Text Mining: Techniques and Applications- Literature Review

Monika chauhan[1] sachin singh[2]
[1]*Scholar, College Of Computing Sciences And Information Technology (TMU)*
[2] *Assistant Professor, College Of Computing Sciences And Information Technology(TMU)*
[1]rajputmonikachauhan@gmail.com
[2]singh.sachin1986@gmail.com

Abstract **:-In today's time, a lot of amount of data exists in digital format. It is also increasing day by day as the use of the electronic media is increasing. This data is however does not exists in a single format, it is stored from the varied different sources. The data stored will be in unstructured or semi-structured format and it can be in any form like text, graphs, tables etc. The data collected from the varied sources should not be necessarily the information so, to extract the useful information from the existing data is a challenging task. There are many text mining tools and techniques that are useful for mining the useful information from the data. This paper presents a review of the text mining techniques and also describes varied applications where these techniques are applied.**

Keywords:-Collected; Applications; Digital format; Extract.

I. INTRODUCTION

Text mining means discovering the useful and meaningful information from large amount of text that exists in varied formats. Text mining study is gaining recently more reputation because the count as well as the format of the data is increasing day by day. Text mining deals with the natural language text which is stored in structured and unstructured format. The problem is pushing aside all the material that currently is not relevant to our needs in order to find the relevant information. Now a days, all the sector like institutions, industries and social networking sites are using the digital media for storing their data and that data not necessarily contains the information which the user want. The data stored needs to be abstracted in such a way that the unwanted data should be removed from the needed information.

This can be done only through text mining. Text mining is also a variant of the Data mining, but data mining cannot be used for extracting the useful information from the unstructured and semi structured data because data mining assumes that the data is present in the relational format. While text mining deals with various forms of data therefore many text mining techniques and algorithms are designed for the data retrieval.

Text mining finds interesting and correct patterns from large amount of data. It is also known as intelligent text analysis. It includes knowledge discovery (KDT) from the text also known as intelligent analysis of the text. Some resources of the data in the world includes news articles, govt. repositories, WWW, digital libraries etc. so, the knowledge discovery from these varied resources has become a research area of the great importance. The various techniques that are used for the text mining includes Information Extraction, Information Retrieval, Natural language processing, Clustering and Text Summarization. There are various areas where text mining techniques are used for the information retrieval like digital libraries, Life science, social media, business intelligence etc. These various areas uses text mining for the retrieval of useful information from large volumes of data.

II. TECHNIQUES FOR TEXT MINING

There are various techniques for the text mining like Information Extraction, Information Retrieval, Natural language Processing, Clustering and Text

summarization. These techniques are used to perform different tasks on the text like text patterns and their mining process on the data which is to be mined.

A. Information Extraction:-

This is the first step in the text mining process for analyzing the text. This Technique identifies the key phrases and relationship within the text. It is very useful when the volume of the text is very huge to analyse. Domain experts specify the attributes and relation according to the domain. Information Extraction systems are used to extract specific attribute and entities to define the relationship in the data. But to perform these tasks, this technique includes tokenization, identification of named entities, sentence segmentation, and part-of-speech assignment. Firstly the parsing of the text is performed and is semantically interpreted then after that the data is saved in the database for further processing.

The most challenging task in the text mining process is that the electronic data is not present in the form of relational databases and is collected from the varied different sources. But, the Information extraction technique has solved this problem by transforming the unstructured text into relational format such that it will be easy to mine the text further.

B. Categorization

This is the second technique for the mining of the text. This technique assigns category to the text that undergoes categorization It uses various input and output examples to categorize the text. The classifier used is trained for the known sources of data such that it can be able to categorize the data which is unknown for it. The categorization technique helps to categorize such that the data can be identified uniquely in the form of classes or the particular category to which it belongs. The categories or the classes that are used to categorize the data are predefined.

This technique also includes various methods such as pre-processing of the data, indexing of the data, reduction on the basis of dimensions and finally the classification. This technique is very important for the mining of the unknown text because the relational data that exists in the database is having relation between the entities and the attributes of the table but still it is difficult to identify the category of the text but Classifier made it possible by assigning predefined classes to the data.

C. Clustering

This technique of the text mining is used to find the text documents which contains the similar content in them. It does not cluster the text into any predefined class or category, it designs various clusters such that each cluster contains many similar type of documents containing the similar text. The clusters are different to each other but the contents of a cluster that are stored in it are similar to each other mainly in context to the content of them. This technique clusters the text but also keep track that any useful document should not be left.

It uses various clustering algorithms that are used to cluster the text. It performs clustering in such a way that if P is the number of the Clusters then d is the number of documents that are stored in one cluster and the d documents will contain the similar type of contents in them. There are various methods for the clustering, some of them are hierarical, distribution density centroid and k-mean.

D. Visualization

This technique for mining the text is used for improving and simplifying the discovering of the relevant information. This technique arranges the text in the form of a visual hierarchy. It is used by the government to identify the terrorist

networks or to find the information about the crimes. The visualization goal mainly includes the following steps to be performed:-

This step includes collecting the data and then prepairing a original data space.

Now the original data that is collected is then analysed and then extraction from that data is performed.

Various mapping algorithms are used to map the original data into the visualization target.

E. Summarization

This technique is used to reduce the length of the detailed text document such that the user can identify by the summary that the information which is needed by the user is present in the long text or not. It will be very time consuming sometimes be worthy for the user to read the whole text document and then identifying that it contains useful information for him.

In past time, the text summarization was done on the basis of count of the words in the textual document but later on many methods were added for the summary. Now the text summarization involves algorithms that are applied for obtaining the summary. The method of the summarizing the text includes the following steps:-

This step includes text preprocessing for obtaining the structured format of the large text document.

After obtaining the text in the structure format the summarization algorithm is applied on the text and the the summary structure is obtained as output of that algorithm.

In this step the summary of the text is obtained from the summary structure.

III. TEXT MINING APPLICATIONS

A. *Life Science*

Life Science and health care industries generated a lot of textual data in varied formats that contains various medicines information, various Patients information, Diseases information and treatment of various diseases according to the symptoms etc. This data is very huge and it is a very challenging task to extract useful information without using the text mining. Life science industries uses text mining tools to extract useful information that is stored in varied unstructured formats. Mining tools in biomedical field provides a simple and helpful way to extract valuable information, their association and inferring relationship among the various diseases, genes and species.

B. *Digital Library*

Digital library that provides facility to a lot of millions of people to search and read a lot of books, magazines, newspapers etc. people can learn easily they do not need to go to the library and then search for books from a huge collection. People can easily search and learn from any place through online library. The online library is providing a lot of facilities to the users so the stored data contains a lot of collection such as text, images, graphs etc. so, to extract useful and valuable data the digital library uses text mining for searching of books and other things online by the users. Digital library uses text mining for the pattern matching of the names of the books , magazines , newspapers etc.

C. *Business* Intelligence

Text mining in business intelligence helps organizations to manage the data of the competitors, customers and employees to take better decisions. There are various text mining tools that provides a deep analysis to the high level officers in the industries to improve their business, to take decisions that will satisfy their customers and to improve demand of their products in the market.

IV. SOCIAL MEDIA

Text mining tools and software packages are too used in the social media for analayzing the plain text from the internet, blogs, emails etc. Text mining is used to analayze the number of posts, likes , followers etc. The data that exists on the social media exits in varied unstructured and structured formats such as text, images. This data can only be handled or analysed through the text mining that's why its softwares are used for the analysis.

V. CONCLUSION

The overall conclusion of this paper is that it describes various techniques that can be used for analysis of the huge amount of text. It describes a brief overview of the various text mining techniques that are used for analyzing or mining the useful and valuable information from the huge unstructured and structured data. This paper also describes various application areas where the text mining is used for the extraction of the valuable information.In future research work, we will focus to design efficient algorithms for the techniques that can be used for the text mining.

REFERENCES

1]] Ramzan Talib, Muhammad Kashif Hanif, Shaeela Ayesha, and Fakeeha Fatima, "Text Mining: Techniques, Applications and Issues", Vol. 7 No. 11, 2016.

2] K. Sumathy and M. Chidambaram, "Text mining: Concepts, applica- tions, tools and issues-an overview," International Journal of Computer Applications, vol. 80, no. 4, 2013.

3] N. Padhy, D. Mishra, R. Panigrahi et al., "The survey of data mining applications and feature scope," arXiv preprint arXiv:1211.5723, 2012.

4] 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.

5] Sonali Vijay Gaikwad, Archana Chaugule, Pramod Patil," Text Mining Methods and Techniques", Volume 85 – No 17, January 2014.

6] 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.

7] Fang Chen, Kesong Han and Guilin Chen (2008), "An approach to sentence selection based text summarization", Proceedings of IEEE TENCON02, 489- 493.

8] Seth Grimes (2005),"The developing text mining market", white paper, Text Mining Summit05 Alta Plana Corporation,,Boston, 1-12.

9] Borko, H. and Bernier, C.L. (1975) Abstracting concepts and methods. Academic Press, San Diego, California.

10] Ian H. Witten," Text mining", Computer Science, University of Waikato, Hamilton, New Zealand email ihw@cs.waikato.ac.nz .

11] 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.

12] Pak Chung Wong, Paul Whitney and Jim Thomas," Visualizing Association Rules for Text Mining", ", International Conference, Pacific Northwest National Laboratory, USA, 1-