CHAPTER 1

INTRODUCTION

1.1 GENERAL

Due to the increasing availability of intelligent storage architecture and powerful computing platforms, the number of electronic documents surpasses the capacity of manual control and management. It is observed that approximately 90% of the world's data is held in unstructured textual formats (Raghavan et al 2004) and hence information-intensive business processes demand from simple document retrieval to knowledge discovery. For example, in the aerospace industry, the lifecycle of a jet engine model, covering up to 50 years of design, maintenance, tests and service data are all documented in textual format, which can easily sum up to several terabytes. Other classic examples of large textual repositories are the biomedical journal repositories published on the Web, serving as important resources for biomedical practitioners aiming to keep abreast with current research. Text documents are considered as a ubiquitous source of information. Representing the information content of a document in a form that is suitable for solving real-world problems is an important task. Topic classification or document classification of large corpus is one of the most crucial techniques to organize the documents in a supervised manner. Topic classification plays an leading role in many application domains where the user are required to access wide ranges of information available in multiple sources.

The automatic classification is considered as the most important and well-proven instrument to organize large volume of on-line text documents such as web pages, e-mails, newsgroups, blog messages, digital libraries, scientific articles and other relevant application domains including email spam filtering, adult content filtering, patent analysis, sentiment analysis and plagiarism detection. The dream of semantic Internet populated with the enormous amount of textual information remains a difficult challenge. In the current scenario, the Web serves as the main source for the text documents where the amount of textual data available is consistently increasing. This increasing availability of the textual data needs efficient text mining methodologies to extract and process meaningful knowledge and patterns existing in the documents. This enables the development of knowledge model for the machine to understand the text information.

To handle and organize the large unstructured heterogeneous document collection, there is a need for automatic techniques to extract and organize the information and assign semantic meaning to it. Automatic text classification is a suitable approach for organizing large amounts of data, providing automated means to categorize documents or text fragments into predefined semantic categories or topics. Semantic based features serve as important criteria for improving the classification accuracy. Most of the current classification systems employ only limited semantics for content management and Information Retrieval.

1.2 KNOWLEDGE DISCOVERY FROM TEXTUAL INFORMATION

Text mining triggers a greater interest in the research community in order to enable a proper exploitation, classification (supervised, unsupervised and semi supervised) and retrieval of textual data. Text mining is a process of

extracting knowledge from unstructured document, whereas data mining is a task of extracting knowledge from structured database.

The text mining, is also referred to as "Text Data Mining" or "Knowledge Discovery from Textual Databases", and can be defined as a new prospect for the analysis and the automatic processing for textual database allowing the discovery of knowledge (Ronen & James 2007). Data mining, Information Retrieval, Statistics and Natural Language Processing (NLP) are the well-established areas related to Text Mining. Hence, the process of the Text Mining is similar to the traditional process of Data Mining, its characteristic lies in the specific steps of preparation of the data due to the semi-structured or unstructured nature of the text documents being processed. From the Information Retrieval (IR) perspective, Text Mining is an elephant among blind researchers which indicates its possible alternative views and several solution strategies.

Web Mining is one of the current research fields interrelated with Text Mining. Depending upon the nature of access, it is categorized as Web structure mining, Web usage mining and Web content mining. The Web structure mining classifies and generates the similarity information of the Web pages by analyzing the hyperlink topology. Web usage mining is the process of identifying the browsing patterns of the user and analyzing the Clickthrough data that reflects the navigational behavior of the user. Web content mining is the process of extracting the information available in the un-structured web content that is mainly intended to provide indexing for efficient information tracking. Hence the Web content mining is closely related to text mining.

1.3 INFORMATION RETRIEVAL SYSTEM

Information Retrieval (IR) is defined as a process of representing, storing, organizing, and providing access to information items. Unlike data retrieval, IR is not concerned about finding precise data in databases with a given structure. In IR systems, the information is not structured; it is contained in the free form like web contents or multimedia content. According to Baeza Yates & Ribeiro (1999), IR emphasis on the retrieval of information not data and focuses on the user information need. The general schema of IR is depicted in Figure 1.1.

Figure 1.1 General Schema of IR

In Figure 1.1, Information resources are denoted as the document repositories. Topic directories serve as an index for easy retrieval of documents. The user queries are represented in the form suitable for mapping it with documents. Document ranking techniques are used to order the documents according to the terms in the query. The potentiality of text classification contributes to build the efficient IR. Information Retrieval techniques are required to deal with lot of text and therefore involve high computational representation.

1.4 MACHINE LEARNING (ML)

Machine Learning is the study of Computer algorithms that improve automatically through experience and make predictions on data. Machine Learning is a computer program that is said to learn from experience E with respect to any class of tasks T and performance measure P, if its

performance at the task improves with the experiences (Mitchell 1997). Machine learning algorithms are either classified as supervised or un-supervised depending upon the labeling of data. Supervised learning is based on the training data with pre-defined label where the class label of each instance is known in advance. Classification is an example for supervised systems that learns the given example with the class label and assigns a correct class label for unknown instances. On the other hand, unsupervised learning is based on unlabeled training data. The patterns are grouped based on similarity and termed as clustering.

1.5 TOPIC CLASSIFICATION (TC) TASK

Topic Classification is considered as a building brick of ML and IR since it shares a number of characteristics with these two fields. Topic Classification(TC) task may also be synonymously referred to as text categorization, document classification, text classification, or topic spotting which is a learning models of categorized collection of documents (Sebastiani 2005). TC always intends to assign document to one or more predefined topics based on their content.

Examples of TC tasks include spam detection where email is classified as either good or spam; Document Classification that automatically decides the topic of news article from a fixed list of topics such as "Technology", "Computer", "Politics' and "Sports"; Word Sense Disambiguation(WSD) that takes decision whether a given occurrence of the word "bank" is used to refer a bank of river or the act of depositing something in a financial institution. TC is viewed as a supervised machine learning task as it is constructed based on training corpora containing the suitable label for each training instance. In this work, TC refers to content-based classification of text documents. For a given input document and a set of predetermined classes, the system seeks the most appropriate topic to this document

according to its contents. The steps involved in document classification are represented in Figure 1.2.

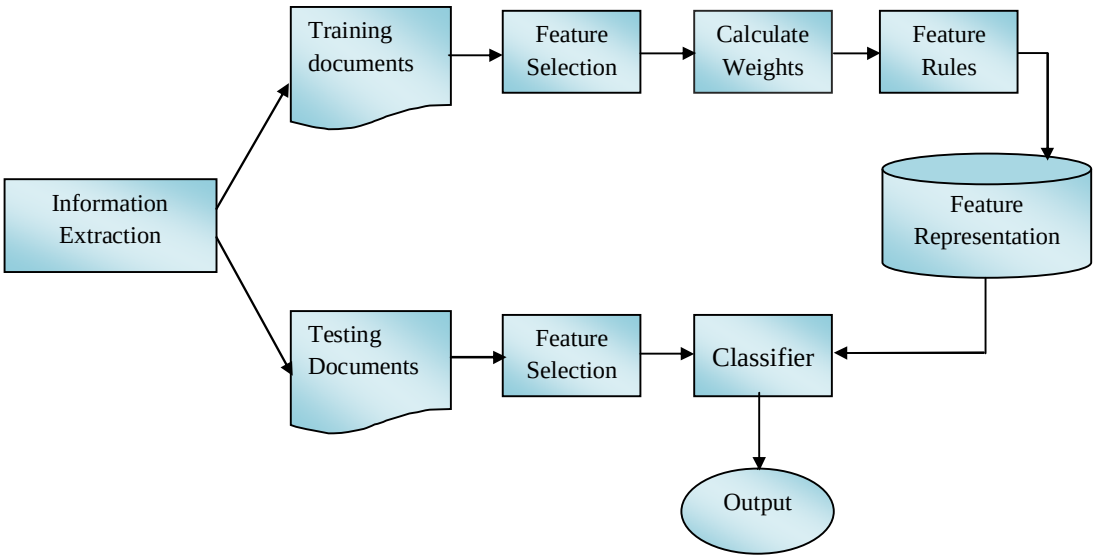

Figure 1.2 Steps in Document Classification

The Topic classification task is bifurcated into two principal phases. The first is document representation, and the second is classification.

1.5.1 Document Representation Model

Text is represented using Bag-of-Words (BOW) technique (Salton et al 1975), where each word being weighted according to how often it occurs in the text. This technique has been the most popular way to represent textual content for Information Retrieval (IR), Clustering and Classification. Even though it is popular, BOW does not bother about the position where the word appears and order of occurrences of each word.

The document representation is the preprocessing step that often reduces the complexity of the documents and makes them easier to process. It involves the mapping of text version of the document to its corresponding document vector suitable for training. Most supervised classification

techniques employs BOW using the Vector Space Model (VSM) as a mapping function to represent text documents. Salton et al(1975) defines Automatic Text Processing as a process of IR in which queries and documents are represented as vectors using VSM provided that each vector are composed of a set of terms. The term elements in each vector are assigned with a weight value either in the form of binary or numeric entailing the importance of the term in the corresponding document.

Similarity between the vectors is computed in order to find the relevance of a document to an input query. Figure 1.3 represents the VSM for IR having Term1 in x-axis and Term2 in y-axis. The document vectors are d1 and d2, where q denotes the query vector, θ and α represents similarity distance as an angle between d1 with q and d2 with q.

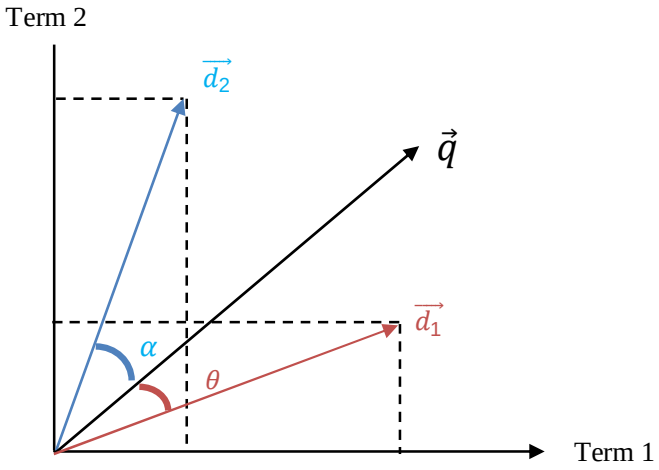

Figure 1.3 Vector Space Model for Information Retrieval

The interpretation behind the model is the documents that are mutually nearer to each other in vector space are contextually similar in meaning. The process of converting documents to a set of features is carried out as BOW where words like ‘Data’ and ‘Mining’ will occur as separate features instead of the discerning single feature ‘Data Mining’.

Dimensionality Reduction (DR) is an important step in TC because irrelevant and redundant features in the representation often degrade the performance of classification algorithms both in terms computational speed and classification accuracy (Durmaz & Bilge 2011). Feature Extraction(FE) and Feature Selection(FS) are the two important techniques contributes for Dimensionality Reduction.

1.5.1.1 Feature extraction

Feature Extraction is a process of alleviating the morphological concern of Vector Space Model by deriving new and informative features. The extracted features are likely to be either syntactic or semantic which lead to the considerable improvement in the performance of the classification. FE also incorporates certain missing information of the standard approach as much as possible. FE tries to eliminate the language-dependent features through the processes such as Tokenization, Stopwords removal and Stemming (Wang & Wang 2005).

In tokenization process, given document is treated as a set of strings, and then fragmented into a list of independent tokens. The stopwords are non-influential words needed to be removed from the document. The common stopwords are ‘it’, ‘can’, ‘an’, ‘and’, ‘by’, ‘for’, ‘from’, ‘of’, ‘the’, ‘to’, ‘with’, etc. Certain stopwords are domain specific which are constructed and maintained separately. Stemming is the process of conflating tokens to their root form by removing the suffix, e.g. networking is conflated to network and extraction, extracted are conflated to extract(Porter 1980).

1.5.1.2 Feature selection

Feature selection is another important preprocessing step in the text classification process contributes for dimensionality reduction. FS removes

the non-informative words from document in-turn increases the classification effectiveness and computational complexity. The main aim of FS is to select the subset of feature from the original documents and contributes to the construction of the vector space of documents, thereby improving the efficiency, scalability and accuracy of the text classifier.

FS process filters the highly weighted words in the document based on their importance (Tao Liu 2003). Term(word) Frequency/Inverse Document Frequency (TF-IDF) measure is used to weight each term in the text document based on the property of uniqueness. Many feature selection measures have been developed and implemented, some of the familiar metrics are Information Gain (IG), Chi-square, Mutual Information and Gini index.

- **Information gain** (Yang & Pedersen 1997) of a word measures the number of bits of information obtained for predicting the category with respect to the presence or absence of the word in a given document.

- **Chi-Square** is a statistical measure the provides an association between the word in the document and their corresponding category (Galavotti et al 2000).

- **Mutual Information** is a statistical measure that signifies the global goodness of an attribute in a feature selection. The mutual information of given attribute a and its category c is defined as

$$MI(a,c) = \frac{p(a,c)}{p(a)p(c)} \tag{1.4}$$

It is a simple but effective feature selection method for text categorization (Yang & Pedersen 1997).

1.5.2 Machine Learning Algorithms

Many supervised machine learning techniques were developed for document classification (Sebastiani 2002), some of the renowned classifiers are Regression model, k-Nearest Neighbor (k-NN), Decision Tree classifier, Naive Bayes (NB), Support Vector Machines (SVM) and Neural Networks. The un-supervised machine learning algorithm is a Clustering that generates groups or clusters of related documents and the similarity among them. This method tries to eliminate the necessity of training documents to be tagged and does not involve a pre-existing class. However, clustering algorithms provide limited support for selecting the appropriate class that is often perceptive to the human users. For this reason, clustering is always intended to work amicably with the above explained supervised learning.

The important learning models or classifier for TC are k-Nearest Neighbor, Naive Bayes, Genetic Algorithm and Neural Network based classifier. The newly emerged learning model for text classification is a Deep Learner. All these classifier models use VSM for text representation.

1.5.2.1 k-Nearest Neighbor (k-NN)

k-NN is employed to perform tests on the degree of likeliness between test documents and k training data. Here certain classification data are stored to determine the category of test documents (Ko & Seo 2000). This method is designated as an instant-based learning algorithm which categorizes the data objects based on the proximity of the feature space in the training set. Here training data are represented in a multi-dimensional feature space. This feature space in turn divided into regions based on the class of the training data. A data point in the feature space is mapped to a particular class if it is the most frequent classes among the k nearest training data. The distance

between the data points are often calculated using the Euclidean Distance measure.

The significant characteristics of this method are the accessibility of a similarity measure for tracking the neighbors of a particular document. In the training phase, the feature vectors and corresponding classes of the training dataset are only stored. In the classification phase, distances from the new vector and stored vectors are calculated by representing the input test documents and k-nearest instances are selected. The annotated class pertaining to a document is predicted based on the nearest point which has been assigned to a particular category. The similarity between test document and each neighbor is calculated and test document is assigned to the class which contains most of the neighbors. In Figure 1.4, G1 and G2 represent the classes.

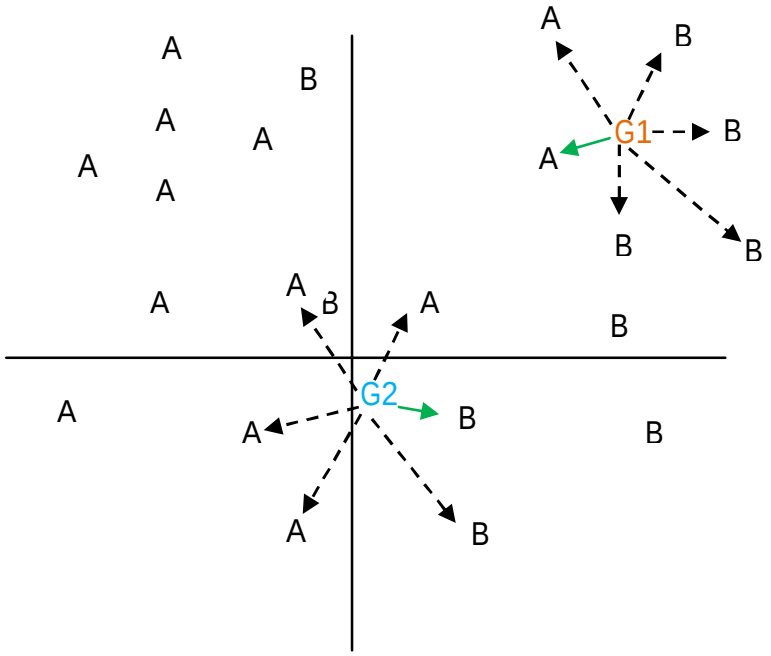

Figure 1.4 k-Nearest Neighbors of Classes G1 and G2

The k-NN classifier is widely used for TC tasks because of its simplicity and ease of efficient training. This technique performs better in the situation where classification tasks with multi-categorized documents are handled. As it involves all the training features for similarity computation, the method is viewed as a computationally intensive one. When the size of the training information increases, the performance of the system is degraded. Moreover the accuracy of k-Nearest Neighbor classifier is severely affected by the presence of noisy or irrelevant features.

1.5.2.2 Naive Bayes (NB)Classifier

Naive Bayes classification is a probabilistic approach that does not require more instances for all possible combinations of attributes. Here, every attribute of interest assumed to be independent of each other. One of the important assumptions behind the technique is that the influence of an attribute is always independent of other attributes for a given class and this assumption is called as class conditional independence (McCallum& Nigam 1998). The joint probability of document features is calculated based on the probability that a new document fit in a specific class is defined in Equation (1.5)

$$P\ (c_i|d') = \frac{P(d'|C_i).P(C_i)}{\sum_{Cj \in C} P(d'|C_i).P(C_j)} \quad (1.5)$$

where P(c_i) is the probability of a given document belongs to the class c_i and P(d'|c_i) is the conditional probability of document d belongs to a s99+9pecific class c_i. Furthermore, Naïve Bayes is used to build an unstructured text classifier with better accuracy.

1.5.2.3 Genetic Algorithm(GA)

Genetic algorithm proposed by Holland (1992) performs better on hybrid problems dealing with both discrete as well as continuous data and on combinatorial problems. Here, three important processing elements such as selection operation, crossover operation and mutation operations are used to generate various learning models. GA is best suited for the problem domain where optimal solution is needed, but the possibilities of algorithm are restricted to stick at local optimal solutions. The computational requirements of learning system are not fulfilled by GA, hence it may not be concerned when there is strong computing power is demanded.

1.5.2.4 Neural Network Based Classifier

A Neural Network (NN) text classifier is a network of processing units called neurons. Each input units represent words of the document, the output unit(s) represent the class or classes of interest, and the weights on the link that connects the processing units denotes dependence relations (Miguel Ruiz & Padmini Srinivasan 1998). To classify a given testing document d_j, the word weights w_{kj} are assigned to the visible units, where w_{kj} denotes the k^{th} term in document d_j. The activations of these input units are propagated forward through the network and the value of the output unit(s) determines the prediction decision.

Another important characteristic involved in training Neural Networks is Back Propagation. Once the word vectors of training documents that are assigned to input units are getting processed and if there is any symptoms of misclassification, the corresponding error is back propagated to the hidden layers. This back propagation allows the learning system to change the attributes of the network in order to eliminate or minimize the network error. Among the other types, the simplest Neural Network classifier is

Perceptron (Dagan et al 1997), which is called as a linear classifier that is extensively used for many applications. Schutze et al(1995) proposed and implemented the linear Neural Network classifiers in the form of logistic decision providing better testing effectiveness.

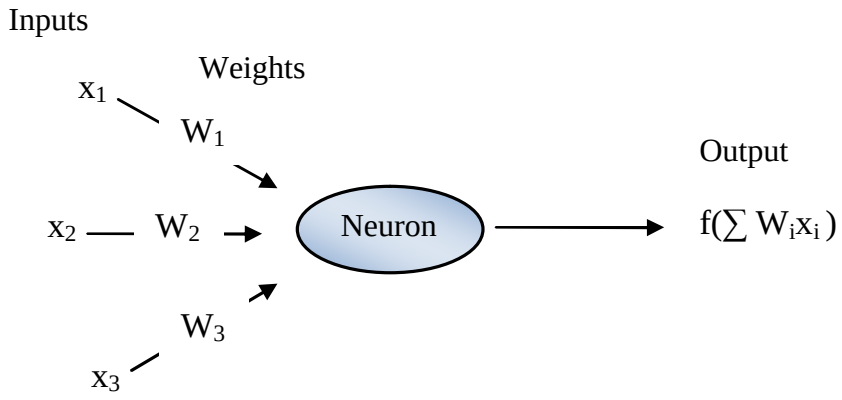

Figure 1.5 Structure of Artificial Neural Networks

The Artificial Neural Networks (ANN) given in Figure 1.5 gets inputs x_i arrives through pre-synaptic connections, Synaptic efficacy is modeled using real weights W_1,W_2 and W_3.The response of the neuron is a nonlinear function $f(\sum W_i x_i)$ of its weighted inputs.

Neural Networks are extensively deployed in the pattern recognition problems and document classification tasks by learning the training vectors to vary the weights between processing nodes. Selamat & Omatu (2004) proposed a Web page classification system that uses a Neural Network with inputs generated by Principal Component Analysis(PCA) and profile-based class features that contain the general words in each class.

1.5.2.5 Deep Learning Approach

According to Hinton(2007), Deep Learning is an advanced class of machine learning techniques that utilizes several layers of non-linear

information processing for both supervised and unsupervised methods of feature extraction and transformation. It is extensively used for applications such as pattern analysis and classification.

Deep Learners are classified into three major categories depending upon the architectures and techniques:

- **Unsupervised Learning with Deep Networks:** Highly intended to capture the high-order associations of the pragmatic or visible data. Suitable for pattern matching tasks and synthesis, especially when there is no sufficient information regarding the target class labels. Unsupervised feature learning specified in the literature often refers to this category of the deep networks. If deep network is employed in the generative mode, then it is possible to categorize the joint statistical distribution of input values and their corresponding classes only when class information is available and hence treat the same as a part of the input data.

- **Supervised Learning with Deep Networks:** Extensively used to directly afford the discriminating power of the learning system for the pattern classification task. Predetermined target class labels of data are always made available either directly or indirectly to fulfill the supervised learning strategy. This technique is otherwise termed as discriminative deep networks.

- **Semi-Supervised Learning with Deep Networks:** Hybrid version of supervised and un-supervised learning architecture. A significant way of assisting the discrimination is provided with the result of unsupervised deep architecture or with generative model. The discrimination is often supported with better optimization and regularization of supervised deep networks. The aim of the deep network also accomplished if discriminating criteria for supervised

learning are used to find the parameters with respect to unsupervised deep networks.

Two basic architectures are used in Deep Learning. They are

- Deep Discriminative Models such as Deep Neural Networks(DNN), Recurrent Neural Networks(RNN) and Convolution Neural Networks(CNN)
- Generative Models include Restricted Boltzmann Machine(RBM), Deep Belief Network(DBN) and Auto Encoders.

The various approaches for text classification are compared in terms of their own advantages and limitations; and are highlighted with help of Table 1.1.

Table 1.1 Comparison of Different Learning Algorithms

Approaches	Advantages	Limitations
Naive Bayes	Simple and quick classification Not sensitive to irrelevant features	The simplified assumption results in low accuracy, assumes independence of features
k- Nearest Neighbor	Cost of learning process is zero. Complex concepts can be learned by local approximation with simple procedure	Increase in the training data causes a decrease in efficiency, curse of dimensionality
Genetic Algorithm	Provides an optimal solution	Expects large amount of data and the algorithm often stick at local optimal solutions.
Neural Networks	Approaches an Expert's classification results	It require sufficient training data, Learning is to slow across multiple hidden layers
Deep Learner	Highly flexible to specify prior knowledge, handle large family of function parameterized with many individual parameters	Involves multiple layers with complex structures

The various learning algorithms mentioned in the Table 1.1 concludes their importance for text classification, out of which Deep Learner prove better functionality for processing the large document corpus.

1.5.3 Evaluation Metrics

Key metrics to determine the effectiveness of the classifier are precision, recall, and accuracy. These metrics are evaluated through Contingency matrix showing the correspondence between actual class and target class. Four cases are used to understand the retrieval performance based on the result of the classifier. The corresponding contingency table is presented in Table 1.2.

- **TP (True Positive)** : the number of documents correctly classified to the respective class.
- **TN (True Negative)** : the number of documents correctly rejected from the class.
- **FP (False Positive)** : the number of documents incorrectly rejected from the class.
- **FN (False Negative)** : the number of documents incorrectly classified to the class.

Table 1.2 Contingency Table

		Predicted Class	
Actual Class		Yes	No
	Yes	True Positive (TP)	False Negative (FN)
	No	False Positive (FP)	True Negative (TN)

Precision (π_i) determines the conditional probability that a document d_i is classified under the class C_m. It denotes the learning ability of the classifier to assign a document to the appropriate class as opposed to all documents assigned in that class which includes both correct as well as incorrect. Equation(1.8) denotes the calculation of precision of the classification.

$$\pi_i = \frac{TP_i}{TP_i+FP_i} \tag{1.8}$$

Recall(ρ_i) determines the conditional probability that, if some document d_i should be classified under category C_m. It is the fraction of relevant document which can be retrieved. Recall is calculated using the formula given in the Equation(1.9).

$$\rho_i = \frac{TP_i}{TP_i+FN_i} \tag{1.9}$$

Accuracy(A_i) is measured as a fraction of correctly classified document in relation to the total number of documents. The formula for finding accuracy is given in the Equation(1.10)

$$A_i = \frac{TP_i+ TN_i}{TP_i+TN_i+FP_i+FN_i} \tag{1.10}$$

1.6 SEMANTIC FEATURE SELECTION

Semantic is defined as the study of meaning in language; and words in the language are semantic units that often convey meaning(Cambridge University Press 2013). Based on their characteristics, the words are classified as polysemous words and synonymous words. Any word with more than one meaning is called as polysemous word and two or more words that have at least one meaning in common are said to be synonymous word(Miller 1995).

A Term is often called as a word or phrase used in the document that pertains to a particular subject. Terms are further divided as a simple or complex terms that are always referred as a concept. In a specific context, concept is defined as an idea or a principle. Astrakhantsev et al (2013) have focused the method in which the concepts under given domain of interest are structured, classified, modeled and represented. Gruber (1995) suggests that the semantic information can be gained from semantic resources with legal agreement and hence can be shared and used if it is consistent with the requirement.

1.6.1 Latent Topic Modeling(LTM)

Latent Topic Modeling is often viewed as a range of statistical techniques that identifies and extracts inherent topics or concepts from documents through the derived lists of co-occurring terms using statistics. Crossno etal (2011) have confronted a hypothesis that the terms constituting topics always appear nearer in meaningful ways and hence by identifying those topics would provide a way for injecting semantics into the list of vocabularies in bag-of-word representation. As this model has a BOW as a initial point, it performs dimensionality reduction process by mapping words to corresponding topics based upon the weighted list of terms for each document.

LTM is further classified as Latent Semantic Analysis(LSA) model, Probabilistic Latent Semantic Analysis (PLSA) model and Latent Dirichlet Allocation (LDA) model.

- **Latent Semantic Analysis**:

LSA model was proposed by Deerwester et al (1990) which utilizes the Singular Value Decomposition (SVD) to discover implicit higher-order structure in the co-occurrences of terms within documents. Here, the Vector

Space Model representing the documents in a large sparse matrix is mapped into smaller subspace having only singular vectors. The mapped subspace is called as Latent Semantic Space. The implementation of the LSA model shows a positive sign to overcome the drawbacks of conventional VSM employing only lexical matching.

- **Probabilistic Latent Semantic Analysis (PLSA):**

The extension of the LSA was proposed by Hofmann (1999) that deals with the variety of probabilistic functions that uses the likelihood principle rather than Single Value Decomposition for reducing the dimensions of conventional VSM. Blei et al (2003) have projected a method to reduce the dimension by mapping each document space to a probabilistic distribution of a fixed set of implicit topics or concepts (Blei et al 2003) that results in a list of the different proportions for topics.

- **Latent Dirichlet Allocation(LDA) Model:**

This model is also a probabilistic based model proposed by Blei et al (2003).It is based on a generative approach and encompasses three hierarchical levels with respect to documents, topics and words in a vocabulary collection. Here the documents are organized as random mixtures of topics. A topic has probabilities of generating other words, by learning the document collection.

Hence, Latent Topic Modeling approaches are considered as a useful method for extracting implied semantics to represent text documents. Because of its inherent nature of un-supervised feature representation, it is not advisable to adapt the models in the situation where supervised document classification is required. Moreover, it affects the efficiency and effectiveness of the supervised learning problems.

1.6.2 Semantic Knowledge Resources

Semantic knowledge resources play a vital role in extracting the semantic information and hence it serves as a knowledgebase for different applications. Some of the well-known semantic resources are WordNet (Miller 1995) and “Yet Another Great Ontology" (YAGO) which is light-weight and extensible semantic knowledgebase that unifies WordNet and Wikipedia (Suchanek et al 2007). There are some interesting domain specific semantic resources like Unified Medical Language Systems(UMLS) for the medical domain (Bodenreider 2004). Wikipedia and Open Directory Project (ODP) are general resources developed under the collaborative projects in achieving and organizing information on Web to focus the target Internet users.

- **WordNet :**

According to Miller (1995), WordNet is a lexical repository for the natural languages especially used in the Computer Science domain. Organizes English words as synsets which denote the synonyms or meaning of the word. Also gives the crisp definitions, intuitive examples and keep track of the number of relationships among the synsets. WordNet resource is normally viewed as a combination of dictionary and thesaurus. In some applications, WordNet serves as an ontology, a specification of a conceptualization that represent knowledge as a set of concepts within a domain. The two types of relationships that exists among the synsets are interpreted as Hypernyms and Hyponyms. These relationships are denoted as specialization relations among conceptual terms.

- **Wikipedia :**

The world largest encyclopedia is a Wikipedia that serves as eligible sources of concept knowledge. It is widely used in various text

mining techniques. Some of the inherent characteristics of Wikipedia are its ability to define the detailed description of each concept and linking facility with rich concepts. Since its inception in 2001, it covers more than 22,000,000 articles in 285 languages, in English alone it has 4,295,594 articles. According to Hartmann et al(1998), the Wikipedia articles include extensive concepts in all branches of knowledge, and hence provide realistic descriptions of the concepts. Moreover hyperlinks are included in the article in order to provide semantic connections among the concepts that is described in the content.

1.7 OBJECTIVES

There are a wide variety of techniques related to semantic representation of text documents and learning models are available, but there exists inherent gaps in the topic classification techniques discussed so far as given below

- Topic Classification model that is developed with BOW representation have to involve the semantic relationship between the word for efficient retrieval of documents.
- Importance should be given to infrequent words in the documents for training the classifier.
- The document representation based on feature extraction and selection techniques have to concentrate on the IE techniques that identify the semantic richness of the document features.
- A novel method for handling the word usage diversity and data sparseness are yet to be designed.
- The appropriate semantic similarity methods have to be employed for finding the relatedness between the document.

- A new learning model that classifies multiple class labels with high discrimination power have to be designed to train the large corpus of documents
- A novel method for handling ambiguities in IR should be devised.
- An adaptive model that incorporates multiple approaches needed to be developed and should be applied in various domains.

To overcome the above discussed gaps, an efficient and intelligent topic classification model is proposed by using semantic information hidden in the textual data.

The main idea of the proposed work is described as "Representation of documents by extracting the semantic features hidden in the text using knowledge-based and corpus-based features; and by proposing a recent learning algorithm for topic classification that can significantly improves the classification accuracy to retrieve the relevant results." Specifically, the objectives of the thesis work are detailed as follows:

- To design a novel Intelligent Topic Classification model with vector representation of text documents that includes efficient Feature Extraction and Feature Selection methods.
- To design a new method for extracting semantic features hidden in the documents using knowledge based methods and inclusion of same in the document representation to efficiently train the proposed model.

- To propose a novel method for selecting significant features using corpus based feature selection methods and finding the semantic similarity between the documents
- To design a new method that deals with ambiguities in the polysemous word representation for efficient Topic Classification.
- To evaluate the performance of the proposed Topic Classification model in terms of accuracy, precision and recall by using standard datasets.
- To investigate the proposed intelligent classification model by deploying it in patent domain especially with electrical patents.

The objectives stated above for the current work is mainly used to overcome the drawbacks of traditional BOW method with the aim to improve the text classification effectiveness.

1.8 ORGANIZATION OF THESIS

The thesis is organized as follows:

Chapter 1 introduces the concepts related to the Topic Classification tasks and semantic feature selection methods. The state of the art classifiers and performance evaluation metrics related to the research work is also discussed. The chapter highlights some of the gaps in the traditional topic classification model and the suitable objectives to fulfill those challenges are clearly stated.

Chapter 2 elaborates the literatures with respect to the document categorization involving semantics; algorithms for topic classification and semantic feature representation; and the various semantic similarity measures.

Based on the literature review, the major issues in the existing work are identified and the motivation of the proposed work is also discussed.

Chapter 3 presents a new method for Semantic feature extraction and indexing for Topic Classification. The basic features and semantic features are extracted from Web contents. The semantic features are extracted in the form of Hypernyms from the knowledgebase and represented as a additional feature for indexing the documents. Radix search Tree is used for the fast retrieval of Hypernyms. TF-IDF weighting technique is used to rank the features. The selected features are trained in a supervised manner with k-NN and Naive Bayes classifier. Experimentation is conducted on the standard WebKB dataset

Chapter 4 presents an enhanced Topic Classification model with additional semantics features by proposing a corpus based feature selection technique known as Semantic Smoothing. A novel Semantic Deep Learner is proposed to train the selected features and it is compared with Neural Network based classifier based on the accuracy. Two standard datasets such as WebKB and 20Newsgroups are used for analysis.

Chapter 5 discusses about the Intelligent Topic Classification approach that deals with the ambiguity problem of polysemous word in a training corpus. A new algorithm known as Decision List based Word Sense Disambiguation is proposed in addition with Semantic Smoothing for efficient feature selection. The Enriched Semantic Deep Learner is proposed to train the selected features and it is compared with Semantic Deep Learner and Neural Network Based Classifier in terms of accuracy. The overall conceptual schema for the proposed Intelligent Topic Classification Model is

presented. Three dataset such as WebKB, 20Newsgroup and Electrical datasets are used for experimentation.

Chapter 6 discusses a method of building a generic model for classifying patents based on Semantic Deep Learner and exhibits a new direction to the patent domain. The experimentation is conducted with IPC dataset provided by World Intellectual Property Organization (WIPO) and shows good classification accuracy with a reasonable number of training and testing corpus.

Chapter 7 summarizes the contributions; and gives the conclusion and future enhancement of the proposed research.