CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

Due to the rapid growth of digital data made available in recent years, knowledge discovery and data mining have attracted a great deal of attention with an imminent need for turning such data into useful information and knowledge. Many applications, such as market analysis and business management, can benefit by the use of the information and knowledge extracted from a large amount of data. Knowledge discovery can be viewed as the process of nontrivial extraction of information from large databases, information that is implicitly presented in the data, previously unknown and potentially useful for users. Data mining is therefore an essential step in the process of knowledge discovery in database (Ning Zhong et al. 2012). Since a textual format is a very flexible way to describe and store various types of information, large amounts of information are stored and distributed as text.

Moreover, the amount of accessible textual data has been increasing rapidly. Such data may potentially contain a great wealth of knowledge. However, analyzing huge amounts of textual data requires a tremendous amount of work in reading all of the text and organizing the content (Nasukawa & Nagano 2001). There are two challenging issues in using pattern mining techniques for finding relevance features in both relevant and irrelevant documents. The first is the low support problem. Given a topic,
long patterns are usually more specific for the topic, but they usually appear in documents with low support or frequency (Yuefeng Li et al. 2007).

If the minimum support is decreased, a lot of noisy patterns can be discovered. The second issue is the misinterpretation problem, which means the measures (e.g., “support” and “confidence”) used in pattern mining turn out to be not suitable in using patterns for solving problems. For example, a highly frequent pattern (normally a short pattern) may be a general pattern since it can be frequently used in both relevant and irrelevant documents. Hence, the difficult problem is how to use discovered patterns to accurately weight useful features (Yuefeng Li et al. 2007). General approaches to document retrieval that have been applied elsewhere (e.g., patents and legal discovery) are based on the construction of semantic networks from the words in the documents. The content in a collection of documents often supports the creation of semantic networks where the document elements (i.e., words and concepts) and their co-occurrence are represented as the nodes and the edges of the semantic network structure, respectively (Taxiarchis Botsis et al. 2015).

There is a continuous growth of textual data recorded in the digital format at heterogeneous sources, such as Internet, Company Intranets, Digital libraries, and Diagnosis repair verbatim. Needless to say, the vast amount of knowledge embedded in such data must be discovered for efficient decision-making in several business functions to detect anomalies, improve Fault Diagnosis (FD), and improve root-cause investigation (Dnyanesh G. Rajpathak 2013). Text mining can work with unstructured or semi-structured data sets such as emails, full-text documents and HTML files and more. Some general approaches about text mining and knowledge discovery in texts can be found. Text mining shares many characteristics with classical data mining, but differs in many ways (Vishwadeepak Singh Baghela & Tripathi 2012).
1.1.1 Record Retrieval

Record Retrieval is defined as the matching of some user stated query against a set of free-text records. The main goal of Information Retrieval System (IRS) is to “finding relevant information or a document that satisfies user information needs (Sneha Lohbare & Ashwini Meshram 2014). There are two basic document retrieval processes. First is browsing or navigation system and another is classical IR system. In former system, User skims document collection by jumping from one document to the other via hypertext or hypermedia links until relevant document found. In later system, also called as question answering system, query answer format is implemented to retrieve document or specific information. A query i.e. question in natural language is placed and answer is directly extracted from text of document collection (Ganga 2014).

Traditional document storage systems allow users to identify documents using metadata such as author, title, keywords. The basic idea is to index every individual word in the document collection. Effectively, documents are represented as a “bag of words”—that is, the set of words that they contain, along with a count of how often each one appears in the document. Many practical systems discard common words or “stop words”, primarily for efficiency reasons. A query is expressed as a set, or perhaps a Boolean combination, of words and phrases, and the index is consulted for each word in the query. A well-developed technology of relevance ranking allows the salience of each term to be assessed relative to the document collection as a whole, and also relative to each document that contains it. These measures are combined to give an overall ranking of the relevance of each document to the query, and documents are presented in relevance order.
1.1.1.1 Why Text Mining?

In times of Information Technology, the volume of data has reached an incredible growth in various fields. For obtaining essential information from abundance of data has become difficult, therefore, text mining has arrived to tackle the needs. Text mining has high potential value in dealing with large and complex unstructured datasets. In addition, unstructured data contains the irrelevant and noisy information which has to be eliminated. Moreover, Text mining is a multidisciplinary field, involving information retrieval, text analysis, information extraction, clustering, categorization, visualization, database technology, machine learning, and data mining. Here, some of them are discussed below.

1.1.2 Information Retrieval

The impact of the growth of the information in every field has made considerable attention in retrieval of essential information. However, it is considered as a biggest challenge of retrieving information. Moreover, information retrieval is a subjective task, and as such various phases of a retrieval process are affected by uncertainty and/or vagueness. For example if the use the search engine for searching for information on the Web, it would have taken much time. In addition, the quality of the results is in general poor, it is difficult to access the appropriate information, and the overall costs for complete satisfaction of the users’ needs are considerably high. Moreover, when using a search engine the query formulation process is subjective, and the user may be uncertain on which keywords to select to properly express her/his needs. In general, the significant amount of time and resources are become the essential parts to acquire relevant and accurate information.
1.1.2.1 Text categorization

Nowadays, text categorization has become the most considerable part in text mining. Due to the tremendous growth of information technology as well as database enhancements cause difficulties in information technology. Therefore, to handle with large amount of data many researchers focused to information extraction (essential content) which made the reason for the presence of text mining in recent days. Text categorization, with the increasing availability of text documents in electronic form, it is of great importance to label the contents with a predefined set of thematic categories.

However, the biggest challenge of text categorization is dealing with learning from high dimensional data. Tens and hundreds of thousands terms in a document may lead to a high computational burden for the learning process, although, some irrelevant and redundant features may hurt predictive performance of classifiers for text categorization.

Therefore, to prevent from the issue of curse of dimensionality and to accelerate the learning processes have become the essential to act upon feature reduction to reduce feature size. A widespread feature reduction approach for text categorization is feature selection. In particular, the simplicity and efficiency of filter approach which makes the filter approach to be used in text categorization. However, the filter approach evaluates the decency of a feature by only utilizing the intrinsic characteristics of the training data without considering the learning algorithm for discrimination, which may lead to an undesired classification performance. Typically, Text classification involves tens of thousands of features, and thus is an ultra-high dimensional problem (Aytuğ Onan et al. 2016).
1.2 MAJOR AND WIDELY USED CLASSIFIERS

There are two widely used classifiers in text classification (Guozhong Feng et al. 2015). They are

- Generative classifiers
- Discriminative classifiers

1.2.1 Generative Classifiers

The generative classifiers learn the combined probability of the features and the label, while discriminative ones directly model the dependence of the label on the feature. The advantage of Generative classifier is considered as it can able to express more complex relationship among the variables (including features and label), and are commonly used in information retrieval and text classification.

1.2.2 Discriminative

Discriminative text classifier is the Stanford Classifier, which is a maximum entropy classifier.

1.2.3 Data Mining

Data mining is the process of discovering insightful, interesting, novel patterns and correlations within large data sets to predict outcomes. Data mining refers to extracting useful information from vast amounts of data and it is commonly agreed that data mining is an essential step in the process of Knowledge Discovery in Databases, or KDD. Normally Data Mining blends traditional data analysis methods with sophisticated algorithms for processing large volumes of data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships
identified (Atif Shahzad & Nasser Mebarki 2012, Lorena Siguenza-Guzman et al. 2015). Data mining consists of five major elements, such as

- Extract, transform, and load transaction data onto the data warehouse system.
- Store and manage the data in a multidimensional database system.
- Provide data access to business analysts and information technology professionals.
- Analyze the data by application software.
- Present the data in a useful format, such as a graph or table.

1.2.3.1 Some techniques in data mining

In data mining, there are several techniques have been used for better performance. Some of them are discussed below such as,

- Artificial Neural Network
- Decision Trees
- Genetic Algorithm

1.2.4 Artificial Neural Network

Artificial neural network ANN is inspired by real networks of neurons. Since the complexity and incomplete understanding of biological neurons, various architecture of artificial neural network had been reported in the literature. The purpose of designing neural network is to act as human ability to adapt to changing in circumstances and the current environment (Zheng Yan & Xuyang Pedrycz 2017, Farman Ali 2016). Moreover, it is a collection of connected input/output units; however, each connection has a
weight present with it. During the learning phase, network learns by adjusting weights so as to be able to predict the correct class labels of the input tuples. Neural networks have the remarkable ability to derive meaning from complicated or imprecise data and can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. These are well suited for continuous valued inputs and outputs. For example handwritten character reorganization, for training a computer to pronounce English text and many real world business problems and have already been successfully applied in many industries. Neural networks are considered as the best tool for identifying patterns or trends in data and well suited for prediction or forecasting needs (Taxiarchis Botsis et al. 2015).

1.2.5 Decision Trees

A decision tree is a structure that includes a root node, branches, and leaf nodes. Each internal node denotes a test on an attribute, each branch denotes the outcome of a test, and each leaf node holds a class label. The topmost node in the tree is the root node. It is one of the most widely used and practical methods for inductive inference over supervised. A decision tree represents a procedure for classifying categorical data based on their attributes data and also efficient for processing large amount of data, so is often used in data mining application. The construction of decision tree does not require any domain knowledge or parameter setting, and therefore appropriate for exploratory knowledge discovery. Each internal node represents a test on an attribute. Each leaf node represents a class.

The benefits of having a decision tree are as follows

- Any domain knowledge is not necessary.
- It is easy to comprehend.
The learning and classification steps of a decision tree are simple and fast.

1.2.6 Genetic Algorithm

Genetic Algorithm is a technique designed by the principle of Darwin’s evolution theory of survival of the fittest in natural genetics. It maintains a population of potential solutions of the candidate problem termed as individuals. By manipulation of these individuals through genetic operators such as selection, crossover and mutation, GA evolves towards better solutions over a number of generations. Genetic algorithm starts with randomly created initial population of individuals that involves encoding of every variable. In Genetic Algorithm, fitness function have been used which is derived from the objective function of the optimization problem to evaluate the individuals in a population. Fitness function is the measure of an individual’s fitness, which is used to select individuals for reproduction. Moreover, selection method in a GA selects parents from the population on the basis of fitness of individuals. High fitness individuals are selected with higher probability of selection to reproduce offspring’s for the next population.

1.2.6.1 Keyword extraction in text mining

Nowadays, the information which is available on the web is progressively expanding, therefore, the bundle of data where cause inefficiency in understanding the essential data. In the field of the research, the keyword extraction has been examined by implementing various techniques into it for better efficiency. Moreover, it is an important research direction in text mining, natural language processing and information retrieval. In the recent days, many applications are utilizing the benefits of keyword extraction process, such as automatic indexing, automatic
summarization, automatic classification, automatic clustering, and automatic filtering. Moreover, keyword extraction has become a significant process of several text mining applications such as search engine, text categorization, summarization, and topic detection. Hence, the number of digital documents available has been progressively expanding and the manual keyword extraction can be an infeasible task. Here, topic detection is considered to be discussed.

1.2.6.2 Topic detection

Due to the heavy amount of information, and data related to a single topic often isolated dispersedly in many different places and appear at different times, which could be thoroughly concluded in difficulties. As a consequence, Topic Detection and Tracking Technology (TDT) have emerged. The dispersive information can be gathered and both the internal and outside connection could be discovered and analyzed efficiently by detecting and tracking topics. TDT technology have its edge in monitoring the source of multilingual information, warning at detecting new topics, which brings it broad prospect in several fields, like information and financial security, industry research etc.

1.2.6.3 Keyword extraction with particle swarm optimization

In times of technology, unstructured data are played vital role every data process, it is data are those that have no prearranged form or structure and are full of textual data. Typically, examples of unstructured systems are emails, reports, telephone or messaging conversations, etc. In recent days, researchers are focused on improving efficiency of keyword extraction method by proposing various methods to tackle their goal. This including Particle Swarm Optimization (PSO) method, researchers are believed this method would help to achieve their goal. The particle swarm optimization
(PSO) algorithm is a randomly determined optimization method in the global search methods environments. Moreover, PSO is a search process based on the idea of swarm intelligence in biological populations, which searches for the global optima by updating its generations. Therefore, Particle Swarm Optimization have been used with recommended documents, it will minimize redundancy in a short list of Keywords and provide accurate result.

1.2.6.4 Benefits of using particle swarm optimization

- Easy-to-implement algorithm
- It has less adjustable parameters
- It is computationally inexpensive in both speed and memory requirements.

1.2.6.5 Text classification in text mining

Since text mining is important for extracting essential information, it has an essential subfield too, text classification. It assigns a text document into one or more predefined classes or categories. Due to tremendous growth of information technology the amount of information also grown in every fields. Therefore, text classification has become the crucial process for obtaining essential data. Nowadays, new articles, digital libraries and Web pages are considered as the important sources of information, which are play vital role text classification. Moreover, many applications of text mining can be modeled as a text classification problem. These applications include news filtering, organization, document organization, retrieval, opinion mining (sentiment analysis), and spam filtering. Here, we briefly discussed about opinion mining.
1.2.6.6 Opinion mining

Nowadays, due to excessive amount of information are sharing between social network causes difficulties in understanding essence of matter presents in it. Therefore, extracting essential information from user comments or opinions in social network regarding specific thing or share their opinion over it by combining techniques from Natural Language Processing (NLP) and computer science, opinion mining or sentiment analysis (Jorge A Balazs & Juan D. Velasquez 2016). Moreover, it is also helpful for various practical applications such as product pricing, competitive intelligence market prediction, election forecasting etc. (Shilang Sun et al. 2017, Kumar Ravi & Vadlamani Ravi 2015).

Sentiment analysis further classified into three primary levels such as of documents, of sentiments, and of entities or aspects. Sentence-level analysis helps to identify whether a particular sentence expresses a positive, negative, or neutral opinion rather than document-level helps to detects the sentimental orientation of the entire text document. Moreover, the ultimate aim of opinion mining is to get some effective conclusion by processing reviews from customer opinions (Lincy Liptha et al. 2006). Subjective text is used for describing main opinion about people and their ideas by the expressive sentences that express positive or negative emotions (Shahab Saquib Sohaii et al. 2016, Zheng Yan & Xuyang Pedrycz 2017). Advanced analytical technology needed for to obtain accurate overview of social opinion from large volume of data in business. In research fields, the aggregation of data mining and opinion has become the serious research topic for to analyze user ratings and reviews. To improve software quality and address missing application features are enhanced by user ratings and reviews. In current situation, several techniques have been used by researchers to classify the
social network reviews such as Naive Byes, Maximum Entropy and SVM techniques (Farman Ali et al. 2016).

1.3 SUPPORT VECTOR MACHINE

The Support Vector Machine (SVM) technique is used for classifying opinion of user in social network. Moreover, the machine learning theory is a major reason for developing SVM models to predict values from very different fields. SVMs are a collection of related supervised learning methods used for classification and regression, and possess the well-known multivariate function for high accuracy. Because of supervised classification technique in SVM shows more promising accuracies than unsupervised requires labeled information provided by user. Moreover, it is an effective classification method with significant advantages such as

- The absence of local minima,
- An adequate generalization to new objects,
- A representation that depends on few parameters.

1.3.1 General Applications of SVM

SVM has been used successfully in many real-world problems

- Text (and Hypertext) Categorization
- Image Classification
- Bioinformatics (protein Classification, cancer Classification)
- Hand-written Character Recognition
1.4 APPLICATIONS AREAS OF OPINION MINING AND SENTIMENT ANALYSIS

- Purchasing Product or Service
- Quality Improvement in Product or service
- Marketing research
- Recommendation Systems
- Detection of “flame”
- Opinion spam detection
- Policy Making
- Decision Making

1.4.1 Purchasing Product or Service

The decision taken by the person for purchasing item helps to predict easily about the opinion, while purchasing a product or service, taking right decision is no longer a difficult task. By this technique, people can easily evaluate other’s opinion and experience about any product or service and also he can easily compare the competing brands. Now people don’t want to rely on external consultant. The Opinion mining and sentiment analysis extract people opinion form the huge collection of unstructured content, the internet, and analyze it and then present to them in highly structured and understandable manner.

1.4.1.1 Quality improvement in product or service

By Opinion mining and sentiment analysis the manufactures can collect the critic’s opinion as well as the favorable opinion about their product or service and thereby they can improve the quality of their product or
service. They can make use of online product reviews from websites such as Amazon and C|Net , RottenTomatoes.com and IMDb.

1.4.1.2 Marketing research

The result of sentiment analysis techniques can be utilized in marketing research. By sentiment analysis techniques, the recent trend of consumers about some product or services can be analyzed. Similarly the recent attitude of general public towards some new government policy can also be easily analyzed. These all result can be contributed to collective intelligence research.

1.4.1.3 Recommendation systems

By classifying the people’s opinion into positive and negative, the system can say which one should get recommended and which one should not get recommended.

Detection of “flame”: The monitoring of newsgroup and forums, blogs and social media is easily possible by sentiment analysis. Opinion mining and sentiment analysis can automatically detect arrogant words, over heated words or hatred language used in emails or forum entries or tweets on various internet sources

Opinion spam detection: Since internet is available to all, anyone can put anything on internet, this increased the possibility of spam content on the web. People may write spam content to mislead the people. Opinion mining and sentiment analysis can classify the internet content into’ spam’ content and ‘not spam’ content.
**Policy Making**: Through Sentiment analysis, policy makers can take citizen’s point of view towards some policy and they can utilize this information in creating new citizen friendly policy.

**Decision Making**: In decision making, people’s opinion and experience play the vital roles; it gives analyzed people’s opinion that can be effectively used for decision making.

1.5 **RESEARCH CHALLENGES IN OPINION MINING AND SENTIMENT ANALYSIS**

The main challenges that are faced by Opinion mining and sentiment analysis are the following:

1.5.1 **Detection of Spam and Fake Reviews**

The spam content in web can damage the efficiency of web; spam content should be eliminated before processing. This can be done by identifying duplicates, by detecting outliers and by considering reputation of reviewer.

1.5.2 **Limitation of Classification Filtering**

There is a limitation in classification filtering while determining most popular thought or concept. For better sentiment classification result this limitation should be reduced. The risk of filter bubble gives irrelevant opinion sets and it results false summarization of sentiment.

1.5.3 **Asymmetry in Availability of Opinion Mining Software**

The cost of opinion mining software is cannot be bearded by small scale industries; it is only affordable for big organization and
government. For utilizing it by everyone around the cost be controlled and reduced.

1.5.4 Incorporation of Opinion with Implicit and Behavior Data

For successful analysis of sentiment, the opinion words should integrate with implicit data. The implicit data determine the actual behavior of sentiment words.

1.5.5 Domain-Independence

In opinion mining and sentiment analysis, the domain dependent nature of sentiment words is considered as biggest challenge. One features set may give very good performance in one domain, at the same time it perform very poor in some other domain.

1.5.6 Natural Language Processing Overheads

In sentiment analysis, the natural language overhead like ambiguity, co-reference, implicitness, inference etc. created hindrance.

1.5.7 Natural Language Processing and Opinion Mining

Natural-language understanding systems convert samples of human language into more formal representations that are easier for computer programs to manipulate. And that is going to be fed to the further process of language processing in the field of Opinion Mining. It develops an in-depth understanding of both the algorithms available for the processing of linguistic information and the underlying computational properties of natural languages.
1.5.8 Text Summarization in Text Mining

Text summarization is a technique which either follows a more intuitive approach without step by step instruction or follows a rule-governed approach which may focus on tasks. In this method, deleting unnecessary or redundant information, identifying or producing topic sentences are the essential processes. Text summarization can be classified as many types they are:

- Abstraction vs. Extraction
- Mono-lingual vs. Multi-lingual
- Single-document vs. Multi-document
- Generic vs. Query-based
- Indicative vs. Informative

**Abstraction vs. Extraction**: Abstractive and Extraction are the two main divisions of text summarization. By extracting the main sentences from original document extractive system produces summary. In addition, they just replicate the original sentence by selecting the highest ranked sentence. Inconsistencies are considered as the major problem with extraction techniques. Therefore, abstraction has emerged for resolving such issue. Moreover, abstractive system generates summary by constructing new sentences like a human being which are short and concise. Summary might contain words and phrases that are not explicitly available in the main text.

**Mono-lingual vs. Multi-lingual**: In general, summarization can be classified in two categories such as mono-lingual system and multi-lingual system.
**Mono-lingual system**: Mono-lingual systems work only on one particular language, such as English.

**Multi-lingual system**: Multi-lingual system works on two or more languages such as English, Spanish, Japanese, etc.


**Generic vs. Query-based**: Sometimes one summary can be used for various types of users. Hence such types of summary are very important and independent to the subject of document. These summaries are called as generic based summary. In contrast, when user wants some specific information from document it called as query-based summary.

**Indicative vs. Informative**: Indicative and informative are the next classification of summarization system. As the name suggest indicative summary gives user only the main idea of text, whereas informative systems present concise information of the original and can be treated as a replacement for the original document.

### 1.6 TEXT MINING IN WIKIPEDIA

Nowadays, the growth of Internet in the modern world is considered as knowledge revolution. With the help World Wide Web information about anything can be updated, it has become the major knowledge repository for various fields. The main advantage of www is with a minimum effort lot of information can be obtained. Wikipedia has emerged as online encyclopedia; therefore, it contains lot of articles and documents.
However, abundance of information causes noise and unwanted features. Text mining plays a crucial role to reduce noise and unwanted features by extracting only essential information rather selecting whole documents.

1.6.1 Data Mining Vs Text Mining vs Web Mining

**Data Mining**

- Data mining systems are used to analyze information as homogeneous and universal.
- It is focused on data-dependent activities such as accounting, purchasing, supply chain, CRM, etc.
- It has been considered a proven, robust and industrial technology for many decades.

**Text Mining**

- Text mining tools have to face major technical challenges such as heterogeneous document formats (text documents, emails, social media posts, verbatim text, etc.), as well as multilingual texts and abbreviations and slang typical of SMS language.
- The complexity of the data processed make text mining projects longer to deploy.
- Text mining was historically thought of as complex, domain-specific, language-specific, sensitive, experimental, etc.

**Web Mining:** Some applications, such as multimedia and e-commerce data have differentiate web mining with text mining. It is wider field than text mining because of elements existing web mining are not there
in text mining. Since the rapid growth of Web is infeasible to predict, Web mining becomes more and more important. Text mining and Web mining are two different fields, although, it must be borne in mind that a great deal of the content on the Web is text-based. Therefore, text mining should also form an important part of Web mining.

1.6.2 Methods and Models used in Text Mining

Traditionally there are so many techniques developed to solve the problem of text mining that is nothing but the relevant information retrieval according to user’s requirement. According to the information retrieval basically there are four methods used

1) Term Based Method (TBM).
2) Phrase Based Method (PBM).
3) Concept Based Method (CBM).
4) Pattern Taxonomy Method (PTM).

**Term Based Method**: Term in document is word having semantic meaning. In term based method document is analyzed on the basis of term and has advantages of efficient computational performance as well as mature theories for term weighting. These techniques are emerged over the last couple of decades from the information retrieval and machine learning communities. Term based methods suffer from the problems of polysemy and synonymy. Polysemy means a word has multiple meanings and synonymy is multiple words having the same meaning. The semantic meaning of many discovered terms is uncertain for answering what users want. Information retrieval provided many term-based methods to solve this challenge.
**Phrase Based Method:** Phrase carries more semantics like information and is less ambiguous. In phrase based method document is analyzed on phrase basis as phrases are less ambiguous and more discriminative than individual terms. The likely reasons for the daunting performance include:

1) Phrases have inferior statistical properties to terms,

2) They have low frequency of occurrence, and

3) Large numbers of redundant and noisy phrases are present among them.

**Concept Based Method:** In concept based terms are analyzed on sentence and document level. Text Mining techniques are mostly based on statistical analysis of word or phrase. The statistical analysis of the term frequency captures the importance of word without document. Two terms can have same frequency in same document, but the meaning is that one term contributes more appropriately than the meaning contributed by the other term (Lincy Liptha et al. 2006). The terms that capture the semantics of the text should be given more importance so, a new concept-based mining is introduced. This model included three components. The first component analyzes the semantic structure of sentences. The second component constructs a Conceptual Ontological Graph (COG) to describe the semantic structures and the last component extract top concepts based on the first two components to build feature vectors using the standard vector space model. Concept-based model can effectively discriminate between non important terms and meaningful terms which describe a sentence meaning (Menaga & Hemapriya 2013). The concept-based model usually relies upon natural language processing techniques. Feature selection is applied to the query concepts to optimize the representation and remove noise and ambiguity.
**Pattern Taxonomy Method:** In pattern taxonomy method documents are analyzed on pattern basis. Patterns can be structured into taxonomy by using is-a relation. Pattern mining has been extensively studied in data mining communities for many years. Patterns can be discovered by data mining techniques like association rule mining, frequent item set mining, sequential pattern mining and closed pattern mining. Use of discovered knowledge (patterns) in the field of text mining is difficult and ineffective, because some useful long patterns with high specificity lack in support (i.e., the low-frequency problem). Not all frequent short patterns are useful hence known as misinterpretations of patterns and it leads to the ineffective performance. In research work, an effective pattern discovery technique has been proposed to overcome the low-frequency and misinterpretation problems for text mining. The pattern based technique uses two processes pattern deploying and pattern evolving. This technique refines the discovered patterns in text documents. The experimental results show that pattern based model performs better than not only other pure data mining-based methods and the concept-based model, but also term-based models.

1.6.3 **Applications of Text Mining**

- **Commercials**
  - Publishing and media.

- **Government sectors**
  - Political institutions,
  - Political analysts,
  - Public administration and legal documents.
- **Entertainments**
  - Telecommunications and other services industries.

- **Business**
  - Information technology sector and Internet.

- **Finance**
  - Banks,
  - Insurance
  - Financial markets.

- **Medical**
  - Pharmaceutical
  - Research companies
  - Healthcare.

1.6.4 **Some Benefits of Text Mining**

- Efficiency
- Unlocking hidden information and developing new knowledge
- Exploring new horizons
- Improved research and evidence base
- Improving the research process and quality
1.7 ORGANISATION OF THESIS

The thesis is organized as follows: Chapter 1 gives an introduction to the efficient text pattern mining, and the various pattern mining approaches for record retrieval. Chapter 2 detailed literature survey is presented. Chapter 3 discusses the objective of our proposed research methodology. Chapter 4 presents limitations of present investigation. Chapter 5 discusses the organization of thesis. Chapter 6 discusses the research methodology and requirement of the research. Chapter 7 work gives the conclusion.