International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-3, January 2020

A Prediction of Different Technologies for the Development of Unstructured Big Data

Parashiva Murthy B M, Sumithra Devi K A

Abstract*: Nowadays, a huge volume of terabytes of data is generated from digital technologies and modern information systems, namely Internet of Things and cloud computing. The extraction of knowledge for making decisions from the analysis of these massive data, leads to requires a huge effort at multiple levels. Hence, the researchers focused on Big Data Analysis (BDA) for better development. Traditional platforms and data techniques are very less efficient in Big Data (BD) context, which shows the lack of accuracy, performance, scalability and slow responsiveness. Several works are carried out to address the complex BD challenges by developing new technologies and different types of distributions. In this research work, various technologies which are developed for BD are described and the impact of open research issues, challenges and tools for processing the BD are discussed. Then, the impacts on key business performances for BD are evaluated. At last, this work presented the four major technical and managerial challenges, which provides a platform for exploring BD at numerous stages.

Index Terms*: Big Data Analysis, Cloud Computing, Decision Making, Internet of Things, Massive Data, Scalability.

I. INTRODUCTION

In general, heterogeneous sources generated a vast amount of data daily, where the sources include marketing, social networks, government, financial and health due to several technology trends such as Internet of Things, cloud computing as well as development of smart devices [1,2]. To improve the competitions and productions of organization, the BDA provides several offers to those organizations [3]. Numerous variety of data are used as input namely, video, blog, social media data and customer-generated data to improve the performance enhancement of BDA [4]. The digitalized data transactions are used nowadays by Amazon, Google, eBay, Facebook for improving their business operations. [5]. The customer behavior, patterns, trends and market conditions are identified on a regular basis by storing the product prices, transaction time, customer credentials and purchase quantities and this storing processing is carried out by these famous firms [6].

Revised Manuscript Received on January 30, 2020.
***** Correspondence Author
Parashiva Murthy B M*, Department of CSE, JSS Science and Technology University, Mysuru, India.
Sumithra Devi K A, Department of CSE, Dayanand Sagar Academy of Technology and Management, Bengaluru, India.

However, the traditional technologies face several challenges such as lack of flexibility, limited storage capacity, shortage of scalability, expensive, rigid management tools and performance in BD context. BD management requires powerful technologies, new methods and significant resources [7]. In addition, a granular access is provided to evaluate the massive data sets by cleaning, processing and analyzing the BD. The new insights are discovered to personalize the services by industries and companies, which will increase the important factors for data analysis [8]. The main goal of BDA is to process the huge data and different records that includes different types of content by using parallel and advanced analytic techniques. The BDA tools provide valuable benefits to the organization to deal with the three types of data, such as structured, semi-structured and unstructured data [9]. However, the conventional database techniques unable to process these challenging data and its only focus on the structured data [10]. To handle the semi-structured and unstructured data, there is no proper methods developed in BD context [11]. Even though BD provides various opportunities for decision makers and industrial areas, it faces the challenges like security and privacy [12]. The analytics tool collects data from various available sources and obtains the security issues during analysis, managing, visualizing, storing and sharing the data to another group of people. Due to the combination and exploration of these specific behavioral data, the Internet users are widely vulnerable to expose their privacy data [13,14]. However, it is to be noted that all data available in the form of BD are not useful for analysis or decision making process. In this research work, an overview of BD, impact of tools and the BDA is discussed. This paper focuses on the challenges in BD and its available techniques. Additionally, open research issues in BD are also stated in this work.

The organization of this paper includes Section 2 provides the taxonomy of BD, the impacts of analysis and tools of BD are presented in Section 3 and 4. The impacts of the challenges of the BDA are discussed in Section 5. The presentation of several recent techniques used in the process of BD is described in Section 6. Finally, the conclusion of this research work is given in Section 7.

II. TAXONOMY OF BIG DATA

The conventional database techniques are unable to handle or process the large volume of heterogeneous data, where these large volumes of data are described by new concepts called BD.

There are various kinds of digital contents present in BD such as structured, semi-structured and unstructured data. These concepts are described below.

A. Structured Data

The structured data are easy to store, visualize, enter, query and process. These data are managed in spreadsheets or relational databases with specific types and sizes and in general, it is also defined as pre-defined fields.

The extraction of useful information is easy due to its rigid structure, since parallel techniques are not required for processing.

B. Semi-Structured Data

This kind of data will not follow only rigid model, which defines the hierarchical description for various fields within the data. Moreover, certain elements are identified by defining the semi-structured data, which consists of different meta-model namely tags and markers. The samples of self-description data are JavaScript Object Notation (JSON) and Extensible Markup Language (XML).

C. Unstructured Data

This kind of data is stored and represented without any pre-defined format. It consists of free form texts such as documents, emails, articles, books and media files. These kinds of data are difficult to define in a rigid form and also process the data, which leads to develop the new processing mechanism NOSQL.

III. ANALYSIS OF BIG DATA

The various computational and traditional intelligence techniques use the data of high velocity, volume, veracity, variety and value, which is the major objective of BDA. Figure 1 shows the BD representation. The analysis of good BD is discussed as follows:

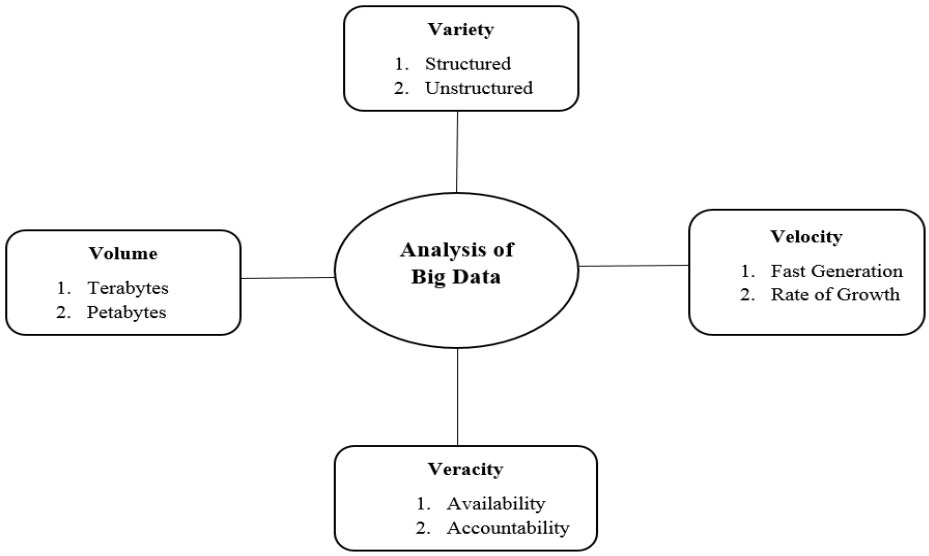

Figure 1: Big Data's Characteristics

A. Volumes

The data are collected and stored in various distributed data storage, which contains a vast amount of content. In order to extract the valuable knowledge, these large data should be available for processing and also scaled to Exabyte. It is more significant for processing, when the volume of data is high, however, it should be respected with four following rules such as value, velocity, variety and veracity.

B. Velocity

The speed of the data is calculated by which data are generated and how fast it is changed, are defined as the velocity. The real-time streams and non-storage data are also used by BD, where it doesn't relate to static records. In case of critical applications, the BD processing should able to handle the generation of large volumes of data and extract the results in few milliseconds or few seconds.

C. Variety

In BD, one of the most significant criteria is content variety, where the data can be external or internal as well as it could be structured or unstructured. The internal resources in the organization are used to collect the internal data, whereas the open sources include web, open BD are used to gather the external data. Therefore, varied information about specific domain are extracted by the processors. But, in order to handle these heterogeneous records, new specific architectures and particular processing techniques should be developed, so the consideration of veracity should be important.

D. Value

The most important features of BD are variety, veracity and volume, however the important thing is to identify the ability of using the extract values in a reasonable time i.e. velocity.

E. Veracity

The validity and accuracy of the collected data are the most important characteristics of the extraction process, but

the large volume of data provides invalid accuracy or not valid, which leads to a false interpretation. Therefore, while guarantee the integrity and the value of the inputs, the vast amount of collected data and its heterogeneity are checked for removing any noises and missing values.

BDA are developed to identify the tools and mechanisms for processing and analyzing the data effectively. These frameworks allow the performance of huge computation on less infrastructure, which are deployed over multiple parallel clusters and nodes. The important tools used in BDA are explained in the next section.

IV. TOOLS FOR PROCESSING THE BIG DATA

Processing for BD done by a number of tools and some of the important tools for evaluating the BD are discussed in this section, where the most important tools include Apache Spark, MapReduce and Storm. The first and most significant tool for processing the BD is known as Apache Hadoop [15]. There are three types of processing for BD is available, those are stream processing, interactive analysis and batch processing. These are mainly concentrated by various important available tools. The sample tools for batch processing includes Dryad and Mahout, stream processing contain Splunk and Strom, whereas interactive analysis preferred Apache Drill and Dremel tools. Among these three processing, stream analysis is used for real time analysis. In real time, the users can able to directly interact for their own analysis by using an interactive analysis process. The BD projects are developed by using these tools, where these lists of tools and techniques are described in [16]. The most important tools are discussed below.

A. Apache Hadoop and MapReduce

MapReduce and Apache Hadoop are considered as one of the most established software platform for analyzing the BD. There are several components presents in this tool such as Apache hive, MapReduce, Hadoop kernel and Hadoop Distributed File System (HDFS), etc. According to divide and conquer method, MapReduce process large dataset, which is also defined as a programming model. To solve the BD problems, the most powerful software tools are used, which includes MapReduce and Hadoop. In addition, it obtains high throughput and helps in fault-tolerant storage, while processing the data.

B. Apache Mahout

The main aim of the Apache mahout is to present the most scalable and commercial machine learning techniques for analyzing the intelligent and large scale data. Through MapReduce, Hadoop function is run by using batch based collaborative filtering. The mahout contains several core algorithms that include evolutionary algorithms, regression, pattern mining, classification, clustering and dimensionality reduction.

C. Apache Drill

The BD are interactively analyzed by using Apache Drill, which can able to support various different types of data sources, formats and query languages. The batch analysis is performed by MapReduce, which is used by drill and also for storing the data, the drill uses the HDFS. In specific, the nested data are exploited by designing the drill. More than 10,000 servers are scaled up to achieve the capacity for processing the trillions of records in seconds and petabytes of data.

D. Apache Spark

The apache spark is considered as one of the open source of BD processing framework, which is used to analyze the process at high speed. There are three components present in Spark such as worker nodes, driver program, and cluster manager. On the spark cluster, the starting point of execution is served by driver program, where the resources are allocated by cluster manager. The data are processed in the form of tasks by using worker nodes. To execute the tasks, every application has an executor, which is defined as the set of processes. The spark applications are deployed in an existing Hadoop cluster, which is the main advantage of these apache sparks.

E. Jaspersoft

The open source software called Jaspersoft package uses the database columns to provide the results for end users. The data visualized on storage platforms such as Redis, MangoDB, Cassandra, etc., where this process are speeded by this package. Without loading, extraction and transformation, the BD are explored quickly by using the main property of Jaspersoft. The package's property is used to extract the vast amount of data from the store directly and also used to build a powerful Hypertext Markup Language (HTML). Anyone from inside or outside the user's organization can able to obtain these generated reports.

F. Splunk

At present, machine generated a huge amount of data, which are exploited by developing the intelligence and real time platform called Splunk. The BD and up-to-the moment cloud technologies are combined by using Splunk, which help the user to analyze, monitor and search their generated data using web interface. Compared to other stream processing tools, Splunk is completely different because it produces the results in intuitive ways namely alerts, graphs and reports. The main applications of Splunk is to diagnose the problems for information technology infrastructures and system, in addition it supports the business operations.

V. PROBLEM STATEMENT FOR BIG DATA ANALYSIS

The below section presents the issues in development and management of BDA, which are obtained from the above discussion. The BD faces the challenges in some important features such as data quality, processing speed, data interpretation, exception handling of BD and visualization. In this research study, four major technical issues such as data security, data quality, data management and privacy are discussed.

A. Data Security

While adopting the BD, the user resistance is created by weak security, which leads to damage and financial loss for firm's reputation. The unauthorized parties can able to access and transfer the confidential information due to improper installation of security mechanisms.

The strong security management protocols are introduced along with several security solutions such as building the firewalls, detection, encryptions and intrusion prevention systems into BD systems to avoid the above security issues.

B. Data Quality

To make the decisions, data quality is the important factors to define the fitness of data along with the specific purpose of usage. The quality of data may reduce due to collection of unstructured data from a wide array of sources. To evaluate the data quality, a new quality metric is needed to develop by a data quality control process, which also used to repair the erroneous data and make a trade-off between assurance gains and costs.

C. Data Management

The streaming social media and sensors generate vast amount of data, but only few firms are interested to invest in the storage for these collected data. The data management issues for firms are reduced by the combination of Hadoop and edge computing. In a distributed environment, the computation of BD and complex transformation are carried out by using Hadoop. But, the streaming analytics and ad hoc data exploration are not done by Hadoop. To overcome these issues and provide the real time responses by developing the edge computing called fog computing. Even though, it is beneficial compare to Hadoop, the development and management for data centers are costlier in edge computing.

D. Privacy

The serious security concerns may arise for governments, individuals and firms on extensive collection of personal data due to the massive growth of BD technologies. Individuals found it is difficult to analyze the BD due to these privacy concerns and decide not to analyze the personal data. The BD enhances the service quality and reducing the cost, even though privacy is an important issue for both firms and end customers. Hence, a balance between privacy and the usage of personal data for services should be considered by both firms and users. However, according to customer service, data type, service types and regulatory environments, the balances are created, but none of the existing works focused on the privacy measures.

VI. Proposed Method

In this section, the suggestion for the proposed model is given according to convolutional neural network (CNN) with two variation models is described. Initially, the big data is collected from the publicly available dataset and the important features are selected by using appropriate feature extraction techniques.

A. Data Collection and Feature extraction

The first module of the proposed framework is in charge of extracting all the data and filtering the ones that do not contain useful information. Then, this data is given as input to the feature extraction techniques. Once the data have been extracted, it would be possible to directly apply any classification method. Collecting such an extensive dataset can be a tedious and very time-consuming task. In addition, the time it requires to train such a classifier would be prohibitively long even if complex computation parallelization techniques and expensive specialized hardware are used. To overcome these limitations, two variations of CNN are developed.

B. Classification of Convolution Neural Network

Convolutional Neural Network is a class in deep neural networks which is most commonly used in classification. The major objective of CNN is to classify data in large streams through learning models. In this work, the training and testing data are given to CNN classifier for prediction. The CNN convolution layer is a network, where data will be convolved with filters to produce feature maps. This kind of feature map is forwarded to the next convolution layer to receive high-level features from the input frame. Between the convolution layers, a non-linearity function and down sampling operation are utilized to add non-linearity and reduce the dimensionality of the data. The max pooling layer is used for down sampling operation that reduces the dimensionality while predicting the dominant features in the feature maps. After the initial layer of Neural Network (NN), the flattening layer is utilized to vectorize the feature maps. In the NN, the flatten input vector is forwarded into the network to produce a number at each output neurons, which shows how much an input vector is classified as a certain activity.

C. Fully Convolution Neural Network

A Fully CNN as FCNN classifier is used to classify the data. The majority of the data we analyzed, contained several relevant information. All the data extracted from a database must be considered when making a prediction of the category as a whole. In this experiments, a simple vote aggregation approach is taken where the final prediction is the most common label among the data. In the context of CNN framework, the application of metric learning is proposed in conjunction with transfer learning to improve the performance of distance-based classification over the features generated by the FCNN. Specifically, metric learning is applied to fine-tune the final layers of the pre-trained FCNN. The reason for using machine learning is to adjust the final layers of pre-trained FCNN, which is twofold. The first reason is that training the final layers of the FCNN will improve the discriminative information present in the generated descriptors. Secondly, the use of metric learning to adapt the final layers will preserve the natural compatibility of feature-extraction FCNN with over-time learning. To classify the data, first triplet loss in DCNN should be calculated.

D. Inception Convolution Neural Network

The proposed architecture (ICNN) builds on several recent developments in deep learning architectures, including Inception Nets and CNNs.

It tries to reduce the number of computational parameters, while providing better segmentation accuracy. The key feature of Inception is that it concatenates the outputs of multiple differently sized convolutional kernels in the inception block. Inception-v4 is a simplified version of Inception-v3 model, using lower rank filters for convolution. Inception-v4 however combines Residual concepts with Inception networks to improve the overall accuracy over Inception-v3.

The outputs of the inception layers are added to the inputs of the Inception-Residual module.

VII. RESULT AND DISCUSSION

In this section, the various experimental analysis are conducted to validate the performance of hybrid techniques in terms of several parameters such as precision, recall, accuracy, f-measure, and Probability distribution and Perplexity (PP). The following subsection contains the experimental validation for hybrid models, discussion of traditional techniques and finally, analysis of state-art-of-the-methods, which are briefly explained.

A. Validation for Suggested Hybrid Model:

In this section, the experiments are carried out for hybrid model includes Decision tree (DT), Naive Bayes (NB), Cart, UN-gram for Mental Depression Health Weightage scheme (MDHW) by means of PP, precision and recall. Table 1 shows the validated results for hybrid techniques. The graphical representation for the hybrid models is represented in Figure 2.

Table 1: Analysis of Hybrid Techniques

Methodology	PP (%)	Precision (%)	Recall (%)
Cart	99.26	93.36	75.10
Decision Tree	99.08	91.08	81.03
Naive Bayes	70.87	85.90	80.15
UN-gram	Not Available	99.36	99.30

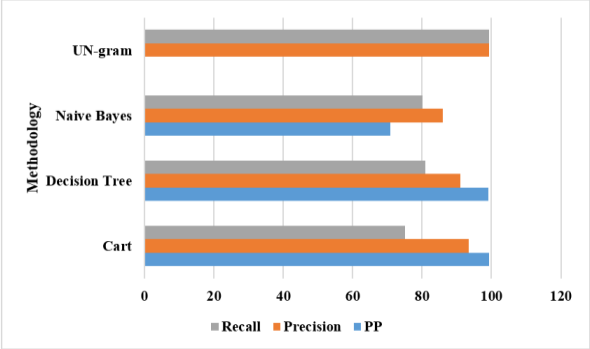

Figure 2: Performance of Hybrid Models

From the Table 1 and Figure 2, it is clearly stated that the hybrid models includes Cart, NB, DT and UN-gram provides better performance by means of PP, recall and precision for MDHW. For the validation of PP, researcher considered only three models, where Cart and DT achieved 99.26% and 99.08%. But, NB achieved poor PP when compared with other two techniques (i.e. 70.87). The experiments were further validated for precision and recall values, where all the models achieved less recall values when compared with precision. The UN-gram achieved 99.35% recall and precision value for MDHW. The NB achieved 85.90% precision and 80.15% recall values, where DT achieved 91.08% precision and 81.03% recall values.

B. Discussion of Existing Techniques

In this section, the survey of recent techniques is presented, which are used in the BDA. Table 2 presents the various techniques used in BD, which consists of its advantage and limitations.

Table 2: Survey of several techniques in BDA

Author with Year	Methodology	Advantage	Limitation	Performance Evaluation
S. Hussain, *et al.*, [17] (2019)	During the conventional partitioning, constraint modeling is developed to avoid the semantic loss of healthcare documents in BD.	According to the scenario, the affected documents are identified and semantic loss is avoided by resolution of incomplete documents.	When the dataset increases, the conflicts of the method also increases, which leads to reduce the ability of having high interoperability.	The performance of the method is tested against CDAR dataset by using conflicts based on the constraint level and resolution of these conflicts are used.

D. Vivek, and P. Balasubramanie, [19], (2019).	The depression phrases predicted by using UN-gram classifier and also the features of unstructured data are extracted by using MDHW.	The exact emotional phrases are identified by continuously splits the sentences by using UN-gram classification technique.	These methods failed to predict the depression level of the individual from these extracted data.	Precision, recall, PP are used to evaluate the effectiveness of MDHW against various existing methods.
P. Shobha Rani, et al., [20] (2017)	Developed the MOUNT a multi-level annotation and integration framework to process the heterogeneous BD.	The categorization of domain information on BD is carried out by employing Yago and Seeds Search technique, which develops the query processing and execution time.	The system obtained the low precision and recall values due to improper format of query, which leads the system to provide irrelevant results.	Execution time, precision and recall are used to validate the performance of the MOUNT technique against Airstore.
M. Chen, et al., [21] (2017)	The chronic disease is effectively predicted from the structured and unstructured data by using Convolutional Neural Network (CNN)-based multimodal algorithm.	The missing data are reconstructed by latent factor model, which is used to address the incomplete data complexity.	When the text feature numbers are relatively small, the vast amount of useful information contained in the texts is unable to describe.	Receiver Operating Characteristic (ROC) curve, accuracy, precision, F1-measure, Area Under Curve (AUC) and recall are used as parameter evaluation.

C. Analysis of CNN and MOUNT techniques

In this section, the validation of existing techniques such as MOUNT and CNN with current study are conducted in heterogeneous BD by means of accuracy, f-measure, precision and recall. Table 3 shows the validated results for CNN, MOUNT against K-nearest Neighbor (KNN), NB and DT and the graphical representation is illustrated in Figure 3.

Table 2: Validation of CNN and MOUNT against Current Study

Methodology	Accuracy (%)	F-measure (%)	Precision (%)	Recall (%)
KNN	52	45	49	48
DT	65	68	65	62
NB	50	65	80	55
CNN	90	91	93	96
MOUNT	89	91.15	87	86
Current Study	95	94.15	96	98

From the Figure 3, it is clearly stated that the current study achieved higher performance than other existing techniques namely KNN, DT, NB, CNN and MOUNT techniques. The existing techniques achieved very low performance in all parameters, however the current study provides better performance. For instance, the KNN and DT achieved nearly 60% accuracy and 50% f-measure, but the current study, CNN and MOUNT techniques achieved nearly 93% accuracy and 93% f-measure. This drastic changes occurred in CNN with its variations and MOUNT is due to the number of increased hidden layers to handle the heterogeneous BD. The existing techniques includes DT and NB are insufficient for handling the BD.

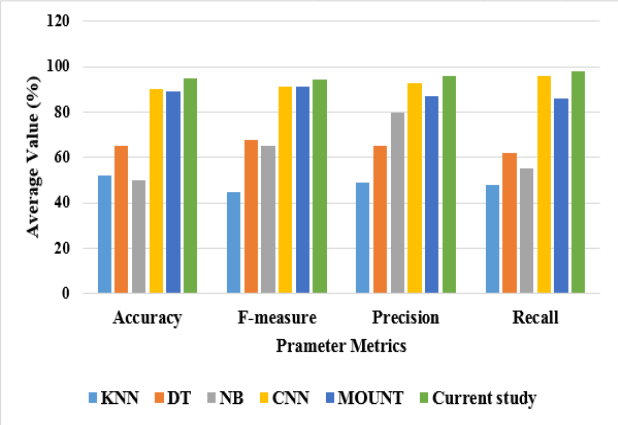

Figure 3: Analysis of Traditional techniques with Current Study

VIII. CONCLUSION

BD serves as a good basis for many organizations and governments in different sectors that intend to automatically process and extract valuable insights in order to help decision makers. However, the fact to collect and compute all possible and varied data could lead to many security and privacy violations. Because of technological growth, the data are generated drastically in recent years, which leads to the fundamental problem that is how to describe the essential characteristics of BD. Analyzing these data are challenging for a researcher.

To solve this issue, the research work presented the various research issues, challenges, and tools, which is used to analyze these BD. From this study, it is understood that every BD platform has its individual focus. Some of them are designed for batch processing, whereas some are good at real-time analysis. Each BD platform also has specific functionality. Different techniques used for the analysis include statistical analysis, machine learning, data mining, intelligent analysis, cloud computing, quantum computing, and data stream processing.

The validation for hybrid techniques, CNN and MOUNT are conducted for various metrics to handle the heterogeneous BD. The results proved that the CNN and MOUNT techniques are sufficient to handle the BD due to increased number of hidden layers. Furthermore, in future research work, the research will pay more attention to these techniques to solve problems of BD effectively and efficiently.

REFERENCES

1. A. Botta, W. De Donato, V. Persico, and A. Pescapé. (2016). Integration of cloud computing and internet of things: a survey. Future generation computer systems, 56. pp. 684-700.
2. J. Q. Li, P. Rusmevichientong, D. Simester, J. N. Tsitsiklis, and S. I. Zoumpoulis. (2015). The value of field experiments. Management Science. 61(7). pp. 1722-1740.
3. J. Li, F. Tao, Y. Cheng, and L. Zhao, (2015). Big data in product lifecycle management. The International Journal of Advanced Manufacturing Technology. 81(1-4). pp. 667-684.
4. S. Maklan, J. Peppard, and P. Klaus. (2015). Show me the money: improving our understanding of how organizations generate return from technology-led marketing change. European Journal of Marketing. 49. pp. 561-595.
5. K. Pousttchi, and Y. Hufenbach. (2014). Engineering the value network of the customer interface and marketing in the data-rich retail environment. International Journal of Electronic Commerce 18(4). pp. 17-42.
6. R. Thackeray, B. L. Neiger, C. L. Hanson, and J. F. McKenzie. (2008). Enhancing promotional strategies within social marketing programs: use of Web 2.0 social media. Health promotion practice. 9(4). pp. 338-343.
7. S. Yin, and O. Kaynak. Big data for modern industry: challenges and trends [point of view]. Proceedings of the IEEE 103(2). pp. 143-146.
8. M. Chen, Y. Zhang, M. Qiu, N. Guizani, and Y. Hao. (2018). SPHA: Smart personal health advisor based on deep analytics. IEEE Communications Magazine. 56(3). pp. 164-169.
9. I. Md Ezazul, I. Md Rafiqul, and A. B. M. Shawkat Ali. (2016). An approach to security for unstructured big data. The Review of Socionetwork Strategies 10(2). pp. 105-123.
10. W. Kun, L. Tong, and X. Xiaodan. (2019). Application of Big Data Technology in Scientific Research Data Management of Military Enterprises. Procedia computer science. 147. pp. 556-561.
11. S. Jogar, P. Naik, V. Vyapari, M. Vaddar, K. Dambal, and B. Hatti. (2019). Chronic Diseases Prediction over Bigdata by using Machine Learning, International Journal of Scientific Research in Computer Science, Engineering and Information Technology. 2019.
12. A. Oussous, F. Z. Benjelloun, A. A. Lahcen, and S. Belfkih. (2018). Big Data technologies: A survey. Journal of King Saud University-Computer and Information Sciences. 30(4). pp. 431-448.
13. P. Grover, and A. K. Kar. (2017). Big data analytics: a review on theoretical contributions and tools used in literature. Global Journal of Flexible Systems Management 18(3). pp. 203-229.
14. A. Kumar, V. Dabas, and P. Hooda. (2018). Text classification algorithms for mining unstructured data: a SWOT analysis. International Journal of Information Technology. pp. 1-11.
15. C. P. Chen, and Z. Chun-Yang. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. Information sciences. 275. pp. 314-347.
16. D. B. K. Kamesh, V. Neelima, and R. Ramya Priya. (2015). A review of data mining using bigdata in health informatics. International Journal of Scientific and Research Publications. 5(3). pp. 1-7.
17. S. Hussain, M. Hussain, M. Afzal, J. Hussain, J. Bang, H. Seung, and S. Lee. Semantic Preservation of standardized healthcare documents in big data. International Journal of Medical Informatics. 2019.
18. M. Tabrez Nafis, and R. Biswas. (2019). A secure technique for unstructured big data using clustering method. International Journal of Information Technology. 1-12.
19. D. Vivek, and P. Balasubramanie. (2019). An Expressive phrases identification supported with feature prediction consuming unstructured data collection. Multimedia Tools and Applications. pp. 1-16.
20. P. Shobha Rani, R. M. Suresh, and R. Sethukarasi. (2017). Multi-level semantic annotation and unified data integration using semantic web ontology in big data processing. Cluster Computing. pp. 1-13.
21. M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang. (2017). Disease prediction by machine learning over big data from healthcare communities. IEEE Access. 5. pp. 8869-8879.

AUTHORS PROFILE

Parashiva Murthy B M is currently working as an Assistant Professor in the Department of CSE at JSS SCIENCE AND TECHNOLOGY, Mysuru from 11/09/2019 to till date Worked as an Assistant Professor at GSSSIETW from 17/07/2013 to 06/09/2019.

Dr. Sumithra Devi K A completed Ph.D in Computer Science Engineering, Specialization in VLSI Partitioning CAD Tool, Avinashlingam Deemed Central University for Women, Coimbatore, M.Tech Master of Engineering, Electronics, University of Visvesvaraya College of Engineering, Bangalore University and B.E Bachelor of Engineering, Electronics and Communication, Malnad College of Engineering, Hassan, Mysuru University.

Currently working as Dean Academics & Head: From Dec 2017 till date, DSATM, Bangalore. 2. Professor and Principal: 15th January 2015- Till Nov 2017 GSSS Institute of Engineering and Technology for Women, Mysore-570016 Achieved Accreditation for 5 UG Engineering Programs under NBA Tier-II, for 3 years 3. Member Evaluator, CSE Board, National Board of Accreditation, New Delhi, India. 4. Professor & Director: 01/09/2009 – 05/01/2015 Department of Master of Computer Applications, R.V. College of Engineering, Bangalore – 59 5. Achieved Accreditation for the program 2 times for 3 years 6. Established increase in intake from 60 to 120 7. Established Research centre Under Visvesvaraya Technological University, Belagavi, Karnataka. 8. Professor and Head: 21/10/2001 – 01/09/2009 Department of Master of Computer Applications, R.V. College of Engineering, Bangalore – 59 a. Faculty to establish the MCA Department 9. Assistant Professor: 01/11/1996 – 21/10/2001 Department of Computer Science and Engineering, R.V. College of Engineering, Bangalore, Karnataka, India – 560059 10. Lecturer: 01/11/1986 – 01/11/1996 Department of Computer Science and Engineering, R.V. College of Engineering, Bangalore, Karnataka, India – 560059 a. Faculty to establish the First batch of CSE Dept in R.V Engineering College.