International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-6, April 2020

Sentiment Analysis of Events on Social Web

Neha Garg, Kamlesh Sharma

Abstract*: These days, Data volume has experienced enormous increase in volume, giving new challenges in technology and application. Data production has been expected at the rate of 2.5 Exabyte (1Ex-abyte=1.000.000Terabytes) of data per day. The main sources of data are: sensors collect climate information, traffic and flight information, social media sites (Twitter and Facebook are popular examples), digital pictures and videos (YouTube users upload 72 hours of new video content per minute), etc. Out of them social media becomes the prominent representative for the data source of big data. Social big data comes from the combination of social media and big data. Here, the data is mostly unstructured or semi-structured. The classical approaches, techniques, tools and frameworks for management of data have become insufficient for processing this huge volume of data and not capable for providing efficient solution to handle the increased production of data. The major challenge in data mining of big data is, its inadequate approaches to analyze massive amount of online data (or data streams). Specially, the field of sentiment analysis and predictive analysis has become so much promising area to place an organization at doom or at boom by provide accurate decision at accurate time.

The current paper provides an insight of machine learning algorithm both supervised and unsupervised method; and the traditional knowledge extraction process. The application field of sentiment analysis, the issues those are faced during data collection and cleaning. This study flourishes a complete picture of recommendation system based on the sentiment analysis of events. The key motivation of the paper is to incorporate the event sentiment analysis and give the feedback and recommendation and illustrate the ongoing researches in the field of sentiment analysis and its application.

Keywords* : *Machine Learning, Predictive Analysis, Sentiment Analysis, Supervise Learning.

I. INTRODUCTION

A huge amount of data is being produced from a variety of users, using heterogeneous sources. Among all the available sources, Social media has turned out to be a common means for online communication; user can generate their own content, share it, and bookmark it and network at an exceptional rate. For example may take account of all networking sites like Facebook, MySpace, Digg, Twitter and JISC listserve on the academic side [18].

Revised Manuscript Received on April 30, 2020.
***** Correspondence Author
Neha Garg*, Deptt of Computer Science Engineering, Faculty of Engineering and Technology, Manav Rachna International Institute of Research and Studies, Faridabad, India. Email: nehagarg.fet@mriu.edu.in
Dr. Kamlesh Sharma, Deptt of Computer Science Engineering, Faculty of Engineering and Technology, Manav Rachna International Institute of Research and Studies, Faridabad, India. Email: kamlesh.fet@mriu.edu.in

Social media is playing a vital role in changing the public opinion and providing them a platform for discussion in community and set trends and agendas in topics that may vary starting from the environmental issue to the politics, to the latest technology and the entertainment industry.

By using this huge amount of information, available on social sites, people can enhance the worth of decision- making by improving the outcomes, extracted from databases [6,24]. If an organization develops its own dataset for mining, and for getting higher profit, an urgent requirement rises to collect extra information from external sources Facebook, MySpace, Digg, Twitter and JISC listserve is generated. Here knowledge extracted from company dataset is referred as 'internal knowledge' while that from Facebook, MySpace, Digg, Twitter and JISC listserve is called 'external knowledge'[7]. This knowledge is getting used by organization to perform sentiment analysis, future forecasting, predictive analysis, supply chain management, employee performance measurements and in other fields by using machine learning algorithms.

II. LITERATURE REVIEW

In this data-driven world decision making has become a promising notation, which provides recognition to the Big Data in a variety of fields. Since the promise of big data appears to be real; there is an ample gap between its actual potential and its realization [8].

The heterogeneity, scalability, timeliness are the major issues that start from the data capture. Because of such a complexity in deciding which data is to be stored, which date is need to be considered, which should be discarded, it's often considered as a new scientific paradigm called data intensive computing [5,9]. The traditional devices like solid state disk, hard disk etc are not appropriate [21] and approaches and SQL can't cope with such a huge amount of heterogeneous data so a new approach NOSQL is introduced [20, 11]. In [2], provide an insight of all available methods, techniques and tools for big data analytics and how to visualize the decisions [15]. In [17], how to do predictive analysis and various methods of analysis have been discussed.

With the increased use of social media and advanced supporting features of WWW, the heterogeneous data from heterogeneous user is generating every day. And to analyze such a huge amount of heterogeneous data and perform predictive analysis has become the current trend for the organizations and to make decision with more precision is the current trend [25]. For increasing the accuracy of results, the researchers are making machine capable to analyze the data and make decision [17]. In continuation, the historical data is also getting used to make the comparison between previous and current situation. The model of comparison, uses certain predefined set of rules (called training models) is defined as supervised machine learning technique and unsupervised model doesn't required such predefined training rules [16].

III. KNOWLEDGE EXTRACTING PROCESS

The process of knowledge extraction is shown through Fig. 1[4]. Some of the major challenges in the knowledge extraction in big data resources are data incompleteness, inconsistence and timeliness, scalability, and data security [8, 9]. The very first constraint in analyzing the data is that data should be well-constructed or well structured. Though, taking it into account a vast range of data sets in Big Data is very huge problem, it is still a challenge for everyone to propose an effective and efficient access, analysis and illustration of semi-structured or unstructured data. A lot of data preprocessing techniques, which includes data cleaning, transformation, integration, and reduction, can be useful to eliminate noise and modify inconsistency. Various issues arise at each step of sub-processes specifically for data-driven applications. In the further subsections, a concise discussion about challenges faced by every sub-process is discussed.

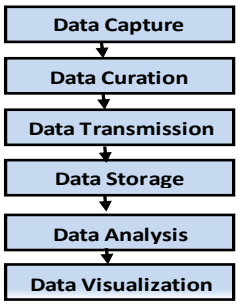

Figure 1: Knowledge Extraction Process

A. Data Capture

Data sets growing in volume rapidly, being received from ubiquitous information-sensing mobile devices, mobile phones, remote sensing, software logs, cameras, microphones, wireless sensor networks, social sites and so on[16]. Everyday approx. 2.5 quintillion bytes of data is being generated, and it's keep increasing exponentially [10, 5]. With such a growth it's become a very difficult task to store all data. Many fields like finance, medical etc. often delete their data in lack of proper storage space. This precious data is being produced and captured at a very high rate, but has to be avoided at last [8].

B. Data Curation

Data Curation focused on searching and recovery of data, maintain the data quality, value addition, reusability and preservation over time. To achieve this, one need to perform a number of sub-tasks such as authentication, archiving, management, preservation, retrieval, and visualization of data [5].

C. Data Transmission

The data can be captured on standalone system as well as in a distributed environment also. Once the data is cured it's transmitted to a storage space. If it is locally captured then there is no need to transmit it but for distributed environment a network is required for transmitting data to the central location for further analysis. Network bandwidth and the data integrity and security are the major issues in distributed systems.

D. Data Storage

Present database management tools are incapable to process and manage Big Data which is increasing exponentially and is very complex. The classical methods for handling structured data have two parts, one is to manage database schema for storing the data values [21], and other is to provide a relational database to facilitate retrieval of data. To organize multi-dimensional data values in a structured manner, data warehouses and data marts are very common methodologies, which are Structured Query Language (SQL) based database systems [3]. A data warehouse supports relational database, for collecting and analyzing data, in-return generate a report and inform the outcome to the end users. The concept of data mart is also based on data warehouse and provides the access to data warehouse for analysis. Before storing the data, preprocessing e.g. cleaning, transformation and sorting of data is performed. After which, the data become accessible for advanced online data mining functions.

NoSQL database [11, 5], more often referred as ''Not Only SQL'', is the recent advancement for huge and distributed database design and management system. NoSQL does not avoid SQL. A few NoSQL systems are completely non-relational; others do not support certain relational functionalities such as static schemas and join operations [11, 20]. The conventional Big Data platforms implement NoSQL to remove inflexibility of normalized RDBMS schemas. The popular example of NoSQL databases is Hbase [20].

Big Data has altered the method to collect and store data, starts from the data storage mediums, to the data storage space architecture (SAN, NAS etc), to the data fetching method and many more things. To beat the rising requirements, the system need advanced storage devices [21] and improved I/O rate to meet the requirements [5], more innovations are required in the field of Big Data storage.

- Firstly, the ease to access Big Data is the highest priority for the knowledge extraction process.
- Secondly current approaches for data storage can't have similar performance to handle both types of I/O, sequential as well as random at the same time.
- Lastly, all the existing storage architectures like SAN, NAS, have many drawbacks and limitations for high-scale distributed systems. High concurrency and per server throughput are the major parameters for the extremely scalable computing clusters, and today's storage systems legging behind in these dimensions. Optimizing data accessibility is a general technique to increase the efficiency of data-intensive computing; which also execute data replication, migration, distribution, and access parallelism [13].

E. Data Analysis

In the last few decades, more research has been done to boost the performance of data analytics algorithms. From the perspective of Big Data analytical techniques, incremental algorithms are scalable, but they are not well suited to all machine learning algorithms [17].

In the recent past, researchers devoted their more time and attentions to increase the performance of analysis algorithms, by considering following things-

- To handle the increasing amount of data and speed of processors by applying the Moore's Law.
- Since the size of data is increasing rapidly in comparison to CPU speeds,

a remarkable shift [8] in processor technology is required—although the frequency of processors' clock cycle is increasing rapidly, using Moore's Law, the clock speeds is still lagging behind. The processors, with growing numbers of cores, are being embedded, taking into account concept of parallel computing [5, 13].

- For real-time applications, like navigation, social networks, finance, biomedicine and internet of thing, where role of Big Data comes into play, on time response is at highest priority. How can one guarantee all the time, timely reply, especially when the amount of data to be processed is very huge or streamlined [9]

F. Data Visualization

The user can better understand the scenarios, the expected outcomes or comparative analysis in the terms of diagrams or pictures. The key purpose of data visualization is to signify knowledge more spontaneously and efficiently with the use of various techniques such as various plots, graphs [12]. To communicate information effectively by giving knowledge buried in the complex and large-scale data sets, visualization and basic functionality both are important parameters [15]. Information, abstracted in some schematic forms, having some attributes or variables, is also important for data analysis. This way of providing information is more sensitive [12] in comparison with classical approaches.

IV. CLASSIFICATION OF BIG DATA ANALYTICS

As big data comprises of unstructured data of various type, so the analysis becomes a critical part where for each type of data there have been a different type of analytic proposed e.g. text analytics, audio analytics, video analytics, sentiment based analytics and predictive analytics [1, 2]. In [1], all the techniques based on the different forms of data have been discussed. This paper emphasize on predictive analysis only.

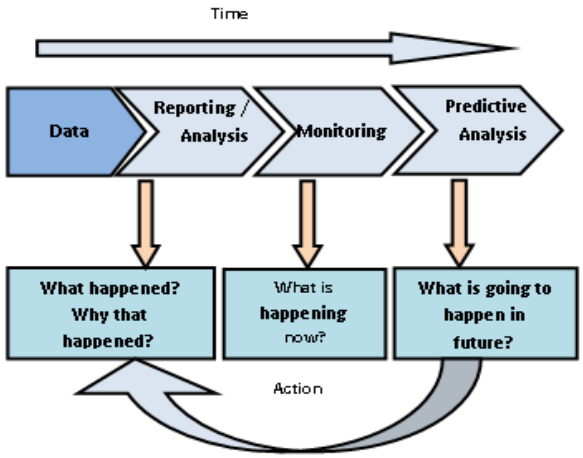

Figure 2: Predictive Analysis

Predictive Analysis comprises various methods that predict future outcomes based on historical data [1]. Data which can be readily used for analysis are structured data like age, name, date of birth; marital status etc. unstructured data are data of any form posted in social media or may be somewhere, which needed to be extracted and other form of data is called semi-structured that is not perfectly organized like structured data but still have certain organizational properties like tags of xml, email etc [26]. Predictive analysis seeks to uncover patterns and find relationships among them [19].

Predictive analysis can be classified into two ways-

A. Regression technique
B. Machine learning technique

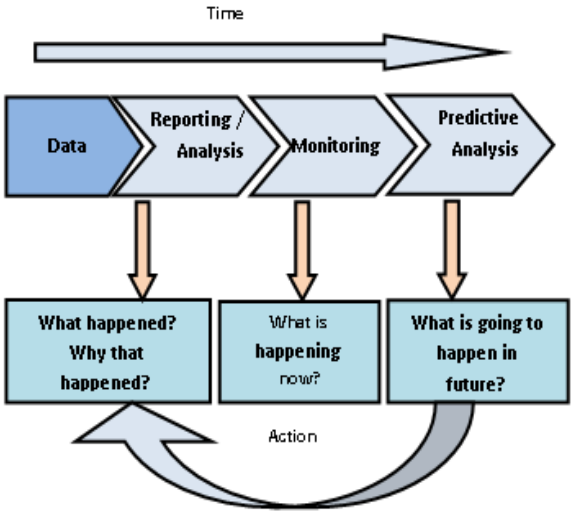

Figure 3: Predictive Analysis

A. Regression Technique

In this technique, the main focus is on developing a mathematical equation for the modelling the relationships among the various variables under concern. Based on the problem in hand, a vast range of models can be developed to perform predictive analytics. Few models are as follows-

i. **Linear regression Technique** analyses the equation between the resultant variable or dependent variable and a set of independent or predictor variables. This equation predicts the resultant variable as a linear function of the parameters. These parameters are defined in such a manner so that result can be optimized.

ii. **Discrete choice models** are used when the dependent variable has discrete values. Some examples of these methods are logistic regression, multinomial logit and probit models. When the dependent variables are binary in nature Logistic regression and probit models are generally used.

iii. **Time series models** are used for forecasting and predicting the future of the variable. These models generally take into account the data points which have some internal correlation over time like weather forecast. Modeling the dynamic path of a variable can enhance the result of forecasting since the predictable components of the series can be estimated into the future. Autoregressive models (AR) and moving-average (MA) models are commonly used.

iv. **Classification and regression trees (CART) a**re non-parametric decision tree technique for predicting continuous dependent variables (regression) and categorical predictor variable (classification) and form a tree accordingly.

Decision trees are formed based on a set of rules to model data set:

- Rules for splitting data set at a node based on variables values.
- Once a rule is decided to split a node into its child, the same process is followed by every node, recursively.
- A stopping rule to decide when CART detects no further splitting can be made or a branch is terminal, or some pre-defined terminating rules are met.

- Every observation falls into precisely single terminal node, and each terminal node is uniquely identified by a set of predefined rules.

A very popular method for predictive analytics is Leo Breiman's random forests.

Multivariate adaptive regression splines (MARS) is a regression method of non-parametric type that develop flexible models by fitting piecewise linear regressions. MARS is based on the concept of Knot, where one local regression model provide path to another one and that's how provide the intersection point between two splines. In multivariate and adaptive regression splines, basis functions, contains information for one or more variable; are used for generalizing the search for knots.

B. Machine learning Techniques

Main aim of these techniques is to make the machines reply, without depending upon the program develops in advance for that particular problem [17]. The role of Artificial Intelligence Researchers is to make the dataset in such a way that makes machines to reply by going through that data. Machine Learning is a Science, which considers previous computations as a model and applies this model on new data resulting into some intelligent response. Some of the machines learning models are as follows-

i. **Supervised Algorithms-** The core of supervised algorithms is to make or model a program on predetermined set of sample data, which helps in finding an approximate output when new data is supplied in place of predetermined dataset [16]. The supervised algorithm can be categorized in multiple type of models have been discussed in the table-

TABLE I: TYPES OF SUPERVISED ALGORITHMS

Algorithm Type	Description	Input Values	Example
Regression Algorithms	• These algorithms formed a model to estimate the relationship between dependent and independent variables [8]. • With the help of multiple such examples, a model/function is developed which help in decision making	dependent and independent variables	• Ordinary Least Squares Regression (OLSR), • Linear Regression, • Logistic Regression, • Stepwise Regression
Instance based Algorithms -	• In this algorithm already available example-data is examined to model the relationship with current targeted value [10]. • So, it's using a similarity measure, finds best match and make prediction. • For example, speller correction, search ranking, etc. can be solved under this algorithm.	Predefined data set	• k-Nearest Neighbor (kNN), • Learning Vector Quantization (LVQ), • Self-Organizing Map (SOM)
Regularization Algorithms	• If the prediction performance is not so good then the issue is called as over fitting problem. This can be removed by Regularization algorithms.	Finite set of variables	• Ridge Regression, • Least Absolute Shrinkage and Selection Operator (LASSO), • Elastic Net, Least-Angle Regression (LARS).
Decision Tree Algorithms	• Decision tree is a binary structure for a set of attributes to be tested in order to predict the output. • The nodes of the tree split the data to find a classifying variable. • This algorithm handles the problem logically and does step-wise execution to get good results.	Set of attributes either having continuous values or binary values.	• Classification and Regression Tree (CART), • Iterative Dichotomies 3 (ID3), • C4.5 and C5.0 (different versions of a powerful approach), • Conditional Decision Trees
Bayesian Algorithms	• Those algorithms which apply Bayes Theorem. • Bayesian method means the one which believes on subjective probability and results in some future reference.	Expected probability	• Naive Bayes, • Gaussian Naive Bayes, • Multinomial Naive Bayes.
Artificial Neural Network Algorithms	• The aim of neural network is to solve the problem as solved by human brain with intelligence. • Hence it is used in ML. • The area is much related to pattern matching problems.	Single neuron with arbitrary number of inputs.	• Preceptor, • Back-Propagation, • Hopfield Network.
Deep Learning Algorithms	• Deep learning algorithms are one of the special cases of Artificial Neural Networks. • They help in building more complex neural networks and work for voluminous datasets.	Use a cascade of multilayer of non-linear processing units. Each layer takes output of previous layers as the input.	• Deep Boltzmann Machine (DBM), • Deep Belief Networks (DBN), • Convolution Neural Network (CNN)

Drawbacks of Supervised Algorithms-

a) Supervised techniques are highly dependent on the input training data set, which are generally human made and can be biased at certain points.

ii. **Unsupervised algorithms-** The core of unsupervised algorithms means that a program is provided with some collection of data,

with no predetermined dataset being available, and one has to make patterns or relations among those data values. Following are different types of un-supervised algorithms:

a. Clustering Algorithms

Clustering algorithms helps in creating cluster/group of similar elements/data/problems. If one analyzes a cluster then there should be some common parameter. Few popular clustering algorithms are k-Means, k-Medians, Expectation Maximization (EM), Hierarchical Clustering.

b. Dimensionality Algorithms

In dimensionality reduction the number of dimensions of internal structure of data is reduced without harming the actual data content.

The data is described using less information; however data summary is kept same. Few of the dimensionality reduction algorithms are Principal Component Analysis (PCA), Principal Component Regression (PCR), and Partial Least Squares Regression (PLSR)

c. Association Rule Algorithms

Association rule is a type of database rule which finds the relationship between dataset and data there. Suppose there are two variables A and B, then the probability that if one rule has A and later on it will require B depends upon the association rule.

Few popular examples for association rule algorithms are a priori algorithm, Éclat algorithm.

Drawbacks of Unsupervised Algorithms-

a) Unsupervised algorithms are computationally complex.
b) These algorithms are less accurate and trustworthy.

V. PROPOSED METHOD

With the advancement in the usage of social network data, the digital world is becoming the replica of real world and it is becoming complex and voluminous day by day.

More research work is going on the area of sentiment analysis to consider various aspects of this data and making analysis more impactful.

In the proposed model, various phases have been introduced to provide the feedback or the prediction related to the events. In broader way, the model is divided into three phases i.e.-

i) *Data Management*- this phase consists of collection of domain specific data, pre-processing it, removing the noise, removing the unwanted information and storing it in a proper format [5].

ii) *Feature Extraction*- in this phase, the data is organized based on the similarities in the keywords [27].

iii) *Recommendations*- here the recommendation or feedbacks are provided based on the sentiment analysis of the collected data. This paper presents the first phase i.e. related data management of social network data.

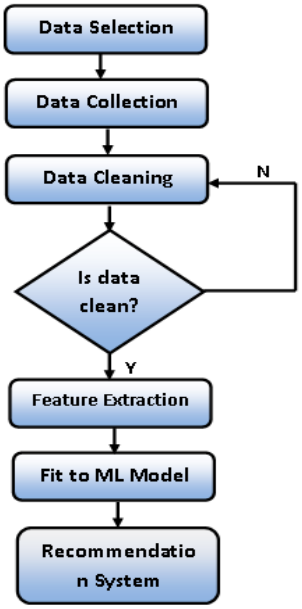

Figure 4: Proposed Method

A. Data Selection

As the users of social networking sites are increasing and the real world activities are being replicated as the events on these social networking sites. Hence the speed and volume of data is increasing and considering such a huge amount of data [1] is very tedious task. Data selection let the user filtering out the domain specific or the event specific data. This selection helps in concentrating in a specific area and increasing the accuracy level.

B. Data Collection

Once the domain of the problem is selected, the data can be collected using API access of the social networking sites with the help of user's credentials or can be taken directly from the online available datasets. Later on the data can be stored in the required form [5].

C. Data Cleaning

The data cleaning phase comprises of data preprocessing; removing punctuations, noise, stop-words, emoticons or other language symbols or words etc [5, 27]. As the proper cleaning may lead to the better results and all the data may or may not require the equal amount of cleaning. Hence in the proposed work the cleaning is done in a stepwise manner, to make sure that unnecessary stages may not increase the complexity of work.

D. Feature Extraction

In the data cleaning stage, the data has discrete features. Based on the problem in hand, the common features should be extracted to smoothen down the model and mapped the data into vector space [27, 28].

E. Machine Learning Model

Based on the problem in hand the machine learning model is opted as discussed in the Table-1. However researchers are generally using SVM and Naïve Bayes because of the simplicity and high accuracy of algorithms [29].

F. Recommendation System

Once the data is clustered or classified, the recommendation is provided to the user in terms of future steps, feedback or the result.

VI. ISSUES

1. For applying classification and clustering technique, the corpus should be carefully maintained [28].
2. The data should be cleaned accurately.
3. Many a times the Python Unicode can't handle the text properly.
4. The emoticon or other language words are treated as a noise, which may affect the results.
5. The same sentence may have different context or different meaning among various users, hence finding part of speech (POS) and bag of words (BOW) is again a tedious task.
6. Abbreviations and inconsistent spellings are also an issue.

VII. OBSERVATIONS & OUTCOMES

The internet is populated with a huge amount of data carrying multiple hash tag. To precise the search, here dataset is populated based on the popular hash tags on twitter e.g. #Chandrayan, #Pulwama Attack, #Deep Learning and #Medicine. Since for an event there are multiple sub-events too which are running in parallel. Here, the dataset is populated with 11,000 approx values where #chandrayan has 702 polarity tweets; #Pulwama Attack comprises of 527 tweets, #Deep Learning the collection of 639 polarity tweets and #Medicine 1323 tweets and remaining are the retweets under the same tags and user-mentions. While mapping these tweets to the polarity map ranges from -1 to +1, the bifurcation of tweets as positive, neutral and negative is found. These observations has been tabularized and recorded in the following table-

Table II: Observation Table

Keywords	No. of Tweets	Positive	Negative	Neutral
Chandrayan	702	179	40	483
Pulwama Attack	527	258	35	234
Deep Learning	639	224	56	359
Medicine	1353	545	61	747
Total	3221			

The outcome from the observation table has been extracted in terms of keywords and the number of tweets based on the polarity values as follows-

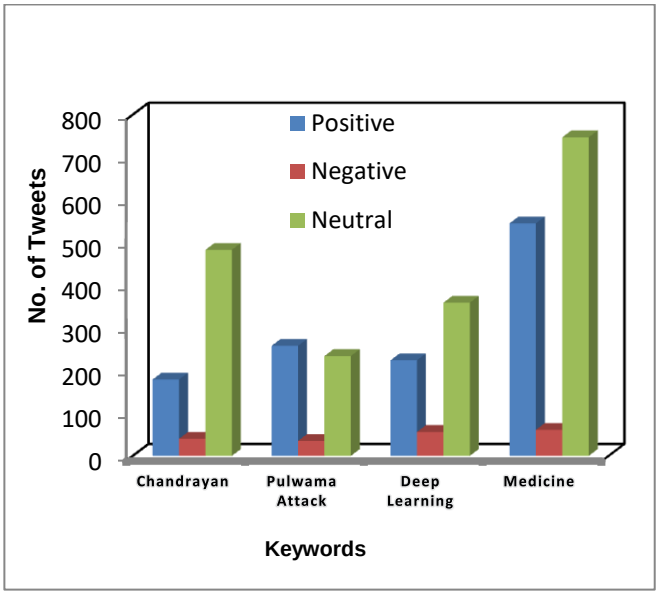

Figure 5: Graph for Observations

As shown in the graph above, hashtags are having more neutral opinions as compared to the positive and neutral approach. However the #medicine is more popular under the consideration of data selection.

VIII. FUTURE SCOPE

The future scope and the application area is very vast. The organizations, government, hospitals etc can use it for the betterment by performing predictive analysis, future forecasting, sequential analysis, political analysis, weather prediction, knowledge management, customer care services and chatbots.

IX. CONCLUSION

In this competitive world, with an increasing usage of digital devices, data is increasing at a very high rate from heterogeneous sources for such an unformatted data it's a tedious task to manage it and provide useful knowledge. Predictive analysis is a way to provide the knowledge on the bases of historic data. The aim of this study is to provide an inside of handling, managing and generating user friendly information in real time manner using predictive analysis.

This paper provides an insight of the complete process of extracting knowledge and the various challenges faced at each level of processing. Further, this paper provides a glance of the various predictive analysis techniques. These techniques and applications are discussed with respect to following aspect- (i) the major technologies and methodologies exists to capture, cure, store and analyze the big data? (ii) How can one analyze big data to find useful information? (iii) What are the challenges faced in predictive analysis of Big Data?

More practically, this survey paper provides an insight of basic processes to extract knowledge from big data, its importance and the predictive methods to analyze it without describing and analyzing the existing and intensively used applications like Facebook, Twitter etc.

REFERENCES

1. Neha Garg, Dr. Kamlesh Sharma:"A REVIEW ON THE STUDY OF BIG DATA AND BIG DATA ANALYTICS", in proceedings of ICCM 2017.
2. Amir Gandomi, Murtaza Haider:"Beyond the hype: Big data concepts, methods and analytics", International Journal of Information Management.
3. H. Liu, H. Lu, J.Yao: "Identifying relevant databases for multidatabase mining", Proceedings of PAKDD '98, 1998, pp. 210–221.
4. Jiawei Han, Micheline Kamber: "Data Mining: Concepts and Techniques", Diane Cerra, second ed., 2000.
5. C.L. Philip Chen, Chun-Yang Zhang: "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data", Information Sciences, www.elsevier.com/locate/ins, January 2014.
6. V. Lesser, B. Horling, F. Klassner, A. Raja, T. Wagner, S. Zhang:" BIG: an agent for resource-bounded information gathering and decision making", Artificial Intelligence Journal, vol. 118, 1–2, 2000, pp. 197–244.
7. Kaile Su a, Huijing Huang b, Xindong Wu c, Shichao Zhang," A logical framework for identifying quality knowledge from different data sources", Decision Support Systems 42 (2006) 1673–1683.
8. Divyakant Agrawal, Philip Bernstein, Elisa Bertino, Susan Davidson, Umeshwas Dayal, Michael Franklin, Johannes Gehrke, Laura Haas, Jiawei Han Alon Halevy, H.V. Jagadish, Alexandros Labrinidis, Sam Madden, Yannis Papakon stantinou, Jignesh Patel, Raghu Ramakrishnan, Kenneth Ross, Shahabi Cyrus, Dan Suciu, Shiv Vaithyanathan, Jennifer Widom: "Challenges and Opportunities with Big Data", CYBER CENTER TECHNICAL REPORTS, Purdue University, 2011.
9. Richard T. Kouzes, Gordon A. Anderson, Stephen T. Elbert, Ian Gorton, Deborah K. Gracio: "The changing paradigm of data-intensive computing", Computer 42 (1) (2009) 26–34.
10. Martin Hilbert, Priscila Lopez: "The world's technological capacity to store", communicate, and compute information, Science 332 (6025) (2011) 60–65.
11. Jing Han, Haihong E, Guan Le, Jian Du: "Survey on nosql database", in: 2011 6th International Conference on Pervasive Computing and Applications (ICPCA), 2011, pp. 363–366.
12. Simeon Simoff, Michael H. Bohlen, Arturas. Mazeika:" Visual Data Mining: Theory, Techniques and Tools for Visual Analytics", Springer, 2008.
13. Yi Cao, Dengfeng Sun: " A parallel computing framework for large-scale air traffic flow optimization", IEEE Trans. Intell. Trans. Syst. 13 (4) (2012) 1855–1864.
14. Agostino Di Ciaccio, Mauro Coli, Angulo Ibanez, Jose Miguel:" Advanced Statistical Methods for the Analysis of Large Data-Sets", Springer, 2012.
15. Simeon Simoff, Michael H. Bohlen, Arturas. Mazeika:" Visual Data Mining: Theory, Techniques and Tools for Visual Analytics", Springer, 2008.
16. Bjon Bringmann, Michele Berlingerio, Francesco Bonchi, Aristides Gionis: "Learning and predicting the evolution of social networks", IEEE Intell. Syst. 25(4) (2010) 26–35.
17. S.MD.MUJEEB, L.KASI NAIDU:" A Relative Study on Big Data Applications and Techniques", ISSN: 2277-3754,ISO 9001:2008 Certified, International Journal of Engineering and Innovative Technology (IJEIT),Volume 4, Issue 10, April 2015.
18. Sagiroglu, S. and Sinanc, D., Big Data: A Review, International Conference on Collaboration Technologies and Systems (CTS), pp.42-47, 20-24, May 2013.
19. https://www.predictiveanalyticstoday.com/what-is-predictive-analytics/
20. http://www.pentaho.com/product/big-data-analytics
21. Vamsee Kasavajhala, Solid state drive vs. hard disk drive price and performance study, Dell PowerVault Tech. Mark. (2012).
22. Shefali Singhal, Neha Garg: Web Page Representation Using Backtracking With Multi-Dimensional Database For Small Screen Terminals, in Studies Comp.Intelligence, Vol. 713, Brajendra Panda et al: INNOVATIONS IN COMPUTATIONAL INTELLIGENCE, 978-981-10-4554-7, 429081_1_En (21), July 2017.
23. Shefali Singhal, Neha Garg: Hybrid Web-page Segmentation and Block Extraction for Small Screen Terminals, In IJCA Proceedings on 4th International IT Summit Confluence 2013 - The Next Generation Information Technology Summit Confluence 2013(2):12-15, January 2014.
24. Tripathy, A., Agrawal, A., and Rath, S.K., 2015. Classification of Sentimental Reviews Using Machine Learning Techniques. Procedia Computer Science, 57, pp.821-829.
25. Elgendy, N. and Elragal, A., 2014. Big Data Analytics: A Literature Review Paper. In Industrial Conference on Data Mining, Springer, Cham, pp. 214-227.
26. Jha, A., Dave, M., & Supriya Madan, D. (2016). A Review on the Study and Analysis of Big Data using Data Mining Techniques (Vol. 6).
27. Garg, N. and Sharma, K., 2020. Machine Learning in Text Analysis. In Handbook of Research on Emerging Trends and Applications of Machine Learning (pp. 383-402). IGI Global.
28. Nabi, J. (2018). Machine Learning—Text Processing, https://towardsdatascience.com/machine-learning-text-processing-1d5a2d638958.
29. Tripathy, A., Agrawal, A., & Rath, S. K. (2015). Classification of Sentimental Reviews Using Machine Learning Techniques. Procedia Computer Science, 57, 821-829. Retrieved from http://www.sciencedirect.com/science/article/pii/S1877050915020529 . doi:https://doi.org/10.1016/j.procs.2015.07.523G.

AUTHORS PROFILE

Ms. Neha Garg is currently working as a Assisstant Professor, MRIIRS, Faridabad, India (around 10 years teaching experience), M. Tech from Banasthali Vidyapith, Near Niwai, Rajasthan and Ph. D. pursuing in Computer Science and Engineering from MRIIRS, Fraidabad, India. She has recently published a book "Analysis and Design of Algorithms- a Beginner Hope" with BPB Publication House. She has supervised and Guided research projects of B.Tech She have published research papers in field of Big data, and Data Mining. Her research interests are in the area of "Big Data Analytics" and "Machine Learning".

Dr. Kamlesh Sharma is currently working as a Associate Professor, MRIIRS, Faridabad, India (more than 14 years teaching experience). MCA, M. Tech from MDU University and Ph. D. in Computer Science and Engineering from Lingaya`s Vidyapeeth, India. is currently Supervising five Ph. D. scholars. She has also supervised and guided research projects of M. Tech, B.Tech and application based projects for different competitions. She is also associated with four Govt. research projects in filed of health recommender system, IOT, Machine Learning, AI and NLP. She has published more than 45 research papers in field of NLP, IOT, Bigdata, Green Computing and Data Miningin reputed Journal (Web of Science, Scopus, UGC, Elsevier) and Conferences (ACM, IEEE). Her research area "Natural Language Processing" is based on innovative idea of reducing the mechanized efforts and adapting the software to Hindi dialect.