Text Mining Techniques for Extraction and Retrieval Information in Research Process

Ms. Rama Chauhan[1], Dr. Hariom Tyagi[2]
[1]M. Tech Scholar, [2]Professor
Department of CSE, RD Engineering College, Ghaziabad, UP, India.

***Abstract* -** The amount of text that is generated every day is increasing dramatically. This tremendous volume of mostly unstructured text cannot be simply processed and perceived by computers. Therefore, efficient and effective techniques and algorithms are required to discover useful patterns. Text mining is the task of extracting meaningful information from text, which has gained significant attentions in recent years. In this paper, we describe several of the most fundamental text mining tasks and techniques including text pre-processing, classification and clustering. Additionally, we briefly explain text mining in biomedical and health care domains.

In this paper we discussed about the text mining techniques and its applications. Text mining is used to extract interesting information or knowledge or pattern from the unstructured texts that are from different sources. It converts the words and phrases in unstructured information into numerical values which may be linked with structured information in database and analyzed with ancient data mining techniques. There are many techniques used in text mining such as information extraction, information retrieval, natural language processing (NLP), query processing, and categorization and clustering.

Keywords - Text mining, Natural Language, Natural Language Processing

I. INTRODUCTION

Data mining technology helps to extract useful information from various databases. Data warehouses turned out to be doing well for numerical information, but unsuccessful when it came to textual information. The 21st century has taken us beyond the limited amount of information on the web. This is good in one way that more information would provide greater awareness, and better knowledge. Text data mining refers to the process of extracting interesting and non-trivial patterns or knowledge from text documents. As text mining is extraction of useful information from text data it is also known as text data mining or knowledge discovery from textual databases. It is challenging issue to find accurate knowledge in text documents to help users to find what they want.

Nowadays most of the information in business, industry, government and other institutions is stored in text form into database and this text database contains semi structured data. A document may contain some largely unstructured text components like abstract additionally few structured fields as title, name of authors, date of publication, category, and so on. Text mining is a variation on a field called data mining that tries to find interesting patterns from large databases. The great deal of studies done on the modeling and implementation of semi structured data in recent database research. On the basis of these researches information retrieval techniques such as text indexing methods have been developed to handle unstructured documents. In traditional search the user is typically look for already known terms and

The text mining studies are gaining more importance recently because of the availability of the increasing number of the electronic documents from a variety of sources. Which include unstructured and semi structured information. The main goal of text mining is to enable users to extract information from textual resources and deals with the operations like, retrieval, classification (supervised, unsupervised and semi supervised) and summarization Natural Language Processing (NLP), Data Mining, and Machine Learning techniques work together to automatically classify and discover patterns from the different types of the documents [1].

A. Challenging Issues - Complexity of natural language is main challenging issue in text mining. The natural language is not free from the ambiguity problem. One word may have multiple meanings and multiple words can have same meaning. The capability of being understood in two or more possible ways means ambiguity. This ambiguity leads to noise in extracted information. Ambiguity cannot be entirely eliminated from the natural language as it gives flexibility and usability. There are various ways to interpret one phrase or sentence thus various meanings can be obtained. Although a number of researches have been conducted in resolving the ambiguity problem, the work is still immature and the proposed approach has been dedicated for a specific domain. It is challenge to answer what user wants as semantic meanings of many discovered words are uncertain.

B. Process Of Text Mining -

i). Document Gathering In the beginning step, the documents are collected that are present in numerous formats. The document could be in form of word, Html, CSS, PDF etc. [17].

ii). Document Pre Processing In this process, the given document is processed for eliminating redundancies, inconsistencies, separate words, stemming and documents are ready for next step, and the stages performed as follows[19], **Tokenization:** The given document is recognizing as a string and identifying single word in document i.e. the given document string is distributed into one unit or token.

Removal of Stop word: In this stage the removal of common words like a, an, and, of, but etc. has been done.

Stemming: A stem may be a natural group of words with very similar meaning. This method defines the base of the particular word. There are two types of stemming method, Inflectional and Derivational [21].

iii). Text Transformation Text document contains a collection of words and their occurrences. There are two ways for representation of such documents is Bag of words and Vector space model.

iv). Attribute Selection This method leads to giving less database space, minimal search technique by removing irrelevant feature from input document. There are two methods in attribute selection, Filtering and Wrapping methods [13].

v). Pattern Selection In this stage the standard data mines process combines with the text mining process. Structured database use the classic data mining techniques that resulted from the previous stage. To identify the correctly interesting patterns representing knowledge based on some interestingness procedures [3].

vi). Interpretation/ Evaluation In this stage measure the result, this result can be put away or it will be used for next set of sequence.

C. Issues In Text Mining - The main challenging issue in text mining arises from the complexity of a natural language itself. The natural language is not free from ambiguity problem. Ambiguity means the capability of being understood in one or more possible ways [4]. In a text document one word can have more than one meanings and one phrase or sentence can be interpreted in many ways which directed to different meanings of statement. Here some of the issues are discussed as follows,

i). Intermediate Form -Intermediate forms with variable degrees of complexity are appropriate for different mining purposes. For a domain-specific knowledge discovery task, it is essential to perform linguistics analysis to derive a sufficiently rich representation to capture the relationship between the objects or concepts defined in the documents [5]. However, this analysis method is computationally expensive and often operates in the order of a few words per second. It remains a challenge to envision how analysis can be made more efficient and scalable for very large amount text.

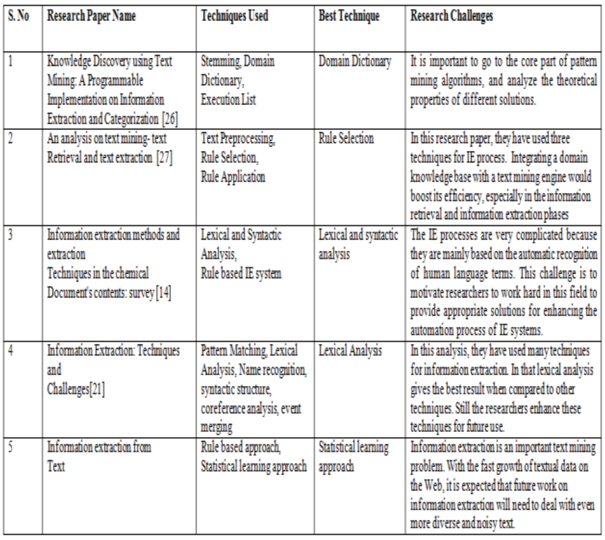
ii). Multilingual Text Refining - Although data mining is basically language independent, text mining comprises a significant language component. It is important to develop text refining algorithms that process multilingual text documents and it produce language-independent intermediate forms [9]. Even though most text mining tools emphasis on processing English documents, mining from documents in other languages permits access to previously unused information and offers a new host of chances [18]

iii). Domain Knowledge Integration - Domain knowledge, not provided for by any current text mining tools, can play an important role in text mining. Specifically, domain knowledge can be used as early as within the text processing stage. It is interesting to explore how one can take advantage of domain information to enhance parsing efficiency and derive an extra compact intermediate form [18]. Domain knowledge can also play a part in knowledge distillation. In a classification or predictive modeling task, domain knowledge helps to enhance learning/mining efficiency and quality of mined knowledge.

II. LITERATURE REVIEW

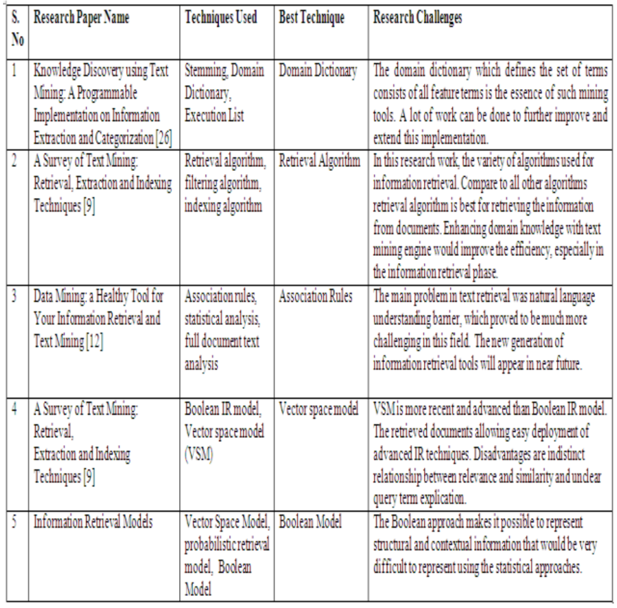
Literature Review

Table 1: Comparative Analysis of Information Extraction Techniques



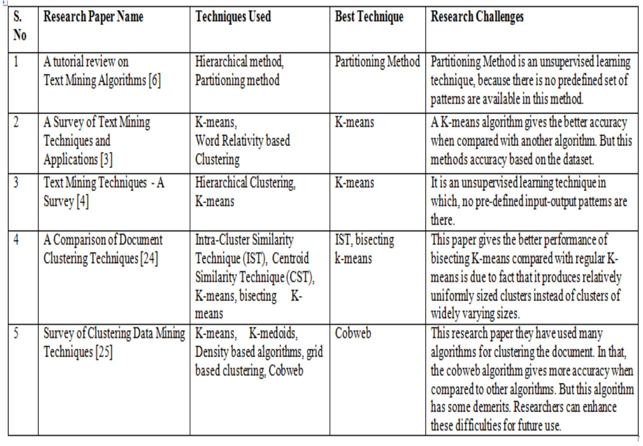
S.No	Research Paper Name	Techniques Used	Best Technique	Research Challenges
1	Knowledge Discovery using Text Mining: A Programmable Implementation on Information Extraction and Categorization [26]	Stemming, Domain Dictionary, Execution List	Domain Dictionary	It is important to go to the core part of pattern mining algorithms, and analyze the theoretical properties of different solutions.
2	An analysis on text mining- text Retrieval and text extraction [27]	Text Preprocessing, Rule Selection, Rule Application	Rule Selection	In this research paper, they have used three techniques for IE process. Integrating a domain knowledge base with a text mining engine would boost its efficiency, especially in the information retrieval and information extraction phases
3	Information extraction methods and extraction Techniques in the chemical Document's contents: survey [14]	Lexical and Syntactic Analysis, Rule based IE system	Lexical and syntactic analysis	The IE processes are very complicated because they are mainly based on the automatic recognition of human language terms. This challenge is to motivate researchers to work hard in this field to provide appropriate solutions for enhancing the automation process of IE systems.
4	Information Extraction: Techniques and Challenges[21]	Pattern Matching, Lexical Analysis, Name recognition, syntactic structure, coreference analysis, event merging	Lexical Analysis	In this analysis, they have used many techniques for information extraction. In that lexical analysis gives the best result when compared to other techniques. Still the researchers enhance these techniques for future use.
5	Information extraction from Text	Rule based approach, Statistical learning approach	Statistical learning approach	Information extraction is an important text mining problem. With the fast growth of textual data on the Web, it is expected that future work on information extraction will need to deal with even more diverse and noisy text.

Table 2: Comparative Analysis of Information Retrieval Techniques



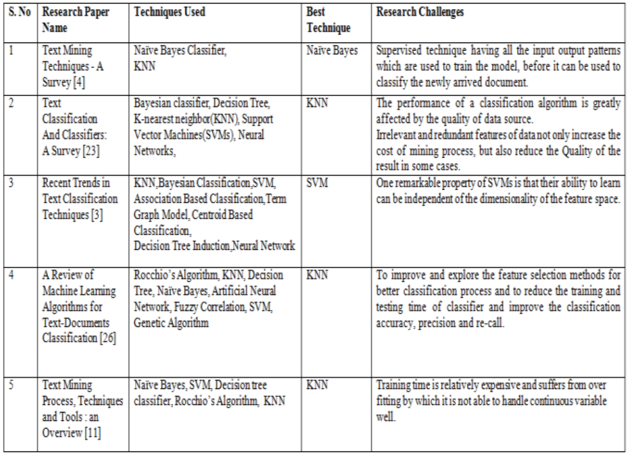
S. No	Research Paper Name	Techniques Used	Best Technique	Research Challenges
1	Knowledge Discovery using Text Mining: A Programmable Implementation on Information Extraction and Categorization [26]	Stemming, Domain Dictionary, Execution List	Domain Dictionary	The domain dictionary which defines the set of terms consists of all feature terms is the essence of such mining tools. A lot of work can be done to further improve and extend this implementation.
2	A Survey of Text Mining: Retrieval, Extraction and Indexing Techniques [9]	Retrieval algorithm, filtering algorithm, indexing algorithm	Retrieval Algorithm	In this research work, the variety of algorithms used for information retrieval. Compare to all other algorithms retrieval algorithm is best for retrieving the information from documents. Enhancing domain knowledge with text mining engine would improve the efficiency, especially in the information retrieval phase.
3	Data Mining: a Healthy Tool for Your Information Retrieval and Text Mining [12]	Association rules, statistical analysis, full document text analysis	Association Rules	The main problem in text retrieval was natural language understanding barrier, which proved to be much more challenging in this field. The new generation of information retrieval tools will appear in near future.
4	A Survey of Text Mining: Retrieval, Extraction and Indexing Techniques [9]	Boolean IR model, Vector space model (VSM)	Vector space model	VSM is more recent and advanced than Boolean IR model. The retrieved documents allowing easy deployment of advanced IR techniques. Disadvantages are indistinct relationship between relevance and similarity and unclear query term explication.
5	Information Retrieval Models	Vector Space Model, probabilistic retrieval model, Boolean Model	Boolean Model	The Boolean approach makes it possible to represent structural and contextual information that would be very difficult to represent using the statistical approaches.

Table 3: Comparative Analysis of Clustering Techniques



S. No	Research Paper Name	Techniques Used	Best Technique	Research Challenges
1	A tutorial review on Text Mining Algorithms [6]	Hierarchical method, Partitioning method	Partitioning Method	Partitioning Method is an unsupervised learning technique, because there is no predefined set of patterns are available in this method.
2	A Survey of Text Mining Techniques and Applications [3]	K-means, Word Relativity based Clustering	K-means	A K-means algorithm gives the better accuracy when compared with another algorithm. But this methods accuracy based on the dataset.
3	Text Mining Techniques - A Survey [4]	Hierarchical Clustering, K-means	K-means	It is an unsupervised learning technique in which, no pre-defined input-output patterns are there.
4	A Comparison of Document Clustering Techniques [24]	Intra-Cluster Similarity Technique (IST), Centroid Similarity Technique (CST), K-means, bisecting K-means	IST, bisecting k-means	This paper gives the better performance of bisecting K-means compared with regular K-means is due to fact that it produces relatively uniformly sized clusters instead of clusters of widely varying sizes.
5	Survey of Clustering Data Mining Techniques [25]	K-means, K-medoids, Density based algorithms, grid based clustering, Cobweb	Cobweb	This research paper they have used many algorithms for clustering the document. In that, the cobweb algorithm gives more accuracy when compared to other algorithms. But this algorithm has some demerits. Researchers can enhance these difficulties for future use.

Table 4: Comparative Analysis of Categorization Techniques



S. No	Research Paper Name	Techniques Used	Best Technique	Research Challenges
1	Text Mining Techniques - A Survey [4]	Naïve Bayes Classifier, KNN	Naïve Bayes	Supervised technique having all the input output patterns which are used to train the model, before it can be used to classify the newly arrived document.
2	Text Classification And Classifiers: A Survey [23]	Bayesian classifier, Decision Tree, K-nearest neighbor(KNN), Support Vector Machines(SVMs), Neural Networks,	KNN	The performance of a classification algorithm is greatly affected by the quality of data source. Irrelevant and redundant features of data not only increase the cost of mining process, but also reduce the Quality of the result in some cases.
3	Recent Trends in Text Classification Techniques [3]	KNN,Bayesian Classification,SVM, Association Based Classification,Term Graph Model, Centroid Based Classification, Decision Tree Induction,Neural Network	SVM	One remarkable property of SVMs is that their ability to learn can be independent of the dimensionality of the feature space.
4	A Review of Machine Learning Algorithms for Text-Documents Classification [26]	Rocchio's Algorithm, KNN, Decision Tree, Naïve Bayes, Artificial Neural Network, Fuzzy Correlation, SVM, Genetic Algorithm	KNN	To improve and explore the feature selection methods for better classification process and to reduce the training and testing time of classifier and improve the classification accuracy, precision and re-call.
5	Text Mining Process, Techniques and Tools : an Overview [11]	Naïve Bayes, SVM, Decision tree classifier, Rocchio's Algorithm, KNN	KNN	Training time is relatively expensive and suffers from over fitting by which it is not able to handle continuous variable well.

III. PROPOSED SYSTEM

A. Text Mining Approaches - Text Mining Approaches Text Mining or knowledge discovery from text (KDT)−first introduced by Fledman et al. refers to the process of extracting high quality of information from text (i.e. structured such as RDBMS data, semi-structured such as XML and JSON and unstructured text resources such as word documents, videos, and images). It widely covers a large set of related topics and algorithms for analyzing text, spanning various communities, including information retrieval, natural language processing, data mining, machine learning many application domains web and biomedical sciences.

i). Information Retrieval (IR): Information Retrieval is the activity of finding information resources (usually documents) from a collection of unstructured data sets that satisfies the information need [44,93]. Therefore information retrieval mostly focused on facilitating information access rather than analyzing information and finding hidden patterns, which is the main purpose of textmining. Information retrieval has less priority on processing or transformation of text whereas text mining can be considered as going beyond information access to further aid users to analyze and understand information and ease the decision making.

ii). Natural Language Processing (NLP): Natural Language Processing is sub-field of computer science, artificial intelligence and linguistics which aims at understanding of natural language using .Many of the text mining algorithms extensively make use of NLP techniques, such as part of speech tagging (POG),syntactic parsing and other types of linguistic analysis)

iii). Text Summarization: Many text mining applications need to summarize the text documents in order to get a concise overview of a large document or a collection of documents on a topic.

iv). Unsupervised Learning Methods: Unsupervised learning meth-ods are techniques trying to find hidden structure out of unlabeled data. They do not need any training phase, therefore can be applied to any text data without manual effort.

v). Supervised Learning Methods: Supervised learning methods are machine learning techniques pertaining to infer a function or learn a classifier from the training data in order to perform predictions on unseen data.

vi). Text Streams and Social Media Mining: There are many different applications on the web which generate tremendous amount of streams of text data. news stream applications and aggregators such as Reuters and Google news generate huge amount of text streams which provides an invaluable source of information to mine.

vii). Opinion Mining and Sentiment Analysis: With the advent of e-commerce and online shopping, a huge amount of text is created and continues to grow about different product reviews or usersopinions.

viii). Biomedical Text Mining: Biomedical text mining refers to the task of text mining on text of biomedical sciences domains.

B. Techniques In Text Mining - The techniques in text mining from different areas such as information extraction, information retrieval, natural language processing (NLP), categorization and clustering. All these stages of text mining process can be combined into a single workflow. Figure 1 shows the text mining techniques.

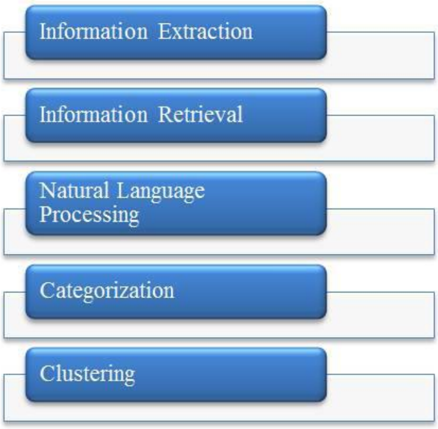


Figure 1: Techniques in Text Mining

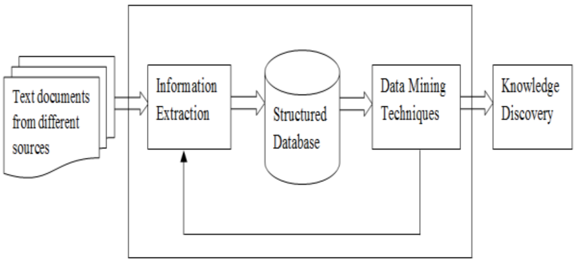


Figure 2: Information Extraction

i). Information Retrieval It is used to identify the relevant documents in a document collection which is matching a user's query. The most important application of information retrieval system is search engine like Google, which identify those documents on the World Wide Web that are relevant to user queries or a set of given words [9]. It also refers to the automatic retrieval of documents from document collection. It deals with crawling, indexing documents and retrieving documents. Information retrieval system used in digital libraries, online document systems and search engine. Information retrieval is deals with entire range of information processing from data retrieval to knowledge retrieval. Figure 3 explains about the information retrieval system.

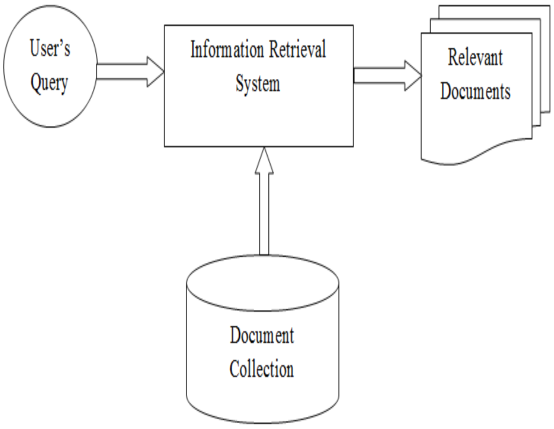


Figure 3: Information Retrieval

ii). Natural Language Processing (NLP) It is concerned with interactions between computer and human (natural) languages. NLP is related to the area of human-computer interaction. NLP is the component of an Artificial Intelligence (AI) [12]. It is used to analyze the human languages so that computers can understand natural languages as humans do. The approaches to NLP is based on machine learning, a type of artificial intelligence that examines and uses the patterns in data to improve a program's own understanding [2]. The role of NLP in Text Mining is to provide the systems in the information extraction phase with linguistic data that they need to perform their task. NLP includes the tasks [11],

- Part Of Speech tagging: It is used to classify the words into categories such as noun, verb or adjective.
- Chunking: It is also called as shallow parsing, used to identify only the main grammatical elements in a sentence such as noun phrases and verb phrases.
- Semantic Role Labeling: It is used to detection of semantic arguments associated with predicate or verb of a sentence.
- Language Model: It assigns a probability of sequence of words by means of probability distribution. It provides context to distinguish between words and phrases that sound similar.
- Semantically Related Words: This is the task of predicting whether two words are semantically related which is measured using the WordNet database.

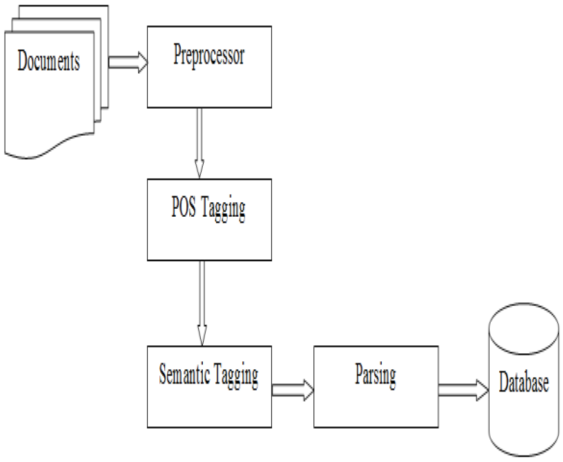


Figure 4: Natural Language Processing

iii). Categorization: The process of categorization is, recognizing, differentiating and understanding the ideas and objects to group them into classes. A category clarifies the relationship between the subjects and objects of information or knowledge [4]. Categorization is crucial in language, prediction, decision making and in all kinds of environmental interaction. It involves identifying the important themes of a document by placing the document into a predefined set of topics. Once the document is categorized, a computer program can typically treat the document as a "bag of words" [1]. It doesn't attempt to process the actual information as information extraction does. Relatively, categorization process only counts the words that appear in the document, and from the counts, it identifies the significant topics that the document covers. Categorization typically depends on a thesaurus for which topics are predefined, and relationships are identified by searching for broad terms, narrow terms, synonyms and related terms. Categorization tools commonly have a method for ranking the documents in order which documents have the most content on a particular topic [23]. It can be used in a variety of application domains. Many businesses and industries provide customer support or have to be compelled to answer queries on a number of topics from their customer. The main goal of categorization is to classify a collection of documents into a fixed number of predefined categories. Each document may belong to more than one class [16].

iv). Clustering - Clustering is a process of partitioning a set of data or objects into a set of meaningful sub-classes, is called clusters. This technique is used to group the similar documents. The benefit of clustering is that documents will seem in multiple subtopics, so ensuring that a useful document will not be omitted from search results. A basic clustering algorithm creates a vector of topic for each and every document and measures the weights of how well the document fits into each cluster [24]. Clustering technology can be helpful in the organization of management information systems, which can contain thousands of documents.

There are many clustering methods available and each of them may give a different grouping of a dataset.

Clustering methods can be classified into two categories [22],

1. Hierarchical Methods
2. Non-Hierarchical Methods

Hierarchical Methods - Hierarchical clustering constructs a cluster hierarchy or, in other words, a tree form of clusters also called as Dendogram. Each cluster node contains child clusters; relation clusters partitioned the points covered by their common parent. Such associate approach permits exploring data on completely different levels of granularity [20]. These methods are divided into Agglomerative (Bottom-UP) and Divisive (Top-Down) methods.

An **Agglomerative** clustering starts with one point clusters and recursively merges two or more with most

Appropriate clusters. A **Divisive** clustering starts with one cluster of all the data points and recursively splits the most appropriate clusters. This process continues until the requested number of clusters is achieved.

Non-Hierarchical Methods -The non-hierarchical methods divide a dataset of N objects into M clusters, with or without overlap. These methods are divided into partitioning methods, in which the categories are mutually exclusive and also the less common stamping methods [17]. Each object may be a member of the cluster with most similar.

C. Proposed Work - Many research works contributed to the field of IE through the use of various techniques. The primary focus of these researches was to determine how different text mining procedures can be utilized as the structured data sets exist in the text document format. This part begins with defining the topic of the research, evaluating previous researches, and then major techniques are applied using information extraction and text mining. In order to determine the topic of each research area and to develop an evolutionary and hierarchical connection between these topics, [35] used the method of text mining. Topics are presented through visualization tools. Moreover, these tools are used in order to show the connection between these topics and to offer interactive functions so that users can effectively find the cross-domain topics and know the trends of cross-domain research.

IV. METHODOLOGY USED

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

A. Term Based Method (TBM).
B. Phrase Based Method (PBM).
C. Concept Based Method (CBM).
D. Pattern Taxonomy Method (PTM).

A. 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[28]. 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.

B. 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[29]. The likely reasons for the daunting performance include:

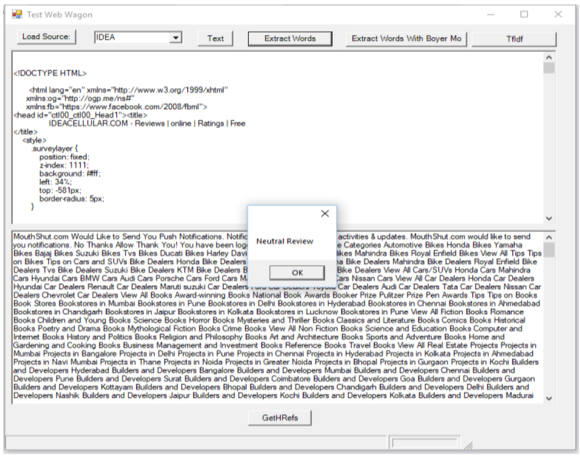
1) Phrases have inferior statistical properties to terms,
2) They have low frequency of occurrence, and
3) Large numbers of redundant and noisy phrases are present among them.

C. 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[34]. 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 [35]. 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.

D. 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[32]. 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[33]. 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.

V. EXPERIMENTAL RESULTS



```
<!DOCTYPE HTML>

     <html lang="en" xmlns="http://www.w3.org/1999/xhtml"
    xmlns:og="http://ogp.me/ns#"
    xmlns:fb="https://www.facebook.com/2008/fbml">
<head id="ctl00_ctl00_Head1"><title>
            IDEACELLULAR.COM - Reviews | online | Ratings | Free
</title>
   <style>
       .surveylayer {
          position: fixed;
          z-index: 1111;
          background: #fff;
          left: 34%;
          top: -581px;
          border-radius: 5px;
       }
```

Neutral Review

```
MouthShut.com Would Like to Send You Push Notifications. Notific activities & updates. MouthShut.com would like to send
you notifications. No Thanks Allow Thank You! You have been log e Categories Automotive Bikes Honda Bikes Yamaha
Bikes Bajaj Bikes Suzuki Bikes Tvs Bikes Ducati Bikes Harley Davi ikes Mahindra Bikes Royal Enfield Bikes View All Tips Tips
on Bikes Tips on Cars and SUVs Bike Dealers Honda Bike Dealers a Bike Dealers Mahindra Bike Dealers Royal Enfield Bike
Dealers Tvs Bike Dealers Suzuki Bike Dealers KTM Bike Dealers B Bike Dealers View All Cars/SUVs Honda Cars Mahindra
Cars Hyundai Cars BMW Cars Audi Cars Porsche Cars Ford Cars M ars Nissan Cars View All Car Dealers Honda Car Dealers
Hyundai Car Dealers Renault Car Dealers Maruti suzuki Car Dealers Car Dealers Audi Car Dealers Tata Car Dealers Nissan Car
Dealers Chevrolet Car Dealers View All Books Award-winning Books National Book Awards Booker Prize Pulitzer Prize Pen Awards Tips Tips on Books
Book Stores Bookstores in Mumbai Bookstores in Pune Bookstores in Delhi Bookstores in Hyderabad Bookstores in Chennai Bookstores in Ahmedabad
Bookstores in Chandigarh Bookstores in Jaipur Bookstores in Kolkata Bookstores in Lucknow Bookstores in Pune View All Fiction Books Romance
Books Children and Young Books Science Books Horror Books Mysteries and Thriller Books Classics and Literature Books Comics Books Historical
Books Poetry and Drama Books Mythological Fiction Books Crime Books View All Non Fiction Books Science and Education Books Computer and
Internet Books History and Politics Books Religion and Philosophy Books Art and Architecture Books Sports and Adventure Books Home and
Gardening and Cooking Books Business Management and Investment Books Reference Books Travel Books View All Real Estate Projects Projects in
Mumbai Projects in Bangalore Projects in Delhi Projects in Pune Projects in Chennai Projects in Hyderabad Projects in Kolkata Projects in Ahmedabad
Projects in Navi Mumbai Projects in Thane Projects in Noida Projects in Greater Noida Projects in Bhopal Projects in Gurgaon Projects in Kochi Builders
and Developers Hyderabad Builders and Developers Bangalore Builders and Developers Mumbai Builders and Developers Chennai Builders and
Developers Pune Builders and Developers Surat Builders and Developers Coimbatore Builders and Developers Goa Builders and Developers Gurgaon
Builders and Developers Kottayam Builders and Developers Bhopal Builders and Developers Chandigarh Builders and Developers Delhi Builders and
Developers Nashik Builders and Developers Jaipur Builders and Developers Kochi Builders and Developers Kolkata Builders and Developers Madurai
```

Figure 5: NaiveBayes

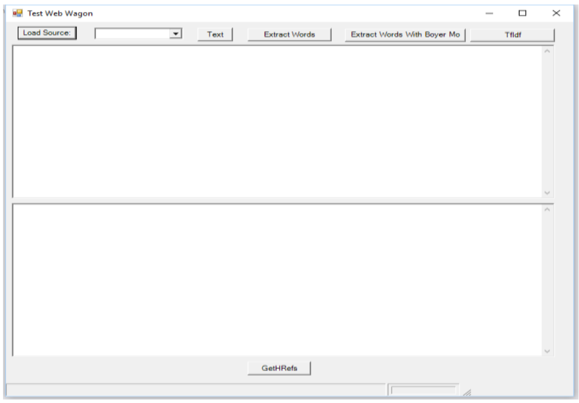


Figure 6: Webwagon

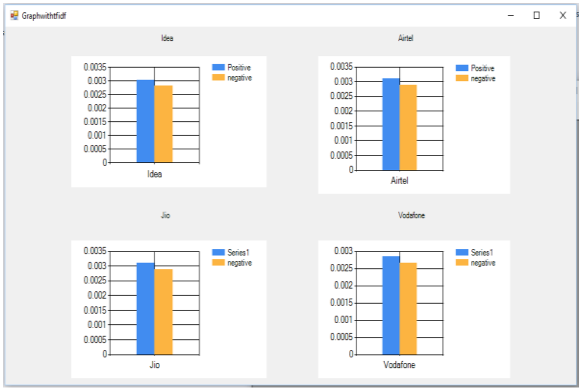


Figure 7: Graph TFIDF

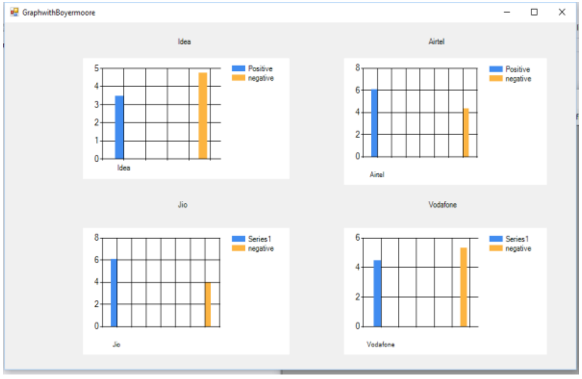


Figure 8: Grpah BOYERMOORE

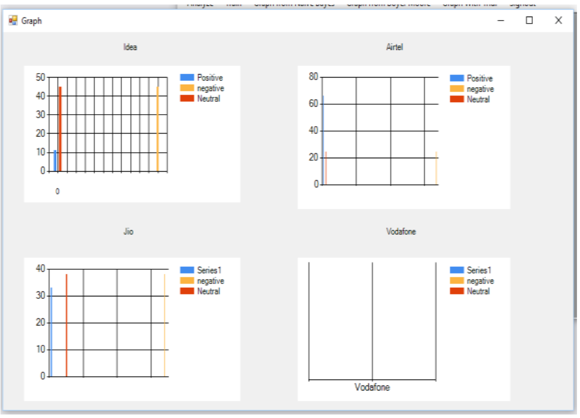


Figure 9: GRAPH WITH NAIVEBAYES

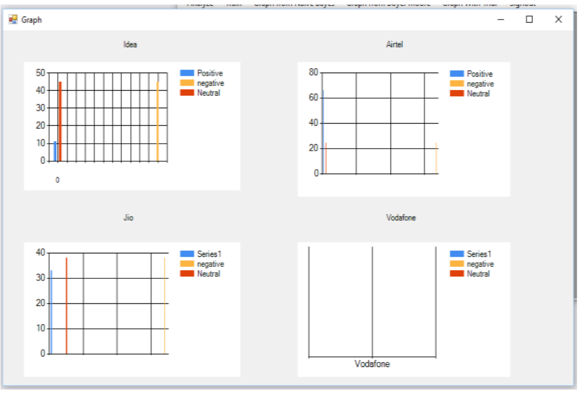


Figure 10: Overall Graph

VI. CONCLUSION

Text mining is the process of extracting non-trivial information from unstructured text. It is an interdisciplinary field involving computational linguistic and natural language processing, information extraction, information retrieval, machine learning and data mining. Text mining process generates features from unstructured text followed by applies mining techniques to discover knowledge. As most information is stored in form of unstructured text, text mining becomes essential to generate hidden useful information and knowledge.

Data Mining is the important as well as active research area helps to extract helpful patterns from the data. These patterns generated facilitate decision making in industries. Text mining is also crucial field that deals with unstructured or semi structured data. In this paper we have delineated the various text mining techniques such as Information Extraction, Information retrieval, Natural Language processing, Categorization and Clustering. And also we have defined text mining processing flow, applications of text mining and issues in text mining. Mining text in different languages may be a major problem, since text mining tools and techniques ought to be able to work with several languages and multilingual languages. Integrating a domain knowledge base with text mining engine would increase its efficiency, especially within the information retrieval and information extraction phase.

VII. REFERENCES

[1]. Varsha C. Pande , Dr. A.S. Khandelwal ,A Survey Of Different Text Mining Techniques, IBMRD's Journal of Management and Research, Online ISSN: 2348-5922, Volume-3, Issue-1, March 2014

[2]. Gobinda G. Chowdhury ,Natural Language Processing

[3]. Vishal Gupta, Gurpreet S. Lehal,A Survey of Text Mining Techniques and Applications, Journal of Emerging Technologies in Web Intelligence, Vol-1,No-1, August 2009

[4]. DivyaNasa, Text Mining Techniques- A Survey,International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 4, April 2012 ISSN: 2277 128X

[5]. Falguni N. Patel, Neha R. Soni,Text mining: A Brief survey, International Journal of Advanced Computer Research, ISSNprint: 2249-7277, ISSN online: 2277-7970 Volume-2 Number-4 Issue-6 December-2012

[6]. Mrs. SayantaniGhosh,, Mr. Sudipta Roy, Prof. Samir K. Bandyopadhyay, A tutorial review on Text Mining Algorithms, International Journal of Advanced Research in Computer and Communication Engineering ,Vol. 1, Issue 4, June 2012 ISSN : 2278 – 1021

[7]. Andreas Hotho, Andreas Nurnberger, Gerhard Paas, A Brief Survey of Text Mining, May 13,2005

[8]. ShaidahJusoh , Hejab M. Alfawareh, Techniques,Applications and Challenging Issues in Text Mining, International Journal of Computer Science Issues, Vol. 9, Issue 6, No 2, November -2012 ISSN (Online): 1694-0814

[9]. R. Sagayam, S.Srinivasan, S. Roshni ,A Survey of Text Mining: Retrieval, Extraction and Indexing Techniques, International Journal Of Computational Engineering Research (ijceronline.com) Vol. 2 Issue. 5, Issn 2250-3005(online), September| 2012

[10]. E. A. Calvillo, A. Padilla, J. Mu˜noz, J. Ponce, and J. T. Fernandez, "Searching research papers using clustering and text mining," in Electronics, Communications and Computing (CONIELECOMP), 2013 International Conference on. IEEE, 2013, pp. 78–81.

[11]. Anna Stavrianou, PeriklisAndritsos, Nicolas Nicoloyannis, Overview and Semantic Issuses of Text Mining, Special Interest Group Management of Data (SIGMOD) Record, September-2007,Vol.36, No.3

[12]. Ronan Collobert, Jason Weston, A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning

[13]. Mahesh T R, Suresh M B, M Vinayababu, Text Mining: Advancements, Challenges and Future directions, International Journal of Reviews in Computing, ISSN: 2076-3328 E-ISSN: 2076-3336

[14]. Anshika Singh , Dr. UdayanGhosh , Text Mining: A Burgeoning technology for knowledge extraction, International Journal of Scientific Research Engineering & Technology (IJSRET) Volume.1 Issue.12 pp 022-026 ,March 2013,ISN 278 – 082

[15]. RanveerKaur, ShrutiAggarwal, Techniques for Mining Text Documents, International Journal of Computer Applications (0975 – 8887), Volume 66, No.18, March 2013

[16]. K.L.Sumathy, M.Chidambaram, Texr Mining: Concepts, Applications, Tools and Issues- An Overview, International Journal of Computer Applications (0975- 8887), Vol-80, No.4, October-2013

[17]. Mr. Rahul Patel, Mr. Gaurav Sharma, A survey on text mining techniques, International Journal Of Engineering And Computer Science, ISSN:2319-7242, Volume 3 ,Issue 5 ,May 2014

[18]. N. Samsudin, M. Puteh, A. R. Hamdan, and M. Z. A. Nazri, "Immune based feature selection for opinion mining," in Proceedings of the World Congress on Engineering, vol. 3, 2013, pp. 3–5.

[19]. PatilMonali S, KankalSandip S, A Concise Survey on Text Data Mining, International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 9, September 2014, ISSN (Online) : 2278-1021,ISSN (Print) : 2319-5940

[20]. K.Nalini, Dr.L.JabaSheela, Survey on Text Classification, International Journal of Innovative Research in Advanced Engineering (IJIRAE),ISSN: 2349-2163, Volume. 1,Issue 6, July2014

[21]. K.Thilagavathi, V.Shanmugapriya, A Survey on Text Mining Techniques, International Journal of Research in Computer Applications and Robotics (IJRCAR), ISSN: 2320-7345, Volume. 2, Issue 10, October 2014

[22]. Lokesh Kumar, ParulKalra Bhatia, Text Mining: Concepts, Process and Applications, Journal of Global Research in Computer Science (JGRCS), Volume 4, No. 3, March 2013

[23]. VandanaKorde, C NamrataMahender, Text Classification and Classifiers: A Survey, International Journal of Artificial Intelligence & Applications (IJAIA), Vol.3, No.2, March 2012

[24]. Michael Steinbach George KarypisVipin Kumar, A Comparison of Document Clustering Techniques, Department of Computer Science and Engineering, University of Minnesota.

[25]. PavelBerkhin, Survey of Clustering Data Mining Techniques, Accrue Software, Inc.

[26]. Atika Mustafa, Ali Akbar, Ahmer Sultan, Knowledge Discovery using Text Mining: A Programmable

Implementation on Information Extraction and Categorization, International Journal of Multimedia and Ubiquitous Engineering

[27]. Umajancy.S,Dr. Antony S Elvadoss Thanamani, An analysis on text mining- text Retrieval and text extraction, International Journal ofAdvanced Research in Computer and Communication Engineering.

[28]. G. Salton and C. Buckley, "Term-Weighting Approaches in Automatic Text Retrieval," Information Processing and Management:An Int'l J., vol. 24, no. 5, pp. 513-523, 1988.

[29]. H. Ahonen, O. Heinonen, M. Klemettinen, and A.I. Verkamo, "Applying Data Mining Techniques for Descriptive Phrase Extraction in Digital Document Collections," Proc. IEEE Int'l Forum on Research and Technology Advances in Digital Libraries (ADL '98), pp. 2-11, 1998.

[30]. W. Lam, M.E. Ruiz, and P. Srinivasan, "Automatic Text Categorization and Its Application to Text Retrieval," IEEE Trans. Knowledge and Data Eng., vol. 11, no. 6, pp. 865-879, Nov./Dec. 1999.

[31]. H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins, "Text Classification Using String Kernels," J. Machine Learning Research, vol. 2, pp. 419-444, 2002.

[32]. S.-T. Wu, Y. Li, Y. Xu, B. Pham, and P. Chen, "Automatic Pattern- Taxonomy Extraction for Web Mining," Proc. IEEE/WIC/ACM Int'l Conf. Web Intelligence (WI '04), pp. 242-248, 2004.

[33]. S.-T. Wu, Y. Li, and Y. Xu, "Deploying Approaches for Pattern Refinement in Text Mining," Proc. IEEE Sixth Int'l Conf. Data Mining (ICDM '06), pp. 1157-1161, 2006.

[34]. S. Shehata, F. Karray, and M. Kamel, "Enhancing Text Clustering Using Concept-Based Mining Model," Proc. IEEE Sixth Int'l Conf. Data Mining (ICDM '06), pp. 1043-1048, 2006.

[35]. S. Shehata, F. Karray, and M. Kamel, "A Concept-Based Model for Enhancing Text Categorization," Proc. 13th Int'l Conf. Knowledge Discovery and Data Mining (KDD '07), pp. 629-637, 2007