption is violated by real-world data and perform very poorly when features are highly correlated and does not consider frequency of word occurrences.

SVM classifier has been recognized as one of the most effective text classification method in the comparisons of supervised machine learning algorithms [74]. SVM capture the inherent characteristics of the data better and embedding the Structural Risk Minimization (SRM) principle which minimizes the upper bound on the generalization error (better than the Empirical Risk Minimization principle) also ability to learn can be independent of the dimensionality of the feature space and global minima vs. local minima, however, the SVM has been found some difficulties in parameter tuning and kernel selection.

If a suitable pre-processing is used with k-NN, then this algorithm continues to achieve very good results and scales up well with the number of documents, which is not the case for SVM [122] [123]. As for naive Bayes, it also achieved good performance with suitable pre-processing. k-NN algorithm performed well as more local characteristic of documents are considered, however the classification time is long and difficult to find optimal value of k.

More works are required for the performance improvement and accuracy of the documents classification process. New methods and solutions are required for useful knowledge from the increasing volume of electronics documents. The following are the some of opportunities of the unstructured data classification and knowledge discovery.

- To improve and explore the feature selection methods for better classification process.
- To reduce the training and testing time of classifier and improve the classification accuracy, precision and recall.
- For Spam filtering and e-mail categorization the user may have folders like electronic bills, e-mail from family, friends and so on, and may want a classifier to classify each incoming e-mail that's automatically move it to the appropriate folder. It is easier to find messages in sorted folders in a very large inbox.
- Automatic allocation of folders to the downloaded articles, documents from text editors and from grid network.
- The use of semantics and ontology for the documents classification and informational retrieval.
- Mining trend, i.e. marketing, business, and financial trend (stock exchange trend) form e-documents (Online news, stories, views and events).
- Stream text quire some new techniques and methods for information management.
- Automatic classification and analysis of sentiment, views and extraction knowledge from it. The sentiments and opinion mining is the new active area of text mining.
- Classification and clustering of semi-structured documents have some challenges and new opportunities.
- An implementation of sense-based text classification procedure is needed for recovering the senses from the words used in a specific context.
- Informational extraction of useful knowledge from e-documents and Web pages, such as products and search results to get meaning full patterns.
- To identify or match semantically similar data from the web (that contain huge amount of data and each website represents similar information differently) is an important problem with many practical applications. So web information, integration and schema matching needs more exploration.

REFERENCES

[1] A. Dasgupta, “Feature selection methods for text classification.”, In Proceedings of the 13th ACMSIGKDD international conference on Knowledge discovery and data mining, pp. 230 -239, 2007.

[2] Raghavan, P., S. Amer-Yahia and L. Gravano eds., “Structure in Text: Extraction and Exploitation.” In. Proceeding of the 7th international Workshop on the Web and Databases(WebDB), ACM SIGMOD/PODS 2004, ACM Press, Vol 67, 2004.

[3] Oracle corporation, WWW,oracle.com, 2008.

[4] Merrill lynch, Nov.,2000. e-Business Analytics: Depth Report. 2000.

[5] Pegah Falinouss “Stock Trend Prediction using News Article’s: a text mining approach” Master thesis -2007.

[6] Sebastiani, F., “Machine learning in automated text categorization” ACM Computing Surveys (CSUR) 34, pp.1 – 47, 2002.

[7] Andreas Hotho “A Brief Survey of Text Mining” 2005.

[8] Shang, W., Huang, H., Zhu, H., Lin, Y., Qu, Y., and Wang Z., “ A Noval Feature Selection Algorithm for text catogorization.” Elsevier, science Direct Expert system with application -2006, 33(1), pp.1-5, 2006.

[9] Montanes,E., Ferandez, J., Diaz, I., Combarro, E.F and Ranilla, J., “ Measures of Rule Quality for Feature Selection in Text Categorization”, 5th international Symposium on Intelligent data analysis , Germeny-2003, Springer-Verlag 2003, Vol2810, pp.589-598, 2003.

[10] Wang, Y., and Wang X.J., “ A New Approach to feature selection in Text Classification”, Proceedings of 4th International Conference on Machine Learning and Cybernetics, IEEE- 2005, Vol.6, pp. 3814-3819, 2005.

[11] Liu, H. and Motoda, ., “Feature Extraction, constraction and selection: A Data Mining Perpective.”, Boston, Massachusetts(MA): Kluwer Academic Publishers.

[12] Lee, L.W., and Chen, S.M., “New Methods for Text CategorizationBased on a New Feature Selection Method a and New Similarity Measure Between Documents”, IEA/AEI, France 2006.

[13] Manomaisupat, P., and Abmad k., “ Feature Selection for text Categorization Using Self Orgnizing Map”, 2nd International Conference on Neural Network and Brain, 2005,IEEE press Vol 3, pp.1875-1880, 2005.

[14] Yan, J., Liu, N., Zhang, B., Yan, S., Chen, Z., Cheng, Q., Fan, W., and Ma, W., “OCFS: Optimal Orthogonal centroid Feature selection for Text Categorization.” 28 Annual International conference on Reserch and Informational reterival, ACM SIGIR, Barizal, , pp.122-129, 2005.

[15] Zi-Qiang Wang, Xia Sun, De-Xian Zhang, Xin Li “An Optimal Svm-Based Text Classification Algorithm” Fifth International Conference on Machine Learning and Cybernetics, Dalian,pp. 13-16 , 2006.

[16] Jingnian Chen a,b,, Houkuan Huang a, Shengfeng Tian a, Youli Qua Feature selection for text classification with Naïve Bayes” Expert Systems with Applications 36, pp. 5432–5435, 2009.

[17] Hiroshi Ogura, Hiromi Amano, Masato Kondo “Feature selection with a measure of deviations from Poisson in text categorization” Expert Systems with Applications 36, -pp 6826–6832, 2009.

[18] Mehdi Hosseinzadeh Aghdam, Nasser Ghasem-Aghaee, Mohammad Ehsan Basiri “Text feature selection using ant colony optimization”, Expert Systems with Applications 36 pp.6843–6853, 2009.

[19] P. Sccuy, G.W.Mineanu “Beyoned TFIDF weighting for text Categorization in the Vector Space Model”, 2003.

[20] E. Youn, M. K. Jeong , “Class dependent feature scaling method using naive Bayes classifier for text datamining” Pattern Recognition Letters , 2009.

[21] G. Forman, E. Kirshenbaum, “Extremely Fast Text Feature Extraction for Classification and Indexing”, Napa Valley California, USA. CIKM’08, October 26–30, 2008

[22] Mostafa Keikha, Ahmad Khonsari, Farhad Oroumchian, “ Rich document representation and classification: An analysis” , Knowledge-Based Systems 22 , pp.67–71, 2009.

[23] D.Fensel, “Ontologies: Silver Bullet for Knowledge Management and e-Commerce”, Springer Verlag, Berlin, 2000.

[24] B. Omelayenko., “learning og ontologies for the Web: the analysis of existent approaches”, in the proceeding of the International Workshop on Web Dynamics, 2001.

[25] OWL Web Ontology Language, viewed March 2008 http://www.w3.org/TR/owl-features.

[26] Sean B. Palmer, “The Semantic Web, an introduction”, 2007.

[27] Lena Tenenboim, Bracha Shapira, Peretz Shoval “Ontology-Based Classification Of News In An Electronic Newspaper” International Conference "Intelligent Information and Engineering Systems" INFOS 2008, Varna, Bulgaria, June-July 2008.

[28] Lewis, D.D., “Naive (Bayes) at forty The independence assumption in information retrieval”, ECML-98, 10th European Conference on Machine Learning, Chemnitz, DE - 1998.

[29] A. Rafael Calvo, Jae-Moon Lee, Xiabo Li, ‘Managin content with automatic document classification”, Journal of Digital Information, 5(2) , Article No.282,2004.

[30] S. Chakrabarti, S. Roy, M. Soundalgekar, “Fast and accurate text classification via multiple linear discriminant projections”, International Journal on Very Large Data Bases 12 (2), pp.170–185, 2003.

[31] Deerwester, S., Dumais, S., Furnas, G.W., Landauer, T.K., Harshman, R; “Indexing by Latent Semantic Analysis”. Journal of the Society for Information Science -1990 41 pp. 391-407, 1990.

[32] Mu-Hee Song, Soo-Yeon Lim, Dong-Jin Kang, and Sang-Jo Lee, "Automatic Classification of Web pages based on the Concept of Domain Ontology", Proc. of the 12th Asia-Pacific Software Engineering Conference, 2005.

[33] Guiyi Wei, Jun Yu, Yun Ling, and Jun Liu, "Design and Implementation of an Ontology Algorithm for Web Documents Classification", ICCSA 2006, LNCS 3983, pp. 649-658. 2006.

[34] Jun Fang, Lei Guo, XiaoDong Wang and Ning Yang “Ontology-Based Automatic Classification and Ranking for Web Documents” Fourth International Conference on Fuzzy Systems and Knowledge Discovery -FSKD -2007.

[35] Alexander Maedche and Ste_en Staab “Mining Ontologies from Text” LNAI 1937, pp. 189-202, 2000. Springer-Verlag Berlin Heidelberg, 2000.

[36] Ching Kang Cheng, Xiao Shan Pan, Franz Kurfess “Ontology-based Semantic Classification of Unstructured Documents”, 2000.

[37] M. Sarnovský, M. Parali “Text Mining Workflows Construction with Support of Ontologies” 6th International Symposium on Applied Machine Intelligence and Informatics- SAMI 2008.

[38] Maciej Janik and Krys Kochut “Training-less Ontology-based Text Categorization” , 2007.

[39] Yi-Hsing Chang, Hsiu-Yi Huang “An Automatic Document Classifier System Based On Naïve Bayes Classifier And Ontology” Seventh International Conference on Machine Learning and Cybernetics, Kunming, July 2008.

[40] G. Wiederhold and M. Genesereth, “The conceptual basis for mediation services”,IEEE Expert / Intelligent Systems, 12(5):38-47, 1997.
[41] S. Staab, J. Angele, S. Decker, M. Erdmann, A. Hotho, A. Maedche, H.-P. Schnurr, R. Studer, and Y. Sure. “Semantic community web portals”, In Proceedings of the 9th International World Wide Web Conference, Amsterdam, The Netherlands, May, 15-19, 2000. Elsevier, 2000.
[42] S. Staab, C. Braun, I. Bruder, A. D¨usterh¨oft, A. Heuer, M. Klettke, G. Neumann, B. Prager, J. Pretzel, H.-P. Schnurr, R. Studer, H. Uszkoreit, and B. Wrenger. Getess, “Searching the web exploiting german texts”, In Proceedings of the 3rd international Workshop on Cooperating Information Agents. Upsala, Sweden, 1999, LNAI 1652, pp. 113-124. Springer, 1999.
[43] http://www.nstein.com/en/tme_intro.php- 2008.
[44] Yah, A.s., Hirschman, L., and Morgan, A.A. “Evaluation of text data mining for databasecuration: lessons learned from the KDD challenge cup.” Bioinformatics 19-(supp.1), pp.i331-i339, 2003.
[45] H.M.Al Fawareh, S.Jusoh, W.R.S.Osman, “Ambiguity in Text Mining”, IEEE-2008.
[46] A.Stavrianou, P. Andritsos, N. Nicoloyannis “Overview and semantic issues of text mining”, SIGMOD Record, 2007, Vol.36,N03, 2007.
[47] Zi-Qiang Wang, Xia Sun, De-Xian Zhang, Xin Li “An Optimal Svm-Based Text Classification Algorithm” Fifth International Conference on Machine Learning and Cybernetics, Dalian, 2006.
[48] H.Kim, and S.S. Chen, “Associative Naïve Bayes Classifier: Automated Linking Of Gene Ontology To Medline Documents” Pattern Recognition doi:10.1016/j.patcog. 2009
[49] Chih-Hung Wu , “Behavior-based spam detection using a hybrid method of rule-based techniques and neural networks”, Expert Systems with Applications, pp. 4321–4330, 2009
[50] Chung-Hong Lee a,, Hsin-Chang Yang , “Construction of supervised and unsupervised learning systems for multilingual text categorization” , Expert Systems with Applications, pp. 2400–2410, 2009.
[51] John Atkinson a,, Anita Ferreira b, Elvis Aravena , “Discovering implicit intention-level knowledge from natural-language texts”, Knowledge-Based Systems -2009.
[52] Shi-jin Wang, Avin Mathew, Yan Chen , Li-feng Xi , Lin Ma, Jay Lee, “Empirical analysis of support vector machine ensemble classifiers”, Expert Systems with Applications, pp. 6466–6476, 2009.
[53] G. Louloudis, B. Gatos, I. Pratikakis2, C. Halatsis , “Text Line and Word Segmentation of Handwritten Documents”, Pattern Recognition doi:10.1016/j.patcog.2008.12.016 ,2009
[54] Bo Yu, Zong-ben Xu, Cheng-hua Li ,“Latent semantic analysis for text categorization using neural network”, Knowledge-Based Systems 21- pp. 900–904, 2008.
[55] Tai-Yue Wang and Huei-Min Chiang “One-Against-One Fuzzy Support Vector Machine Classifier: An Approach to Text Categorization”, Expert Systems with Applications, doi: 10.1016/j.eswa.2009.
[56] [Duoqian Miao , Qiguo Duan, Hongyun Zhang, Na Jiao, “Rough set based hybrid algorithm for text classification”, Expert Systems with Applications -2009 .
[57] Ming Li, Hang Li , Zhi-Hua Zhou “Semi-supervised document retrieval” Information Processing and Management - 2008 .
[58] Wen Zhang a, Taketoshi Yoshida a, Xijin Tang “Text classification based on multi-word with support vector machine” , Knowledge-Based Systems 21 -pp. 879–886, 2008
[59] Youngjoong Ko a, Jungyun Seo, “Text classification from unlabeled documents with bootstrapping and feature projection techniques”, Information Processing and Management 45 -,pp. 70–83, 2009
[60] Matthew Changa, Chung Keung Poon_, “Using Phrases as Features in Email Classification”, The Journal of Systems and Software ,doi: 10.1016/j.jss, 2009.
[61] Willian W. Cohen and Yoram Singer, “Context-sensitive learning method for text categorization”, SIGIR’ 96, 19th International Conference on Research and Develeoement in Informational Retrieval, pp-307-315, 1996.
[62] Eui-Hong (Sam) Han, George Karypis, Vipin Kumar; “Text Categorization Using Weighted Adjusted k-Nearest Neighbor Classification”, Department of Computer Science and Engineering. Army HPC Research Centre, University of Minnesota, Minneapolis, USA. 1999.
[63] Russell Greiner, Jonathan Schaffer; AIxploratorium - Decision Trees, Department of Computing Science, University of Alberta,Edmonton,ABT6G2H1, Canada.2001. URL :http://www.cs.ualberta.ca/ ~aixplore/ learning/ DecisionTrees
[64] Chidanand Apte, Fred Damerau, Sholom M. Weiss.; “Towards Language Independent Automated Learning of Text Categorization Models”, In Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, pp. 23-30. 1994.
[65] Chidanand Apte, Fred Damerau, Sholom M. Weiss; “Automated Learning of Decision Rules for Text Categorization”, ACM Transactions on Information Systems (TOIS), Vol. 12 , Issue 3, pp. 233 – 251. 1994.
[66] Tam, V., Santoso, A., & Setiono, R. , “A comparative study of centroid-based, neighborhood-based and statistical approaches for effective document categorization”, Proceedings of the 16th International Conference on Pattern Recognition, pp.235–238, 2002.
[67] Bang, S. L., Yang, J. D., & Yang, H. J. , “Hierarchical document categorization with k-NN and concept-based thesauri. Information Processing and Management”, pp. 397–406, 2006.
[68] Trappey, A. J. C., Hsu, F.-C., Trappey, C. V., & Lin, C.-I., “Development of a patent document classification and search platform using a back-propagation network”, Expert Systems with Applications, pp. 755–765, 2006 .
[69] Que, H. -E. “Applications of fuzzy correlation on multiple document classification.Unpublished master thesis”, Information Engineering epartment, Tamkang University, Taipei, Taiwan-2000.
[70] YiMing Yang, Xin Liu; “A Re-examination of Text Categorization Methods, School of Computer Science”, Carnegie Mellon University. 1999.
[71] Wu W, Gao Q, Wang M “An efficient feature selectionmethod for classification data mining” WSEAS Transactions on Information Science and Applications,3: pp 2034-2040. 2006.
[72] Y.Yang; “An evaluation of statistical approaches to text categorization”, Information Retrieval, Vol.1, No.1, pp. 69-90, 1999.
[73] T.H.Ng, W.B.Goh, and K.L.Low, “Feature selection, perception learning and a usability case study for text categorization”, Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Philadelphia, pp.67-73, 1997.
[74] Y.Yang,and X.Liu, “An re-examination of text categorization”, Proceedings of the 22nd Annual International ACM

SIGIR Conference on Research and Development in Information Retrieval, Berkeley, pp.42-49,August 1999.
[75] Rocchio, J; "Relevance Feedback in Information Retrieval", In G. Salton (ed.). The SMART System: pp.67-88.
[76] Ittner, D., Lewis, D., Ahn, D; "Text Categorization of Low Quality Images", In: Symposium on Document Analysis and Information Retrieval, Las Vegas, NV .pp. 301-315, 1995
[77] Balabanovic, M., Shoham Y.: FAB; "Content-based, Collaborative Recommendation", Communications of the Association for Computing Machinery 40(3) pp. 66-72, 1997.
[78] Pazzani M., Billsus, D; " Learning and Revising User Profiles", The Identification of Interesting Web Sites. Machine Learning 27(3) pp. 313-331, 1997.
[79] Joachims, T; "Text Categorization With Support Vector Machines: Learning with Many Relevant Features", In: European Conference on Machine Learning, Chemnitz, Germany 1998, pp.137-142 , 1998.
[80] Kim, J., Lee, B., Shaw, M., Chang, H., Nelson, W; "Application of Decision-Tree Induction Techniques to Personalized Advertisements on Internet Storefronts", International Journal of Electronic Commerce 5(3) pp.45-62, 2001.
[81] Wang Xiaoping, Li-Ming Cao. Genetic Algorithm Theory, Application and Software[M].XI'AN:Xi'an Jiaotong University Press, 2002.
[82] ZHU Zhen-fang, LIU Pei-yu, Lu Ran, "Research of text classification technology based on genetic annealing algorithm" IEEE,, 978-0-7695-3311-7/08, 2008.
[83] Dino Isa, Lam Hong lee, V. P Kallimani, R. RajKumar, " Text Documents Preprocessing with the Bahes Formula for Classification using the Support vector machine", IEEE, Traction of Knowledge and Data Engineering, Vol-20, N0-9 pp-1264-1272, 2008.
[84] Dino Isa,, V. P Kallimani Lam Hong lee, "Using Self Organizing Map for Clustering of Text Documents", ", Elsever , Expert System with Applications-2008.
[85] Sang-Bum Kim, Kyoung-Soo Han, Hae-Chang Rim, and Sung Hyon Myaeng, "Some Effective Techniques for Naive Bayes Text Classification", IEEE Transactions On Knowledge And Data Engineering, Vol. 18, No. 11, , Pp-1457- 1466 ,November 2006.
[86] P. Domingos and M. J. Pazzani, "On the Optimality of the Simple Bayesian Classifier under Zero-One Loss," Machine Learning, vol. 29, nos. 2/3, pp. 103-130, 1997.
[87] Thiago S.Guzella, Walimir M. Caminhas " A Review of machine Learning Approches to Spam Filtering", Elsever , Expert System with Applications-2009.
[88] M. Ikonomakis, S. Kotsiantis, V. Tampakas, "Text Classification Using Machine Learning Techniques", Wseas Transactions on Computers, issue 8, volume 4, pp. 966-974, 2005.
[89] Bao Y. and Ishii N., "Combining Multiple kNN Classifiers for Text Categorization by Reducts", LNCS 2534, , pp. 340- 347, 2002.
[90] Bi Y., Bell D., Wang H., Guo G., Greer K., "Combining Multiple Classifiers Using Dempster's Rule of Combination for Text Categorization", MDAI, 2004, 127-138, 2004.
[91] Sung-Bae Cho, Jee-Haeng Lee, "Learning Neural Network Ensemble for Practical TextClassification", Lecture Notes in Computer Science, Volume 2690, Pages 1032– 1036, 2003.
[92] Nardiello P., Sebastiani F., Sperduti A., "Discretizing Continuous Attributes in AdaBoost for Text Categorization", LNCS, Volume 2633, , pp. 320-334, 2003
[93] "Cheng Hua Li , Soon Choel Park, "An efficient document classification model using an improved back propagation neural network and singular value decomposition" Expert Systems with Applications 36 ,pp- 3208–3215, 2009.
[94] S. M. Kamruzzaman and Farhana Haider; "Hybrid Learning Algorithm For Text Classification", 3rd International Conference on Electrical & Computer Engineering ICECE 2004, 28-30 December 2004, Dhaka, Bangladesh.
[95] Alex Markov and Mark Last, "A Simple, Structure-Sensitive Approach for Web Document Classification", Springer, AWIC 2005, LNAI 3528, pp. 293–298, 2005.
[96] Cheng Hua Li, Soon Cheol Park , "Combination of modified BPNN algorithms and an efficient feature selection method for text categorization.", Information Processing and Management 45, 329–340, 20009.
[97] Ming Li , Hang Li , Zhi-Hua Zhou ,"Semi-supervised document retrieval", Information Processing and Management 45, pp, 341–355 -2009.
[98] Wenqian Shang, Houkuan Huang, Haibin Zhu, Yongmin Lin Youli Qu, and Hongbin Dong "An Adaptive Fuzzy kNN Text Classifier", Springer, ICCS 2006, Part III, LNCS 3993, pp. 216 – 223, 2006.
[99] Heide Brücher, Gerhard Knolmayer, Marc-André Mittermayer; "Document Classification Methods for Organizing Explicit Knowledge", Research Group Information Engineering, Institute of Information Systems, University of Bern, Engehaldenstrasse 8, CH - 3012 Bern, Switzerland. 2002.
[100] Andrew McCallum, Kamal Nigam; "A Comparison of Event Models for Naïve Bayes Text Classification", Journal of Machine Learning Research 3, pp. 1265-1287. 2003.
[101] Irina Rish; "An Empirical Study of the Naïve Bayes Classifier", In Proceedings of the IJCAI-01 Workshop on Empirical Methods in Artificial Intelligence. 2001.
[102] Irina Rish, Joseph Hellerstein, Jayram Thathachar; "An Analysia of Data Characteristics that affect Naïve Bayes Performance", IBM T.J. Watson Research Center 30 Saw Mill River Road, Hawthorne, NY 10532, USA. 2001.
[103] Pedro Domingos, Michael Pazzani; "On the Optimality of the Simple Bayesian Classifier under Zero-One Loss, Machine Learning", Vol. 29, No. 2-3, pp.103-130. 1997.
[104] Sang-Bum Kim, Hue-Chang Rim, Dong-Suk Yook, Huei-Seok Lim; "Effective Methods for Improving Naïve Bayes Text Classification", 7th Pacific Rim International Conference on Artificial Intelligence, Vol. 2417. 2002.
[105] Susana Eyheramendy, Alexander Genkin, Wen-Hua Ju, David D. Lewis, and David Madigan; "Sparce Bayesian Classifiers for Text Categorization" , Department of Statistics,RutgersUniversity.2003.
[106] Miguel E. Ruiz, Padmini Srinivasan; "Automatic Text Categorization Using Neural Network",In Proceedings of the 8th ASIS SIG/CR Workshop on Classification Research, pp. 59-72. 1998.
[107] Petri Myllymaki, Henry Tirri; "Bayesian Case-Based Reasoning with Neural Network", In Proceeding of the IEEE International Conference on Neural Network'93, Vol. 1, pp. 422-427. 1993.
[108] Hwee-Tou Ng, Wei-Boon Goh, Kok-Leong Low; "Feature Selection, Perceptron Learning, and a Usability Case Study for Text Categorization, In Proceedings of the 20th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, pp. 67-73. 1997.
[109] Vladimir N. Vapnik, "The Nature of Statistical Learning Theory" , Springer, NewYork. 1995.
[110] Thorsten Joachims, "Text Categorization with Support Vector Machines: Learning with Many Relevant Fea-

tures" ECML-98, 10th European Conference on Machine Learning, pp. 137-142. 1998.
[111] Saurav Sahay, "Support Vector Machines and Document Classification"URL:http://www-static.cc.gatech.edu/~ssahay/sauravsahay7001-2.pdf
[112] Soumen Chakrabarti, Shourya Roy, Mahesh V. Soundalgekar;, " Fast and Accurate Text Classification via Multiple Linear Discriminant Projection" , The International Journal on Very Large Data Bases (VLDB), pp. 170-185. 2003.
[113] Yi Lin, "Support Vector Machines and the Bayes Rule in Classification", Technical Report No.1014, Department of Statistics, University of Wiscousin, Madison. 1999.
[114] Wikipedia Ensembles of classifiers, http://en.wikipedia.org/ wiki/ Ensembles_of_ classifiers, 2008.
[115] Monica Rogati , Yiming Yang "High-Performing Feature Selection for Text ClassificationMonica Rogati, Monica Rogati",CIKM'02, November 4–9, 2002, McLean, Virginia, USA., 2002.
[116] Fabrice Colas and Pavel Brazdil,"Comparison of SVM and Some OlderClassification algorithms in Text Classification Tasks" ,"IFIP International Federation for Information Processing", Springer Boston Volume 217, Artificial Intelligence in Theory and Practice, pp. 169-178, 2006.
[117] Hanuman Thota , Raghava Naidu Miriyala , Siva Prasad Akula, .Mrithyunjaya Rao , Chandra Sekhar Vellanki ,Allam Appa Rao, Srinubabu Gedela , "Performance Comparative in Classification Algorithms Using Real Datasets", JCSB/Vol.2 February 2009
[118] Bo Yu a,, Zong-ben Xu b , "A comparative study for content-based dynamic spam classification using four machine learning algorithms", 2008 Elsevier , Knowledge-Based Systems 21 ,pp. 355–362,2008.
[119] Youn and Dennis McLeod, " A Comparative Study for Email Classification, Seongwook Los Angeles" , CA 90089, USA, 2006.
[120] Kang Hyuk Lee, Judy Kay, Byeong Ho Kang, and Uwe Rosebrock, "A Comparative Study on Statistical Machine Learning Algorithms and Thresholding Strategies for Automatic Text Categorization", pp. 444-453, 2002. Springer-Verlag Berlin Heidelberg 2002.
[121] How, B. C. and Kiong, W. T. (2005). An examination of feature selection frameworks in text categorization. In AIRS. 558–564.
[122] Pingpeng Yuan, Yuqin Chen, Hai Jin, Li Huang "MSVM-kNN: Combining SVM and k-NN for Multi-Class Text Classification"978-0-7695-3316-2/08,2008,IEEE-DOI 10.1109/WSCS.2008
[123] Fabrice Colas and Pavel Brazdil, "Comparison of svm and some older classification algorithms in text classi_cation tasks", Artificial Intelligence in Theory and Practice (2006), pp. 169-178, 2006.

Aurangzeb khan received BS-Degree in Computer Science from Gomal University DIKhan, Pakistan and Master Degree in Information Technology From University of Peshawar, Pakistan and is currently a PhD student at the Department of Computer and Information Sciences, Universiti Teknologi PETRONAS, Malaysia. He is an assistant professor at University of Science and Technology Bannu (USTB) Pakistan. (on study leave). His current research interests include data mining, sentiment analysis and text classification through AI techniques.

Baharum Baharudin received his Masters Degree from Central Michigan University, USA and his PhD degree from University of Bradford, UK. He is currently a Senior Lecturer at the Department of Computer and Information Sciences, Universiti Teknologi PETRONAS Malaysia. His research interests lies in Image Processing, Data Mining and Knowledge Management.

Lam Hong Lee received the bachelor's degree in computer science from University Putra, Malaysia, in 2004 and his PhD from the Faculty of Engineering and Computer Science, University of Nottingham, Malaysia Campus. He is an Assistant Professor at the Faculty of Science, Engineering and Technology of Universiti Tunku Abdul Rahman, Perak Campus, located in Kampar, Malaysia. His current research interest lies in improving text categorization using AI techniques.

Khairullah khan received BS-Degree in Computer Science form Gomal University DIKhan, Pakistan and Master Degree in Computer Science from University of Science and Techniligy Bannu, Pakistan. and is currently a PhD student at the Department of Computer and Information Sciences, Universiti Teknologi PETRONAS, Malaysia. He is Senior Lecturer at University of Science and Technology Bannu (USTB) Pakistan. His current research interests include data mining, opinion mining and text classification through AI techniques