# Text Mining Framework, Methods and Techniques

*Yugandhara Bapurao Dasri[1], Bhagyashree Vyankatrao Barde [2], Nalwade Prakash Shivajirao [3], Anant Madhavrao Bainwad [4]

[1]*(Dept. of Computer Science and Engg, SGGS IE and T, Nanded, Maharashtra, India 431-606* )
[2]*(Dept. of Computer Science and Engg, SGGS IE and T, Nanded, Maharashtra, India 431-606* )
[3]*(Dept. of Computer Science and Engg, SGGS IE and T, Nanded, Maharashtra, India 431-606*)
[4]*(Dept. of Computer Science and Engg, SGGS IE and T, Nanded, Maharashtra, India 431-606*)
*Corresponding Author: Yugandhara Bapurao Dasri*

***Abstract :*** *Nowadays growth of digital data is increasing rapidly. Textual documents created and distributed on the Internet are changing in various forms. All these data or documents are not efficiently useful. We have to analyze large amounts of data in an effort to find correlations, patterns, and insights that is nothing but Data mining. To discover relationship between two or more variables in data we require Data Mining. Data mining have attracted huge attention with coming up need for turning such data into knowledge and useful information. Text mining is an important data mining technique. Text mining is one of the recent area for research, is defined as the process of extracting data from large amount of texts. It allows structuring and categorizing the textual contents which are initially unstructured. Text mining includes the most successful technique to extract the effective patterns. In this paper we will discuss framework of text mining, techniques and methods to give effectiveness over information extraction in text mining*.

***Keywords:*** *Data mining, Text mining, Framework, Methods, Techniques*

Date of Submission: 08-07-2017 Date of acceptance: 15-07-2017

## I. Introduction

Text Mining comes under the domain of data mining. Data mining technology helps to extract useful information from various databases. Due to the quick growth of digital data [2] made available in recent years, data mining and knowledge discovery have attracted a great deal of attention with forthcoming need for turning such data into useful information and knowledge [5]. Data mining consist of five major elements
1) Extract, transform and low transaction data onto the data ware house system.
2) Store and manage the data in the multidimensional database system.
3) Data access to business analysis and information technology professionals.
4) Analyze the data by Application software.
5) Present the data in useful formats such as graph or table.

So all techniques which are used in Data Mining are also comes in text mining in addition to it more methods and techniques are used in Text Mining. Text Mining is Textual data. Text analytics helps to analyze, extract meanings, Patterns and structures which are hidden in unstructured textual data. Nowadays business uses data and text mining to examine customer and competitor data to improve competitiveness. The pharmaceutical industry mines research articles to improve drug discovery within academic research, mining and analytics of huge data sets are delivering efficiencies and new knowledge in areas as diverse biological science. In this paper, we focus on methods and techniques which are used in text mining [3].

Text mining plays vital role in

- Customer profile analysis
- Banks, insurance and financial markets News stories classification
- Web search
- Trend analysis
- Information filtering and routing Telecommunications
- Energy and other service industries Event tracks
- Patent analysis
- Publishing and media
- Company resource planning
- Political institutions, political analysis
- Public administration and legal documents Information Technology sector and Internet

## II. Framework Of Text Mining

Text Mining can be visualized as consisting of two phases Text refining and Knowledge filtering as shown in fig. 1 Text refining phase transforms the textual documents into a chosen intermediate form and knowledge filtering deduce patterns or knowledge from intermediate form[3]. Intermediate form can be document based where each entity represents a document or concept based where each entity represents an object or concept of interests in specific domain. Mining a document based intermediate form derives patterns and relationships across documents [6]. Document clustering, categorization, visualization, summarization are examples of mining from a document based.

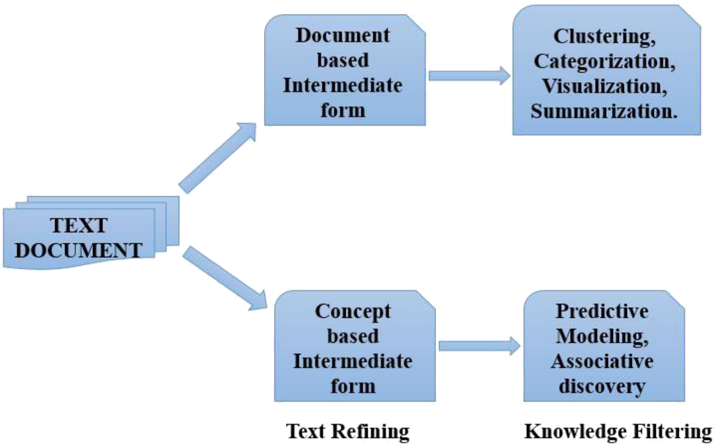
Document based Intermediate form
Clustering, Categorization, Visualization, Summarization.
TEXT DOCUMENT
Concept based Intermediate form
Predictive Modeling, Associative discovery
Text Refining
Knowledge Filtering

**Fig. 1**. Framework for Text Mining

## III. Steps Involved In Text Mining

The steps which are involved in Text Mining process shown in fig. 2 are as follows

### A. Text preprocessing

The text preprocessing [1] is further divided into

**1) Tokenization:** Text document is collection of statements. This step divide the whole statements into words by removing blank spaces, commas etc.

**2) Stop word removal:** This step involves removing of HTML, XML tags from web pages. There is list of stop words such as ’a’, ’is’, ’of’, ’an’ and so on. According to these words stop word removal process removes words from documents.

**3) Stemming (stem word removal):** Stemming is the process to identify the root of certain words such as presented, presenting, presentation gets convert into original word present. The most commonly used algorithm is porters’ algorithm for stemming [6].

### B. Techniques used in Text Mining

Data mining methods such as clustering, classification information retrieval etc. can be used for text mining [5].

**1) Information Extraction:** Information extraction is the process of analyzing unstructured textual data and finding of Knowledge or key terms from data and relationship within text.

**2) Categorization:** Text categorization [1] is also known as text classification is the task of sorting a set of documents into categories from a predefined set of documents. It assigns labels to each document. It is based on supervised leaning. Classification techniques like Nearest Neighbor classifier, Naive Bayesian classifier, Decision Tree, and Support Vector Machines can be used to categorize text.

**3) Clustering:** Clustering is the process to find groups of documents with similar content. The output of clustering gives number of clusters P and each cluster contains number of documents d. The contents of the documents within one cluster are similar and between the clusters are dissimilar. Even though clustering technique used to group similar documents it differs from classification because in clustering documents are clustered on the fly instead of use of predefined set of documents. It is based on unsupervised learning. In data mining, K-means clustering is frequently used clustering algorithm, in text mining field it also gives good results.

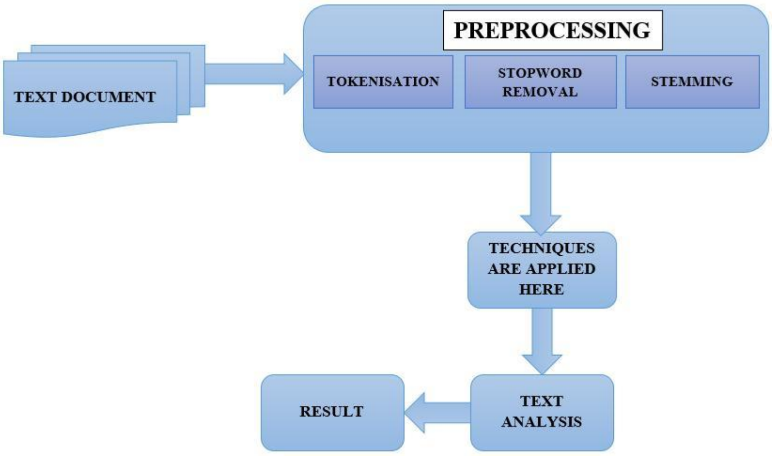

**Fig. 2.** Steps involved in text mining process

**4) Visualization:** Visualization methods can improve and simplify the discovery of relevant information. Visual text mining puts large textual sources in a visual hierarchy. To represent individual documents or groups of documents text flags are used to show document category. The user can interact with the document by zooming and scaling. Information visualization is applicable to government to find information about crimes. Information visualization divided into three steps:
a) Data preparation step decide and obtain original and form original data space.
b) The process of analyzing and extracting visualization data which is needed from original data and to form visualization data space is called as data analysis and extraction.
c) Visualization mapping step employ certain mapping algorithms to map visualization data space to visualization target.

**5) Summarization:** It is the process to reduce the length and detail of a document while retaining general meaning of the document. Summarization process include following steps:
a) Preprocessing obtain a structured representation of the original textual data.
b) To transform summary structure from textual structure algorithm is applied in next processing step.
c) Then final summary is obtained from the summary structure.

## IV. Methods Used In Text Mining

There are so many techniques developed to solve the problem of text mining that is nothing but the relevant information retrieval according to users' 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)

### A. Term Based Method

In this method everything is defined in TERM. Term in document is the word having semantic meaning. Document is analyzes on the basis of term. Term based methods have problems of polysemy and synonymy. Polysemy means a word has multiple meanings and synonymy means multiple words having the same meaning. Information extraction provided many term-based methods to solve this challenge.

### B. Phrase Based Method

Phrases are less ambiguous than the term and it carries more semantics like information. In this method the document analysis on the basis of phrases. Following are the reasons which intimidating performance:
1) Phrases have low frequency of occurrences.
2) Phrases have low quality statistical properties to terms.
3) Large numbers of redundant phrases and noisy phrases are included in them.

### C. Concept Based Method

In this method terms are analyzed on sentence basis. Text Mining techniques are mostly based on statistical analysis of phrase or word. Two terms can have same occurrence in same document, but the meaning of one term contributes more appropriately than the meaning contributed by the other term. The terms that capture the semantics of the text should be given more importance so, a new concept-based mining is introduced [1]. This model included three components. 1) Analyzes the semantic structure of sentences. 2) Constructs a conceptual ontological graph (COG) to describe the semantic structures. 3) Extract top concepts based on the first two components to build feature vectors using the standard vector space model. This model can effectively differentiate between unimportant terms and meaningful terms which describe a sentence meaning. The concept-based model usually depends upon natural language processing techniques. Feature selection is applied on the query concepts to optimize the representation and remove noise and ambiguity.

### D. Pattern Taxonomy Method

In this method documents are analyzed on pattern basis. Patterns can be structured into taxonomy by using is-a relation. Patterns can be discovered by using data mining techniques like closed pattern mining, sequential pattern mining, frequent item set mining and association rule mining. Use of discovered patterns in the field of text mining is very difficult and use-less, because some useful long patterns with high specificity lack in support this is called low-frequency problem. Here not all frequent short patterns are useful hence known as misinterpretations of patterns and it leads to the useless performance [2]. 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 (PDM) 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 data mining-based methods and the term-based models, but also concept-based model.

## V. Conclusion

At last we conclude that, Text mining is also known as Text Data Mining or Knowledge-Discovery in Text [4]. Many data mining techniques have been proposed in the last decade for mining useful information and knowledge. Nowadays growth of digital data is increasing rapidly, knowledge discovery and data mining have attracted huge attention with coming up need for turning such data into knowledge and useful information. In general text mining consists of analyzing large amount of text documents by extracting terms, phrases, concepts, pattern and prepare the text processed for further analysis with data mining techniques. In this paper an overview of Data mining framework, methods, techniques of text mining is presented to give the researchers to carry it to the next level.

## References

[1] Sonali Vijay Gaikwad, Archana Chaugule, Pramod Patil, Text Mining Methods and Techniques, International Journal of Computer Applications (0975 8887) VOL. 85, No. 17, JANUARY 2014

[2] Ning Zhong, Yuefeng Li, and Sheng-Tang Wu, Effective Pattern Discovery for Text Mining, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 1, JANUARY 2012.

[3] K.L.Sumathy, M.Chidambaram, Text Mining: Concepts, Applications, Tools and Issues An Overview, International Journal of Computer Applications (0975 8887) VOL. 80, No.4, OCTOBER 2013

[4] Haralampos Karanikas, Babis Theodoulidis Manchester, (2001), Knowledge Discovery in Text and Text Mining Software, Centre for Research in Information Management, UK.

[5] 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

[6] Mustafa M. Shaikh, Ashwini A. Pawar, Vibha B. Lahane, Pattern Discovery Text Mining for Document Classification, International Journal of Computer Applications (0975 8887) VOL. 117, NO. 1, MAY 2015

[7] Jiaqi Zhu, Member, Kaijun Wang, Yunkun Wu, Zhongyi Hu, and Hongan Wang, Member, IEEE Mining User-Aware Rare Sequential Topic Patterns in Document Streams, IEEE transactions on knowledge and data engineering, vol. 28, no. 7, july 2016

[8] M Ozaki, Y. Adachi, Y. Iwahori, and N. Ishii, Application of fuzzy theory to writer recognition of Chinese characters, International Journal of Modelling and Simulation, 18(2), 1998, 112-116.

IOSR Journal of Computer Engineering (IOSR-JCE) is UGC approved Journal with Sl. No. 5019, Journal no. 49102.

Yugandhara Bapurao Dasri. "Text Mining Framework, Methods and Techniques." IOSR Journal of Computer Engineering (IOSR-JCE) 19.4 (2017): 19-22.