# A SURVEY TAXONOMY ON TEXT MINING TECHNIQUES

**Ch.S.K.V.R.Naidu[1],T.Y.Ramakrushna[2],T.Chandana Gouri[3],K.Komali[4]**
[1]Associate Professor in Dept. of CSEEmail:,naiduch@iaitc.in
[2]Associate Professor in Dept. of CSE,Email:tyramakrushna@iaitc.in
[3] Assistant Professor in Dept. of CSE, Email:tchandana@iaitc.in
[4]Assistant Professor in Dept. of CSE,Email:komalik@iaitc.in
[1,2,3,4]Indo American Institutions Technical Campus, Anakapalle, Affiliated to JNTUK

**ABSTRACT:**

In today's reality, the measure of put away data has been massively expanding step by step which is for the most part in the unstructured shape and can't be utilized for any handling to extricate helpful data, so a few methods, for example, rundown, grouping, bunching, data extraction and perception are accessible for the same which goes under the classification of text mining. Text Mining can be characterized as a procedure which is utilized to extricate intriguing data or learning from the text records. In this work, an exchange over structure of text mining with the methods as above with their professionals and cons furthermore uses of Text Mining is finished. Also, concise exchange of Text Mining advantages and restrictions has been displayed. Text mining is an extremely energizing examination range. The issue of text mining, i.e. finding helpful information from unstructured text, is drawing in expanding consideration. It tries to find learning from unstructured writings. These writings can be found on a PC desktop, intranets and the web. Average text mining assignments incorporate text order, text grouping, idea/element extraction, generation of granular scientific classifications, feeling investigation, record synopsis, and element connection displaying (i.e., learning relations between named substances). Text investigation includes data recovery, , labeling/annotation, data extraction, information mining strategies including join and affiliation examination, representation, and prescient examination. The all-encompassing objective is, basically, to transform text into information for investigation, by means of use of regular dialect handling (NLP) and systematic strategies. The point of this paper is to give an outline of text mining in the connections of its systems, application spaces and the most difficult issue. The attention is given on essentials systems for text mining which incorporate common dialect having and data extraction. This paper likewise gives a short audit on areas which have utilized text mining.

**Index Terms—** text mining, information extraction, applications, benefits and limitations, classification.

## I. INTRODUCTION

Text Mining [1] is the revelation by PC of new, already obscure data, via naturally separating data from diverse composed assets. A key component is the connecting together of the removed data together to frame new certainties or new speculations to be investigated further by more ordinary method for experimentation. Text mining is not quite the same as what are acquainted with in web seek. In pursuit, the client is regularly searching for something that is as of now known and has been composed by another person. The issue is pushing aside all the material that as of now is not significant to your requirements keeping in mind the end goal to locate the applicable data. In text mining, the objective is to find obscure data, something that nobody yet knows thus couldn't have yet recorded. Text mining is a minor departure from a field called information mining [2] that tries to discover fascinating examples from expansive databases. Text mining, otherwise called Intelligent Text Analysis, Text Data Mining or Knowledge-Discovery in Text (KDT), alludes for the most part to the procedure of extricating fascinating and non-trifling data and learning from unstructured text. Text mining is a youthful interdisciplinary field which draws on data recovery, information mining, machine learning, measurements and computational semantics. As most data (more than 80%) is put away as text,

text mining is accepted to have a high business potential worth. Learning might be found from numerous wellsprings of data, yet, unstructured writings remain the biggest promptly accessible wellspring of information. The issue of Knowledge Discovery from Text (KDT) [6] is to extricate express and verifiable ideas and semantic relations between ideas utilizing Natural Language Processing (NLP) methods. Its point is to get bits of knowledge into vast amounts of text information. KDT, while profoundly established in NLP, draws on routines from insights, machine learning, thinking, data extraction, information administration, and others for its revelation process. KDT assumes an inexorably critical part in developing applications, for example, Text Understanding. Text mining [1] is like information mining, aside from that information mining devices [2] are intended to handle organized information from databases, yet message mining can work with unstructured or semi-organized information sets, for example, messages, full-message records and HTML documents and so forth. Thus, message mining is a vastly improved answer for organizations. To date, nonetheless, most innovative work endeavors have focused on information mining endeavors utilizing organized information. The issue presented by text mining is self-evident: regular dialect was produced for people to correspond with each other and to record data, and PCs are far from understanding normal dialect. People can recognize and apply semantic examples to text and people can without much of a stretch overcome deterrents that PCs can't undoubtedly handle, for example, slang, spelling varieties and logical importance. Then again, in spite of the fact that our dialect capacities permit us to fathom unstructured information, we do not have the PC's capacity to process text in substantial volumes or at high speeds. Figure 1 on next page, delineates a non-specific procedure model [3] for a text mining application. Beginning with a gathering of records, a text mining instrument would recover a specific archive and preprocess it by checking design and character sets. At that point it would experience a text investigation stage, in some cases rehashing procedures until data is extricated. Three text examination procedures are appeared in the illustration, however numerous different blends of strategies could be utilized relying upon the objectives of the association. The subsequent data can be set in an administration data framework, yielding an inexhaustible measure of information for the client of that framework.

## II. RELATED WORK

This paper investigates late endeavors and commitments on text mining systems. In this way various exploration article and examine papers and their commitments are set in this segment. Numerous information mining methods have been made arrangements for mining significant examples in text archives. In any case, how to effectively utilize and overhaul uncovered examples is still an open examination issue, particularly in the area of text mining. Following most existing text mining routines received term-based methodologies, they all experience the ill effects of the inconveniences of polysemy and synonymy. This paper introduces an imaginative and significant example revelation method which incorporates the procedures of example sending and design developing, to propel the viability of utilizing and upgrading found examples for finding proper and intriguing data. Considerable tests on RCV1 information accumulation and TREC points show that the proposed arrangement accomplishes empowering execution [2]. The ―helpfulness‖ normal for online client surveys offers purchasers some assistance with dealing with data over-burdens and encourages choice making. Be that as it may, numerous online client audits require adequate support votes in favor of different clients to survey their actual supportiveness level. Text mining strategies are utilized to expel semantic qualities from audit writings. Our discoveries additionally prompt that surveys with solid conclusions get more graciousness votes than those with blended or nonpartisan feelings. This paper reveals insight into the obliging of online clients' accommodation voting exercises and the configuration of an upgraded support voting component for online client survey frameworks [3]. Learning bases and controlled rundowns are having vital impact in numerous applications,

for example, text outline, question replying, paper evaluating, and semantic hunt. Albeit, numerous frameworks (e.g., DBpedia and YaGo2) offer unfathomable learning bases of such outlines, they all experience the ill effects of deficiency, irregularities, and unsoundness. These inconveniences can be tended to and quite upgraded by joining and coordinating distinctive learning bases, yet their extensive sizes and their dependence on differing wordings and ontologies make the undertaking exceptionally troublesome. In this demo, we will display a framework that is making great progress on this assignment by: i) utilizing accessible interlinks in the present information bases (e.g. outside Link and divert joins in DBpedia) to consolidate data on individual elements, and ii) utilizing generally accessible text corpora (e.g. Wikipedia) and our IBminer text mining framework, to create and check organized data, and settle phrasings crosswise over distinctive information bases. We will likewise express two apparatuses intended to manage the mix process in close cooperation with IBminer. The primary is the InfoBox Knowledge-Base Browser (IBKB) which offers organized synopses and their provenance, and the second is the InfoBox Editor (IBE), which is intended to prompt applicable properties for a client determined subject, whereby the client can undoubtedly enhance the learning base without requiring any information about the inward wording of individual frameworks [4]. Breaking down substantial literary accumulations has grow progressively difficult given the span of the information existing and the rate that more information is being made. Point based text outline routines combined with agreeable perceptions have offered promising ways to deal with location the test of assessing extensive text corpora. As the text corpora and vocabulary develop bigger, more points require to be made in direction to catch the noteworthy dormant subjects and subtleties in the corpora. On the other hand, it is intense for a large portion of late subject based perceptions to speak to expansive number of themes without being scattered or unintelligible. To empower the representation and route of countless, we offer a visual examination framework – Hierarchical Topic (HT).User associations are conveyed for clients to make varieties to the theme pecking order in view of their mental model of the point space we have additionally guided a client study to quantitatively compute the impact of various leveled subject.

## III. TEXT MINING FRAMEWORK

Definition: Text Mining is the process of extracting interesting information or knowledge or patterns from the unstructured text that are from different sources. As the text is in unstructured form, it is quite difficult to deal with it. Finding ―nuggets‖ of interesting information from the natural language text is the purpose of text mining. The Text Mining Process is shown in Fig. 1:g

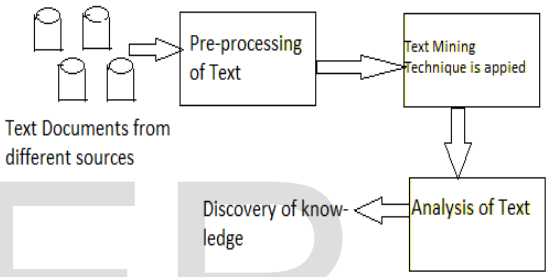


Fig. 1 Text Mining Process

Stage-I: Pre-processing Text: Mining from a pre-processed text is easy as compare to natural languages documents. So, pre-processing of documents that are from different sources is an important task during text mining process before applying any text mining technique.

As Text documents can be represented as ―bag of words‖ on which different text mining methods are based. Let Ω be the set of documents & W= {w1, w2, ---- wm} be the different words from the document set. In order to reduce the dimensionally of the documents words, special methods such as filtering and stemming are applied. Filtering methods remove those words from the set of all words, which do not provide relevant information; stop word filtering is a standard filtering method. Words like prepositions, articles, conjunctions etc. are removed that contain no informatics as such. Stemming methods: are used to produce the root from the plural or the verbs. For e.g. Doing, Done, Did may be represented as ‗Do'. After this method is applied, every word is represented

by its root word. Originally it was proposed by M. poster [4].
Stage II- Text Mining Technique is applied: This is an important stage in which the selected algorithm is applied on text in order to process the text. The algorithm such as clustering, classification, summarization, information extractions or visualizations which are explained next could be used.
Stage III - Analysis of Text: Here the outputs are analysed for discovering the knowledge. Various tools such as link discovery tool can be used or the outputs can be visualised so that the users could navigate through in order to achieve the perspective.

## IV. TEXT MINING TECHNIQUES

There are different kinds of techniques available by which the text pattern analysis and mining is performed. Some of the essential techniques are discussed in this section.

### A. Information Extraction

A starting point for computers to examine unstructured text is to use information extraction. Information extraction software identifies key phrases and relationships within text. The software infers the relations between all the identified people, places, and time to deliver the user with significant information. This technology can be very helpful when dealing with large volumes of text. Traditional data mining assumes that the information to be ―mined‖ is previously in the form of a relational database. Unfortunately, for many applications, electronic information is only obtainable in the form of free natural language documents rather than structured databases. Since IE addresses the difficulty of transforming a corpus of textual documents into a extra structured database, the database constructed by an IE module can be provided to the KDD module for advance mining of knowledge as illustrated in Figure 2.

### B. Topic Tracking

A topic tracking system mechanism by custody of user profiles and, based on the documents the user views, guess other documents of interest to the user. Yahoo offers free topic tracking tool that permits users to choose keywords and informs them when news relating to those topics becomes existing. Topic tracking methodology have its own limitations, however. For example, if a user sets up an alert for ―text mining‖, s/he will receive numerous news stories on mining for minerals, and very few that are really on text mining. Some of the improved text mining tools let users select specific categories of interest or the software routinely can even infer the user's concern based on his/her reading history and click-through information.

### C. Summarization

Text summarisation is enormously helpful for trying to figure out whether or not a extensive document meets the user's needs and is worth reading for advance information. With huge texts, text summarization software procedures and précises the document in the time it may take the user to read the first paragraph. The key to summarisation is to decrease the extent and feature of a document while retaining its main points and overall meaning. The challenge is that, although computers are able to recognize people, places, and time, it is still complex to teach software to analyze semantics and to interpret meaning.

### D. Categorization

Categorization engage identifying the main themes of a document by placing the document into a pre-defined set of topics. When categorizing a document, a computer program will often delight the document as a ―bag of words.‖ Rather, categorization only calculate words that emerge and, from the counts, identifies the main topics that the document covers. Categorization often relies on a vocabulary for which topics are predefined, and relationships are recognized by looking for broad terms, narrower terms, synonyms, and related terms. Categorization utensils normally have a technique for grade the documents in order of which documents have the most content on a specific topic.

E. Clustering

Clustering [7] is a technique used to group similar documents, but it differs from categorization in that documents are clustered on the fly instead of through the use of predefined topics. Another advantage of clustering is that documents can emerge in multiple subtopics, thus ensuring that a useful document will not be absent from search results. A basic clustering algorithm generates a vector of topics for each document and determines the weights of how well the document fits into each cluster. Clustering technology can be useful in the organization of management information systems, which may contain thousands of documents.

F. Concept Linkage

Concept linkage tools [3] attach related documents by identifying their commonly-shared idea and help users find information that they perhaps wouldn't have establish using conventional searching methods. It promotes browsing for information rather than searching for it. Concept linkage is a valuable idea in text mining, especially in the biomedical fields where so much study has been done that it is impossible for researchers to read all the material and make organizations to other research. Ideally, concept linking software can identify links between diseases and treatments when humans cannot. For example, a text mining software solution may easily identify a link between topics X and Y, and Y and Z, which are well-known relations. But the text mining tool could also detect a potential link between X and Z, something that a human researcher has not come across yet because of the large volume of information s/he would have to sort through to make the connection.

G. Information Visualization

Visual text mining, or information visualization [3], puts large textual sources in a visual hierarchy or map and provides browsing capabilities, in addition to simple searching. DocMiner as shown in figure12, is a tool that shows mappings of large amounts of text, allowing the user to visually analyze the content. The user can interact with the document map by zooming, scaling, and creating sub-maps. Information visualization is useful when a user needs to narrow down a broad range of documents and explore related topics. The government can use information visualization to identify terrorist networks or to find information about crimes that may have been previously thought unconnected. It could provide them, with a map of possible relationships between suspicious activities so that they can investigate connections that they would not have come up with on their own.

H. Question Answering

Another application area of natural language processing is natural language queries, or question answering (Q&A), which deals with how to find the best answer to a given question. Many websites that are equipped with question answering technology, allow end users to ―ask‖ the computer a question and be given an answer. Q&A can utilize multiple text mining techniques. For example, it can use information extraction to extract entities such as people, places, events; or question categorization to assign questions into known types (who, where, when, how, etc.). In addition to web applications, companies can use Q&A techniques internally for employees who are searching for answers to common questions. The education and medical areas may also find uses for Q&A in areas where there are frequently asked questions that people wish to search.

I. Association Rule Mining

Association rule mining (ARM) [33] is a technique used to discover relationships among a large set of variables in a data set. It has been applied to a variety of industry settings and disciplines but has, to date, not been widely used in the social sciences, especially in education, counseling, and associated disciplines. ARM refers to the discovery of relationships among a large set of variables, that is, given a database of records, each containing two or more variables and their respective values, ARM determines variable-value combinations that frequently

occur. Similar to the idea of correlation analysis (although they are theoretically different), in which relationships between two variables are uncovered, ARM is also used to discover variable relationships, but each relationship (also known as an association rule) may contain two or more variables.

## V. APPLICATIONS

Text mining application uses unstructured textual information and examines it in attempt to discover structure and implicit meanings hidden within the text. Through text mining, we can uncover hidden patterns, relationships, and trends in text. text mining enables organizations to explore interesting patterns, models, directions, trends, rules, contained in text in much the same way that data mining explores tabular or "structured" data.

### BIOINFORMATICS

Research work for IE has grown dramatically in a bioinformatics domain, where biomedical journal articles have become an important application area in the recent years. In the bioinformatics domain, biomedical research literature has been a target for text mining. The first textbook on biomedical text mining with a strong genomics focus appeared in 2005. The goal of text mining in this area is to allow biomedical researchers to extract knowledge from the biomedical literature in facilitating new discovery in a more efficient manner. In evaluating biomedical text mining, Hersh, claimed that, most research in text mining still focuses on the development of specific functions or algorithms. Although some text mining systems have been developed, such as MedScan andTextpresso

### BUSSINESS INTELLIGENCE

Of the major concerns in any business is to minimize the amount of guessing work involved in decision making. The risk of making wrong prediction should be reduced. Most of the data mining techniques are created to deal with prediction. The problem with data mining is that it can help only up to a certain point, since most of data are available in texts. Data mining and text mining techniques can complement each other,LIKE, data mining techniques may be used to reveal the occurrence of a particular event while text mining techniques may be used to look for an explanation of an event.

### NATIONAL SECURITY

The use of text mining tool in national defence security domain has become an important issue. Government agencies are investing considerable resources in the surveillance of all kinds of communication, such as email, chats in chat rooms. Email is used in many legitimate activities such as messages and documents exchange. Unfortunately, it can also be misused, for example in the distribution of unsolicited junk mail. Thus automatic text mining tools offer a considerable promise in this area. Text mining technology is becoming an emergence technology for national security defence. Example of text mining system is COPLINK system. It was done at University of Arizona in Tucson to help the police to discover the link between agencies.

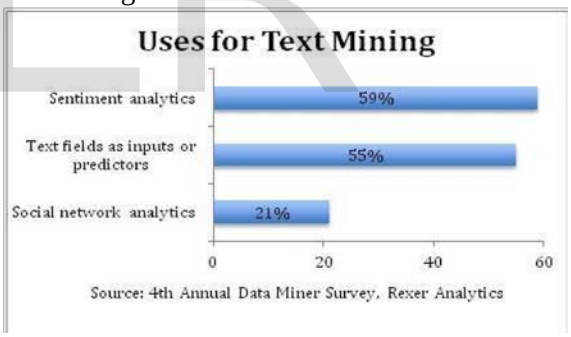


**Figure 2:** Uses of text mining

## VI. CHALLENGING ISSUES

The major challenging issue in text mining arise from the complexity of a natural language itself. The natural language is not free from the ambiguity problem. 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. Most of the IE systems that involve semantic analysis exploit the simplest part of the whole spectrum of domain and task knowledge, that is to say, named entities. IE does a more limited task than full text understanding. However, the growing need for IE application to domains such as functional genomics requires more text understanding. the ambiguity is still the major "world problem" in text mining applications.

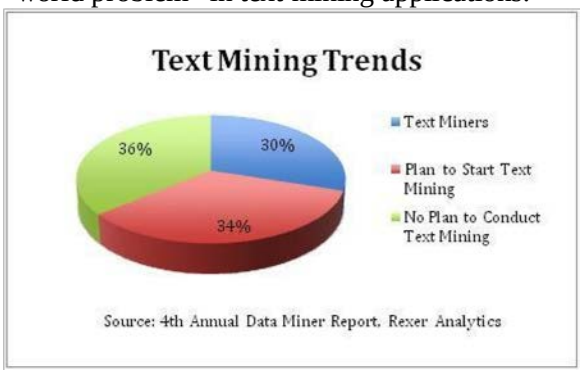


**Figure 3:** Trends in text mining

## VII. CONCLUSION

This paper has presented overview techniques, applications and challenging issue in text mining. This paper has presented overview techniques, applications and challenging issue in text mining. The focus has been given on fundamental methods for conducting text mining. The methods include natural language processing and information extraction. In this paper, we showed that text-mining systems can be developed relatively rapidly and evaluated easily on existing IE corpora by utilising existing Information Extraction (IE) and data mining technology. We presented an approach of using an automatically learned IE system to extract structured databases from a text corpus. Due to variability and diversity in natural-language data, some form of soft matching based on textual similarity is needed when discovering rules from text. The paper also addressed the most challenging issue in developing text mining systems.

## REFRENCES

[1]Http://en.wikipedia.org/wiki/Text_mining#Security_applications

[2]http://people.ischool.berkeley.edu/~hearst/text-mining.html

[3]http://sitecore.jisc.ac.uk/publications/briefingpapers/2008/bptextminingv2.aspx

[4] M. Grobelnik, editor. Proceedings of IEEE International Conference on Data Mining (ICDM-2001) Workshop on Text Mining (TextDM'2001), San Jose, CA, 2001.

[5] M. A. Hearst. What is text mining? http://www.sims.berkeley.edu/˜heast/text-mining.html,Oct. 2003.

[6] HaralamposKaranikas and Babis Theodoulidis Manchester, (2001), "Knowledge Discovery in Text and Text Mining Software", Centre for Research in Information Management, UK

[7] Liritano S. and Ruffolo M., (2001), " Managing the Knowledge Contained in Electronic Documents: a Clustering Method for Text Mining", IEEE, 454-458 , Italy.

[8] Brin S., and Page L.(1998), " The anatomy of a largescale hyper textual Web search engine", Computer Networks and ISDN Systems, 30(1-7): 107-117.

[9] Kleinberg J.M., (1999), "Authoritative sources in hyperlinked environment", Journal of ACM, Vol.46, No.5, 604-632.

[10] Dean J. and Henzinger M.R. (1999), "Finding related pages in the world wide web", Computer Networks, 31(11-16):1467-1479.