An Empirical Study on Text Analytics in Big Data

R.Merlin Packiam
Assistant Professor of Computer Science
Cauvery College For Women, Tiruchirapalli, India
rmerlinpackiam@gmail.com

Dr. V. Sinthu Janita Prakash,
Head, Assistant Professor of Computer Science
Cauvery College For Women, Tiruchirapalli, India
sinthujanitaprakash@yahoo.com

***Abstract*— Today's world is flooded with unstructured information. Big data is not just a description of raw volume but it has to real issue of usability. The major part of information retrieval is giant experience in big data. The real challenge is identifying or developing most cost effective and reliable methods for extracting value from all the terabytes and petabytes of data now available. That's where big data analytics become necessary. Conventional analytics focused on structured data but these methods are not appropriate for large volume of unstructured data in order to extract knowledge. Text analytics is the way to extract significance from the unstructured text to find out patterns and transformations. The importance of text analytics is increased more in social media and business intelligence. This study reveals that big data text analytics can breed new insight to the world of text information and discusses various researches carried out in text analytics.**

Keywords— Big data, unstructured data, text analytics.

I. INTRODUCTION

Today's data is innumerable and vast as wikibon blog reveals that the big data in today's business and technology environment, 2.7 zeta bytes in digital universe, face book stores, accesses and analyzes 30plus petabytes, in 2008 , Google was processing 20 petabytes (20,000 terabytes) a day and rapid growth of unstructured data is also enormous level . Data production will be 44 % greater in 2020 than it was in 2009 and marketer's challenges have become a big data. Increasing in data leads to the progressive demand for hard disk drives(HDDs) and the role of Big Data in the current environment of enterprise and technology[1]. So retrieving information gets more challenge because of the large volume of heterogeneous data that is big data. Most of the big data researches are at an early stage in their journey.

The rest of the paper is organized as follows. Section II explains fundamental concept of big data and its characteristics: Section III discusses the text mining; Section IV discusses the related work done in text mining; Section V describes the summary and concludes the paper.

II. BIG DATA

A. Big Data and its Characteristics

Big data is the capacity to manage a huge volume of dissimilar data at the precise speed and within the right time frame to allow real-time analysis and reaction[2]. Big data is characterized by three dimensions volume, variety and velocity. These are reasonable basic dimensions to quantify big data and take into account the typical measures around 3V's volume, variety and velocity dimension, which is a key compounding factor. Volume means quantity of data; variety means a variety of data and velocity means processing speed of data. Now, more V's are introduced other than the basic 3V's such as value , variability and veracity. Big data incorporates all data, including structured data and unstructured data from e-mail, social media, text streams, and more.

B. Big Data Analytics

The big data challenges include analysis, capture, search, sharing, storage, transfer, visualization, and privacy infringement [3].

Big data analytics is the process of examining a large data to expose concealed patterns, unfamiliar correlations , market trends, customer preferences and other valuable information which is useful to create better decisions. These findings can guide to more effective marketing, new proceeds, better customer service, enhanced operational efficiency, competitive advantages over competitor organizations and other business benefits. Big data analytics used a wide variety of advanced analytics [4], as listed in Table 1.

Table 1. Analytics Spectrum[4]

Big data analytics	Techniques
Sql Analytics	Count, Mean, OLAP.
Descriptive Analytics	Univariate distribution, Central tendency, Dispersion.
Data Mining	Association rules, Clustering, Feature extraction.
Predictive Analytics	Classification, Regression, Forecasting, Spatial, Machine learning, Text analytics.
Simulation	Monte carlo, Agent-based modeling, Discrete event modeling.
Optimization	Linear optimization, Non-linear optimization.

C. Text Analytics

Big data is set to transform business, but text analytics will take part in a enormous role in creating transformation. The unstructured data is textual in nature so it is a big pact to the text analytics in big data. The ability to

978-1-4799-7849-6/15/$31.00 ©2015 IEEE

extract meta data from unstructured is major responsibility of text analytics and plays a huge role in transformation. Text analytics is the process of analyzing unstructured text, extracting relevant information and transforming it into structured information that can then be leveraged in various ways[2].

III. TEXT MINING

Mining is discovering patterns in data. Text Mining is a subfield of data mining, focussing on Knowledge Discovery from unstructured text data. From the unstructured data, the process of extracting information and the facts is called text mining or knowledge discovery in text (KDT). This is the rising field of information recovery, statistics, machine learning and computational linguistics. It leads useful applications such as loan default analyses, sentiment analyses, opinion mining, medical diagnostics and e-discovery and so on. Text mining can assist to receive potentially valuable industry insights from text-based content such as word documents, email and postings on social media streams like Face book and Twitter. Because of the incoherent data, text mining with natural language processing (NLP), statistical modeling and machine learning techniques can be difficult.

On a functional level, text mining systems follow the general model provided by classic data mining applications. Preprocessing tasks and core mining operations are the two most important areas for any text mining system and typically describe serial processes within a generalized view of text mining system architecture[5], as shown in Figure 1.

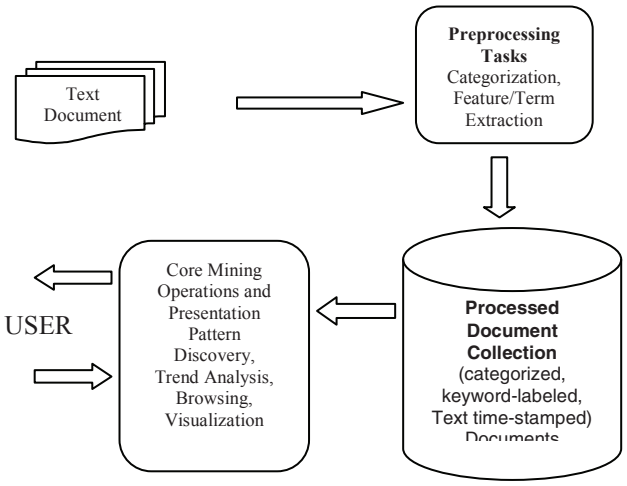

Figure 1. Text Mining Functional Architecture

A. Information Extraction

Information extraction is the task of finding structured information from unstructured or semi-structured text. It is an vital task in text mining and has been widely studied in various research communities including natural language processing, information retrieval and Web mining. Basic measures[6] for text retrieval are precision and recall.

B. Text Mining Approaches, Tools and Techniques

There are many approaches to text mining, which can be classified from different perspectives, based on the inputs taken in the text mining system. In general, the major approaches are the keyword-based approach and the information extraction approach. Text analytics solutions use a combination of statistical and content analysis to extract information from unstructured data. Statistical analysis is carried on text at various dimensions such as term frequency, document frequency, term proximity, document length. Content Analysis on text at different levels such as

1) Lexical or syntactic processing : Recognizing tokens, normalizing words, language constructs that is sentence, parts of speech and paragraph.

2) Semantic processing : Extracting meanings, name entity extraction(categorization, summarization, query expansion and text mining.

3) Extra semantic feature : Identify feelings or sentiments(feelings, emotions, and mood).

4)Goal : Dimension reduction.

Big data infrastructure deals with Hadoop. It is a framework of Java based program that supports the processing of large data set in a distributed computing environment. Hive is a data warehousing infrastructure built on top of apache hadoop. MapRaduce is a framework by google and it is used to distributed system. Mahout is a Machine language and other related software like Storm, HPCC, GridGain. Now more tools are available for big data analytics.

Various types of data sets are available such as records, graphs (web, social science and network, molecular structure), ordered data (sequence of image, transaction sequences, genetic sequence data), spatial image (map) and multimedia (image, video) data. The data bases are also existing in open source such as Cassandra by facebook , hbase by apache, mongoDB, neo4j, couchDB, flockDB by twitter and hypertable by Nosql and so on. The open software like kttcoder, Corrot2, Natural language Toolkit and GATE are joining together with big data analytics to meet the necessity of unstructured data.

C. Application of Text Mining

General applications of text mining areas are Relationship Analysis, Trend analysis , Mixed applications and business application of text mining areas such as Decision Support in CRM, knowledge management and Personalization in E-Commerce. There are many opportunity and challenges in Big Data text analytics, the most often used in the following areas: Sentiment analysis, Search access of unstructured data, Social media monitoring, Competitive intelligence, E-discovery, records management, Scientific discovery especially life science, Publishing and media, Pharmaceutical, research companies and healthcare.

IV. RELATED WORK

The business needs to apply the information to take action that can transform production outcomes. In social

media analytics, a lot of visibility recently and in fact, is helping to drive the text analytics market. These are some of the examples of how text analytics can be used to help gain insight into data.

In this chapter some of the related works are discussed here. Table 2. explains the brief details of some of these work which is based on tools , techniques , findings and data source.

Table 2. Some of the Related Work Performance

Author	Techniques / Tools	Data Source	Extracting Information from data source	Findings
He et al., [7]	SPSS Clementine text mining tool and Nvivo 9.	Tweeter and Face book	Customer-generated content of the three largest pizza chains.	This case study made a contribution by using text mining to perform competitive analysis for the user-generated data on Twitter and Facebook in three major pizza chains
Jonnagaddala et al., [8]	Text mining, Apache Ruta.	Electronic Health Record (EHR)	Risk factor data Framingham risk score (FRS).	A rule-based text mining system was developed to extract risk factors to calculate 10-year CAD FRS
Xiang et al., [9]	Opinion mining and PivotTable function in Microsoft Excel, Factor analysis.	Online customer reviews and Expedia.com	Guest experience and satisfaction.	This study applies text analytics to classify a large amount of online customer reviews, assess the quality of these data, as well as identify inherent relationships between two domains of variables in hotel management
Ishikiriyama et al.,[10]	Text mining and Frequency distribution Integration and querying R-project, version 3.1.2 and packages tm.	Business Intelligence or Business intelligence or business intelligence in a Brazilian academic search engine.	Relevant papers BI	Small sample of what is possible to achieve by analyzing text data from academic papers by using the software R-project.
Bhaskar et al., [11]	Speech and text (Sentiment) mining and SVM classifier, Feature Extraction.	Semval -2007 and eNTERFACE'05 EMOTION Database.	News headlines, wav form of audio of corresponding Dataset.	New classification method to detect the emotions in utterances of human speech. This method exploits both audio and textual features corresponding to it.
Agarwal et al., [12]	Sentiment analysis and mRMR feature selection technique, feature extraction.	Benchmark movie review dataset provided by Cornel University and product review data sets on books,DVDs, and electronics.	Unstructured natural language text.	Novel feature extraction method that uses the dependency relation between words to extract features from text.
Weichselbraun et al., [13]	Opinion mining and Naïve Bayes approach, SenticNet, ConceptNet query.	Structured resources, WordNet and ConceptNet.	Electronics and software product reviews from Amazon as well as reviews from the IMDb categories comedy, crime, and drama.	A novel method to extend sentiment lexicons with concept knowledge, which aims to increase the lexicons' coverage and derive concept information for subsequent opinion mining.

V. SUMMARY AND CONCLUSION

The text analytics research in big data is very vast, new and the challenges is also outsized. Some of that challenges are studied and presented in the Table 2. Maximum number of research area is in social media with sentiment analysis and the business intelligence like hotel management and marketing. The data bases are taken from the open source database and social network like twitter and facebook. Recently, fast developments in genomics and proteomics have generated a huge amount of biological data it gives new challenging and opportunity of text analytics.

To develop the data quality and the analysis results understanding the technique by which data can be preprocessed is essential. Datasets are often very large , and they originate from heterogeneous sources. Hence, current real-world databases are highly susceptible to inconsistent, incomplete, and noisy data. Therefore, numerous data preprocessing techniques, including data cleaning, integration, transformation, and reduction, should be applied to remove noise and correct inconsistencies [4]. Big data challenges such as data capture, storage, searching, distribution, analysis, and visualization. These challenges must be overcome by new innovations. Thus, further research is needed to deal with these issues and improve the analysis and storage of Big Data.

Text analytics is being used in all sorts of analysis from predicting to fraud and to social media analytics. This paper presents insights and depict the main concerns and the main challenges for the future with some of the research papers. From the taken information we could not reach out a proper way to describe the text analytic techniques because of the configuration which is used by the tools. We need to improve the algorithms to overcome zero probability in naïve bays approach, scalability of algorithms and usage of tools in cloud with cost effective manner.

Big Data is the new boundary for the scientific data research and for business applications, as one lives in the a new age where Big Data text analytics will facilitate us to determine knowledge that no one has discovered before with cost effective. This paper highlights some of the rising research areas in text analytics that gives new inventive ideas to the further research studies.

REFERENCES

[1] Khan et al., "Big Data: Survey, Technologies, Opportunities, and Challenges", Hindawi Publishing Corporation the Scientific World Journal, Vol. 2014, Article ID : 712826 ,http:// dx .doi org. Http://dx.doi.org/10.1155/2014/712826.

[2] Judith Hurwitz, Alan Nugent , Dr. Fern Halper, Marcia Kaufman " Big Data For Dummies", Published by John Willey & Sons, inc.,

[3] Vikas Upadhyay, Insha Shaikh, " Big Data Analytics", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 5, Iss. 6, 2015.

[4] Michael Minelli, Michele Chambers, Ambiga Dhiraj, " Big Data Big Analytics", Wiley Cio Series, ISBN 978-81-265-4469-1.

[5] Ronen, Feldman, James Sanger," The Text Mining Handbook – Advanced Approaches in Analyzing Unstructured Data", Cambridge University Press.

[6] Han and Micheline Kamber, "Data Mining Concepts and Techniques" Second Edition.

[7] Wu He, Shenghua Zha, Ling Li, "Social media competitive analysis and text mining : A case study in the pizza industry",international Journal of Information Management , vol. 33, pp. 464-472, 2013.

[8] J. Jonnagaddala et al., "Coronary artery disease risk assessment form unstructured electronic health records using text mining", J. biomed Inform 2015, http://dx.doi.org/10.1016/jjbi.2015.08.003

[9] Z. Xiang, X. Schwartz, J.H. Gerdes, M. Uysal, "What can big data and text analytics tells us about hotel guest experience and satisfaction?", International Journal of Hospitality Management, vol. 44,nov 2015, pp. 120–130, DOI: 10.1016/jijhm 2014.10.13

[10] Celia Satiko Ishikiriyama, Diego Miro, Carlos Francisco Simoes Gomes, " Text Mining Business Intelligence : a small sample of what words can say", Procedia Computer Science, vol. 55, pp. 261 – 267,2015

[11] Jasmine Bhaskar ,Sruthi K, Prema Nedungadi , "Hybrid Approach for Emotion Classification of Audio Conversation Based on Text and Speech Mining", Procedia Computer Science, vol. 46, pp. 635 – 643, 2015. International Conference on Information and Communication Technologies(ICICT)2014.

[12] Basant Agarwal, Soujanya Poria, Namita Mittal, Alexander Gelbukh, Amit Hussain, " Concept – Level Sentiment Analysis with Dependency –Based Semantic Parsing : Novel Approach ", Springer Science and Business Media New York 2015, DOI 10.1007/s12559-014-9316-6

[13] A. Weichselbraun, S. Gindl, A. Scharl," Enriching semantic knowledge bases for opinion mining in big data applications", Knowledge-based systems,vol.69,pp.78-85,2014,