ROLE OF TEXT MINING FOR DATA ANALYSIS

[1] **Devendra Kumar Mishra**
[1]Assistant Professor
[1]Computer Science and Engineering,
[1]Amity University Madhya Pradesh , Gwalior, India

Abstract : The base of tod**ay's** world is data. companies are realizing the importance of using more data to make decision for their planning to improvement. The data that are generating due to development of technology may be structure ,unstructured or semi structured. Data typically existing in relational database is called structured . This data smoothly map into pre-formatted fields. In other hand unstructured data is different from relational data.it doesn't fit in any kind of models that are predefined . The unstructured data files contain multimedia and text content, messages in E-mail, videos , pages on internet, audio data files, pictures, presentations data , and other business documents may be come in category of unstructured data.The data deal by any organization is 85 to 90 percent is of unstructured types. The term text mining indicate the process of retrieve result of appealing query or unknown knowledge from unstructured text.Text mining is an multidisciplinary field related to information extraction,use machine learning approach,analysis tools on statistics , and data mining. text mining contain some step that involve preprocessing of ducuments ,classification process ,clustering on the basis of feature ,information retrieval and finally, visualization.

***IndexTerms* - Bigdata analytics, Supervised Machine learning, Unstructured Data.**

I. Introduction

Today is the era of internet, internet represents a big space where large amounts of data are added every day. This huge amount of digital data and interconnection exploding data.

This large group of data is called big data.Big Data term is introduced by Roger magoulas in 2005.they define it as large amount of data that is not process and manage due to complexity and size by traditional data management techniques[1][2]. **Big Data mining** have the capability to retrieving useful information in large datasets or streams of **data.** Analysis can also be done in a distributed environment .the framework needed for analysis to this large amount of data must support statistical analysis and data mining. The framework should be design in such a way so that big data and traditional data can be combine.so results that comes analyzing new data with the old data[3][4] . Traditional tools are not sufficient to extract information those are unseen . Machine Learning approach contain algorithms based on statistic methods those are capable for analysis of big volume data in real time. there are so many learning approaches available to solve specific problems,but supervised and unsupervised learning commonly used. K-means clustering is one of the example of unsupervised learning[5].

In case of supervised learning labeled training data available that guide to calculate the value of given input.example includes handwriting recognition, classification of e-mail messages[6]. There are lots of algorithms available to create learners, for example Support Vector Machines , Naive Bayes Classifiers ,and Neural Networks. In case of Unsupervised learning no guide is available to make sense of data, this approach is generally used for clustering purpose[7][8] .

II. AREAS OF TEXT MINING

we can divide Text Mining into seven highly interrelated practice areas by their unique feature .

a) **information retrieval (IR) and Searching:** This area include indexing of documents, extraction of documents from large databases and search the desire outcome on the basis of keyword queries.

b) **Document clustering:** This area deal with clustering using algorithms to prepare similar documents into a group.

c) **Document classification:** Document Classification use classification methods of data mining for categorizing object. that are based on trained models.

d) **Web mining:** web mining have large amount of data that are available on web.

e) Web data generally present in a structured way of text format that include links on pages. It focus on text mining on the Internet.

f) **Information extraction (IE):** The aim of this area is to identify and extract of facts and relationships .it focus on build structured data from semistructured or unstructured text.

g) **Natural Language Processing (NLP):** This area moved focus on text mining as a tool for extracting meaningful information for text mining.

h) **Concept extraction:** It deal with Extracting concepts by Combining of words and phrases in groups on the basis of semantic.

III. APPLICATION OF TEXT MINING

Text mining application examines unstructured textual information to find structure and meanings those are available but hidden within the text. In this , we try to uncover patterns, relationships among object , and trends present . Applications of text mining can be define into a number of domains on the basis of analysis type or by business function. text mining application categories include:

- Social media monitoring
- Natural Language/Semantic Toolkit
- Sentiment Analysis Tools
- Enterprise Business Intelligence
- Search/Information Access
- E-Discovery
- Scientific discovery, especially Life Sciences
- Listening Platforms
- Publishing
- Automated ad placement
- National Security
- Competitive Intelligence

IV. PROBLEMS IN TEXT MINING

This can be seen as a practice of numericizing text.all words found in the input documents will be indexed and counted to construct a table of documents and words ,after this process we have a matrix that showing frequencies of each word in documents. once a table of words from documents derived ,all standard statistical and data mining methods can be apply to predict outcome of interest.

Some other problems that arises while text mining are:-

(a)**Stop List:-** prepare Stop list is major issue in text mining.stop list contains high frequency words such as the,a, to, of etc. that should be ignored from the text.

(b) **Word Sense disambiguation:-** Meaning of word should be clear.

(c) **Stemming:-** this is also known as lemmatization it deals with reduce the words to their stems.

(d) **Noisy data:-** Text data should be clear from noisy data.

(e)**Tagging:-**it specify data annotation or characteristics of speech.

(f) **Collocations:-**Specify how we deal compound or technical terms.

(g) **Grammar/syntax:-** we should make syntactic or grammatical analysis.

(h)**Tokenization:-** For tokenize which method is using .

(i)**Text Representation:-** How we represent text? which model we are using for text representation?

(j) **Automated learning:-** is our approach having self learning capability?

Following problems are identified in text mining process:-

a) How analysis is do for patterns existing in logs with improve security ?

b) How Effectively analyze information from call centers to get pattern and improve customer satisfaction?

c) How efficiently analyze content of social media to improve services and products?

d) How easily and accurately detect fraud in the online transactions?

e) How effectively analyze information coming from financial market for risk assessment?

V. CONCLUSION

Presently Researchers are trying to develop new methodology for text mining like rule based approach,analysis tools based on statistical ,machine learning concept .on the other hand for solving text mining problem NLP and IE technique are more suitable.the NLP emphases on text processing and IE on retrieve information from textual data. after information extraction that is stored in database for query purpose , mining of data,and for summarization.

REFERENCES

[1]Dursun Delen and Asil Oztekin ,2014,Introduction to Data ,Text and Web Mining for Managerial Decision Support Mini –track" 47th Hawaii International Conference on System Science(HICSS),IEEE,768.

[2] Lu Gao and Neil Eldin 2014. Employers Expectations:A Probabilistic Text Mining Model" Procedia Engineering ,Vol.85,pp-175-182.

[3] Renaud Richardet,Jean Cedric Chappelier, Shreejoy Tripathy and Sean Hill, 2015,Agile Text Mining with Sherlok",International Conference on Big Data,IEEE,pp-1479-1484.

[4] Guo Aizhang and Yang Tao,2015. Based on rough sets and the associated analysis of KNN text classification research 14th International Symposium on Distributed Computing and Applications for Business Engineering and Science,IEEE,485-488.

[5] Celia Satiko Ishikiriyama, Diego Miro, and Carlos Francisco Simoes Gomes,2015.Text Mining Business Intelligence:A small Sample of what words can say",Procedia Computer Science,vol.55,261-267.

[6]Xuan Lv,and Nora El-Gohary,2016.Text analytics for supporting stakeholder opinion mining for large scale highway projects" Procedia Engineering 145,518-524.

[7] Wonchul Seo, Janghyeok Yoon, Hyunseok Park, Byoung-youl Coh, Jae-Min Lee,and Oh-Jin Kwon,2016 ,Product opportunity identification based on internal capabilities using text mining and association rule mining " Technological Forecasting and Social Change,Vol.105,94-104.

[8] Jieun Kim, Mintak Han, Youngjo Lee,and Yongtae Park, 2016.Futuristic data driven scenario building :incorporating text mining and fuzzy association rule mining into fuzzy cognitive map",Expert Systems with Applications,Vol.57,311-323.