International Journal on Emerging Technologies 10(2): 50-59(2019)

ISSN No. (Print) : 0975-8364
ISSN No. (Online) : 2249-3255

Recent Trends in Sustainable Big Data Predictive Analytics: Past Contributions and Future Roadmap

Syed Muzamil Basha[1], Dharmendra Singh Rajput[2], S. Bharath Bhushan[3], Ravi Kumar Poluru[4], Rizwan Patan[5], R. Manikandan[6] and Ambeshwar Kumar[6]

[1]*Assistant Professor, Department of Information Technology, SKCET, Coimbatore, Tamil Nadu 641008, India.*
[2]*Associate Professor, School of Information Technology and Engineering, Vellore Institute of Technology, Vellore. 632014, India.*
[3]*Associate Professor, Department of Computer Science and Systems Engineering, Data Analytics Research Lab, Sree Vidyanikethan Engineering College, Tirupathi (Andhra Pradesh), India*
[4]*Research Scholar, School of Computer Science & Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu India*
[5]*School of Computing Science and Engineering, Galgotias University, Greater Noida (Uttar Pradesh), India*
[6]*School of Computing, SASTRA Deemed University (Tamilnadu), India*

(Corresponding author: R. Manikandan*)*
(Received 03 April 2019, Revised 01 June 2019 Accepted 05 July 2019)
(Published by Research Trend, Website: www.researchtrend.net)

ABSTRACT: As the vast amount of digital data is available and generated by most of the industries. To make use of such vast amount of data in critical decision making, Predictive analytics needs to perform on it. In the recent years Big Data Predictive Analytics (BDPA) is being a popularly used Technology to extract knowledge from huge data, addressing the many dimensions in all the industries. At this point of view, an attempt is made to understand the things happening around BDPA and its impact shown on businesses. This paper contributes in investigating the research carried out by observing current and past trends on BDPA from the last ten years and applying Machine Learning Algorithms in BDPA. Additionally, a standard reference model is developed. To provides a way to research in BDPA, finally list out the few challenges and issues of BDPA. The research carried out throughout the paper helps in providing the road map to the researchers in the area of BDPA.

Keywords: Big Data, Predictive Analytics, Machine Learning Algorithm, Decision Making.

I. INTRODUCTION

Big data is often cut down to a few varieties of data generated by machines, people, and organizations. Machine-generated data referred to the data generated from real-time sensors and industrial machinery or vehicles, personal health trackers, among many other sense data sources. Human-generated data referred to the vast amount of social media data, status updates, tweets, photos, and videos. Whereas Organization generated data, referred to more traditional types of data, including transaction information databases and structure data often stored in data warehouses. Big data can be structured, semi-structured, and unstructured. The real value from big data will come from integrating different types of data sources and analyzing them at scale. Modelling, Managing, and integrating diverse streams can improve business and add value to big data even before analyzing it. Modelling and Managing big data is focusing on the dimensions of scale availability and considering the challenges associated with these dimensions. Volume, variety, and velocity are the main dimensions of big data. Huge amounts of data in different formats and varying quality which must be processed quickly, veracity refers to the biases, noise, and abnormality in data, or the immeasurable certainty is in the truthfulness and trustworthiness of data, and valence refers to the connectedness of big data. The goal of data modelling is to explore the nature of data formally so that one can figure out what kind of storage is needed, and what kind of processing one can do on it. The goal of data management is to figure out what kind of infrastructure support does one need for the data. For example, does an environment need to keep multiple replicas of the data? Does it need to do statistical computation with the data? These operational requirements help anyone to choose the right system. The process involved behind producing all the products of different companies used in daily life into goods and services is called operations. Most of the organizations invest lots of revenue on human resource to perform such operations. Operation Management (OM) is critical for any organization, as it involves in effective decision making in the areas like strategic operation management, product design, supply chain management, Quality Management. As we all know that, in recent years, Big Data Predictive Analytics (BDPA) is being a popularly used Technology to extract knowledge from massive data, addressing the many dimensions in all the industries. At this point of view, an attempt is made to understand the things happening around BDPA and its impact shown on businesses.

The amount of data generated from different objects is increasing exponentially. All of this data can provide new insights into consumer understanding perceptively. Today the consumers can make use of an online platform like Face-book, Twitter, and micro blogs to give their feedback on a particular product. Which is in unstructured format? This data helps in getting a 360-degree view of a customer. The patterns and behaviour of the customer across social media are analyzed and make it ready for a specific need. BDPA can bring data into the hands of decision-makers very quickly. The real-

time data are enabling the marketers in interacting with the customers. This Technology can be easily adaptable to the industry faster. The Different dimensions of Big Data. The Fig.1 addresses different dimensions of data like, Volume: Large range and pure volume of data is itself a big challenge. Variety: Data captured in diverse forms like text, email, tweets, blogs, and business transactions. Veracity: data quality, Velocity: rate of growth in data.

Variability: constantly and rapidly changing, Visualize: representing knowledge more instinctively and effectively, interactively. Value: high-valued data, Volatility: To determine the point of the data, from which it is no longer relevant to the present analysis. It is necessary to address a different kind of technologies and the associated prototypes involving text data, from the text data [1]. Extracting features like scenario concepts from futuristic data in textual documents by applying the algorithm, text mining, and latent semantic analysis [2]. A knowledge-based solution for automatic schema mapping can manage the data heterogeneity [3] using Automatic ontology extraction and semantic inference for novel Big Data analytics (BDA).

Fig. 2 represents the amount of research carried out on dimensional of Big Data concerning the number of publications year rise from 1996 to 2016. Big Data attributes are classified into five-dimensional densities [4]. They are: Sending density, Content density, transferring density, purpose density, and receiving density with a different perspective. Based on data behaviour, big data models are classified into five different models, the dimensions are as follows: $ToD = \{tod_1, tod_2, tod_3 \ldots todn\}$ represents when the data occurred (Time of Data) timestamp for each data incident, $SoD = \{sod_1, sod_2, sod_3..sod_n\}$ represents where the data come from (Source of Data) Twitter, Facebook etc, $CoD = \{cod_1, cod_2, cod_3 \ldots cod_n\}$ represents what data contain (content of Data) Likes, comments from Facebook, Tweets from Twitter, $MoT = \{mot_1, mot_2, mot_3 \ldots mot_n\}$ represents how the data is transferred (Mode of Transfer) by Internet, phone, email, $EoD = \{eod_1, eod_2, eod_3 \ldots eod_n\}$ represents why data occurred (Event of Data) posting photo, posting comment, $RoD = \{rod_1, rod_2, rod_3 \ldots rod_n\}$ Represents who received data (Receiver of data) follower on Twitter, Face book friend. Thereby each data incident can be defined as a node:

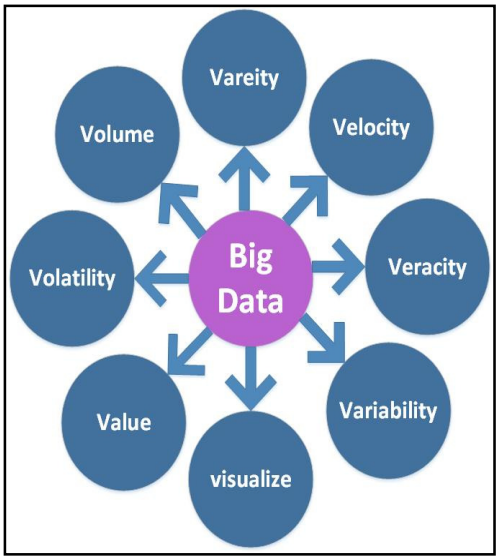

Fig. 1. Dimensions of Big Data.

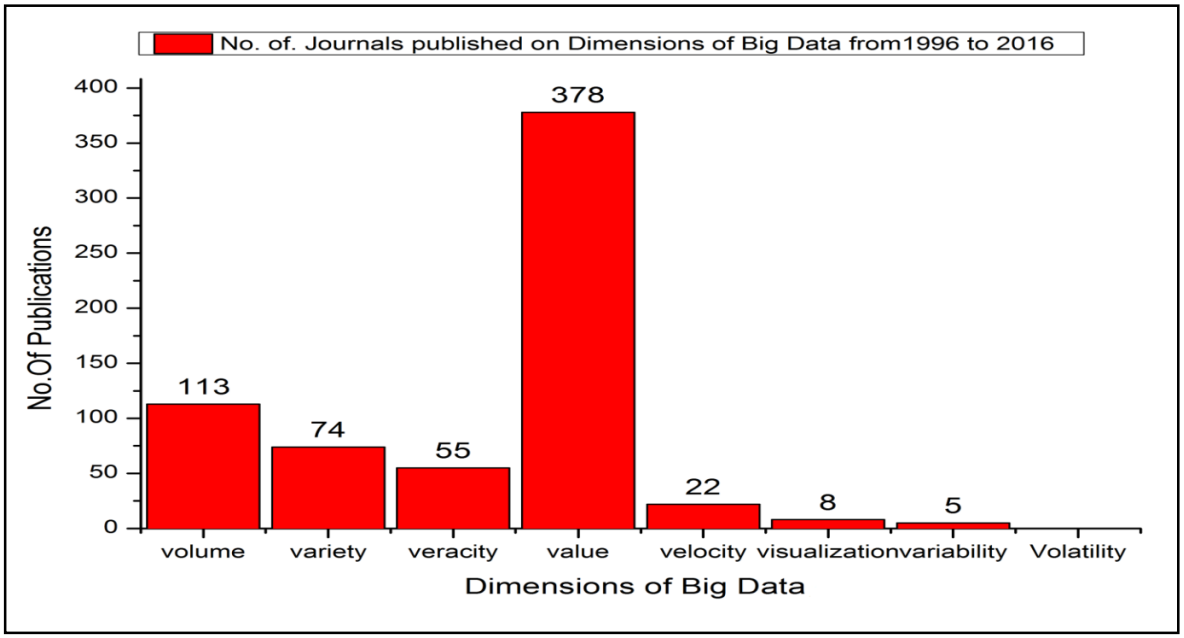

Fig. 2. Amount of research carried out on Big Data Dimensions.

$$f(tod, sod, mot, eod, rod) \quad (1)$$

Therefore, all data incidents in equation 1 in the T time slot are represented as a set F, which contains number n incidences, and so can be defined as:

$$F = (f_1, f_2, .f_n) \quad (2)$$

F contains all the incident nodes within a certain time period. For example, during the 2014 (ToD) FIFA World Cup Final (EoD) between Germany and Argentina, there were 280 million Facebook interactions including posts, comments (CoD) and likes across 88 million Face book (SoD) users. Twitter users also sent 618,725 messages (MoD) per minute at the moment of Germany's victory [5]. For a particular incident node where *sod =d, cod = a, mot = b, eod = g and rod = e,* the incident node can then be represented as *f(t, sod(d),cod(a),mot(b),eod(g), rod(e)).* A subset *F(t,d,a,b,g,e)* that contains all the particular incident nodes *f(t, sod(d),cod(a), mot(b), eod(g), rod(e))* is therefore using equation 1 and equation 2.

$$F_{(t,d,a,b,g,e)} = \left\{ f \in F \middle| \begin{matrix} f(t,sod,cod,mot,eod,rod) \\ ,t,sod=d,cod=a,mot=b, \\ eod=g,rod=e \end{matrix} \right\} \quad (3)$$

In the next stage of the data, to prove the big data showed positive impact [6]. The Big Data phenomenon has shown an impact on business intelligence [7]. In our study, discussion continues on the importance of Deep learning and the problems of it concerning Big Data Analytics. Big data analytics with deep learning and also addressed how the data acquisition, storage, management, and analysis are related with performing analytics on big data using Deep Learning [8]. Utilizing Deep Learning concept for addressing some crucial problems: like Memory and input-output overhead in Big Data Analytics. To choose the right platform for Big Data Analytics listed out and classified Big Data Analytics Platforms accordingly to their level of support for personalization [9].improvement in the decision-making capabilities in Big Data Analysis and to understand the concurrency workflow. The 3As Data Quality-in-Use model is to understand the nature of data in a distributed environment [10]. Similarly, a real-time and energy-efficient, resource scheduling, and optimization framework also needed [11]. The Excellent visualization of an unlimited number of concurrent jobs and its nature of workflows in a distributed environment address in [12]. The researchers and decision makers are using most productive manner a systematic review on "Big Data" [13].

This paper contributes to investigating the research carried out by observing current and past trends on BDPA from the last ten years in the area of modern operations management (OM). Additionally, a standard reference model is developed to provide a way to research BDPA in OM, finally listing the few challenges and issues alone with applications of BDPA in OM.

II. METHODOLOGY

In the way of propose a common framework for deriving knowledge from social networks using Big Data Analytics. A reference architectural framework for Big Data Analytics proposed that helps in understanding the role of machine learning in analytics [14]. Similarly, a framework for data mining based big data analytics [15]. For evaluating the Business data science model , use cases are necessary in big data solution reference architecture [16].

To understand the steps involved in Big Data Analytics, an extensive review on different steps of Big Data Analytics, like how the data can be stored, managed, processed, analyzed, visualized and verified the quality of data [17]. Here five ways to collect and drive your data to get more value out of it: Social Network Analysis, Click stream Data, Video Analytics, Machine Data, and Transactional Data. In this paper, we concentrate on Social Network Analysis using BDPA [18]. In Figure 3, a standard reference model is proposed based on the literature review made. The details of the proposed Methodology are discussed in the next sections.

A. Data Collection

In which, we need to collect data from different sources. Possible approaches for data collection which is accepted by the technology vendors in the industry [19]. All the Data centres are considered as Nodes and their respective communication paths as edges. Next step is to find out a common model to retrieve data [20] from different data sources, optimizing the communication path between different big data machines and sources. A novel semantic-based approach to support the data collection at different speed using Data retrieval model [21].

B. Business Requirements

There are a variety of reasons predictive analytics campaigns fail to meet their goals: G1: Lack of realistic and measurable goals. G2: Inadequate infrastructure. G3: inconsistent data preservation, usually caused by poor communication. G4: Failure to maintain strict user permissions. So, the possible solutions addressed to meet the goals in Big Data predictive Analytics are S1: Structuring Big Data Properly. S2: Communicating Data Strategy to your Team. The details steps involved in achieving our solutions are as follows:

C. Structuring Big Data Properly

The process of Structuring the Big Data properly involves three necessary steps. They are as follows:

Step1: one should rank data according to its importance [22]. Several keys attributed, that one should evaluate in selecting the best measure for the particular domain. Step2: Data needs to be preserved in a structure that allows for scalability [23]. Family of Hash-based scalable, distributed data structure. Step3: Data need to be easily extracted in real-time [24]. Data processing procedures for the analysis of quantitative real-time data are needed. **Communicating Data Strategy to Your Team**: The final process of setting up is to predictive analytics system is getting everyone on board, as the truth is that data is structured around human users. It should be presented in a format that everyone can understand and make the best use of it. A model that translating the analysis into decision making for intelligent business [25].

Data Preprocessing. In the pre-processing stage, the huge data needs to be cleaned, before loading it into the cluster of computers. Different steps for data processing, such as data capture, data storage, data analysis, and data visualization [26] is necessary in capturing the

insight from the regular data patterns. A change point analysis, correlation analysis, and Monte Carlo simulation can be used in removing the unwanted data in the pre-processing stage [27].

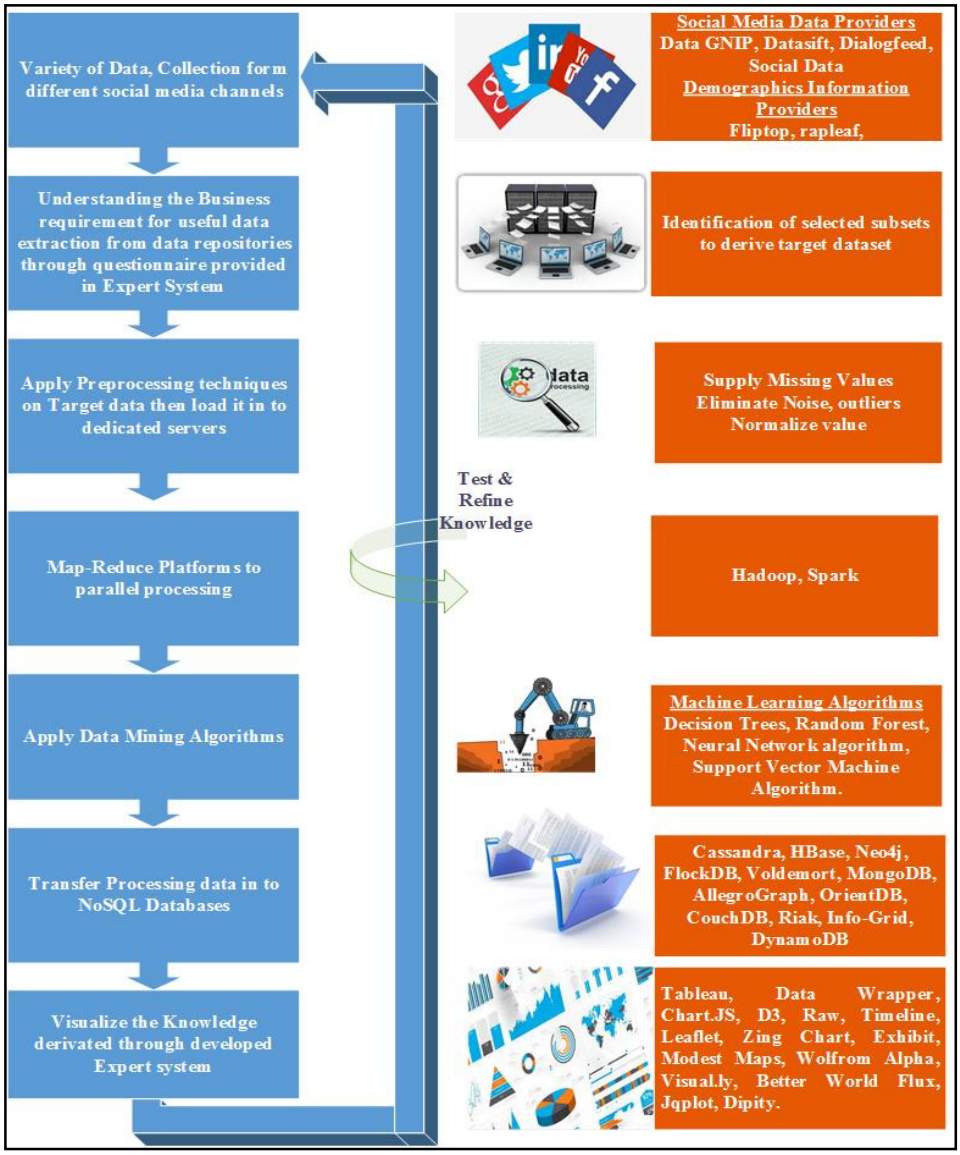

Fig. 3. A Methodology.

D. Big Data Analytics Platform

Data processing is a do-or-die requirement for businesses today. There are many Big Data software out there. Many of them promise to save you money, time, and help you find never-before-seen insights. Although that may be true, navigating the internet for the possible best software can be difficult when there are numerous choices to select from diverse sources. The following are some of the most popular Big Data software we found: (Note: * in Table 1 represents, rating of the platform concerning specific domain) Scalability: Is a measure of the system handling a growing amount of workload capably. Reliable: Is a measure of the user to show the degree of dependency on data. Fault Tolerance: A system can work efficiently even when one or more components get failed to function. User-Friendly: Degree of human willingness to work with the environment, Data Latency: Measure of Delay in terms of data processing. Analytics: Level of support for decision making on massive data. Visibility: Is a measure of the degree of support provided for Data Visualization. It is not possible to qualify the significance of particular software with a general scenario. It is only possible with a specific application. From Table 1, one can select the Big Data platform. Table. 2 helps the researchers for selecting Map Reduce platform based on attributes like an Analytical DBMS: which data structure is used to represent data, In-memory DBMS: Where the active data resides either in memory or I/O, Stream-analysis is supported or not, Hadoop distribution is implemented or not, which plays a prominent role in performing Predictive Analytics on Big Data. The selection of technologies for Big Data Analytics, [28] facilitating architecture design. To rank the Big Data

Platforms based on latency. A review on current researches on low latency at different system levels in Big Data Platforms [29]. A deep Comparison of the major machine learning frameworks listed in Table 2 [30]. For test the feasibility and efficiency of proposed system architecture, Implementation on Hadoop single node setup on UBUNTU 14.04 LTS core™i5 machine with 3.2 GHz processor and 4 GB memory. To understand the way the task is processed in Big Data platforms [31]. A distributed K-means algorithm based on Map-Reduce model, where the big vocabulary tree is divided into many sub-trees [32]. A hybrid algorithm for many-tasks computing used in Big Data processing platforms [33]. An algorithm Meta-Map-Reduce helps in reducing the time complexity in exchanging the intermediate vital values when the number of machines is increased compared to single machine [34].

Table 1: Survey on Big Data Platforms.

Name	Scalable	Reliable	Fault tolerance	User Friendly	Data Latency	Analytics	Visibility
Apache Hadoop	*****	****	**	**	*	*	*
Apache Spark	***	***	****	***	**	*	*
Apache Storm	****	***	****	****	**	*	**
Apache Flink	***	***	*****	***	****	**	**
Attivio	***	***	***	***	****	****	*****
Splunk	***	***	***	***	****	***	*****
Samza	***	***	***	***	****	***	*****
DataPlay	***	***	***	***	****	***	*****
Sap Hana	***	***	***	***	****	*****	*****
Bottlenose	***	***	***	***	****	***	*****

E. Machine Learning Algorithms

The next step in our study is to find different ways, how to apply Machine Learning Algorithms in Big Data Predictive Analytics. Supporting to that, we have reviewed research work [35]. In which, a survey on Machine learning techniques like Deep learning and Active learning. Need to have Comparisons with two distributed learning algorithms Apriori algorithm and Frequent Pattern Growth algorithm focused on improvements of associative classification of Big Data accuracy [36]. A series of random forest models in training the new instance of data using K Fold cross-validation method [37]. Artificial Neural Network model was accomplished with the back-propagation approach in predicting bead geometry using MAT LAB programs [38]. In a way to understand the implementation details of applying Machine Learning algorithms, we started with Support Vector Machine, which is used for classification. Next, list out the limitations and find the appropriate algorithms which better suits for real-time internet data.

Table 2: Survey on Platforms of Big Data Analytics.

Analytics Platform	Analytical DBMS	In-Memory DBMS	Stream-Analysis	Hadoop Distribution
Cloudera	Cloudera Impala	Apache Spark	Cloudera Standard, Cloudera Enterprise	Hadoop include Storm
Amazon	Amazon Relational Database Service	Altibase, SAP Hana, ScaleOut	Amazon Elastic MapReduce	Amazon Kinesis
Horton -works	SQL querying on top of Hadoop	Apache Spark	HDP for Windows, Hortonworks Sandbox	Hadoop include Storm
Pivotal	Greenplum	GemFire	Hadoop and Spring XD	Pivotal HD
SAP -Hana	SAP Hana, SAP IQ	SAP Hana	Event Stream Processing	Cloudera and MapR

Support Vector Machine: Given a training sample set of input-output pairs

$$(x_1, y_1),.....,(x_n, y_n), x_i \in R^m \quad (4)$$

The 2D point in the hyper-plane is described as equation 4 when Binary Classification has used the values of y coordinates may be as equation 5.

$$yi \in \{1,-1\}, i = 1,...n \quad (5)$$

Limitations of Support Vector Machine:

1. The final decision separating hyper-plane is located right in the middle of the parallel hyper-planes, which leaves the prior information within the classes.
2. Do not take the compelling advantage of prior data distributed information.

Overcome the limitation of Support Vector Machine, A novel algorithm called Structural Regularized Support Vector Machine [39].

F. Data Mining Techniques Currently Used In Social Network Analysis

To try, on the process of data analysis knowledge discovery in databases from the platform Data mining

[40]. Apriori algorithm is the well known best association rule mining procedure to mine involvement rules, which utilize a breadth-first search plan to count the support of item sets and take advantage of the downward closure assets of sustain [41]. To bring in a multiple walk plans to process the raw "Big Data" into dense time series that is superiorly matched for regression and causality scrutiny. Finally, attempts to apply the relationship rule and Link forecast related to triadic copyright filed during the era from 1955 to 2011 [42]. A tool that assists in terms mining from unstructured data storages [43]. To search the appropriate Data Mining Techniques currently used in Social Network Analysis [44].

Fig. 4 is listing out all the techniques used to extract knowledge in Social Area Network.

G. NoSQL Databases

In the selection of NoSQL Databases to be considered for Analytics, an investigative study on BASE features of some of NoSQL databases., HBase, Voldemort, Cassandra, and MongoDB [45]. Different NoSQL databases used by social networking sites Twitter, Facebook, Whatsup, and Orkut. The NoSQL database sublevels are used in social networking sites [46]. which are listed in Table [3].

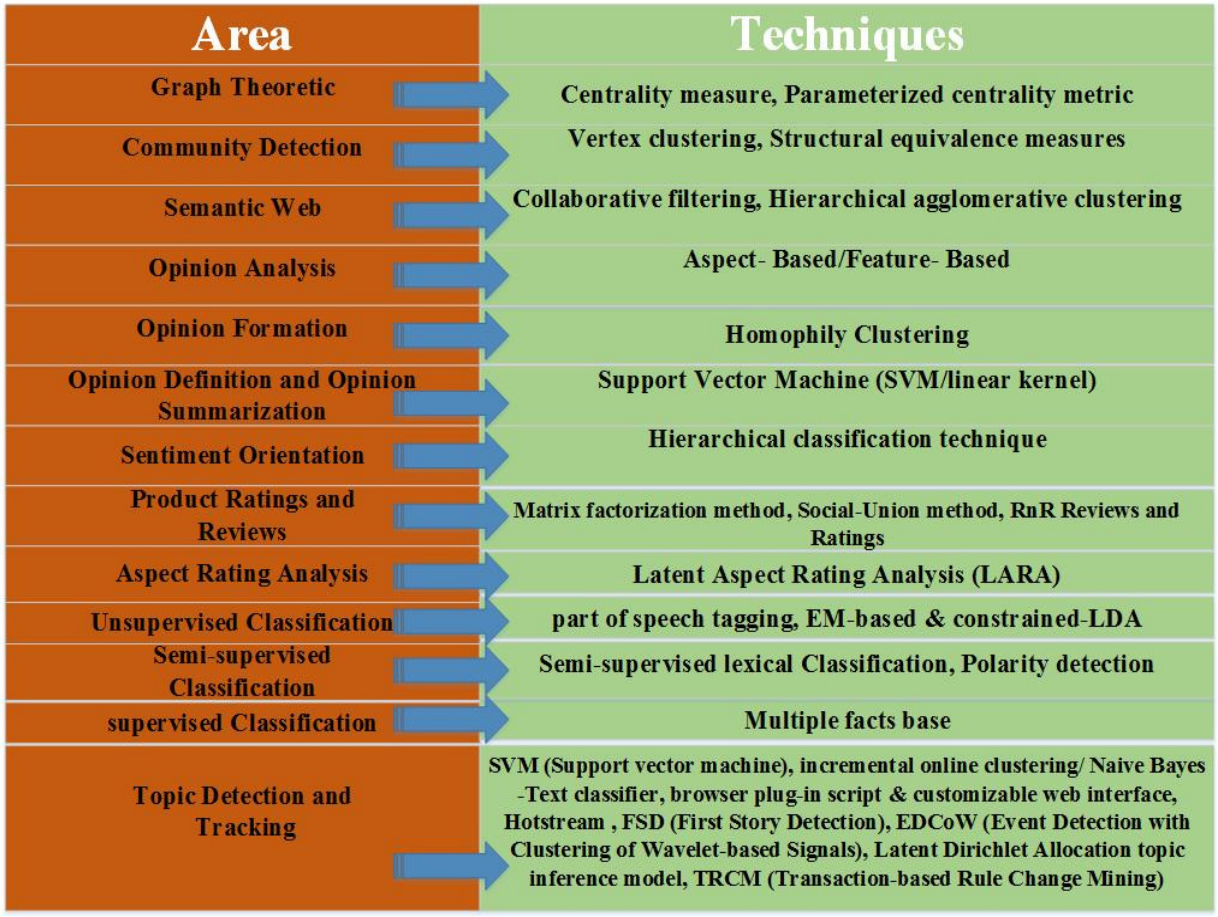

Fig. 4. Data Mining Techniques in Social Network Analysis.

Table 3: Survey on No-SQL Databases.

Social Networking Areas	NoSQL Databases
Myspace	MongoDB (MDB)
Twitter	Cassandra (CSD), HBase (HB)
LinkedIn	MDB, HB
Flickr	MDB
Friendfeed	HB, CSD
Foursquare	MDB, CSD
Facebook	HB, CSD

H. Knowledge Extraction

In the Knowledge Discovery process, we focus more on extracting valuable data that lead to better analytics. It is essential to have a higher degree of relationship between the selected independent attributes. As the start of our research work, we focus on Static forecasting Methods.

$$D_S = (L+T)^* SF \tag{6}$$

L=Estimate of Level at *t=0*, *T*=Estimate of Trend, *SF*= Seasonal Fluctuations, D_s = Actual demand observed in period *t*, F_t = forecast of demand for period *t*. In a static forecasting method, the forecast in period *t* for demand in period *t +1* is given by

$$F_{t+1} = [L+(t+1)T]_{t+1} \tag{7}$$

Considering, Periodicity (P) is the number of periods after which the seasonal cycle. We derive the following equations.

$$\overline{D}t = \frac{\left[D_{t-(p/2)} + D_{t+(p/2)} + \sum_{i=t+1-(p/2)}^{t-1+p/2} 2D_i\right]}{2p} \text{ for } p \text{ even} \tag{8}$$

Equation 8 represents Demand when p is odd.

$$\overline{D}t = \frac{\left[D_{t-(p/2)} + D_{t+(p/2)} + \sum_{i=t-[(p-1)/2]}^{t-1+p/2} 2D_i\right]}{p} \text{ for } p \text{ odd} \tag{9}$$

Equation 9 represents Demand when p is even. With the help of equation 8 and equation 9, we can forecast the demand for a particular product by considering time. Next is to focus on how to perform analytics on online consumer reviews. Sentiment mining approach for big data analytics for Sorting and Classification of online consumer reviews data [47].

I. Data Visualization

The final step is to find out the right tool to Visualize predictive data obtained from the previous step. The ability of augmented reality and virtual reality could be applied to the field of BDV. As part of our study work, we try to plot the rules formed using Apriori algorithm on data-set having 13 attributes and 45,000 observations on R Platform using packages (rules and rules viz) [48].

To interpret Fig. 5 we give: Support(s) is an indication of how frequently the item-set appears in the database, Confidence is an indication of how often the rule is true, confidences as

$$confidence(x \Rightarrow y) = \frac{s(x \cup y)}{s(x)} \tag{10}$$

and lift as

$$Lift(x \Rightarrow y) = \frac{s(x \cup y)}{s(x) \times s(y)} \tag{11}$$

where X and Y are attributes of Data-set.

To choose the right tool for data Visualization based on their objective with its specific advantages. Although equipped with six kinds of the chart, open-source library Chart.js is the ideal BDV tool. They are using HTML 5 canvas elements to make charts. Raw lets users make vector-based BDV. Dy graphs extremely adjustable works in major browsers, and you can even touch to zoom on mobile devices. Zing Chart offers over 100 chart types to fit our data. Wolfram Alpha is excellent at astutely put on view charts in response to data queries without the requirement for any pattern.

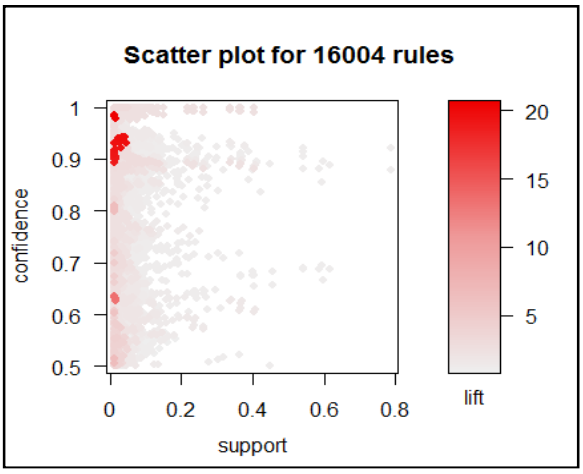

Fig. 5. Visualization with R.

Visually proposes a simple test for building dramatic data representations. Fusion Charts approach with a full JavaScript API that makes it simple to put together it with any AJAX application or JavaScript structure. These visualizations are incredibly interactive, customizable, and work across all devices and platforms. jqPlot is an excellent solution for line and point charts. R A statistical package used to parse large data sets. Weka is a tool used for classifying and clustering data based on features that help to explore data.

J. Challenges

Challenges and Issues related to Big Data Predictive Analytics make all the researchers be aware and grow along with the problems. This is associated with obtaining data from diverse sources and storing. A conservatory to the data warehouse model, knowledge warehouse (KW) designed that will facilitate in the coding of knowledge, also improve the recovery and allocation of knowledge across the businesses [49]. Data Mining and cleaning this is associated with taking out and clean data from a collected group of the extensive range of unstructured data. As Big-Data has often been strident, unreliable, varied, dynamic; in this context, these kindnesses do not apply to non-relational, databases without schema. The introduction to using time data, analysis, and modelling is the key elements considered in the development of transportation preparation methodologies [50]. This makes data understandable for the end-users. The data analysis and modelling results are obtainable to the decision-makers to understand the findings for extracting sense and information. BD masquerade big privacy worry and how to protect privacy in the digital era is a great challenge. Significant investments are made in BD projects to provide privacy at the agreement level. It was not appropriately addressed, then the occurrence of BD will not receive many receipts globally. Securing BD has distinctive challenges that are not deeply different from the current data. As BD is continually growing, organizations have their plans to perceive data governance as a possible advance to collateral the data quality, humanizing, and vital information, maintaining its value as a means of the organizational asset.

Data and Information Sharing needs to be unbiased and controlled to capitalize on its effect, as this will facilitate organizations in establishing secure connections and harmonization with their business partners. The continually increasing data in all dissimilar forms have led to an increasing insist for BD processing in complicated data centres. Need of evaluating the contact of protection plan and the networks component of BD and its availability on the operational cost [51].

K. Applications of Big Data

Different application areas, where BDPA can be applied successfully are, Businesses, Internet of Things, Health care, Science and Engineering, other fields in the future.

Business: The origins of big data applications and its current trends with Quantitative study of trends in publications related to analytical, to help firms build the strategy to combine knowledge from both marketing and big data domains [52]. A new approach, which is used text mining to afford further detail on key sector product cells [53]. Identified and visualized trends and patterns that can have insightful results on complete business performance [54].

Health-care: The potential of big data in emergency medicine research is identified and detailed survey [55] helps the researcher in identifying the recent trends in the field of Health care. Similarly, historical health data, is used in assesses the medical Quality of service using a new healthcare-specific analytical [56]. Considered health care as an application area and made a review of Big data and one health [57].

Agriculture: The threat to food security for each people and essential needs if increasing food productivity by conventional agricultural methods [58]. The road map to get rid of poverty [59] helps to get ready for the food poverty period. The food production needs to be doubled in the next decade to solve the poverty problem [60]. Conventional agriculture methods in providing food security for each people [61].

Operational Management: A Data Mining methodology in identifying risk factors of incident narratives. In which, Latent semantic analysis (LSA) was used in transforming narrative to contemporary weights based on patterns of co-occurrences in the narratives by using A data streaming model in training the classifiers [62]. In the experiment performed, the classifier generates 88% of accuracy in identifying the risk factors in order to reduce the impact of data quality on economic and social factor. A Framework can improve the quality of data, for the consumer perspective by capturing the aspects of data quality [63].

In the research of operational Management (OM), making use of Google Scholar searches using primary keywords such as "big data, data-driven, data analytics," supplemented by secondary keywords like "operations," "operational," "management," "marketing," "optimization." In making the human activities automation in the field of OM, Why not Data science can be combined with all the activities of OM [64], focused on five factors, i.e., (Data Collection, Data Storage, Data Processing, Data Analysis, and Reporting) in making the OM activities automation. Similarly, the services of the Big Data can also be transformed into OM. model to enhance the experience of customer-interaction in providing services in data collection from different sources [65]. Attempt a discussion on Big data innovation cycle [66].

All the key operation challenge is being facing in network-based professional service organizations, towards building effective network capabilities in a global context. Identified three case analyses include [67]:

a) Deploying dispersed resources
b) Integrating network activities
c) Knowledge management.

User comments on online social media could help the firms in enhancing various applications such as stock investment, acquisition, customer relationship management (CRM) and proposed a parallel aspect-oriented sentiment analysis algorithm capable in mining consumer sentiments from social media [68].

Others. Developing Construction waste management by mining the 2,212,026 ravage disposal records produced from 5764 projects in two successive years of 2011 and 2012 [69]. Similarly, a study to develop a BD based platform to classify, collect, and store data about workers unsafe behaviour that is derived from a metro construction project [70].

CONCLUSION AND FUTURE SCOPE

The contributions made in the paper are: recognized the study carried out in the field of Big Data Predictive Analytics (BDPA), and the researcher in future should be aware of these and interpret the information provided along with the context of limitation. This research work can provide the right direction to the researcher, towards understanding the challenges in handling the BigData. There are other challenges left out that needs to be addressed in the next publication. As a Future scope, a connection needs to be established between the theoretical concept and different associated areas of research related to BDPA.

Conflict of Interest: Nil

REFERENCES

[1]. Chen, J., Tao, Y., Wang, H., & Chen, T. (2015). Big data-based fraud risk management at Alibaba. *The Journal of Finance and Data Science*, **1**(1): 1-10.

[2]. Kim, J., Han, M., Lee, Y., & Park, Y. (2016). Futuristic data-driven scenario building: Incorporating text mining and fuzzy association rule mining into fuzzy cognitive map. *Expert Systems with Applications*, **57**(1): 311-323.

[3]. Esposito, C., Ficco, M., Palmieri, F., & Castiglione, A. (2015). A knowledge-based platform for Big Data analytics based on publish/subscribe services and stream processing. *Knowledge-Based Systems*, **79**(1): 3-17.

[4]. Guo, K., Zhang, R., & Kuang, L. (2016). TMR: towards an efficient semantic-based heterogeneous transportation media Big Data retrieval. *Neurocomputing*, **181**(1): 122-131.

[5]. Lorenzetti, V., Solowij, N., Fornito, A., Ian Lubman, D., & Yucel, M. (2014). The association between regular cannabis exposure and alterations of human brain morphology: an updated review of the literature. *Current pharmaceutical design*, **20**(13): 2138-2167.

[6]. Aye, K.N., & Thein, T. (2015). A platform for big data analytics on distributed scale-out storage system. *IJBDI*, **2**(2): 127-141.

[7]. Larson, D., & Chang, V. (2016). A review and future direction of agile, business intelligence, analytics and data science. *International Journal of Information Management*, **36**(5): 700-710.

[8]. Yin, L., Cheng, Q., Wang, Z., & Shao, Z. (2015). 'Big data'for pedestrian volume: Exploring the use of Google Street View images for pedestrian counts. *Applied Geography*, **63**(1): 337-345.

[9]. Lopes, C., Cabral, B., & Bernardino, J. (2016, July). Personalization using Big Data Analytics Platforms. In *Proceedings of the Ninth International C* Conference on Computer Science & Software Engineering* (pp. 131-132). ACM.
[10]. Merino, J., Caballero, I., Rivas, B., Serrano, M., & Piattini, M. (2016). A data quality in use model for big data. *Future Generation Computer Systems*, **63**(1): 123-130.
[11]. Sun, D., Zhang, G., Yang, S., Zheng, W., Khan, S. U., & Li, K. (2015). Re-Stream: Real-time and energy-efficient resource scheduling in big data stream computing environments. *Information Sciences*, **319**(1): 92-112.
[12]. Kempa-Liehr, A. (2015). Performance analysis of concurrent workflows. *Journal of Big Data*, **2**(1): 10.
[13]. Frizzo-Barker, J., Chow-White, P. A., Mozafari, M., & Ha, D. (2016). An empirical study of the rise of big data in business scholarship. *International Journal of Information Management*, **36**(3): 403-413.
[14]. Colombo, P., & Ferrari, E. (2015). Privacy aware access control for big data: A research roadmap. *Big Data Research*, **2**(4): 145-154.
[15]. Bhatnagar, V. (2013). Data mining-based big data analytics: parameters and layered framework. *International Journal of Computational Systems Engineering*, **1**(4): 265-276.
[16]. Geerdink, B. (2015). A reference architecture for big data solutions-introducing a model to perform predictive analytics using big data technology. *International Journal of Big Data Intelligence*, **2**(4): 236-249.
[17]. Hou, S., Huang, X., Liu, J.K., Li, J., & Xu, L. (2015). Universal designated verifier transitive signatures for graph-based big data. *Information Sciences*, **318**(1), 144-156.
[18]. Bello-Orgaz, G., Jung, J.J., & Camacho, D. (2016). Social big data: Recent achievements and new challenges. *Information Fusion*, **28**(1): 45-59.
[19]. Perrons, R.K., & McAuley, D. (2015). The case for "n «all": Why the Big Data revolution will probably happen differently in the mining sector. *Resources Policy*, **46**(1): 234-238.
[20]. Esposito, C., Ficco, M., Palmieri, F., & Castiglione, A. (2015). A knowledge-based platform for Big Data analytics based on publish/subscribe services and stream processing. *Knowledge-Based Systems*, **79**(1): 3-17.
[21]. Guo, K., Zhang, R., & Kuang, L. (2016). TMR: towards an efficient semantic-based heterogeneous transportation media Big Data retrieval. *Neurocomputing*, **181**(1): 122-131.
[22]. Tan, P. N., Kumar, V., & Srivastava, J. (2002, July). Selecting the right interestingness measure for association patterns. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 32-41). ACM.
[23]. Litwin, W., Neimat, M. A., & Schneider, D. (1994, September). Rp*: A family of order preserving scalable distributed data structures. In *VLDB*, Vol. **94**, pp. 12-15.
[24]. Schmittgen, T.D., & Livak, K.J. (2008). Analyzing real-time PCR data by the comparative C T method. *Nature protocols*, **3**(6): 1101.
[25]. Tanev, S., Liotta, G., & Kleismantas, A. (2015). A business intelligence approach using web search tools and online data reduction techniques to examine the value of product-enabled services. *Expert Systems with Applications*, **42**(21): 7582-7600.
[26]. Yuan, J.S., Reed, A., Chen, F., & Stewart, C.N. (2006). Statistical analysis of real-time PCR data. *BMC bioinformatics*, **7**(1): 85.
[27]. Jeon, S., & Hong, B. (2016). Monte Carlo simulation-based traffic speed forecasting using historical big data. *Future generation computer systems*, **65***(1)*, 182-195.
[28]. Pääkkönen, P., & Pakkala, D. (2015). Reference architecture and classification of technologies, products and services for big data systems. *Big data research*, **2**(4): 166-186.
[29]. Tian, X., Han, R., Wang, L., Lu, G., & Zhan, J. (2015). Latency critical big data computing in finance. *The Journal of Finance and Data Science*, **1**(1): 33-41.
[30]. Landset, S., Khoshgoftaar, T.M., Richter, A.N., & Hasanin, T. (2015). A survey of open source tools for machine learning with big data in the Hadoop ecosystem. *Journal of Big Data*, **2**(1): 24.
[31]. Ahmad, A., Paul, A., & Rathore, M. M. (2016). An efficient divide-and-conquer approach for big data analytics in machine-to-machine communication. *Neurocomputing*, **174**(1): 439-453.
[32]. Duan, H., Peng, Y., Min, G., Xiang, X., Zhan, W., & Zou, H. (2015). Distributed in-memory vocabulary tree for real-time retrieval of big data images. *Ad Hoc Networks*, **35**(1): 137-148.
[33]. Bittencourt, L. F., & Madeira, E. R. M. (2011). HCOC: a cost optimization algorithm for workflow scheduling in hybrid clouds. *Journal of Internet Services and Applications*, **2**(3): 207-227.
[34]. Yang, J., & Yecies, B. (2016). Mining Chinese social media UGC: a big-data framework for analyzing Douban movie reviews. *Journal of Big Data*, **3**(1): 3.
[35]. Qiu, J., Wu, Q., Ding, G., Xu, Y., & Feng, S. (2016). A survey of machine learning for big data processing. *EURASIP Journal on Advances in Signal Processing*, **2016**(1): 67.
[36]. Bechini, A., Marcelloni, F., & Segatori, A. (2016). A MapReduce solution for associative classification of big data. *Information Sciences*, **332**, 33-55.
[137]. Guo, S.Y., Ding, L.Y., Luo, H.B., & Jiang, X.Y. (2016). A Big-Data-based platform of workers' behavior: Observations from the field. *Accident Analysis & Prevention*, **93**: 299-309.
[38]. Sarkar, A., Dey, P., Rai, R.N., & Saha, S. C. (2016). A comparative study of multiple regression analysis and back propagation neural network approaches on plain carbon steel in submerged-arc welding. *Sādhanā*, **41**(5): 549-559.
[39]. Chen, D.Q., Preston, D.S., & Swink, M. (2015). How the use of big data analytics affects value creation in supply chain management. *Journal of Management Information Systems*, **32**(4): 4-39.
[40]. Andrejevic, M. (2014). Big data, big questions| the big data divide. *International Journal of Communication*, **8**(1): 17.
[41]. Geetharamani, R., Revathy, P., & Jacob, S.G. (2015). Prediction of users webpage access behaviour using association rule mining. *Sadhana*, **40**(8): 2353-2365.
[42]. Lee, H.L. (2018). Big data and the innovation cycle. *Production and Operations Management*, **27**(9): 1642-1646.
[43]. Lomotey, R.K., & Deters, R. (2013). RSenter: terms mining tool from unstructured data sources. *International Journal of Business Process Integration and Management*, **6**(4): 298-311.

[44]. Stahl, F., Gaber, M.M., & Adedoyin-Olowe, M. (2014). A survey of data mining techniques for social media analysis. *Journal of Data Mining & Digital Humanities*, *2014*.
[45]. Chandra, D.G. (2015). BASE analysis of NoSQL database. *Future Generation Computer Systems*, **52**: 13-21.
[46]. Mathew, A.B., & Kumar, S.M. (2015). Novel research framework on SN's NoSQL databases for efficient query processing. *International Journal of Reasoning-based Intelligent Systems*, **7**(3-4): 330-338.
[47]. Salehan, M., & Kim, D.J. (2016). Predicting the performance of online consumer reviews: A sentiment mining approach to big data analytics. *Decision Support Systems*, **81**: 30-40.
[48]. Olshannikova, E., Ometov, A., Koucheryavy, Y., & Olsson, T. (2015). Visualizing Big Data with augmented and virtual reality: challenges and research agenda. *Journal of Big Data*, **2**(1): 22.
[49]. Nemati, H.R., Steiger, D.M., Iyer, L.S., & Herschel, R.T. (2009). Knowledge Warehouse: An Architectural Integration of Knowledge Management, Decision Support, Data Mining and Data Warehousing. *University of North Carolina at Greensboro*.
[50]. Kitamura, R., Fujii, S., & Pas, E.I. (1997). Time-use data, analysis and modeling: toward the next generation of transportation planning methodologies. *Transport Policy*, **4**(4): 225-235.
[51]. Machuca, C. M., Moe, O., & Jäger, M. (2008). Impact of protection schemes and network component's availability on operational expenditures. *Journal of Optical Networking*, **7**(2): 142-150.
[52]. Xu, Z., Frankwick, G.L., & Ramirez, E. (2016). Effects of big data analytics and traditional marketing analytics on new product success: A knowledge fusion perspective. *Journal of Business Research*, **69**(5): 1562-1566.
[53]. Nathan, M., & Rosso, A. (2015). Mapping digital businesses with big data: Some early findings from the UK. *Research Policy*, **44**(9): 1714-1733.
[54]. Palanimalai, S., & Paramasivam, I. (2016). Big data analytics bring new insights and higher business value-an experiment carried out to divulge sales forecasting solutions. *International Journal of Advanced Intelligence Paradigms*, **8**(2): 207-218.
[55]. Taylor, R. A., Pare, J.R., Venkatesh, A.K., Mowafi, H., Melnick, E.R., Fleischman, W., & Hall, M.K. (2016). Prediction of in-hospital mortality in emergency department patients with sepsis: a local big data–driven, machine learning approach. *Academic emergency medicine*, **23**(3): 269-278.
[56]. Batarseh, F.A., & Latif, E.A. (2016). Assessing the quality of service using big data analytics: with application to healthcare. *Big Data Research*, **4**: 13-24.
[57]. Asokan, G.V., & Asokan, V. (2015). Leveraging "big data" to enhance the effectiveness of "one health" in an era of health informatics. *Journal of epidemiology and global health*, **5**(4): 311-314.
[58]. Biggs, E. M., Bruce, E., Boruff, B., Duncan, J. M., Horsley, J., Pauli, N., & Haworth, B. (2015). Sustainable development and the water–energy–food nexus: A perspective on livelihoods. *Environmental Science & Policy*, **54**: 389-397.
[59]. Prasad, R., Bhattacharyya, A., & Nguyen, Q. D. (2017). Nanotechnology in sustainable agriculture: recent developments, challenges, and perspectives. *Frontiers in microbiology*, **8**, 1014.
[60]. Waldron, A., Garrity, D., Malhi, Y., Girardin, C., Miller, D. C., & Seddon, N. (2017). Agroforestry can enhance food security while meeting other sustainable development goals. *Tropical Conservation Science*, **10,** 1940082917720667.
[61]. Altieri, M., Nicholls, C., & Montalba, R. (2017). Technological approaches to sustainable agriculture at a crossroads: an agroecological perspective. *Sustainability*, **9**(3): 349.
[62]. Shi, D., Guan, J., Zurada, J., & Manikas, A. (2017). A data-mining approach to identification of risk factors in safety management systems. *Journal of management information systems*, **34**(4): 1054-1081.
[63]. Wang, R.Y., & Strong, D.M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of management information systems*, **12**(4): 5-33.
[64]. George, G., Osinga, E.C., Lavie, D., & Scott, B.A. (2016). Big data and data science methods for management research. *Academy of Management Journal*, **59**(5): 1493-1507.
[65]. Demirkan, H., & Delen, D. (2013). Leveraging the capabilities of service-oriented decision support systems: Putting analytics and big data in cloud. *Decision Support Systems*, **55**(1): 412-421.
[66]. Lee, W.S., Han, E.J., & Sohn, S.Y. (2015). Predicting the pattern of technology convergence using big-data technology on large-scale triadic patents. *Technological Forecasting and Social Change*, **100**: 317-329.
[67]. Peng, J., Quan, J., Zhang, G., & Dubinsky, A.J. (2016). Mediation effect of business process and supply chain management capabilities on the impact of IT on firm performance: Evidence from Chinese firms. *International journal of information management*, **36**(1): 89-96.
[68]. Lau, R.Y.K., Zhang, W., & Xu, W. (2018). Parallel aspect-oriented sentiment analysis for sales forecasting with big data. *Production and Operations Management*, **27**(10): 1775-1794.
[69]. Wang, J., Wu, H., Tam, V.W., & Zuo, J. (2019). Considering life-cycle environmental impacts and society's willingness for optimizing construction and demolition waste management fee: An empirical study of China. *Journal of cleaner production*, **206**: 1004-1014.
[70]. Tong, R., Zhang, Y., Cui, P., Zhai, C., Shi, M., & Xu, S. (2018). Characteristic analysis of unsafe behavior by coal miners: Multi-dimensional description of the pan-Scene data. *International journal of environmental research and public health*, **15**(8): 1608.

How to cite this article: Basha, S.M., Rajput, D.S., Bhushan, S.B., Poluru, R.K., Patan, R., Manikandan, R. and Kumar, A. (2019). Recent Trends in Sustainable Big Data Predictive Analytics: Past Contributions and Future Roadmap. *International Journal on Emerging Technologies*, **10**(2): 50-59.