BIG DATA: DATA MINING METHODS, APPLICATIONS AND BEYOND

Dr.Pulipati Nageswar, Assistant Professor in computer Applications,
Govt.City College, Nayapul, Hyderabad,
tejanagesh@gmail.com

Abstract:-

Big data has recently become commonplace in a wide range of research fields. Big data is defined as "datasets that are larger than the capacity of traditional database software program equipment to capture, store, manage, and analyze". Nowadays, large data sets enable educational institutions to conduct organizational analytics and carry out new business intelligence using a learning control system. This data visualization allows you to assess overall performance indicators in coaching, management, and research. Actual-time analytics provide the ability to tune people and provide interventions to improve learning by reshaping and personalizing learning experiences. The education sector has faced several challenging situations in terms of coaching effectiveness, student acquisition and retention, and ineffectiveness in storing, processing, or studying data. The goal of this study is to examine the challenges of implementing big data technology (BDT) in higher education institutions (HEIs). This study provides a foundation for future research and highlights new insights and guidelines for the successfuluse of big data in education.

Keywords:- Big Data Technology, Higher Education Institutions, Big Data Analytics, Data Mining

Introduction

The inquiry that tends to make a conclusion stash away aspects and patterns is important to business. Overtime, big data investigation aids business visionaries by physically investigating the knowledge in order to create beneficial examples that are sought out. The foundations, strategy developers, educationalists, overseers, and pupils all have various freedoms, according to large data analysis. The chances include improved information flow and learning success throughout an organization, cross-joint effort over the foundations becomes acceptable and learning viability is improved, cost reduction by coordinating monetary execution becomes feasible, and scholastic risk is reduced. Enormous measurements are an essential part of advancement, which has, nowadays, won fundamental interest from all educators and professionals.

When we refer to Big Data, we mean the combination of structured, semi-structured, and unstructured data collected by Organizations and used in various projects in combinations with predictive modeling tools and advanced Big Data analytics applications. The classifications of data referred above are very important to understand due to the rapid increase of semi-structured and unstructured data nowadays on the one hand, and the advanced development of tools that make managing and analyzing these classes of data on the other hand.

Structured Data:-Structured data can be created by machines and humans having a pre-defined (fixed) data model, format, structure where a database designer can create in a way that entities can be grouped together to form relations. This makes structured data easy to store, analyze and search.A relational database is a representative example of structured data where tables are linked together using unique IDs and query language to interact with the data. Today the estimated amount of structured data accounts for less than 20 percent of all data whereas a much bigger percentage of all the data is unstructured data in our world.

Unstructured data. –The unstructured data has no inherent structure, cannot be contained in a row-column database and doe does not have an associated data model. The unstructured data is usually stored as different types of files for instance text documents, PDFs, photos, videos, audio files, social media content, satellite imagery, websites, and call center transcripts/recordings. Compared to structure datawere stored in spreadsheets or relational databases the unstructured data is usually stored in NoSQL databases, applications, and data warehouses. Plethora of information in unstructured data can be automatically processed with artificial intelligence algorithms today.

Semi-structured data. –The semi-structured data basically is a mix between structure and unstructured data, has some defining or consistent characteristics with some structure but does not conform to a data model. The semi-structured data lacks a fixed or rigid schema, cannot be stored in a form of rows and columns in Databases but contain tags and elements in the form of Metadata which is used to group data and describe how the data is stored. Examples of semi-structured Data sources are the E-mails, XML and other markup languages, binary executable, TCP/IP packets, Zipped files, and Web pages

Thinking about the meaning of the preparation area, the state-of-the-art propensity is moving nearer to breaking down the situation of enormous measurements on this area. Up until this point, many examinations have been performed to comprehend the utility of enormous measurements in exceptional fields for assorted purposes. Nonetheless, a total evaluation remains missing in enormous measurements in preparing. Subsequently, this investigation focuses on conducting a logical evaluations on enormous measurements to prepare you to find the patterns, arrange the examinations subjects, and spotlight the limitations and suitablepropositions to determine guidelines inside the space. The utilization of large information in education will be inspected in this review. Likewise, how much information can be utilized and separated to make something helpful will be inspected, helping the business to expand their incomes.

BIG DATA DIMENSIONS

The concept of Big Data gained momentum in the early 2000s where Gartner analyst Doug Laney articulated the definition of Big Data analyzing the Volume, Velocity and Variety dimensions the so called three (Vs). According to that, there are three significant dimensions of the Big data–Figure 1.

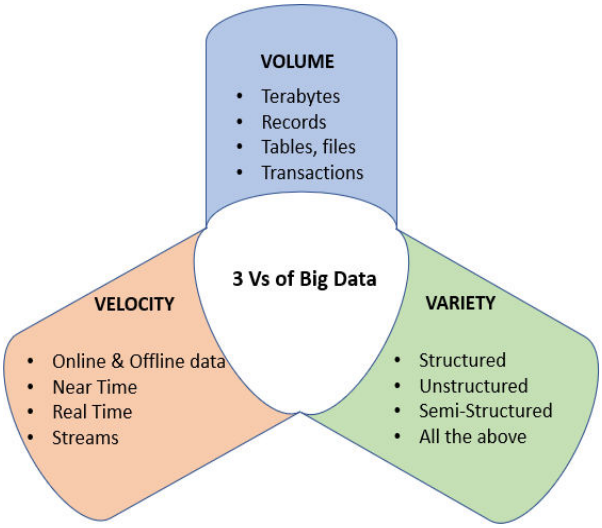

Figure1: The 3 (Vs) of Big Data

Nowadays, we all know that Big Data has penetrated in every industry accepting that is a prevailing driving force for every Organization to succeed across the globe. The –Big Data‖ as a terminology refers to huge and complex data that it is difficult to process them by using traditional methods compared to old fashioned data. It is fine whenever business is dealing with data using excel sheets and databases, however when the data cannot be fitted with such tools, then we think about Big Data and Analytics.

Volume. – When we refer to volume, we mean the size of huge amount of data setslying between terabytes and zettabytes, from variety of Sources. Sources such as business transactions, smart Internet of Things (IoT) sensor devices, social media, and other e-commerce platforms where get real- time, structured, and unstructured data.It is estimated that 2.5 quintillion bytes of data is created each day. According to McAffe and Brynjolfsson more data crosses the internet every second than the total amount of data stored online 20 years ago.

Velocity. – Broadly speaking Velocity refers to the speed of generating, processing, and analyzing the data. Nowadays, it is crucial for the Organization to have the information quickly as close to real-time more possible in the sense of paying much more importance to Velocity than to volume giving to Organizations bigger comparative advantage. The appropriate business decisions are strongly dependent to the data availability at the right time since after a couple of hours there may be useless under certain circumstances. For instance, in a machine learning service running in a social media platform with billions of users who post and upload messages or photos and videos, there is a continues transactions of petabytes of data that is being transferred from millions of devices. As we can understand the rate of the volume data that inflows per second is very high defining the velocity of the data. A representative example of data generated with such a high velocity will be Twitter messages and Facebooks posts. Another example of velocity is the sensor data with the Internet of Things (IoT) evolution where the connected sensors are taking off at a dramatic rate with data being transmitted at a near constant rate. Another example of velocity is the packet analysis for cybersecurity, where unfortunately threatening payloads can be hidden in a data flow passing through the firewall. Those data must be investigated and analyzed for patterns of suspect behavior and the situation is getting harder as more data is protected using encryption and the malware payloads are inside the encrypted packets.

Variety. – Variety refers to different data types of formats, namely, the diversity of data types and data sources, from structured numeric data stored in traditional databases to unstructured data types such as text documents, PDFs, photos, videos, audio files, social media content, XML and so on. This kind of heterogeneous data set possess a big challenge for big data analytics and requires distinct processing capabilities and specialist algorithms. A typical example of high variety of data sets would be the Closed- circuit television (CCTV) audio and video generated in a surveillance area in a city. More than 80 % of the data in the world today is unstructured and at first look does not show any clue of relationships. Moreover, when it comes to BigData, two additional dimensions are under consideration, the Veracity, and the Value, –Figure 2.

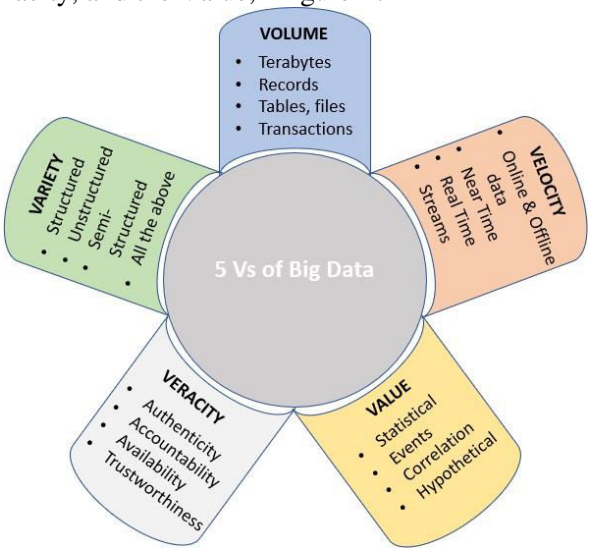

Figure2: The 5 (Vs) of Big Data

Veracity. – Veracity refers to the quality, the accuracy, and the reliability of the collected data since data comes from so many different sources. The first side of the Veracity in Big data it is not just the quality itself but how trustworthy are the data type, the data source considering abnormalities, inconsistencies, duplication as well. The second side of data veracity involves the processing method of the data and the adequate output to objectives based on business needs.

Value. – Value refers to an organization's ability to transform those huge amounts of data into real business since accurate data enables businesses as a steppingstone to get closer to their customer needs and expectations. Namely, Value denotes the *added value* for companies where huge

amounts of data (Volume) from highly diverse sources (Variety) with different quality (Validity) are used to quickly make vitalbusiness decisions to gain comparative advantage.

DATA MINING METHODS

When we refer to data mining, we mean the process of finding potentially useful patterns by using huge data sets. During this process, Machine Learning, Statistics, and Artificial Intelligent (AI) is used to extract information about the probability of future events.The diversified aspects of data mining comprise data classification, data integration, data transformation, data discretization, and pattern evaluation and more.Data mining techniques are used to discover hidden and unsuspected relationships amongst the data and used for marketing, sales, fraud detection, scientific discoveries,product development, healthcare, and education. Moreover, data mining techniques are used by the Organizations to solve business problems such as increasing revenues, acquiring new customers, improving cross-selling and up-selling, increasing Return of Investment (ROI) from marketing campaigns. As a result, the Organizations deliver consistent results that keep businesses ahead of the competition.

Association Rule Learning

In data science, the association rules technique is used to discover correlations between seemingly independent relational and transactional databases and datasets,and to observe frequently occurring patterns. The constraints on various measures of significance and interest are used, so that to select the suitable rules among the set of all possible rules.An association rule has always two parts, the antecedent (if) and a consequent (then) where an antecedent is something that is found in data, and a consequent is an item that is found in combination with the antecedent. The two primary patterns that association rules use is support andconfidence which are user defined measures of interestingness.

Support. –Support is the measure of how frequent an itemset appears in the dataset where for a given rule, *itemset* is the listof all the items in the antecedent and the consequent.

BIG DATA METHODS:-

Context

The context of the current study was virtual locationsin which we studied human action in social media. However, the language of the social media posts as a result of query words in Bahasa Indonesia located online activities taking place in this country. In other words, this reflected PE in Indonesia. In addition, the time frame of the study was from January to December 2020. Most of the school days during the year were dominated by distance learning due to the COVID-19 pandemic. Although the extracted posts did not always reflect what had happened during this time period, the condition of the pandemic became an important context in which the social media users engaged in online activities.

Data Collection

We began data collection with big data analytics by determining a set of query words in Bahasa Indonesia. The English translation included “physical education”, “health” and “physical education”, “PE”, and “HPE”. The set was also made of ordinary terms commonly associated with the subject such as “physed”, “gym class”, “sports learning”, and “physed/sports teachers”. Since the query words were in Bahasa Indonesia, the results of all extracted contents were also in the same language. Additionally, our query words also limited the location of the content posted to mostly being from Indonesia. Our initial query has extracted posts from the four major social media platforms.

Data Analysis

The data analysis took advantage of the machinery analysis performed by big data analytics. This included machine learning algorithms to extract the social media data and process the data through indexing and issue ranking. We carefully selected data that were produced from this simple processing such as the number of posts and engagement activities, location detections, and issue ranking. In other words, we did not take into consideration the complex analysis involving semantic processing, natural language detection, and relationship detection in which the validity and reliability have still been undetermined. Furthermore, the extraction data was broken down into visualization entities such as demographic information, activities, users' locations, topics, and engagements and then described narratively.

Results and Discussion

Results

The big data analytics produced data informing topics of physical PE from posts. The machine learning algorithm also ranked these topics based on the identification of the most discussed issues. Figure 1 displays the top topics regarding PE. Among these topics, the top three included PE teachers (47.6%), PE learning (16.6%), and students (8.2%). Other issues were shared in almost equal values (approximately 7%). It is noteworthy that these were the top topics and there might be other topics that were not displayed by the big data analytics because of their insignificant counts.

Who were the contributors and how they contributed to these representational topics? The big data analytics extracted posts being contributed by social media accounts. Some of these have been identified their gender with male contributors were slightly more than the females. Almost a half of them were married. They were also relatively young with more contributors being below 30 years old. Slightly more than half of the contributors had high school diplomas and the remaining half of them had higher education degrees, particularly some college diplomas. The demographic data ofsome contributors to the topic of PE.

By analyzing learner behavior, activities, and processes, as well as organizational and curricular procedures, workflows, and resources, the education industry has begun to embrace big data analytics for the creation of learning and academic activities. Let us look at the educational sector's big data analytics demands, prospects, and challenges.

Demand and Possibilities

Big data analytics has made it possible to improve students' learning outcomes and help them meet their academic objectives. Tests, exams, and other traditional means of evaluating student achievement are not required. The data trail of each student may be tracked and analyzed in real time to discover their strengths and weaknesses, as well as their average answer time for various types of questions and topic areas, questions skipped, and other academic skills. This information can be used by teachers and mentors to provide feedback, extra help, and tutelage to students who require it, as well as to establish an environmentin which children can thrive.

Examination has made it practical to further develop educational plans, showing procedures, and cycles. Instructors, tutors, and educational program originators can figure out what works and what doesn'tas far as educational programs, course materials, associations, and strategies and make changes depending on the situation. Big data analytics enables each student's program to be customized. Students have access to materials and training that are targeted to their specific learning levels and needs. The hybrid strategy used by MOOCs today, where online learning is self-paced and offline/online coaching from lecturers is also provided, enables for personalization even with tens of thousands of students.

Learning adequacy can be bettered through both managerial/educator level intercessions and self-estimationby students with the assistance of enormous information examination.

Cost decrease is conceivable through further developed adequacy and productivity of projects, eliminating study halls, using time more productively, lessening weakening, and so forth.

Cross coordinated effort and correlation among various organizations and courses should be possible effortlessly with the assistance of enormous information examination.

Challenges Existing

1) Guaranteeing the information stream is significant for large information investigation. Helpless web availability and ineffectively incorporated information frameworks make it hard to get to information and guarantee an information stream. It will be counterproductive if low quality and erroneously arranged information are utilized for instructive examination.

2) Instructing and preparing teachers are more major tasks and tedious tests for the utilization of large information to the schooling area. Indeed, even to get all instructors and tutors to collaborate and show energy are major achievements.

3) With such large benefits and freedoms to utilize huge information investigation, more establishments and associations are endeavoring to direct in front of difficulties and accepting it for accomplishing better results.

4) Huge information is the capacity, estimation, and investigation of enormous informational collections. These informational indexes are tremendous, to the point that it is beyond the realm of possibilities to expect to chip away at them utilizing customary information investigation apparatuses. The informational collections may likewise be unstructured in nature, which further makes them complex to manage utilizing conventional informational indexes. Nowadays, organizations are putting resources into enormous information to empower their association's dynamics and improve their proficiency. They enlisthuge information examination experts to chip away at huge informational indexes and will pay cutthroat compensations to them. These experts acquire their information examination confirmation before they canbe employed by associations.

5) Significant tech goliaths, like Google, Facebook, Amazon and so forth, utilize progressive, large information techniques and apparatuses to deal with their information and assemble applicable client bits of knowledge. They use it to upgrade the client experience on their sites and applications. The

unmanageable information of the past can now be able to be successfully overseen and perceived on accountof enormous information.

6.Major monetary organizations store tremendous volumes of information identified with their clients. They work on razor-thin flimsy edges and need to zero in on making of helpful bits of knowledge for better results. Large information can furnish them with greater benefit. McKinsey and associates say that large information examination is one of the main five impetuses that will drive work market development and will help the US economy constantly 2021.

CONCLUSION:-

Exploring the meaning of PE will be advanced from taking social media research into consideration.Borrowing the classic social constructionist framework, this study can be an important part of inquiries into the meaning of PE in which the members of a society conjointly construct what PE means to them and what assumes to be the reality of PE. Through social media platforms, people apparently co-constructed PE that was centralized around the topics about PE teachers, PE learning, and students. Further analysis also showed typical tones within the representation of PE such as physicality and masculinity. The current pandemic has also become a global context served as a backdrop in which PE has locally been implementing. One way or another, activities and engagement in social media platforms reflected this condition.

Technically speaking, the overall presentation of the results has been enabled by big data analytics. Surely, we deal with such a large dataset at the outset. Analytical features within big data processes have enabled data extraction and presentation informing the central topics, users' profiles, their locations, and times of engaging with social media. We would not be able to work with such enormous data in traditional ways of data processing if we might not have utilized big data analytics. When this became the channel for a qualitative inquiry, it can promise alternative perspectives toward what we have understood all this time, or even new knowledge and

understanding in educational literature.

First, let us say that big data, machine learning, and data science have a huge impact on education. This is a very fascinating theme! MOOCs are much more exciting in the context of MOOCs as they have the potential to provide free (or very affordable) education to everyone in the world. Improving automatic scoring may impact education. MOOCs are currently the subject of criticism because there is no actual assessment of how well students are absorbing the curriculum. With hundreds of thousands of students in a class, scoring becomes difficult. Peer grading solves this problem significantly. Peer grading can be very subjective, and machine learning helps to make grades less subjective by identifying hard / luxury marks. With protection and security issues, the utilization of huge information in instruction is on the ascent. Sincehuge information centers around digitizing information, there are no obstructions to handling, putting away,and getting to understudy learning information with regards to shielding it from being abused or manhandled.

Big data helps firms stay ahead of the competition by assisting in the enhancement of organizational procedures. Big data analytics is becoming a must-have expertise for each business. It serves as a significantpoint of distinction between firms and their competitors. Many firms are still catching up to big data analytics, which is still in its infancy. Organizations, on the other hand, have seen how it may help them outperform their competition.

This study demonstrates how big data affects the learning experience. Information protection, information security, ineffective options, and the inability to capture, access, or preserve information are just some of the challenges faced by educators.

Future Research

Big data, machine learning, and data science have influenced all industries, including education. The scopeof change in the education sector is currently limited, but the future looks bright! The basic goal of using data science and machine learning in education is to adapt learning so that both students and teachers can understand what needs to be done to improve the quality of education. Among the most significant problems in education today is that it is perfectly uniform. This standardization allows many subjects to be passed on to many future generations, but it is inadequate in two ways: People who do not fit the educationsystem perfectly; that is, many people cannot fully realize their potential. People do not actually realize their potential, but "just overcome it." Some children do not fit into a pre-formatted education system. Thereis no other problem. You are abandoned. They artificially lower the average when they cannot find a job that works independently. You can achieve a lot with the right data.

Future research should focus on theory-based precision instruction, cross-disciplinary application, and effective use of educational technology. The government should focus on encouraging lifelong learning, providing teacher education programs, and safeguard personal information. In order to improve academia- industry collaboration, reciprocal and mutually beneficial ties should be formed in the education industry.

REFERENCES

[1] Kirk D. (1992). *Defining physical education: The social construction of a school subject in postwar Britain.* London: Routledge Falmer.

[2] Hall S. (1997). The work of representation. In S. Hall (ed), *Representation: Cultural representations and signifying practices* (pp. 13-74). London: SAGE Publication Ltd.

[3] Hyndman B. P., Harvey S. (2020). Preservice teachers' perceptions of Twitter for health and physical education teacher education: A self- determination theoretical approach. *Journal of Teaching in Physical Education* 39, 472-480. DOI: https://doi.org/10.1123/jtpe.2019-0278

[4] Walton-Fisette J. L., Walton-Fisette T. A., Chase L.
F. (2017). Captured on film: A critical examination of representations of physical education at the movies. *Physical Education and Sport Pedagogy*22, 1-12. DOI: http://dx.doi.org/10.1080/17408989.2017.1294670

[5] McCullick B., Belcher D., Hardin B., Hardin, M. (2003). Butches, bullies and buffoons: images of physical education teachers in the movies. *Sport, Education, & Society* 8, 3-16. DOI:

http://dx.doi.org/10.1080/1357332032000050033

[6] Brook C., McMullen, J. M. (2020). Using social media: One physical education teacher's experience. *Journal of Teaching in Physical Education* 39, 464- 471. DOI: https://doi.org/10.1123/jtpe.2020-0005

[7] Goodyear V. A., Casey A., Kirk, D. (2014). Tweet me, message me, like me: using social media to facilitate pedagogical change within an emerging community of practice. *Sport, Education & Society* 19, http://dx.doi.org/10.1080/13573322.2013.858624

[8] Richard K A., Killian C. M., Kinder C. J., Badshah K., Cushing C. (2020). Twitter as a professionaldevelopment platform among U.S. physical education teachers. *Journal of Teaching in Physical Education* 39, 454-463. DOI: https://doi.org/10.1123/jtpe.2020-0001

[9] Bopp T., Vadeboncoeur J. D., Stellefson M, Weinz M. (2019). Moving beyond the gym: A content analysis of YouTube as an information resource for physical literacy. *Int. J. Environ. Res. Public Health*16, 1-18. DOI: https://doi.org/10.3390/ijerph16183335

[10] Goodyear V., Armour K. M., Wood H. (2018). Young people and their engagement with health-related social media: New perspectives. *Sport, Education, & Society* 24, 673-688. DOI: https://doi.org/10.1080/13573322.2017.1423464

[11] Quennerstedt M. (2013). PE on YouTube – investigating participation in physical education practice. *Physical Education and Sport Pedagogy* 18, 42-59. http://dx.doi.org/10.1080/17408989.2011.631000

[12] Harvey S., Pill S. (2018). Exploring physical education teachers 'everyday understandings' of physical literacy. *Sport, Education & Society* 24, 841-854. DOI: https://doi.org/10.1080/13573322.2018.1491002

[13] Kitchin R. (2017). Big data – Hype or revolution? In L. Sloan, A. Quan-Haase (eds), *The SAGE handbook of social media research methods* (pp. 27-39). Los Angeles: SAGE Publication Ltd.

[14] Campbell D., Gray S., Kelly J., Maclsaac S. (2018). Inclusive and exclusive masculinities in physical education: a Scottish case study. *Sport, Education,& Society* 23, 216-228. DOI:https://doi.org/10.1080/13573322.2016.1167680