International Journal of Engineering Technologies and Management Research

IJETMR

A Knowledge Repository

CHALLENGES IN TEXT MINING FOR BUSINESS INTELLIGENCE

Devendra Kumar Mishra[*1]
[*1] Department of Computer Science and Engineering, Amity School of Engineering and Technology, Amity University Madhya Pradesh - 474020, India

Abstract:
Today is the era of internet; the internet represents a big space where large amounts of data are added every day. This huge amount of digital data and interconnection exploding data. 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 combined, so results that come analyzing new data with the old data. Traditional tools are not sufficient to extract information those are unseen.

Keywords: *Big Data; Business Intelligence; Machine Learning; Unstructured Data.*
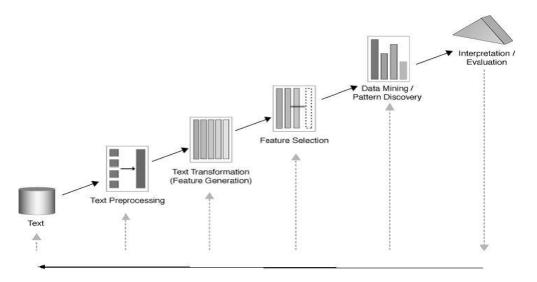
1. Introduction

Large amount of data are generating on WWW and can be used not only for browsing information but also to obtain useful knowledge. This large group of data is called big data. Big Data term is introduced by Roger magoulas in 2005.They define it as a large amount of data that is not process and manage due to complexity and size by traditional data management techniques[1]. Text analytics refers to the analysis of text data. This analysis is done for retrieving information. It is also known as text mining. Social network feeds, survey responses, emails, news, blogs, call center logs, online forums, and corporate documents. Text analysis has statistical tools for analysis, processing tools, and machine learning approach. Text mining is a technique for finding useful information from unstructured text data [2].Business Intelligence (BI) indicates approaches that Accumulate, store and analyze business data that finally used to make a decision [3]. BI helps to make better decisions for an organization it involves computer-based technology for the specifying present, past and future trends of the organization. BI has an important place in decision support system, online analytical processing, business performance management and predictive analysis. BI provides competitive intelligence by taking the data of competitors. BI also helps in knowledge management that is helpful to make better strategies. BI analyze enterprise data for decision purpose but all data is not available in a structured form that is easy to understand, data also exists in the semi-structured or unstructured form that is more time taking process for interpretation. So, in this circumstance decisions making is very complex. Text mining is a technique that extracts valuable information from the large volume of

unstructured text. In other word we can say that, unidentified information is derived for future predictions [4][5].

Text mining process indicates a system that analyzes the huge amount of text data by parsing it and find out lexical or linguistic patterns to extract correct information [6]. Text mining process looking for patterns in the text by automatic extraction. This is a flexible technique for management of information, research, and textual data analysis and interpretation [7].

2. Text Mining Process

Text Mining can be used for descriptive or predictive purposes. The step those are taken for text mining is shown in figure.

Text- This involves following -

a. Document Clustering: In this textual data, decomposes and transformed into the quantitative representation   that is suitable for analysis and decision- making purpose.
b. Text Characteristics:-The text used for mining should allow dependency of words and phrases and allow different input modes for human or automated consumers.

Text Pre-processing- After document identification pre-processing will start. Text pre-processing contain following steps:-

a) **Text Cleanup-** This process includes deletion/removal of advertise data from the web and normalize data so that processed text is free from redundancy and fake text generation.
b) **Tokenization-** This is a parsing technique in which text is split into tokens.
c) **Parts of Speech Tagging-** In this tagging process, grammar rules used on the text according to parts of speech.
d) **Word-sense Disambiguation-** This process finds different senses of a word that are used in many places in different situations.
e) **Semantic structures-** it allows studying the semantics of text stored in the document warehouse.

Text Transformation/ Attribute Generation- In this step labels generates from the text. it contains following steps-

a. **Text Representation-** It allows representation of text by features and their occurrences using approaches of "Bag of words" and "Vector space" where each word is represented as an individual variable having numeric weight.

b. **Feature Generation-** this process selects features of a document so as to improve its representation which may be earlier misleading or redundant. It uses approaches of selection before use or selection based on use.

c. **Feature Selection-** This step is deal with dimension and irrelevant attributes reduction. It improves the representation of text by choosing **a** subset of features.

d. **Data Mining-** This step is also known as Knowledge Discovery and Data Mining (KDD). It extracts useful patterns from the database.

e. **Interpretation/ Evaluation-** This is the last step in the process which terminated if well-suited results are achieved for business intelligence or the process could be finally iterated if the results are not up to the mark or are used as a part of further inputs.

3. Challenges With Mining

a) Moving text analytics from a standalone application into a component of other applications.

b) Incorporating social media into text analytics in ways that better benefit businesses.

c) Handling data (or at least text) volumes are increasingly important.

d) How exactly do text analytics tools determine sentiment and deal with variables such as irony or sarcasm?

e) How do text analytics deal with the slang/vernacular/abbreviations?

f) Is there a danger that because of re-tweets and forwarded emails, in that case these tools analyzing the same text numerous times?

4. Conclusion

This paper provided a brief introduction to data mining for business intelligence. Businesses need quick decision for improvement of their business process. The data that can make effect on business process are coming from different sources and mostly in text form. So effective text mining technique is required. There are so many challenges and issues involve with text mining while extracting information from text.

References

[1] Jeffrey Friedlin, Malika Mahoui, Josette Jones and Patrick Jamieson "Knowledge Discovery and Data Mining of Free Text Radiology Reports" First IEEE International Conference on Healthcare Informatics, Imaging and Systems Biology, IEEE, pp-89-96, 2011.

[2] Naoya Otsuka and Mitsunori Matsushita" Constructing Knowledge using Exploratory Text Mining", SCIS &ISIS, Kitakyushu, Japan, IEEE, pp-1392-1397, 2014.

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

[4]   Shiqun Yin, Yuhui Qiu and Jike Ge, "Research and Realization of Text Mining Algorithm on web" International Conference on computational intelligence and security workshops, IEEE, pp-413-416,2007.

[5]   Suan Lee,Namsoo Kim and Jinho Kim, "A Multi-Dimensional Analysis and Data Cube for Unstructured Text and Social Media" Fourth International Conference on Big Data and Cloud Computing, IEEE, pp-761-764, 2014.

[6]   W.Sunayama,Y.Takama,Y.Nishihara, T.Kajinami ,M.Kushima and H.Tokunaga "Practical application in development and use of mining tools with total environment for text data mining" Journal of the Japanese Society for Artificial Intelligence, vol.29, no.1, pp-100-112, 2014.

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

*Corresponding author.
E-mail address: dkmishra@ gwa.amity.edu