ISSN: 0976-3031

*Available Online at http://www.recentscientific.com*

**CODEN: IJRSFP (USA)**

*International Journal of Recent Scientific Research*
*Vol. 9, Issue, 5(G), pp. 26959-26962, May, 2018*

**International Journal of Recent Scientific Research**

**DOI: 10.24327/IJRSR**

**Research Article**

# TEXT MINING: A COMPREHENSIVE SURVEY

**Ruchi Rautela., Priyanka Dilip Huilgol and Sunit Pravin Kajarekar**

Department of MCA, VESIT, India

**DOI: http://dx.doi.org/10.24327/ijrsr.2018.0905.2158**

**ARTICLE INFO**

***Article History:***

Received 10th February, 2018
Received in revised form 6th March, 2018
Accepted 24th April, 2018
Published online 28th May, 2018

***Key Words:***

Structured data, unstructured data, text mining, and text mining techniques

**ABSTRACT**

Text mining sometimes also referred to as Text analytics is an Artificial Intelligence (AI) technology that extracts meaningful quantitative and functional/workable data from natural language text. Text mining is the discovery by computer of new previously unknown information, by automatically extracting hidden information from different written resources to evolve into semi-structured data. This involves using various techniques that vary depending on the application need, efficiency and performance. This paper will introduce the techniques involved in text mining and also discuss its applications and issues.



## INTRODUCTION

Day by day data is growing at an ever-increasing rate. A large number of organizations, institutions and professions rely on this data to work in an efficient and timely manner. This data is scattered as structured and unstructured data, approximately 90% of which is held in unstructured format. Thus, there arises a need to extract meaningful data from this unstructured pool. Text mining overcomes this issue and helps serve this purpose effortlessly. This technology deals with automatically extracting text from different written sources. It differs from data mining since the former works with unstructured data while the latter handles structured data. The main goal of text mining is to identify information that is not previously known. Application areas like search engines, customer relationship management system, filter emails, product suggestion analysis, fraud detection, and social media analytics use text mining for opinion mining, feature extraction, sentiment, predictive, and trend analysis [10]. Text mining techniques have been widely used in variety of major fields such as web applications, internet, Business Intelligence, content enrichment among others [11].

### *Background*

#### *Unstructured Data*

Data that does not have a recognizable structure or is loosely structured is termed as "Unstructured data". It is unorganized and raw and can be in the form of text or non-text but is usually text heavy [12]. Textual unstructured data may include number and dates as well as facts while non-textual unstructured data is commonly generated in the form of media that includes MP3 audio files, JPEG images and Flash video files, etc. For example, email includes time, date, recipient, sender details and subject etc., but an email body remains unstructured [13].


**Figure 1** Unstructured data types

#### *Semi-Structured Data*

Semi structured data is the data that is not regulated into database or other specialized repositories but contains associated information, such as metadata, that serves useful to processing than raw data. Neither is it raw data, nor typed data in a conventional database system but rather structured data not organized in a rational model. Since it falls between structured and unstructured it has certain aspects that are structured and others that are not. Information contained in semi structured

**Corresponding author:* **Ruchi Rautela**
Department of MCA, VESIT, India

data is usually related to a database schema and hence is sometimes referred to as "self-describing". Examples include data found on the Web, XML and other markup languages.

### *Text Mining*

Text mining or Text analytics is the process where text in high quality is usually extracted from data that is in unstructured or semi-structured format. This data may involve patterns or relevant information from different sources. The natural framework of text mining is comprised of two different components (1) Text refining that is responsible for transforming text documents in free-form into an intermediate form and (2) knowledge distillation responsible for deducing patterns/knowledge from the intermediate form. Semi structured form such as conceptual graph representation or structured such as the relational data representation can be considered as Intermediate form (IF) [3]. This technology is concerned with evolution of numerous pattern recognition, mathematical and statistical approaches for analysing unstructured information and extracting important and essential data from it [3].

## METHODS

### *Term Based Method*

Documents have term that contain useful information especially semantic meaning. Hence, documents need to be analysed on term basis. Each term is associated with a value known as weight. This method has two problems –

1. Polysemy i.e. a term having many possible meanings.
2. Synonymy i.e. multiple words having the same meanings.

This method is emerging as a result of some Information retrieval and machine learning communities [15].

### *Phrase Based Method*

Phrases carry more information than a single term because they are a collection of semantic terms. They are more descriptive and less ambiguous than

TERM because in phrase-based method document is analysed on phrase basis.

Some reasons which avert the performance are:

1. Inferior statistical properties to terms
2. Low frequency of occurrences.
3. Redundant phrases and noisy phrases [15].

### *Concept Based Method*

In Concept based method the terms are predicted or guessed at a sentence or a document level. Rather than a single term analysis, this model tries to analyse a term on a document or sentence level by finding a significant matching term aptly.

This model contains three components:

1. Examining the semantic construction of sentences.
2. Building a conceptual ontological graph to describe the semantic structures.
3. Extracting top concepts based on the first two components to build feature vectors using the standard vector space model [15].

### *Pattern Taxonomy Method*

In pattern-based model document is analysed on pattern basis i.e. pattern of document is formed by analysing is-a-relation between terms to form taxonomy. Taxonomy is tree like structure. The pattern-based approach can improve the accuracy of system for evaluating term weights because discovered patterns are more specific than whole documents. Patterns can be discovered by using data mining techniques like closed pattern mining, sequential pattern mining, frequent item set mining and association rule mining. The pattern-based technique uses two processes pattern deploying (PDM) and pattern evolving. This technique refines the discovered patterns in text documents [15].

### *Text Mining Process*

1. The textual data from various sources that is in semi-structured or unstructured format is collected to perform text mining.
2. Pre-processing involves cleaning the data that is collected.
3. Various techniques used in text mining which are discussed later are then applied to extract meaningful information.
4. The data obtained is analyzed to extract knowledge and meaning out of it.
5. Finally, the required knowledge is obtained and can then be used for further analysis.

The process is illustrated in Figure 2.

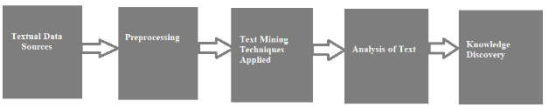


**Figure 2Text** Mining Process

### *Techniques in Text Mining*

It was in the 1980s that text mining techniques were first used [14]. These manual techniques were labor intensive and proved to be expensive requiring too much time to process the growing quantity of information. This gave rise to some techniques that were beneficial over manual ones with respect to cost, time as well as efficiency. These techniques include:

### *Information Extraction*

Automatically extracting structured definite information from unstructured or semi-structured data by means of Natural Language Processing and which is in the form of text is known as Information extraction (IE) [8]. It identifies the extraction of entities for example names of persons, organization, location and relationship between entities, attributes, events and relationships from text. Extracted information is well-organized (structured) and stored in database like patterns and is then available for further use. IE systems are used to extract specific attributes and entities from the document and establish their relationship. Process used to check and evaluate the relevance of results on the extracted data is 'Precision and Recall' [8].

### *Information Retrieval*

Information retrieval (IR) refers to collecting and accessing relevant information from a variety of resources [1]. It is also referred to as finding material that are usually documents in unstructured format fulfilling information needs from within

extensive collections. Thus, it can be defined as a set of methods and approaches for methodically developing information needs of the users in form of queries which are used to fetch document from a collection of databases [4].
IR helps to extract relevant and associated patterns according to a given set of words or phrases [5].

***Text Categorization***

This technique involves designating pre-decided categories to free-text documents responsible for presenting visionary aspect of document collections and has significant transactions in the real world [3]. The main purpose of text classification/text categorization is to increase detection of information to allow and assist in better decision [9]. For example, news stories are typically organized by subject categories (topics) or geographical codes; academic papers are often classified by technical domains and subdomains; patient reports in health-care organizations are often indexed from multiple aspects, using taxonomies of disease categories, types of surgical procedures, insurance reimbursement codes and so on. Another widespread application of text categorization is spam filtering, where email messages are classified into the two categories of spam and non-spam, respectively [4].

***Document Clustering***

This technique is used in order to find groups of documents with similar content. It makes use of descriptors and descriptor extraction that are essentially sets of words that describe the contents within the cluster [6]. A process which is unsupervised is responsible for classifying objects into groups called clusters. Dividing similar text into same cluster forms the basis of this method. Each cluster consists of number of documents. Any labels associated with objects are obtained solely from the data. Ensuring that a useful and essential document will not be missing from search results since documents can emerge in numerous subtopics is an advantage of this technique [2]. For example, if clustering is performed on a collection of news articles, it can be discovered that the similar documents are closer to each other and lie in the same cluster.

***Text Visualization***

Text Visualization is a technique that represents large textual information into a visual map layout which provides enhanced browsing capabilities along with simple searching. In text mining visualization methods can improve and simplify the discovery of relevant information [15]. Text flags are used to show document category to represent individual documents or groups of documents and colours are used to show density. Visual text mining puts large textual sources in an appropriate visual hierarchy which helps the user to interact with the document by scaling and zooming [14]. Visualization technique is even used by the government to identify terrorist networks and to find information about crimes in particular areas as well.

There are 3 steps involved in visualization process [15].

1. Data preparation - decides and obtains the original data of visualization and form original data space.
2. Data analysis and extraction-This process includes analysing and extracting visualization data needed from original data and to form visualization data space.
3. Visualization mapping - includes certain mapping algorithm to map visualization data space to visualization target.

***Applications***

***Digital Libraries***

Digital libraries are of great significance to researches contributing immensely to the field of research and development. These libraries have huge collection of documents that can be accessed online. Automatically extracting useful information from these documents proves to be of great use and saves time. For example, Green-stone international digital library that support multiple languages and multilingual interfaces provide a method for extracting documents that handle multiple formats, i.e., Microsoft word, pdf, postscript, HTML, scripting languages and e-mail messages and also supports extraction in the form of audio visual and image formats. Some of the tools for text mining in digital libraries include GATE, Net Owl and Aylien are frequently used tools for text mining in digital libraries.

***Education***

Student's interests and employment ratio is useful to analyse specific educational trends. Various text mining techniques are applied here for this purpose. One such techniques includes use of K-Means clustering for identifying attributes of relevant information. For example, how different attributes affect selection of subjects by students can be examined or student's performance in different subjects can also be analysed.

***Life sciences and Health Care***

Hospitals and industries in this field generate a huge amount of textual information pertaining to the patients, medicines, diseases, symptoms among others. Filtering relevant information and taking decision from this ever-growing repository is a big challenge. Moreover, this information contains records of varying nature, complex and vocabulary making it even more difficult to extract relevant information. Use of appropriate text mining tools in medical field help to evaluate the effectiveness of medical treatments that show effectiveness by comparing different diseases, symptoms and their course of treatments. One such example of text mining tool in this field is Linguamatics text mining platform I2E. The tool is useful for gene-disease mapping & target identification, biomarker discovery etc.

***Social Media***

Monitoring and analysing the online plain text from internet news, blogs, email, posts, followers, likes etc. to examine people's reaction to certain post, news is heavily used. Text mining tools help serve this purpose showing similarity or variation for a post/news by people belonging to different age groups. Tools include Parallel Dots and barnd 24.

***Business Intelligence***

Text mining in Business Intelligence (BI) has a lot of significance since it helps organizations in analysing their customers as well as competitors so that they can take better strategic decisions. Tools like IBM text analytics, Rapid miner,

GATE help to take decisions about the organization that generate alerts about good and bad performance, market changeover that help to take remedial actions.

***Resume Filtering***

Organizations receive hundreds and thousands of resume every day in different formats. The recruiter is on the lookout for job title, educational qualification, buzzwords, employment history and other personal information. Automatically extracting this information from resume will thus help conserve time and energy. Text mining tools help the recruiter by automating this task.

***Issues***

Text mining although having applications in a variety of fields has some issues to deal with. The information to be processed is in unstructured formats and is multilingual. Multilingual extraction is not supported by many tools thereby affecting the effectiveness of text mining process. Another issue is integration of domain knowledge since it helps to extract the exact information needed. Tools available do not support this integration. Synonyms, antonyms and abbreviation make it difficult for the tools categorize documents. Problems in natural language such as two words having the same spelling but different meaning e.g. tear (in the eye) and tear (rip apart) could also influence the extraction/refinement process [16]. Grammatical rules according to context and humans could also differ thus causing further problems.

## CONCLUSION

Text mining has become an important field in recent times. The vast information available on the web as well as other sources prove to be beneficial to organizations from various domains. This information is not available in readable format or needs some pre-processing i.e. it has to undergo the text mining process before it can be used for analysis. Techniques right from information extraction, information retrieval to visually representing the information are essential part of Text Mining process. These different techniques are used for varying needs. When a simple query is the input, Information retrieval is the best choice whereas when complex information is needed from unstructured or semi-structured data, it involves building models that automatically extracts the information. On the other hand, simple technique such as visualization also proves to be useful in streamlining the search for relevant information. Text mining is beneficial to a variety of fields. From education and health care to business organizations, text mining is being used extensively for different purposes. Despite having a few issues, text mining has definitely automated tasks thereby saving a lot of time.

## References


1. N. Venkata, L. Padmasree and N. Mangathayaru, "Survey of Text Mining Techniques, Challenges and their Applications", *International Journal of Computer Applications*, vol. 146, no. 11, pp. 30-35, 2016.
2. P. Shinde and S. Govilkar, "A Systematic study of Text Mining Techniques", *International Journal on Natural Language Computing*, vol. 4, no. 4, pp. 54-62, 2015.
3. M. Sukanya and S. Biruntha, "Techniques on text mining", *2012 IEEE International Conference on Advanced Communication Control and Computing Technologies (ICACCCT)*, 2012.
4. S. Niharika, V. Sneha Latha and D. Lavanya, "A Survey On Text Categorization", *International Journal of Computer Trends and Technology* (IJCTT), vol. 3, no. 1, 2012.
5. C. Uma, S. Krithika and C. Kalaivani, "A Survey Paper on Text Mining Techniques", *International Journal of Engineering Trends and Technology*, vol. 40, no. 4, pp. 225-229, 2016.
6. V. Gupta and G. Lehal, "A Survey of Text Mining Techniques and Applications", *Journal of Emerging Technologies in Web Intelligence*, vol. 1, no. 1, 2009.
7. W. Goffman, "A searching procedure for information retrieval", *Information Storage and Retrieval*, vol. 2, no. 2, pp. 73-78, 1964.
8. Chia-Hui Chang, M. Kayed, M. Girgis and K. Shaalan, "A Survey of Web Information Extraction Systems", IEEE Transactions on Knowledge and Data Engineering, vol. 18, no. 10, pp. 1411-1428, 2006.
9. Y. Yang and T. Joachims, "Text categorization", Scholarpedia, vol. 3, no. 5, p. 4242, 2008.
10. W. He, "Examining students' online interaction in a live video streaming environment using data mining and text mining", *Computers in Human Behavior*, vol. 29, no. 1, pp. 90-102, 2013.
11. "Text mining application: 10 examples improving our today life", Expertsystem.com, 2018. [Online]. Available: http://www.expertsystem.com/10-text-mining-examples/. [Accessed: 24- May- 2018].
12. N. Padhy, "The Survey of Data Mining Applications and Feature Scope", *International Journal of Computer Science, Engineering and Information Technology*, vol. 2, no. 3, pp. 43-58, 2012. "
13. What is unstructured data? - Definition from WhatIs.com", SearchBusinessAnalytics, 2018. [Online]. Available: https://searchbusinessanalytics.techtarget.com/definition/unstructured-data. [Accessed: 24- May- 2018].
14. L. Kumar and P. Bhatia, "TEXT MINING: CONCEPTS, PROCESS AND APPLICATIONS", *Journal of Global Research in Computer Science*, vol. 4, no. 3, pp. 36-39, 2013.
15. S. Gaikwad, A. Chaugule and P. Patil, "Text Mining Methods and Techniques", *International Journal of Computer Applications,* vol. 85, no. 17, pp. 42-45, 2014.]
16. R. Talib, M. Kashif, S. Ayesha and F. Fatima, "Text Mining: Techniques, Applications and Issues", *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 11, 2016.



**How to cite this article:**

Ruchi Rautela *et al.*2018, Text Mining: A Comprehensive Survey. *Int J Recent Sci Res.* 9(5), pp. 26959-26962.
DOI: http://dx.doi.org/10.24327/ijrsr.2018.0905.2158