AN OVERVIEW OF TEXT MINING- APPLICATIONS AND TECHNIQUES

R.Prema[1], MrV.P.Muthukumar[2]

Research scholar[1], PG and Research department of computer science and applications Vivekanandha College of arts and sciences for women,

Assistant professor[2], PG and Research department of computer science and applications, Vivekanandha College of arts and sciences for women [autonomous], , Tiruchengode

ABSTRACT

In the modern world, Data mining is the major issue to deal with huge dataset.it always deals with the text. So that, mining become essential to develop the techniques and methods to extract the data from unstructured /semi structured data**.** Text mining involves a series of activities to be performed in order to efficiently mine the information. It process in linguistic processing or natural linguistic processing (NLP) and it involves in both supervised learning and unsupervised learning. Text mining broadly used by government sector, research institution, medical care, security, business sector, education areas, netting, magazines and daily needs etc. Thus, it has become essential to develop techniques and algorithms to get useful information.

Keywords-Text Mining, linguistic processing, Information Extraction.

INTRODUCTION

Data mining is used extract extraction of patterns and knowledge from large amounts of data not extracting data itself. It is the process method of discovering patterns in massive knowledge sets involving ways at the intersection of computing, machine learning, statistics and info systems [1]. Text Mining is a new area that searches to take out meaningful data from text language. It can be designed as the flow of analyzing text to separate information that is needful for a specific purpose. Examining with the type of data stored in databases, text is not designed, unclear, and hard to process. Yet, in today's society, text is the most commercial way for the formal exchange of information. Text mining is same as data mining, except the data mining tools [2] are designed to use structured data from databases, also text mining can work in fields with unstructured or semi-structured data sets like emails, text documents and HTML files, social media, etc. As a result, text mining has a extreme better solution. Text mining is a process to get interesting and important patterns to explore knowledge from textual databases [3].

Text mining is used in many areas like risk management, cyber security management, fraud detection, Contextual Advertising, Business intelligence, Content enrichment, Spam filtering, Social media data analysis, Knowledge management. Text mining is the common process of structuring the input text data (which usually includes the methods like parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a database), deriving patterns within the structured data, final correction and interpretation of the output**.** The techniques are classified as categorization, entity extraction, sentiment analysis and others, text mining extracts the useful information and knowledge hidden in text content

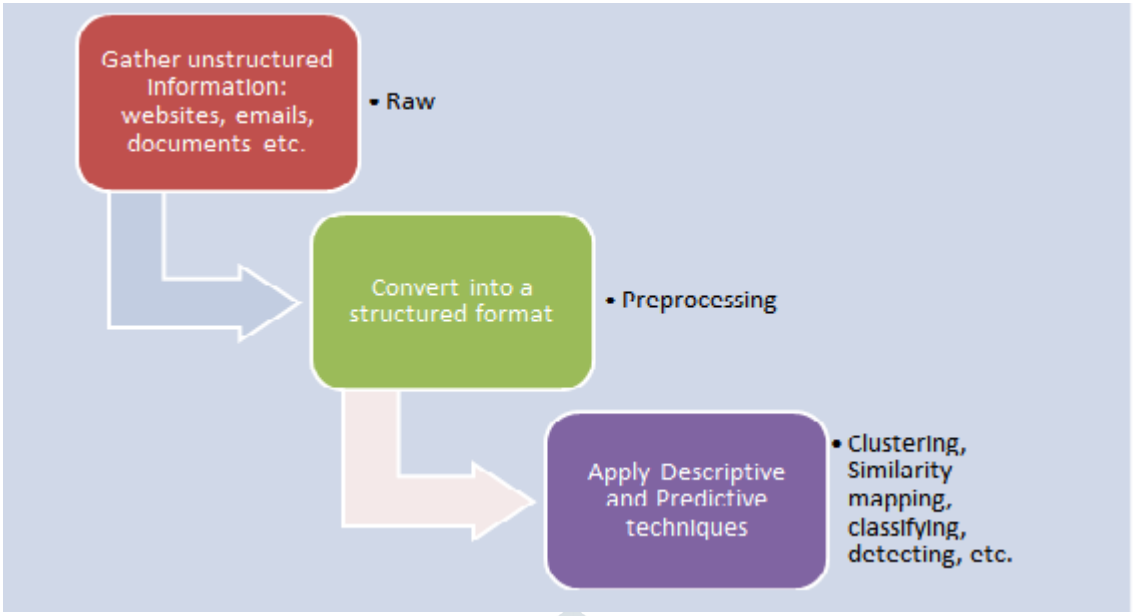

Fig.1: Basic Process of Text Mining

AREAS OF TEXT MINING

Text analysis involves

1. Sentiment analysis
2. Information extraction
3. Natural language processing
4. Data mining
5. Machine learning

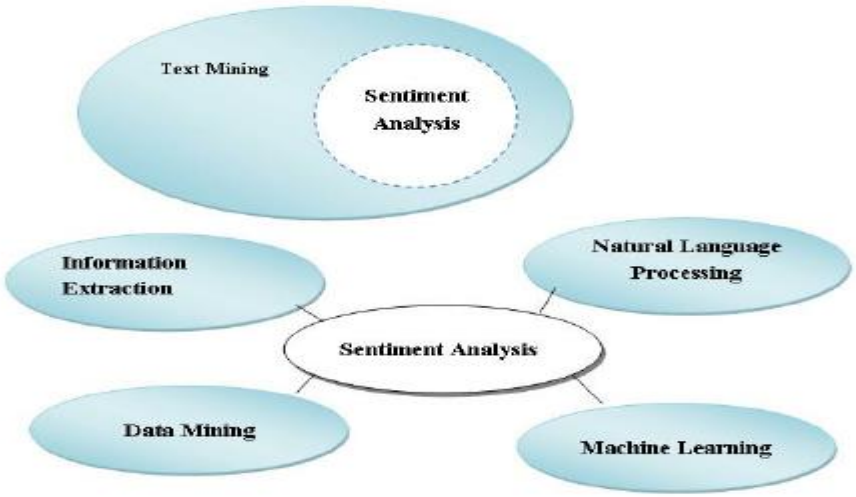

Fig.2: Text mining areas

TEXT PRE PROCESSING

Text preprocessing is tedious stage in the text mining. Text preprocessing used to extract the useful information and little valuable from unstructured text data. The text preprocessing involves in removal of noisy data and identifies the root word and decrease the length of the text. There are three steps in preprocessing namely tokenization, stop word removal and stemming.

Tokenization

It process of fragment the words, phrases, symbols and other meaning words is called token. Challenges in tokenization it affects writing system and structure of the words.

Stop Word Removal

Stop removal words is removing irrelevant data from the data. It can be done in two methods one list out the pre-defined stop removal words and using it are creating by us.

Stemming

Stemming is the process used to reduce words into their roots. It applicable for words with greater than three times occurrences.

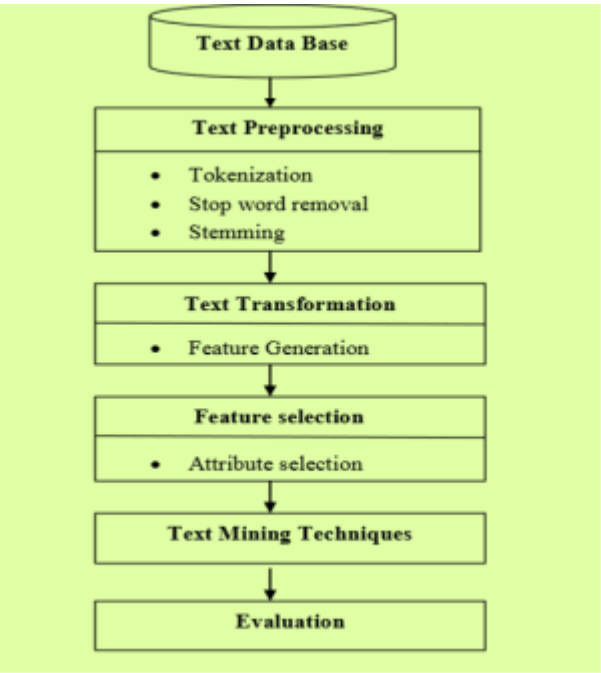

Fig.3: Text preprocessing

Text Transformation (Attribute Generation)

A text document is representing by the words (features) it contain and their occurrence. Two approaches of document representation are

a) Bag of words

b) Vector Space

Feature Selection (Attribute Selection)

Feature selection also called as variable selection, the selecting a subset of important features for use in model creation. The main assumption when using a feature selection technique is that the data contain many superfluous or irrelevant features. Redundant features are the one which provides no additional information. Irrelevant features provide no helpful or related information in any context.

Evaluation it verify the result for the exactness, after the corrections the result can be omitted or the generated result can be used as an input for the next set of string.

APPLICATIONS IN TEXT MINING

1. Text categorization-it categorization is based on specific domain.
2. Text clustering-grouping similar text in the document.
3. Sentiment analysis-it is based on the emojis activities.
4. Concept/entity extraction-it extracts the essential information from unstructured or semi structured data.
5. Document summarization-it summarizes the original document.
6. Learning relations-it is based on the personal interest.
7. Parts-of-speech of the language-such as subject, noun, and verb etc., in a text document.

CONCLUSION

Text mining can be used huge volume of data**.** Now days, more number of text document are used by people in their day to day life. Text mining glowing in a height and more number of techniques and algorithm. This paper gives a brief detail about text mining process. Selection and use of correct methods and tools according to the domain will help to make the text mining method more easy and efficient. Domain knowledge based integration, concepts based granularity, multilingual text of refinement, and using natural language processing ambiguity are major issues and risks that arise during text mining techniques. In the medical field, text mining tools is more effectively helpful for calculating the value of medical treatment and compare with various diseases and symptoms.

REFERENCES

1.G.Kesavaraj; S.Sukumaran, "A study on classification techniques in data mining" Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), IEEE *Xplore*, DOI **:** 10.1109 /ICCCNT.2013.6726842, 4-6 July 2013, Pp. 1-7

2.Vishal Gupta, Gurpreet S. Lehal, 2009. ―A Survey of Text Mining Techniques and Applications‖ in Journal of Emerging Technologies in Web Intelligence, Vol. 1 No. 1.

3. C. Ding and H. Peng, ―Minimum redundancy feature selection from microarray gene expression data,‖ Journal of bioinformatics and computational biology, vol. 3, no. 02, pp. 185–205, 2005.

4.W. Fan, L. Wallace, S. Rich, and Z. Zhang, ―Tapping the power of text mining,‖ Communications of the ACM, (vol. 49, no. 9, pp.76–82, 2006).

5.S. M. Weiss, N. Indurkhya, T. Zhang, and F. Damerau, Text mining:predictive methods for analyzing unstructured information.(Springer Science and Business Media, 2010.)

6.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.)

7.IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN: 2278-8727 PP 46-51 www.iosrjournals.org Next Generation Computing Technologies 46 | Page Sankara College Of Science And Commerce A Survey Paper on Text Mining - Techniques, Applications And Issues *Mrs.B.Meena Preethi1 , Dr.P.Radha2