CHAPTER 2

LITERATURE SURVEY AND MOTIVATION

2.1 INTRODUCTION

The Web contains a treasure trove of information spanning different domains. Document classification and topic discovery play a crucial role in information retrieval process and content management. Topic classification using different text mining methods is one of the most important techniques to process the documents in a supervised manner. Dealing with semantic features has special requirements that standard classification approaches cannot handle. In semantic based Information Retrieval, more efforts have to be taken to improve the performance and accuracy of the topic classification process. Hence, a new system of methods and solutions should be formulated to gain the meaningful knowledge from the increasing capacity of the text corpus. Since many Information Retrieval tasks rely on the precise classification, study of the advanced classification methods will promote systems that search and organize information on the Web.

One of the central challenges in text classification is bridging the gap between information in text databases and their organization within structured topics. The proposed work develops a methodology that assigns documents to semantically meaningful topics by utilizing a supervised learning approach in which the learning model using pre-classified documents is first constructed and then the model is deployed to classify new documents.

2.2 DOCUMENT CATEGORIZATION

Document Categorization is a machine-learning task that reveals the construction and study of model which can learn from and make predictions on data. It is divided as a supervised learning which is termed as a Document Classification and un-supervised learning which is termed as a Document Clustering. Document classification is the process of assigning a document into more than one pre-defined document classes (Antonie & Zaiane 2002). Another method is document clustering that splits many documents into groups according to the similarity between documents. Similarity is measured by evaluating key representing attributes and features among documents. Both document classification and document clustering extract and use the features of the document. The main distinction between the two techniques is that document classification compares document features with pre-defined class features and selects the most suitable document class whereas document clustering divides a set of documents into groups without using pre-defined classes.

The traditional document classification is to classify documents by experts within a specific domain by constructing rules. Since experts are costly and vary in capabilities, the result obtained by the classification is not accurate and reliable. In order to limit the role of domain experts to write and process the rules, most of the manual classifications are replaced with rule-based expert system. The implementation and maintenance of rule-based system causes a labour-intensive and time consuming task(Manning et al 2008) which in turn leads to the development of supervised document classification that employs trained examples in training corpus need to learn the classification rules. Thus, many techniques for supervised classification emerged with an objective to assign text documents to their appropriate classes using the characteristics of rule-based system that mimics the domain experts.

Mohammad Khabbaz et al (2012) have presented the XML document classification by considering both structural as well as content-based features of the documents. The set of informative feature vectors is constructed to represent structural and textual aspects of XML documents. Here the soft clustering of words and feature reduction are integrated into the process. To acquire structural information, an existing frequent tree-mining algorithm was employed in addition with an information gain filter to retrieve the most informative substructures from XML documents. Also, in order to extract content information, soft clustering of words using each cluster as a textual feature was introduced. Many experiments were conducted based on a benchmark dataset, namely 20NewsGroups, and an XML document dataset given in LOGML that illustrates the web-server logs of user sessions. While the classifier composed by utilizing the proposed textual feature, the outcome proved that it is not only performed better than Support Vector Machine (SVM)based classifier but also better than Information Retrieval Classifier (IRC). Again, by applying SVM and decision tree algorithms using the proposed feature vector representation of the XML documents dataset, the classification accuracy obtained were 85.79% and 87.04%, respectively, which are much higher than accuracy achieved by XRules, which is a renowned structural-based XML document classifier.

Nikos & George (2011) proposed a method for the classification of documents in terms of their content. A dual system which shares a two-level architecture which includes a word map created via unsupervised learning that functions as a document representation module and a supervised multilayer-perceptron based classifier. Two methods of creating word maps such as Hidden Markov Models (HMMs) and the Self-Organizing Map(SOM) are proposed and compared. Fewer numbers of experiments were performed on many datasets having text based documents, which was written either in

Greek or in English. The comparison with established methods, such as the Support Vector Machine (SVM) is carried out to show the effectiveness of the systems.

Jemma Wu (2012) has introduced a complete framework for XML document classification where knowledge representation technique related to typed higher logic format is used. Due to this representation method, an XML document is represented as a higher order logic term where both its contents and structures are captured. A decision-tree learning algorithm inspired by Precision/Recall breakeven point which is denoted as Precision/Recall Decision-Tree(PRDT) was introduced to create comprehensible theories. Ultimately, a semi-supervised learning algorithm is presented, based on the PRDT algorithm and the co-training framework. The experimental outcome shows that the framework is capable to achieve better performance in both supervised and semi supervised learning with addition to produce comprehensible learning theories.

A document classification and search method based on neural network technology, which assists companies to manage patent documents more accurately has been suggested by Amy Trappey et al (2006). The classification process is initiated by acquiring key phrases within the document set by means of automatic text processing and determining the importance of key phrases according to their frequency in the text. To maintain a constructive number of independent key phrases, correlation analysis is utilized to evaluate the similarities between key phrases. Phrases which are consisting of higher correlations are synthesized into a smaller array of phrases. The back propagation network model is used as a classifier. The work was examined by using patents related to the design of power hand-tools. Similar patents are automatically classified by applying neural network models. In the prototype system, two modules are applied for patent

document management. Moreover, the automatic classification module assists the users to classify patent documents and the search module helps to find relevant and related patent documents. Finally, the experimental outcome signifies an improvement in document classification and identification than other previous theories of patent document management.

Pei-Yi Hao et al (2007) have suggested a hierarchical classification technique which generalizes Support Vector Machine learning that is related to the outcome of support vector clustering approach, and are structured in a way that mirrors the class hierarchy. If compared to earlier non-hierarchical SVM classifier and famous documents categorization systems, the proposed hierarchical SVM classification is efficient in classification accuracy with standard Reuters corpus. Duoqian Miao et al(2009)have introduced a hybrid algorithm based on variable precision rough set in order to add-up the efficiency of both k-NN and Rocchio methods. An experimental evaluation of different theories on two usual text corpora, i.e., the Reuters-21578 collection and the 20Newsgroups collection were conducted. Thus, the experimental outcomes showed that the developed algorithm acquired a significant performance enhancement.

Erdem Alparslan et al(2011) have introduced a mixed procedure by involving Support Vector classifier and adaptive neuro-fuzzy classifier. The method suggested the preprocessing tasks required for document classification with natural language processing. To denote term-document relations, term frequency and inverse document frequency were selected to compose a weight matrix. Agglutinative nature of Turkish documents was processed by Turkish stemming algorithms. Ultimately, a few experiments outcomes and success metrics with high accuracy were shown.

2.3 SEMANTIC FEATURE REPRESENTATION

Document representation plays a major role in TC task. The choice of the document features influences the performance of the document classification. Feature Extraction (FE) and Feature Selection (FS) are the important flavors of document representation. FE is a process of building a new set of features from the original feature set and FS is a process of selecting a subset of the original feature set. The terms in the document are weighted according to how unique it is. FE and FS for BOW representation provides limited functionality for efficient IR. Therefore, incorporating semantics features that are hidden in the documents enable the classifier to perform prediction.

A context is a semantically significant document set, and earlier studies suggest that contexts are determined through, atleast two strategies: neighboring terms identification of the keyword; and identification of terms indicating documents scope and semantics. Semantic methods exploit relationship among document's words to evaluate the semantic relevance. Xinhui Tu et al (2010) & Yan Liu et al (2012) have focused on computing semantic using methods such as language modeling approach to IR and multi-document summarization that are either knowledge-based or corpus-based. Knowledge bases include lexical resources, ontologies such as WordNet (Miller 1995) to measure term distance and Wikipedia resources. Corpus-based methods include Latent Semantic Analysis (LSA) and Explicit Semantic Analysis (ESA) for identifying and representing the document’s topic and content. Executing Topic classification task, Naive Bayes(NB) classifier estimates a conditional probability (the subsequent probability in which document b belongs to class c), classifying the test document to the class resulting highest condition probability. When calculating conditional probability, NB composes word independence assumptions and decomposes

conditional probability into a product of individual word probabilities. Also, Unigram Language Modeling (ULM) is a common choice because of its effectiveness and computational efficiency. NB with ULM only considers frequencies of words in a class; hence it may get affected by the problems of data sparseness and word usage diversity resulting in lower performance for TC. So, to cope up with the data sparseness (zero probability) problem of ULM, many well-established methods like, Laplace and Jelinek-Mercer smoothing techniques has been introduced.

The Latent Topic Model(LTM) utilizes a set of latent topics to decompose the relationships between documents and classes (Jianping Zeng et al 2012 & Lili Yang et al 2013). Some of the LTM includes Latent Semantic Analysis(LSA),Probabilistic Latent Semantic Analysis(PLSA) and Latent Dirichlet Allocation(LDA). For these methods, classification scores are not evaluated directly based on the frequencies of the words, but are computed based on a set of latent topic with the possibility that each class generates the respective topics. The application of latent topics manages the word usage diversity problem for ULM and performs TC in a concept matching manner. Again, NB with LTM methods have considered the semantic information, the documents level semantic cues are not directly incorporated into the TC job.

Most of the proposed approach involves the semantics in the process of text classification at different steps of processing. Many works argued the utility of semantics at text representation step (Caropreso et al 2001; Liu et al 2004; Seaghdha 2009; Aseervatham et al 2009; and Li et al 2009). Bashar Tahayna et al(2010) have introduced and explained the effectiveness of the topic-based document classification techniques. In this method, the Wikipedia, a large scale Web encyclopedia which has top-quality and large-scale article and a category system is used. Here Wikipedia has

been utilized by applying an N-gram method in order to transform the document from being a "container of words" to a "container of concepts". Based on this transformation, a concept-based weighted scheme (denoted as Conf.idj) to index the text with the flavor of the traditional tf.idf indexing scheme was proposed. Also, a genetic algorithm-based support vector machine optimization technique is utilized for featuring subset and instance selection. The experimental results proved that the introduced weighting scheme performed better than orthodox indexing and weighted scheme.

Semantics Based Feature Vector using Part of Speech(POS), to extract a concept of terms using WordNet, co-occurring and associated terms was proposed by Khan et al (2010) which was applied on small documents dataset showing that it outperformed the Term Frequency/ Inverse Document Frequency (TF-IDF) with BOW feature selection for text classification. Discrete Cosine Transform (DCT) and feature selection with Proportion of Variance was proposed by Durmaz & Bilge (2011) to ensure more effective classification results and short classification time. WebKB and R8 datasets in Reuters-21578 were used in experiments. Using DCT classification, success was ensured while using Proportion of Variance.

A model for document preprocessing through removal of stop words; stemming with Porter stemmer algorithm; WordNet thesaurus was applied to maintain a relationship between important terms and global unique words (Patil & Atique 2013).Frequent word sets are generated and next data matrix is formed. Finally, terms were extracted from documents using term selection approaches such as term frequency and document frequency based on the minimum threshold value. Evaluation was on Reuters Transcription Subsets such as trade, wheat, grain, money and ship. An efficient index, IR-tree with a top-k document search algorithm facilitating four tasks in document searches was proposed by Li et al(2011). The tasks include textual

filtering, spatial filtering, relevance computation, and fully integrated document ranking. Additionally, IR-tree permitted searches to adopt varied weights on documents, textual/spatial relevance at runtime and thus catering to varied applications. Comprehensive experiments over a range of scenarios were conducted with the results demonstrating that IR-tree outperformed state-of-the-art for geographic document searches approaches.

A demonstration of Word-Class Lattices (WCLs) multilingual generalization, a supervised lattice-based model is constructed to identify textual definitions and extracted Hypernyms from them (Faralli & Navigli 2013). Lattices were learned from an automatically-annotated Wikipedia definitions dataset. A Java API was released for programmatic use of multilingual WCLs in English, French and Italian and also a Web application to define Hypernym extraction from user-provided sentences. A new machine learning model incorporating ontology information was introduced by Sofia & Nyberg (2011) for learning a classification model and enhancing it with ontology information in a case study for Finnish National Archives and a set of digital documents were manually classified. The ontology enhanced model was applied to data and most likely documents classes learnt. The results showed that the model's classification accuracy increased when ontology information was added.

David Bracewell et al (2009) have approached two algorithms for topic analysis of new articles. Topic analysis expects category classification and topic discovery and classification. Dealing with news, needs some requirements that standard classification methods cannot accomplish. The algorithms introduced by them are so efficient that they can execute the online training for both category and topic classification, at the same time can focus on new topics as they arise. Both algorithms are related to a keyword extraction algorithm that is applicable to any language that has basic

morphological analysis tools. Also, both the category classification and topic discovery and classification algorithms can be utilized at ease by several languages. According to the experimentation, the algorithms are proved to have high precision and recall in tests on English and Japanese.

2.4 PROSPECTS OF WORDSENSE DISAMBIGUATION (WSD)

In the last few years, concentration has been paid to the semantic analysis of domain texts at the lexical level, a task called Domain Word Sense Disambiguation and it works based on the Semantic Model Vector (SMV) (Roberto Navigli et al 2011). Word Sense Disambiguation is the mission of allocating to each occurrence of an unclear word in a text to one of its feasible senses (Massimiliano Ciaramita et al 2003). WSD is employed to know the meaning of the polysemous word in the given context. A word sense is a commonly accepted meaning of a word. When the system searches the knowledge based resources for a polysemous word, it may find matches with different meaning. For example, the word "register" has many senses in English, such as book or memory device. To face this problem, state of the art approaches for conceptualization proposes multiple strategies to deal with ambiguities (Bloehdorn et al 2006). Three possible solutions are given below:

- **Every concept:** This method involves all the concepts as it matches for the considered word.
- **Initial Concept:** This method accepts the most frequently used concepts among the different members according to language statistics.
- **Contextual concept:** This method accepts the concepts that have the most similar semantic context, as compared to the original word's context in the document (Aronson et al 2010 & Bai et al 2010).The context of a concept is related to its

definition or its descriptive words in the semantic resource or to its textual context in a text corpus.

The Graph-based Word Sense Disambiguation of biomedical documents has been elucidated by Eneko Agirre et al (2010). Word Sense Disambiguation (WSD) is the process that recognizes the meaning of ambiguous words in context and serves as a significant stage in text processing. The method does not need any labeled training data and was unsupervised. Unified Medical Language System (UMLS) was the main knowledge base used in their approach where the Meta thesaurus was symbolized as a graph. A state-of-the-art algorithm, Personalized Page Rank was employed to carry out WSD. In terms of the content, a method for the classification of documents was introduced (Nikos & George 2011).

Sergio Greco et al (2011) have made a clear method for collaborative clustering of XML documents and addressed the difficulty of clustering XML documents in a joint distributed framework. First, XML documents were represented based on semantically cohesive subtrees, next modeled as transactional data that embedded both XML structure and content information. This clustering framework uses a centroid-based partitioned clustering method that was proposed for a peer-to-peer network. Each peer in the network was permitted to calculate a local clustering solution over its own data, and to replace its cluster representatives with other peers. The replaced representative was employed to calculate representatives for the global clustering solution in a joint way. Damiano et al (2013) proposed a method for discovering filter keywords to handle disambiguation in the microblog services such as twitter where the minimum contextual information is available. The accuracy of the algorithm is estimated with the range 75% to 79% with the 30% to 60% of tweets.

2.5 SEMANTIC SIMILARITY MEASURES

Semantic similarity is one of the key components behind the magnitude of NLP, IR, TC and WSD especially for finding semantic relationship between terms, concepts and documents. Two types of semantic similarity measures are knowledge-based measure and corpus-based measure. Knowledge -based measure includes Lexical resources and ontologies such as WordNet. Corpus-based measure includes a Latent Semantic analysis that employs probabilistic approach to decode the semantics of the words and have an ability to introduce similarity between two words as long as they appear in the corpus used for training.

Many measures have been proposed. On the whole, the commonly used measures are path length based measures, information content based measures and feature based measures (Lingling et al 2013). Definition of related concepts in the following measures is given as:

$len(c_i,c_j)$ - the length of the shortest path from synset c_i to synset c_j in WordNet.

$lso(c_i,c_j)$ - the lowest common subsumer of c_i and c_j

$depth(c_i)$ - the length of the path to synset c_i from the global root entity, and depth(root)=1.

deep_max - the max $depth(c_i)$ of the taxonomy

2.5.1 Path-based Similarity Measures

The idea of path-based measures is that the similarity between two concepts is the function of the length of the path linking the concepts and the position of the concepts in the taxonomy.

- **Shortest Path based Measure**

The length of c_1 and c_2 are taken into consideration, where c_1 and c_2 represents the concepts. It assumes that the sim(c_1, c_2) depends on how close the two concepts are in the taxonomy. This measure is a variant of the distance method (Radha et al 1989 & Bulskov et al 2002). It is based on two observations. One is that the behavior of conceptual distance resembles the metric. The other is that the conceptual distance between two nodes is proportional to the number of edges separating the two nodes in the hierarchy (Verelas et al 2005)

$$sim_{path}(c_1, c_2) = 2 * deep\_max - len(c_1, c_2) \quad (2.1)$$

From Equation (2.1), it is noted that deep_max is a fixed value. The similarity between two concepts (c_1, c_2) is the function of the shortest path len(c_1, c_2) from c_1 to c_2. If len(c_1, c_2) is 0, sim_{path}(c_1, c_2) gets the maximum value of 2*deep_max. If len(c_1, c_2) is 2*deep_max, sim_{path}(c_1, c_2) gets the minimum value of 0. Thus, the values of sim_{path}(c_1, c_2) are between 0 and 2*deep_max.

- **Wu & Palmer's Measure**

Wu and Palmer introduced a scaled measure (Wu& Palmer 1994). This similarity measure takes the position of concepts c_1 and c_2 in the taxonomy relatively to the position of the most specific common concept lso(c_1,c_2) into account. It assumes that the similarity between two concepts is the function of path length and depth in path-based measures.

$$sim_{WP}(c_1, c_2) = \frac{2*depth\ (lso(c_1,c_2))}{len(c_1,c_2) + depth\ (lso(c_1,c_2))} \quad (2.2)$$

From Equation (2.2), it is noted that the similarity between two concepts (c_1,c_2) is the function of their distance and the lowest common subsume(lso(c_1,c_2)).If the lso(c_1,c_2) is root, then depth(lso(c_1,c_2))=1, sim_{WP}(c_1,c_2) >0; if the two concepts have the same sense and lso(c_1,c_2) are the same node, then len(c_1,c_2)=0, sim_{WP}(c_1,c_2) = 1; otherwise 0<depth(lso(c_1,c_2)) < deep_max, 0<len(c_1,c_2)<2*deep_max, 0<sim_{WP}(c_1,c_2) < 1. Thus, the values of sim_{WP} (c_1, c_2) are specified within 0 and 1.

2.5.2 Information Content-based Measures

It is assumed that each concept includes much information in WordNet. Similarity measures are based on the Information content of each concept. It is assumed that if two concepts share the more common information, then there exists more similarity between them.

- **Resnik's Measure**

Resnik (1995) in his work proposed information content-based similarity measure. The assumption of the measure is that, for any two given concepts, similarity depended on the Information Content(IC) that subsumes them in the taxonomy.

$$sim_{Resnik}(c_1, c_2) = -\log p\big(lso(c_1, c_2)\big) = IC\big(lso(c_1, c_2)\big) \quad (2.3)$$

From Equation (2.3), it is noted that the values only rely on concept pairs' lowest subsumer in the taxonomy.

- **Lin's Measure**

Lin (1995) proposed a method for similarity measure. It uses both the amount of information needed to state the commonality between the two concepts and the information needed to fully describe these terms.

$$sim_{Lin}(c_1, c_2) = \frac{2*IC(lso(C_1,C_2))}{IC(C_1)+IC(C_2)} \quad (2.4)$$

From Equation(2.4), it is noted that the measure has taken the information content of compared concepts into account respectively. As IC(lso(c_1,c_2))<=IC(c_1) and IC(lso(c_1,c_2)) <=IC(c2), therefore the values of this measure vary between 1 and 0.

2.5.3 Feature-based Measure

This measure differs from all the above presented measures in various aspects. It is independent on taxonomy, the subsumers of the concepts and attempts to exploit the properties of the ontology for extracting the similarity values. It is based on the assumption that each concept is described by a set of words indicating its properties or features, such as their definitions or “glosses” in WordNet. It is based on the more common characteristics of two concepts, less non-common characteristics they have and how far the concepts are similar.

One classical measure is Tversky’s model, which argues that similarity is not symmetric. Features between a subclass and its super class have a larger contribution to the similarity evaluation than those in the inverse direction. It is defined by Tversky (1977)

$$\mathrm{sim}_{\mathrm{Tversky}}(c_1, c_2) = \frac{|C_1 \cap C_2|}{|C_1 \cap C_2| + k|C_1/C_2| + (k-1)|C_2/C_1} \quad (2.5)$$

where c_1, c_2 corresponds to description sets of concept c_1 and c_2 respectively, k is adjustable and k∈[0,1]. From the Equation(2.5), it is noted that, the values of $\mathrm{sim}_{\mathrm{Tversky}}(c_1,c_2)$ vary from 0 to 1. $\mathrm{sim}_{\mathrm{Tversky}}(c_1,c_2)$ increases with commonality and decreases with the difference between the two concepts.

2.5.4 Concept-based Measure

Point-wise Mutual Information (PMI) is one of the concept-based measure proposed by Blanco et al (2008) is a corpus-based measure used to calculate the semantic similarity between two words w_1 andw_2.

$$PMI(w_1, w_2) = \log_2 \frac{C(w_1 near w_2)}{C(w_1) * C(w_1)} \quad (2.6)$$

This indicates the statistical dependency between w1 and w2, and can be used as a measure of the semantic similarity of two words. c(w_1 near w_2) represents the number of times that word w_1 appears near word w_2. For this co-occurrence count, a window of length l is used, that is the co-occurrence of the words are counted only within this window. For a word, PMI(w,w) = 1, therefore max_{Sim}(w,D) is 1 if w appears in document D.

2.6 DEEP LEARNING

Deep learning has emerged as a new area of machine learning research since 2006. Deep learning is otherwise called as feature learning or representation learning. It is a set of machine learning algorithms which attempt to learn multiple-layered model of inputs and to train complex and deep models on large amounts of data, in order to solve a wide range of text mining and natural language processing (NLP) task (Glorot et al 2011 & Yan Liu et al 2012).

Restricted Boltzmann machines (RBMs) have been used as generative models of many different types of data, including labeled or unlabeled images (Hinton et al 2006), windows of mel-cepstral coefficients that represent speech (Mohamed et al 2009), bag of words that represent documents (Salakhutdinov & Hinton2009), and user ratings of movies (Salakhutdinov et al 2007). The greedy layer-wise unsupervised pre-training

is based on training each layer with an unsupervised learning algorithm, taking the features produced at the previous level as input to the next level. Finally, the set of layers with learned weights could be stacked to initialize a deep supervised predictor, such as a neural network classifier, or a deep generative model, such as a Deep Boltzman Machine. Xavier Glorot et al (2011) proposed a deep learning approach that learns to extract a meaningful representation of review in an un-supervised fashion and address the problem of domain adaptation for sentiment classifiers. Sentiment classifiers trained with this high-level feature representation outperform state-of-the-art methods on a benchmark dataset composed of reviews of 4 types of Amazon products such as Kitchen, Electronics, DVDs and Books. The metrics such as transfer ratio and in-domain ratio are evaluated with reasonable results.

RBMs provide a self-contained framework for deriving competitive non-linear classifiers (Hugo & Yoshua 2008). An evaluation of different learning algorithms for RBMs proves that RBMs aim at introducing a discriminative component to train and improve their performance as classifiers. This approach is simple in that RBMs are used directly to build a classifier, rather than as a stepping stone. The demonstration on how the discriminative RBMs can also be successfully employed in a semi-supervised setting.

Classification RBM (ClassRBM), a variant on the RBM adapted to the classification setting (Hugo Larochelle et al 2012). Different strategies for training the ClassRBM are studied and proved that competitive classification performances can be reached when appropriately combining discriminative and generative training objectives. Finally a method to adapt the ClassRBM to two special cases of classification problems, namely semi-supervised and multitask learning. Li Dong (2014) proposed Adaptive Multi-Compositionality (AdaMC) layer of recursive neural models. The basic idea

is to use more than one composition functions and adaptively select them depending on the input vectors. The general framework to model each semantic composition as a distribution over these composition functions was designed. The proposed AdaMC was integrated into existing recursive neural models and have conducted extensive experiments on the Stanford Sentiment Treebank. The results illustrated that AdaMC significantly outperforms state-of-the-art sentiment classification methods. It helps to push the best accuracy of sentence-level negative/positive classification from 85.4% up to 88.5%

2.7 ISSUES IDENTIFIED IN THE ABOVE TOPIC CLASSIFICATION METHODS

There are wide varieties 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. Hence, to reduce the gaps and to fulfill the above requirements, the following challenges were identified.

- BOW representation for IR having a single words feature faces many inherent deficiencies that affect the accuracy of Topic Classification learning algorithms. The major drawbacks are the poor quality of the training data, lack of discriminating power of the classifier, limited semantic richness of the features to represent documents (Manning & Schutze 2000)

- It is also observed that the computational cost involved in building a classifier for huge number of relevant features is expensive. Due to the weak and strong frequency of certain terms, it is difficult to construct reliable models initiating from some instances in the training set. Moreover, it is examined

that most of the frequent terms have not brought important information since those terms present in all documents (David Bracewell et al 2009).

- Data sparseness and word usage diversity problems in Language model greatly influence the performance of TC task and moreover the lack of document-level semantic cues affect the underlying TC (How Jing 2013)
- Computation of semantic relatedness by incorporating knowledge-based and corpus-based features is a challenging issue in performing TC. (Carmen Banea et al 2014)
- Another important issue needed to be addressed is ambiguity in IR that greatly affects the retrieval of relevant documents, while different words which represent the same concept can cause the retrieval system perform poorly to find all of the relevant documents and model is orthogonal because it ignores relations between words and treats them independently (Huang et al 2012).

2.8 MOTIVATION

The motivation of the proposed research is to develop an Intelligent Topic Classification model by investigating the semantic representation of text in order to overcome the drawbacks of the conventional BOW method with the aim to improve the text classification effectiveness. The behavior of the proposed model is investigated through several application scenarios where semantics play a dominant role.

The main goal of the proposed model is twofold:

- To design and implement a novel method for extracting the semantic features hidden in the text to make the efficient topic classification.
- To propose a novel Learning algorithm with good classification accuracy that retrieves the relevant results.

Literature review presented in the previous section reveals the state-of-the-art methods that investigated about the influence of semantics on supervised text classification and other relevant methods pertaining to IR domain. Web content represents a universal repository of knowledge. The semantic feature extraction and significant feature selection remain the backbone for building a reliable learning model.

Even though, there are many techniques and approaches for handling the semantic information in the text for effective classification, there is a room for proposing more solutions to address the growing demands of knowledge sources. In this way, the proposed research work involves in the development of Semantic Smoothing Model which serves as a mathematical model for feature extraction along with un-supervised weighted clustering, a Semantic Deep Learner(SDL) which is a recent latent topic modelling for IR to train the features and predict the category (topic) of the documents from different domain. In order to make the prediction more accurate, a Decision List based Word Sense Disambiguation (DLWSD) is proposed in addition with semantic smoothing to extract the actual meaning of the polysemous words with respect to the context thereby providing more precise results.

The implementation of the proposed work allows us to classify documents from different domains. Initially, the proposed model is investigated by classifying the news articles where news emerging from heterogeneous sources are classified into their appropriate topics such as

sports, computers, politics, health, etc., that are useful for effective IR. The second application allows the xml documents from different universities to be mapped to the suitable category, such as project, students, staff, course, etc., thereby enabling the collaborative academic activities. The third one is the application of our proposed model to new field such as patent analysis where semi-structured patents are grouped semantically based on their category. By achieving the patent analysis the organizations are benefited to identify intellectual property and technological competitiveness. The fourth one is the classification of electrical documents which allows the electrical engineers to derive the subject knowledge and calibrations.