BIG DATA ANALYSIS: FOUNDATIONS, TECHNIQUES, AND RESOURCES

B.Rani
Assistant Professor, Department of Computer Science & Engineering, Malla Reddy (MR) deemed to be University, Hyderabad. Email id.: rani@mrec.ac.in
Vemula Nikitha
Assistant Professor, Department of Computer Science & Engineering, Malla Reddy (MR) deemed to be University, Hyderabad. Email id.: nikitha479@gmail.com
Dikshendra Daulat Sarpate
Professor, Department of Artificial Intelligence & Data Science, ZEAL College of Engineering & Research, Pune, Email id -dikshendra@gmail.com
B. Sankaraiah
Assistant Professor, Department of Computer Science & Engineering, Malla Reddy (MR) deemed to be University, Hyderabad. Email id.: Shankar61186@gmail.com
Dr. Syed Umar
Professor, Department of Computer Science & Engineering, Malla Reddy (MR) deemed to be University, Hyderabad. Email id.: syedumar@mrec.ac.in

Abstract— The revolution in technology is, in many respects, contributing to the rapid increase in the amount of data that is being generated and captured simultaneously. As a consequence of this, the term "big data" is today the most widely used keyword in each and every research and engineering field. Small data is comparable to big data, with the exception that the volume of big data is substantially higher. The term "big data" refers to information that is exceeding a certain threshold in terms of either petabytes or terabytes in size. While it is realistic to anticipate that technological advancements will continue, it is also reasonable to anticipate that the size of data sets, which will be referred to as Big Data, will also continue to rise. On a daily basis, the amount of data is increasing at an exponential rate, and as a consequence, the amount of data that is currently available has beyond the capacity of traditional databases to store it. Big Data platforms, such as Hadoop, make it possible to manage and analyze massive datasets that are not compatible with traditional databases. This is made possible by the way Hadoop works. This study's objective is to investigate the introduction, characteristics, and multiple instruments that are utilized in the process of managing and analyzing massive amounts of data. The storage, processing, and analysis of enormous volumes of data has become a new challenge as a result of the rapid rise of data that has occurred as a result of the development of social networks and cloud computing. The development of a big data platform is required since traditional technologies are no longer enough for the processing of massive amounts of data. It is indisputable that big data platforms provide users with assistance in the development of analytical services in an effective manner. Nevertheless, the procedure of data collection, the creation of algorithms, and the provision of analytics services all require time.

Index Terms—Big Data, Big data environment, Velocity, Traditional Data, Sources, Hadoop, Enterprise electronic documents, HDFS.

I. INTRODUCTION

Big data refers to a dataset that is so vast that it is difficult to store, analyze, and administer using standard database software systems. An assertion that Big Data is larger than a particular number of petabytes or terabytes is not something that we are able to make. It is our expectation that the size of data sets that are regarded to be big data will increase in tandem with the progression and development of technology. When it comes to data volume and processing velocity, big data can also be defined as a continual shift or expansion in both of these measurements. Hadoop is used by a variety of organisations, including Facebook, Twitter, Linkedin, Google, and retail websites like Amazon, amongst others, to manage enormous amounts of data. The use of big data can be found in the manufacturing industry, healthcare, and public or private clouds [1]. In hospitals, big data analysts have the potential to save lives, improve care, and save costs by identifying patterns and trends, as well as discovering correlations and linkages between patients and their conditions. It is necessary to utilize the term "Big Data" in order to handle the data that is already available because it is currently expanding and changing at an alarming rate. In order to derive useful information from this ever-expanding and ever-changing data, it is necessary to employ a variety of processing strategies [2].

Big data is described as "data of extremely large scale, generally to the point where its modification and handling offer substantial logistical issues" [3]. This definition comes from the Oxford English Dictionary (OED), which is a reference to the English language.

"Statisticians have been dealing with massive volumes of data in domains as diverse as astronomy, genomics, and data mining [4] for years," Supriya (2022) states in his article. "Big Data" is a rather unusual phenomenon due to the fact that data is generated on a vast scale by numerous online interactions between individuals, transactions between individuals and systems, and devices that are equipped with sensors' capabilities.

This is according to Naveen (2024): - "I define Big Data as un-sampled data that is distinguished by the building of databases from electronic sources with a main objective other than statistical inference" [5].

Table 1: shows the distinctions between Big Data and Traditional Data

S. No.	**Parameter**	**Traditional Data**	**Big Data**
1	Size	GB	Continually increasing
2	Data production rate	Per hour, day	Rapid increase
3	Data type	Structured	Semi or unstructured
4	Approach	Centralized	Distributed
5	Data model	Fixed Schema	Schema less
6	Data store	RDBMS	Hadoop, NoSQL

II. 3V'S OR 3V MODEL OF BIG DATA

It was in the year 2001 that Doug Laney, an industry analyst, first presented the concepts of volume, diversity, and velocity [6]. For the purpose of defining Big Data, these three variables are utilized. On the basis of these V's, the 3V model of Big Data was developed.

The Big Data system receives data input from a variety of sources, including social networks, bank transactions, the content of web pages, GPS traces, financial data, and other sources of data output. Does it even make sense to assume that all of these are the same thing?

When attempting to make sense of the volume of Big Data, it is usual practice to characterize the various types of Big Data using the terms variety and velocity.

Through the course of the 19th century, the characteristics of the data continued to expand on a daily basis. This indicates that the quantity of v's is likewise fast increasing. Three distinct types can be distinguished amongst data properties when viewed from the perspective of the application domain. data domain, business intelligence domain, and statistics domain are the three domains in question.

When it comes to the data domain, searching for suitable patterns is included. Because of this, the Gartner 3 V's fall into this group [7]. The volume, the velocity, and the variety are the three v's. In this case, volume is the dominant vector. This refers to the size of the data set.

Additional three characteristics of data are introduced to the business intelligence domain. These characteristics are value, visibility, and verdict. When it comes to the application level, data scientists are required to view the necessary data. When seen from the perspective of business intelligence (BI), visibility offers insight, hindsight, and foresight regarding a problem and the corresponding solution to that problem. Value refers not just to the value of data but also to the value of business intelligence (BI) for the purpose of issue solving. A verdict is a potential or possible choice or decision that should be made by the decision maker as a result of the breadth of the problem, the resources that are accessible, and the computational capability that is available.

An additional set of v's is introduced in the area of statistics, and it is based on problems that are associated with statistics. Validity, truthfulness, and variability are the three components. veracity is the process of determining whether or not a body of data is trustworthy and certain [8]. Verifying that the quality of big data is logically significant is what is meant by the term "validity." The intricacy of the data and the variety both have a connotation of variability.

It is possible to create a combine venn diagram and determine the relationship between the three aspects of the data once all 3^2 v's characteristics have been specified from the three aspects of the data. The diagram that follows provides an illustration of all the different facets of big data.

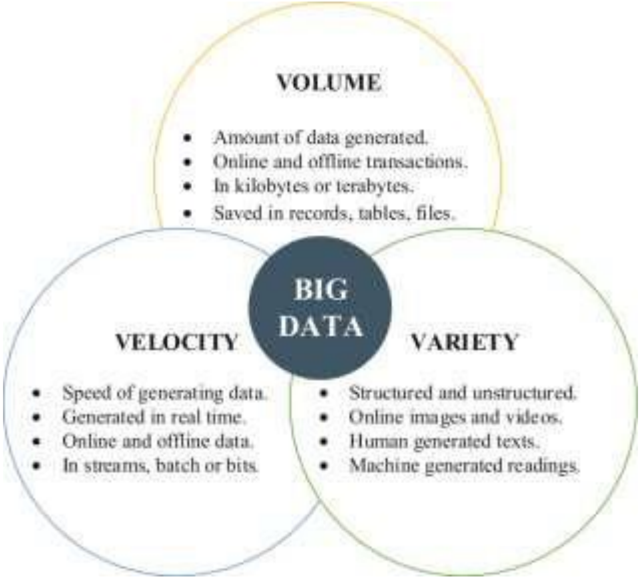
VOLUME
Amount of data generated.
Online and offline transactions.
In kilobytes or terabytes.
Saved in records, tables, files.
BIG DATA
VELOCITY
Speed of generating data.
Generated in real time.
Online and offline data.
In streams, batch or bits.
VARIETY
Structured and unstructured.
Online images and videos.
Human generated texts.
Machine generated readings.

Fig 1: The 3V's of Big Data

A. Volume:-

Volume refers to size of data which is increasing day by day typically starting at tens of terabytes. As an estimation2.5 Quintillian data is generated every day. Facebook generates 500+ terabytes data daily. Twitter generates 8TB data per day. To handle such amount of increasing data Big Data term is used.

B. Velocity:-

Velocity refers to speed at which data is being generated and changes. Today data is created and processed rapidly.

C. Variety:-

This defines the fact that data came from different sources in different formats and structures with different data types. These data types can be structured, semi structured and unstructured. Big Data consists of semi structured or unstructured data.

D. Different Types of Data:

Structured Data:

Structured data is the kind of data used by traditional database software where data is stored in defined relations or tables which can be easy to search, store, and retrieve based upon some conditions and rules [8]. Structured data is handled by SQL commands. Structured data of class students is shown in figure 2.

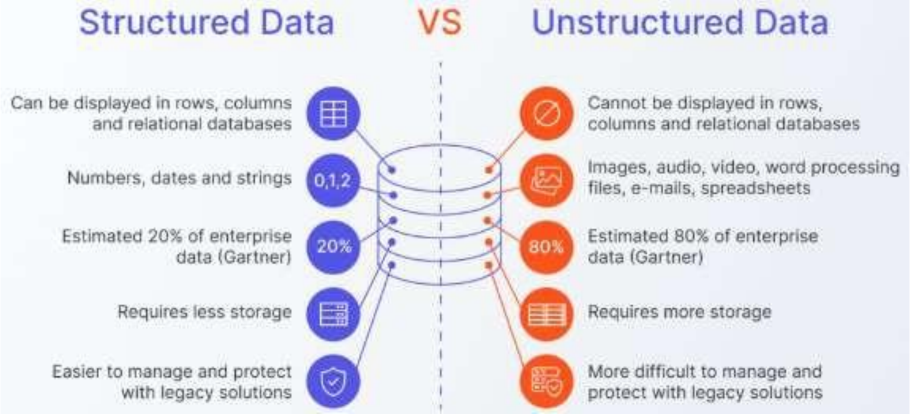
Structured Data
VS
Unstructured Data
Can be displayed in rows, columns and relational databases
Numbers, dates and strings
0,1,2
Estimated 20% of enterprise data (Gartner)
20%
Requires less storage
Easier to manage and protect with legacy solutions
Cannot be displayed in rows, columns and relational databases
Images, audio, video, word processing files, e-mails, spreadsheets
Estimated 80% of enterprise data (Gartner)
80%
Requires more storage
More difficult to manage and protect with legacy solutions

Fig 2: Structured and Unstructured Data [9]

Untructured Data:

This type of data has no predefined format. This data cannot be stored in databases or other architectures. This type of data can be textual or non textual. Data generated by media like e-mail messages, word documents falls under textual category.JPEG images, audio files are examples of non textual unstructured data.

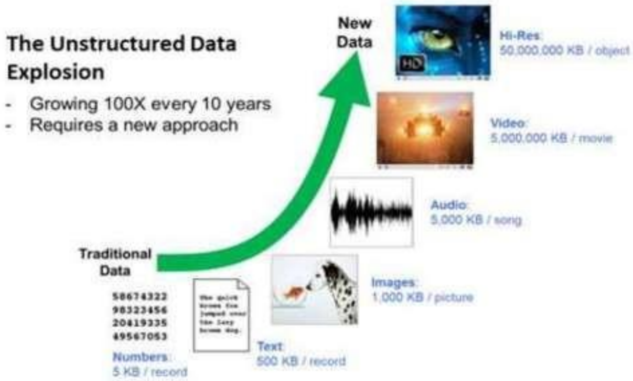
The Unstructured Data Explosion
- Growing 100X every 10 years
- Requires a new approach
New Data
HD
Hi-Res: 50,000,000 KB / object
Video: 5,000,000 KB / movie
Audio: 5,000 KB / song
Images: 1,000 KB / picture
Traditional Data
58674322
98323456
20419335
49567053
Numbers: 5 KB / record
Text: 500 KB / record

Fig 3: Unstructured Data [9]

Semi-structured Data:

This type of data is the combination of the two discussed above. This data cannot be organized in databases. It is structured data but cannot be stored in relational model. XML, EDI (Electronic Data Interchange),web server log and data collected by remote sensors are examples of semi structured data [9].

WHY BIG DATA REQUIRED

1. To increase storage capacities
2. To increase processing power
3. To store different types of data (structured, unstructured)
4. By a survey performed in 2012, more than 950 million users, Facebook ingested 500+ terabytes of new data into their database every day [10].
5. In every second, 6000 tweets are tweeted on twitter which corresponds to 350,000 tweets in one minute, 500 million tweets in one day and 200 billion tweets in one year. Based upon this estimation twitter generates 8TB data daily.
6. A survey performed by IBM shows that today's 90% of stored data was generated in previous two years [11].
7. A report says that Data from the U.S. healthcare system alone reached in 2011,150 exabytes [1]. At this rate of growth, Big Data for U.S.healthcare will soon reach zettabytes (1024 Exabytes) not longer after yottabytes(1024 Zettabytes).

aa8. In every second, 6000 tweets are tweeted on twitter which corresponds to 350,000 tweets in one minute, 500 million tweets in one day and 200 billion tweets in one year. Based upon this estimation twitter generastes 8TB data daily [12].

III. VRIOURS DATA GENERATION SOURCES

1. Data Collected From Sensors:

The next phase in the development of information technology is the analysis of sensor data. Companies in the manufacturing industry are incorporating sensors into their products and machinery in order to assess the production and quality of their work. When we take into consideration the data from sensors, it is expected that after a few years we will be required to communicate in Bronto Bytes. The data that sensors collect includes things like log data, geolocation data, temperature data, and utilization of the central processing unit [13]. When it comes to diversity, we can observe that remote sensing data is made up of several sources (such as lasers and radars), multiple temporal sources (data from various locations), and multiple geographical resolutions. When it comes to velocity, in the field of remote sensing, velocity does not only refer to the speed at which data is generated, but it also refers to the efficiency with which data is processed and analyzed. In other words, we can say that data should be analyzed in a reasonable amount of time in order to finish the task at hand. For instance, in the event of natural disasters such as earthquakes and tsunamis, a few seconds can save the lives of hundreds of thousands of people.

2. Point of Sale:

POS (Point of Sale) and inventory control systems include the following: The e-commerce sector in the United States accounts for approximately $400 billion, which is less than 8% of the total retail industry worldwide. In an e-commerce company such as Amazon, each and every search, purchase, and visit is mined, and the information is then used to improve the shopping experience for the customer.

3. Internet Websites:

User information can be most effectively disseminated through the usage of websites. In the world, there are around one billion websites, and the number of websites is growing at a rate of one billion each minute. Social networking services such as Facebook, Twitter, and Google Plus generate an enormous amount of data every single day, and this number is only going to continue to grow. Two hundred goods from an online store were purchased in a single second. The creation of data by online websites amounts to a total of 2.5 quintillion. That is a massive amount of material to examine.

4. Public, Private, Community Clouds:

Computing in the cloud is a network-based approach to resources that is on demand. SaaS is for software as a service, PaaS stands for platform as a service, and IaaS stands for infrastructure as a service. Cloud computing may be broken down into these three categories.

5. Data Generated By Government Agencies:

Public sector is large area of the global economy facing challenges to improve its productivity. Govt. has rights to approach this digital data but it is difficult for them to take advantage of this data in efficient way.

6. Bank / Credit Card Transactions:

Customer walk-ins, emails, internet banking, voice calls, social media, websites, and other forms of communication are all examples of data sources in banks. Big data is being adopted by HDFC, which is the first bank to do so. As a general rule, banks produce petabytes of data, and the banking industry is not an exception to this rule. The percentage of responders from each data source is displayed in Figure 4, which may be found here.

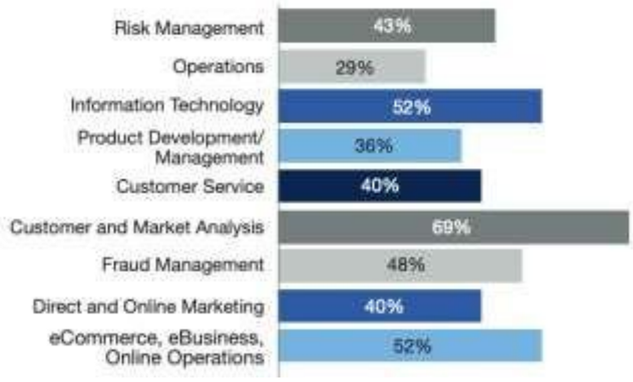

Fig 3: Percentage of Respondents [2]

IV. BIG DATA IN DIFFIRENT COMPANIES

Big Data at Google:
Not only has Google had a tremendous impact on the way in which we are now able to analyze big data (which includes Hadoop, MapReduce, and BigQuery, amongst others), but they are also more responsible than anybody else for making it a part of our everyday life. During the year 2007, Google introduced its universal search, which is capable of collecting information from a vast array of sources, such as weather forecasting, language databases, financial and historical information, and many more [14]. Knowledge graphs, which display information on a subject from a variety of sources and are included into search results, were made available in the year 2012. Google Query was introduced in 2010 with the purpose of processing, storing, and analyzing large amounts of data on cloud systems [15]. The self-driving cars that Google is working on as part of its Big Data initiative will collect and generate data from various sensors, cameras, and tracking devices, among other things.

Big Data at Amazon:
Amazon is the one of the Big Data generation company in these days. Amazon pioneered e-commerce in many ways, but possibly one of its greatest innovations was the personalized recommendation system which, of course, is built on the big data it gathers from its millions of customer transactions.

V. BIG DATA TECHNIQUES AND TOOLS

A) Hadoop: A software platform known as Hadoop is utilized for the purpose of running and writing applications that process massive data collections. The processing of enormous amounts of data can be done in a manner that is scalable, efficient, reliable, and cost-effective [16]. The Hadoop technology is being utilized by companies such as Yahoo, AOL, and Facebook. There are approximately 100,000 central processing units (CPUs) in over 40,000 servers that are running Hadoop, with the largest Hadoop cluster that Yahoo has operating 4,500 nodes. This is a significant amount of data, and it is about four times greater than the largest Hadoop cluster held by Facebook [17,18]. Big Data is broken down into a number of units via Hadoop, each of which is capable of being processed and analyzed simultaneously. A user-friendly function called MapReduce, which was invented by Google in the early 2000s and implemented using Hadoop Distributed File System (HDFS), is responsible for its implementation. Through the use of MapReduce, the task is broken down into smaller parts and then processed in parallel. There are two stages in the MapReduce process:

Map stage: Map’s responsibility is to process input data, divide into small units and assign to worker nodes. Worker node then do it again that corresponds to multi-level tree structure [19].

Reduce stage: Master node receives the answers from all sub units and add them to form output. It is the combination of shuffle and reduce stages basically.

B) HDFS (Hadoop Distributed File System): HDFS is a block structured distributed file system which follows client-server paradigm has Name Node and many Data Nodes, Name Node stores the meta data [20]. When there is a failure in Name Node then Hadoop cannot do automatic recovery.

C) NoSQL: A related new form of databases known as NoSQL (Not Only SQL) has arisen to process enormous volumes of multi-structured data, similar to how Hadoop does it. In contrast to the traditional method of storing and retrieving data in a database, NoSQL offers an alternative method for storing and retrieving data. In order to simplify the process of application creation, NoSQL Database offers drivers for Java, C, Python, and node.js, in addition to a REST API. Late in 2015, Oracle NoSQL Database 12.1.3.5.2 was made available to the public. Integration with Kerberos, which allows for external authentication, is one of the major features. This new application programming interface (API) enables the user to do Bulk Put operations for both rows and Key/Value inserts in a single API call when using the Bulk Put API [21].

VI. CONCLUSION

An acronym known as "Big Data" has been developed. In the not-too-distant future, every organization will be required to update their outdated databases with Big Data technologies in order to analyze and store large datasets. During the process of data analysis, there are a great deal of challenges to overcome. Big Data is utilized by a variety of businesses, including but not limited to Google, Facebook, Twitter, eBay, and others. The use of big data offers many benefits, including the reduction of costs and the acceleration and improvement of decision-making skills. Through the utilization of Big Data technologies, we might be able to gather the information that is the most pertinent and correct from a wide variety of independent sources in order to improve our comprehension and decision-making abilities. You will gain a better grasp of the principles of Big Data, as well as the reasons why it is necessary, as well as the tools and strategies that are utilized to manage and analyze Big Data, with the help of this research.
.

REFERENCES

[1] Batko, K., Ślęzak, A. The use of Big Data Analytics in healthcare. J Big Data, Springer Open, 9, 3 (2022).

[2] Darlan Arruda and Nazim H. Madhavji, The Role of Big Data Analytics in Corporate Decision-making, Scientific and Technology Publications, International Conference on Data Science, Technology and Applications,2017.

[3] Habeeb, M. S., & Babu, T. R. (2022). Network intrusion detection system: a survey on artificial intelligence-based techniques. Expert Systems, 39(9), e13066.

[4] RS Supriya Khaitan, Divya Rohatgi, Sana Nalband, Tejali Mhatre, Shweta Patil, "Enhancing Essay Grading Efficiency and Consistency through Two-Layer LSTM Models and Attention Mechanisms", Journal of Information Systems Engineering and Management 10 (2), 191-202.

[5] Naveen Sai Bommina, Uppu Lokesh, Nandipati Sai Akash, Dr. Hussain Syed, Dr. Syed Umar, "Optimizing AI-Driven Security Protocols in IoT Networks Using Metaheuristic Algorithms", International Journal of Intelligent Systems and Applications in Engineering, IJISAE, 2024, 12(23s), 3339–3347.

[6] Habeeb, M. S. (2024). Predictive analytics and cybersecurity. Intelligent Techniques for Predictive Data Analytics, 151-169.

[7] Jai Prakash Verma, Smita Agrawal, Bankim Patel and Atul Patel " Big data analytics: challenges and applications for text, audio, video, and social media data", International Journal on Soft Computing, Artificial Intelligence and Applications (IJSCAI), Vol.5, No.1, February 2016.

[8] Naveen Sai Bommina, Nandipati Sai Akash, Uppu Lokesh, Dr. Hussain Syed, Dr. Syed Umar, "A Hybrid Optimization Framework for Enhancing IoT Security via AI-based Anomaly Detection", International Journal on Recent and Innovation Trends in Computing and Communication, (2023) ISSN: 2321-8169 Volume: 11 Issue: 3.

[9] Lee I, Mangalaraj G. Big Data Analytics in Supply Chain Management: A Systematic Literature Review and Research Directions. Big Data and Cognitive Computing. 2022; 6(1):17

[10] Naveen Sai Bommina , Nandipati Sai Akash, Uppu Lokesh , Dr. Hussain Syed , Dr. Syed Umar, "Multi-Objective Genetic Algorithms for Secure Routing and Data Privacy in IoT Networks", International Journal of Communication Networks and Information Security (IJCNIS), (2020), 12(3), 632–643.

[11] M.S.Arun Kumar, R.S.Soundariya, M.Nivaashini, P.S.Dinesh, S.Iniya Shree,

[12] Applications of Big Data Analytics In Healthcare: A Research Perspective, International Journal Of Scientific & Technology Research Volume 9, Issue 02, February 2020

[13] Naveen Sai Bommina , Nandipati Sai Akash, Uppu Lokesh , Dr. Hussain Syed , Dr. Syed Umar, "Privacy-Preserving Federated Learning for IoT Devices with Secure Model Optimization", International Journal of Communication Networks and Information Security (IJCNIS), (2021), 13(2), 396–405.

[14] Ahmad, Z., Khan, A. S., Aqeel, S., Julaihi, A. A., Tarmizi, S., Annuar, N., & Habeeb, M. S. (2022, May). S-ADS: spectrogram image-based anomaly detection system for IoT networks. In 2022 Applied Informatics International Conference (AiIC) (pp. 105-110). IEEE.

[15] K Sankar, Divya Rohatgi, S Balakrishna Reddy, "COX Regressive Winsorized Correlated Convolutional Deep Belief Boltzmann Network for Covid-19 Prediction with Big Data", Grenze International Journal of Engineering & Technology (GIJET), Grenze ID: 01.GIJET.9.1.547, © Grenze Scientific Society, 2023.

[16] R. K. Chawda and G. Thakur, "Big data and advanced analytics tools," 2016 Symposium on Colossal Data Analysis and Networking (CDAN), Indore, India, 2016, pp. 1-8, doi: 10.1109/CDAN.2016.7570890. 7. S. Demigha, "The impact of Big Data on AI," 2020 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA, 2020, pp. 1395-1400.

[17] Divya Rohatgi, Dr. Tulika Pandey, "Regression Test Selection Framework for Web Services", INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH VOLUME 9, ISSUE 03, MARCH 2020.

[18] Umar, Syed, Bommina Naveen Sai, Nagineni Sai Lasya,Doppalapudi Asutosh, and LohithaRani. "Machine Learning based Sentiment Analysis of Product Reviews Using DeepEmbedding." Journal of Optoelectronics Laser 41, no. 6(2022): 108-113.

[19] R. Gnanakumaran, Divya Rohatgi, A K Sampath, Nidhi Nagar, D. Amuthaguka, Raj Kumar Gupta, "Robust Extreme Learning Machine based Sentiment Analysis and Classification", 2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT), (2023), DOI: 10.1109/ICSSIT55814.2023.10061017.

[20] Naveen Sai Bommina, Uppu Lokesh, Nandipati Sai Akash, Dr. Hussain Syed, Dr. Syed Umar, "Optimized AI Models for Real-Time Cyberattack Detection in Smart Homes and Cities", International Journal of Applied Engineering & Technology, Vol. 4 No.1, June, 2022.

[21] Zhi-Hua Zhou, Nitesh V. Chawla, Yaochu Jin, and Graham J. Williams, Big Data Opportunities and Challenges: Discussions from Data Analytics Perspectives IEEE Computational Intelligence Magazine, Vol. 20, NO. 10, 2020.