Abstract—Text mining also known as knowledge discovery in text (KDT) is an emerging technology and it focuses on discovering a text from unstructured text. The reason for text mining is stored information is increasing day by day, so discovering a text from unstructured or semi-structured data is important one. Therefore several techniques and algorithms are required for extracting useful information. In our work, we discussed about text mining approach, various techniques and applications.

Keywords—KDD, Stemming, Semi structured data, Accuracy, Filtering.

I. INTRODUCTION

Text mining is the process of extracting interesting, non-trivial extraction of implicit, previously unknown and potentially useful information from large amount of textual data (from different written resources). With the help of the practical example, we can understand the need for text mining. The Research and Development (R&D) Industry research scholar focuses on research paper -80% of information is only in research paper (unstructured data).

If a research scholar manually read 50 research papers, 10% of those data are useful. But online databases like IJARCSSE adds more research papers are using Text mining and then the research scholar has the chance to read only their relevant papers. Thus the performance of useful information is increased when we use text mining. It presents the need for Text mining.

Data mining is deals with structured data only but text mining deals with un-structured or semi structured text. In this paper we discussed about text mining process, technologies, applications and challenges.

II. TEXT MINING PROCESS

Text mining process includes text pre-processing, Text Transformation, Feature Selection, Pattern Discovery and Evaluation. Figure 1 shows the process of Text mining.

Text Preprocessing
- Text Preprocessing transforms a text into a structured format. The text pre processing is achieved into the following steps:
  - Tokenization: It removes all punctuation, white spaces, spaces and commas etc.
  - Filtering: A standard filtering method is stop word filtering, It remove content information like html or xml tags, articles, conjunctions, prepositions etc.
  - Lemmatization: It maps verb forms to the infinite tense and nouns to the singular form.
  - Stemming: It build the basic forms of words, E.g. like, liked, like, liking, dislike, likes belongs to like.

Text Transformation
- Text transformation also known as feature generation, it converts the text document into bag of words or Vector space document model notation.
- Use of this step: This can be used for further effective analysis task.

Feature Selection
- Feature selection also known as attribute selection, it performs removing features that are considered irrelevant for mining purpose.
- Use of this step: This procedure gives advantage of smaller dataset size, less computations and minimum search space required.

Pattern discovery
- There are different text mining methods as in Data mining had been proposed such as: Clustering, Classification, Information retrieval, Topic discovery, Summarization, Topic extraction.
This phase includes Evaluation and Interpretation of results in terms of calculating Precision and Recall, Accuracy, F measure etc.

III. TEXTMINING TECHNOLOGIES

In present days humans can replaced by computers for analysing the text mining tasks. But previously Text Mining used manual techniques. Manual technique was expensive and too much time. Text mining techniques plays an important role, because it is a continuing process. Currently the text mining technologies typically using following methods, in this section we cover text mining technologies.

A. Information Extraction

In general information extraction is the process of text simplification. Information extraction is the task of extracting structured information from semi structured/unstructured data. It uses the pattern matching method. It matches with the predefined text with user text. The overall goal of IE is create a more easily machine-readable text to process the sentences.

B. Text Summarization

After IE, Text summarization takes placed. In every organization no one have the time to read all the documents and makes decision. For that purpose text summarization used for grouping the documents and gives the basic idea behind the every document. It is the process of creating a compressed version of a given text that provides useful information for the user. The text summarization classified as extractive summarization and abstractive summarization. In extract summarization it gets the essential contents of the documents from a paragraph or documents. Extractive methods work by selecting a subset of existing words, phrases, or sentences in the original text to form the summary. In abstractive summarization gets the main idea behind the single document or group of documents. It gives an understanding the main concepts of documents. In abstractive methods build an internal semantic representation and then use natural language generation techniques to create a summary that is closer to what a human might generate. Such a summary might contain words not explicitly present in the original. Figure 2 represents the overall process of text summarization.

C. Topic Tracking

Topic tracking mainly used in radio and news broadcasts. The idea behind the topic tracking is it maintains users profile based on their search interest and predicts another documents of interest to the user. In user profile it just maintains what the user previously searched.

D. Classification

Classification is the process of analysing a document and adding metadata 'tags' that describe that document which is sourced from taxonomy or other form of controlled vocabulary. Text classification has many direct applications such as: Figure 3 shows the basic classification model.

E. Clustering

It is an unsupervised technique, no input and output labels are pre defined. In terms clustering dividing the same text into one group and makes cluster. Clustering is a method to group the similar documents but differ from categorization. Figure 4 shows that the basic concept of clustering.

F. Concept linkage

Concept linkage is most valuable concept in text mining. Concept linkage tools provide the facility to connect the related documents. When going with concept linkage it
provides the facility to browse the documents instead of searching. Concept linkage is very useful in bio medical field, criminal’s identification and etc. Now a day’s online doctor is an emerging field, the concept linkage in online doctor’s field maps the disease and treatments. Also government can use concept linkage to maps the present criminal activity with their previous activity so the government get idea about possible relationships with criminals and crimes.

G. Information visualization

Information visualization also known that visual text mining, the scope of information visualization is giving the visual representations of abstract data. Data preparation, Data Analysis & extraction, visualization mapping are the three main steps of information visualization. The target for information visualization is crime analysis, market studies and etc. Figure 5 represents that the technique of visualization.

Figure 5: Visualization

H. Natural Language processing(NLP)

Natural language processing concerned with the interactions between computers and natural (human) languages. In terms of text mining it gives summarization after knowing the meaning of the document for command and queries understanding also the analysis purpose.

IV. TEXTMINING APPLICATIONS

Text mining is an emerging technology. Now a day’s many business domains uses the applications of text mining. Some of them listed below.

1. Customer profile analysis

Now a day’s Customer profile analysis is a critical business area. The customer profile analysis focuses on marketing plan and company’s business plan. Companies use text mining to find out the occurrences, events and complaints from articles, forums, web pages and etc. From the source they gathered the unstructured or semi structured text, then applying the necessary pre processing technologies. Get the necessary format and they made decision based on the mining

2. Security Applications

As technology grows the crimes are differ from earliest days. In recent days the terrorism spread via. Blogs, Facebook, Twitter and personal websites etc., Text mining used for the security applications. It just monitors and analyse the online plain documents. In market the text mining software for security applications are available

3. Bio Medical Application

There is an increasing interest in text mining and information extraction strategy applied to the biomedical and molecular biology literature due to the increasing number of electronically available publications stored in databases such as PubMed


Mining a company’s reports and correspondences for activities, status, and problems reported, so its resource status and problems reported can be handled properly and future action

5. Patent analysis

Patent analysis is a management technique for addressing the strategic management of the firm’s technology and product or service development process. Text Mining translating patent data into competitive intelligence allows the firm to gauge its current technical competitiveness, to forecast technological trends, and to plan for potential competition based on new technologies. Also Text Mining Analyzing patent databases for major technology players, trends, and opportunities

6. Open-ended survey responses

When the user buy a product the user gives the feedback or complaint about the product. Now a days the feedbacks and complaints collected via. Online and analyzing the things with the help of text mining. As per response of customers, industry takes the advantage of this for marketing

7. Competitive Intelligence

Enabling companies to organize and modify the company strategies according to present market demands and the opportunities based on the information collected by the company about themselves, the market and their competitors, and to manage enormous amount of data for analyzing to make plan

8. Customer Relationship Management

The product review, compliance, feed back everything is now a day collected from the customers. So all these text are summarized and the knowledge is fetch from that source using text mining technique. This knowledge builds the customer-owner relationship


In organizations the HRM board is very important one. It focuses on staff’s opinions and monitoring the level of employee to improve the growth of the particular firm. Also selecting the CV for interview process also done by with the help of text mining process

10. Analysis of the junk Emails

Another common application for text mining is in automatic analysis of the junk E-mails which are undesirable. Classification technique of text mining can be used to classify such mails on the basis of pre-defined frequently occurring terms.

V. MERITS, DEMERITS AND CHALLENGES OF TEXT MINING

1. Merits

A recent study indicates that 80% of a company’s information is stored as text such as emails, memos, customer correspondence, responses and feed back etc.;
from the text source getting knowledge is very challenge. Text mining solves this problem easily whether it is a semi structured or unstructured text. Text mining process that documents and the knowledge is obtained. Also the development of net everything is electronically documented. The size is growing day by day, so the information could not store in the database. Because the database can store less information. Text mining solves this problem easily. Every electronic document is pre processed and gathered as information.

2. Demerits and Challenges

There is no specific text mining process for non English data also Ambiguity is the biggest challenge in text mining. Ambiguity means the capability of being understood in two or more possible senses or ways. Ambiguity gives a natural language its flexibility and usability, and consequently, therefore it cannot be entirely eliminated from the natural language. One word may have multiple meanings. One phrase or sentence can be interpreted in various ways, thus various meanings can be obtained. So avoiding the ambiguity is the biggest challenge in text mining.

VI. CONCLUSION

Text mining also known as Knowledge Discovery in Text (KDT) is a technique which is used to find the knowledge form semi structured or unstructured text. Here in this work briefly discussed about text mining process like text pre processing, feature selection and etc., Technologies including classification, clustering summarization and etc., applications such as customer profile analysis, Security, Spam mails and Marketing etc., Also merits, demerits and the challenges of text mining discussed. The future directions of text mining tools are text mining tools not only for trained specialists. It is for all because text mining tools can be used to management team from the organization too.

REFERENCES

2. Haralampos Karanikas and Babis Theodoulidis Manchester, “Knowledge Discovery in Text and Text Mining Software”, Centre for Research in Information Management (2001), UK.
4. Divya Nasa “Text Mining Techniques- A Survey”, USICT , GGSIPU.