International Journal of Education and Science Research Review
E-ISSN 2348-6457
Jan-Feb- 2019, Volume-6, Issue-1 P-ISSN 2349-1817
www.ijesrr.org Email- editor@ijesrr.org

CHALLENGES AND OPPORTUNITIES OF BIG DATA IN EDUCATION

Mrs. Vishakha Abhay Gaidhani,
Assistant Professor, MBA Department,
Sir Visvesvaraya Institute of Technology, Nashik, Maharashtra

Mr. Abhay Raghunath Gaidhani,
Assistant Professor, Computer Engineering Department,
Sandip Institute of Technology and Research Center, Nashik, Maharashtra

ABSTRACT

The data generated at an exponential rate has resulted in Big Data. This data has many characteristics and consists of structured, unstructured, and semi-structured data formats. The emergence of big data in educational contexts has led to new data-driven approaches to support informed decision making and efforts to improve educational effectiveness. Digital traces of student behavior promise more scalable and finer-grained understanding and support of learning processes, which were previously too costly to obtain with traditional data sources and methodologies. This synthetic review describes the affordances and applications of microlevel (e.g., clickstream data), macrolevel (e.g., text data), and macrolevel (e.g., institutional data) big data. Institutional data are often used to improve student and administrational decision making through course guidance systems and early-warning systems. Furthermore, this chapter outlines current challenges of accessing, analysing, and using big data. Such challenges include balancing data privacy and protection with data sharing and research, training researchers in educational data science methodologies, and navigating the tensions between explanation and prediction. We argue that addressing these challenges is worthwhile given the potential benefits of mining big data in education. Big data analytics if followed by big data analysis process plays a significant role in generating meaningful information from big data. Big data analysis process consists of data acquisition, data storage, data management, data analytics, and finally data visualization. However, it is not simple and brings many challenges that need to be resolved. This paper presents the challenges and opportunities related to big data, prominent characteristics of big data, big data analytics, big data analysis process, and technologies used for processing the massive data.

Keywords: — big data, big data analytics, big data processing, big data processing technologies, big data analysis

I. INTRODUCTION

The purpose of this position paper is to present current status, opportunities, and challenges of big data in education. The work has originated from the opinions and panel discussion minutes of an international conference on big data in education (The International Learning Sciences Forum, 2019), where prominent researchers and experts from different disciplines such as education, psychology, data science, AI, and cognitive neuroscience, etc., exchanged their knowledge and ideas. This article is organized as follows: we start with an overview of recent progress of big data and AI in education. Then we present the major challenges and emerging trends. Finally, based on our discussions of big data in education, conclusion and future scope are suggested. Rapid advancements in big data technologies have had a profound impact on all areas of human society including the economy, politics, science, and education. Thanks in large part to these developments, we are able to continue many of our social activities under the COVID- 19 pandemic. Digital

tools, platforms, applications, and the communications among people have generated vast amounts of data ('big data') across disparate locations. Big data technologies aim at harnessing the power of extensive data in real-time or otherwise. The characteristic attributes of big data are often referred to as the four V's. That is, volume (amount of data), variety (diversity of sources and types of data), velocity (speed of data transmission and generation), and veracity (the accuracy and trustworthiness of data). Recently, a 5th V was added, namely value. Because of intrinsic big data characteristics (the five Vs), large and complex datasets are impossible to process and utilize by using traditional data management techniques. Hence, novel and innovative computational technologies are required for the acquisition, storage, distribution, analysis, and management of big data.

Big data analytics commonly encompasses the processes of gathering, analyzing, and evaluating large datasets. Extraction of actionable knowledge and viable patterns from data are often viewed as the core benefits of the big data revolution. Big data analytics employ a variety of technologies and tools, such as statistical analysis, data mining, data visualization, text analytics, social network analysis, signal processing, and machine learning.
As a subset of AI, machine learning focuses on building computer systems that can learn from and adapt to data automatically without explicit programming. Machine learning algorithms can provide new insights, predictions, and solutions to customize the needs and circumstances of each individual. With the availability of large quantity and high-quality input training data, machine learning processes can achieve accurate results and facilitate informed decision making.

These data-intensive, machine learning methods are positioned at the intersection of big data and AI, and are capable of improving the services and productivity of education, as well as many other fields including commerce, science, and government.

Regarding education, our main area of interest here, the application of AI technologies can be traced back to approximately 50 years ago. The first Intelligent Tutoring System "SCHOLAR" was designed to support geography learning, and was capable of generating interactive responses to student statements. While the amount of data was relatively small at that time, it was comparable to the amount of data collected in other traditional educational and psychological studies. Research on AI in education over the past few decades has been dedicated to advancing intelligent computing technologies such as intelligent tutoring systems, robotic systems, and chatbots. With the breakthroughs in information technologies in the last decade, educational psychologists have had greater access to big data. Concretely speaking, social media (e.g., Facebook, Twitter), online learning environments [e.g., Massive Open Online Courses (MOOCs)], intelligent tutoring systems (e.g., AutoTutor), learning management systems (LMSs), sensors, and mobile devices are generating ever-growing amounts of dynamic and complex data containing students' personal records, physiological data, learning logs and activities, as well as their learning performance and outcomes. Learning analytics, described as "the measurement, collection, analysis, and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs", are often implemented to analyze these huge amounts of data. Machine learning and AI techniques further expand the capabilities of learning analytics. The essential information extracted from big data could be utilized to optimize learning, teaching, and administration. Hence, research on big data and AI is gaining increasing significance in education and psychology. Recently, the adoption of big data and AI in the psychology of learning and teaching has been trending as a novel method in cutting-edge educational research.

II. CHALLENGES

Though data mining offers numerous potential benefits for education research, there are also many challenges to be overcome to achieve those benefits. We summarize them below in three main areas: accessing, analyzing, and using big data.

ACCESSING BIG DATA

Educational data exist in a wide array of formats across an even wider variety of platforms. In almost all cases, these platforms were developed for other purposes, such as instruction or educational administration, rather than for research. Many commercial platform providers, such as educational software companies, have no interest in making their data available publicly. Other companies make their data available in a limited way but have not invested resources to facilitate access to data for research. Only a small number of platforms, such as Cognitive Tutor and Assessment's, have made high-quality data broadly available. By contrast, Google makes available the API (Application Programming Interface) of its widely used Google Docs program so that third-party companies can create extensions and other products that use or integrate with the software. It also allows users to view the history of their writing process in individual documents they have written or collaborated on down to 4-second increments; these documents can also be shared with others, who can also view those histories. The combination of open API and document history should, in theory, allow users to analyze metadata from large sets of writing data, for example, all documents written by students and teachers in a school district under a Google Docs site domain. In principle, though, writing the software to extract and analyze the data is a hugely complicated task. Some university and commercial groups have taken small steps in this direction, including the Hana Ohana research lab at University of California, Irvine, which has developed tools for analyzing collaboration history on individual Google Docs, and the private company Hapara (2019), which mines school district data for patterns related to time and amount of student writing, but these are very partial solutions to what largely remains an out-of-reach treasure of student writing data. In addition, even platforms that make their data available may require programming skills to extract the data. Though many education researchers are familiar with statistical software such as R or Stata, far fewer know programming languages superior for data extraction, such as Python.

Finally, and most important, the availability of data is complicated by privacy issues. Parents, educators, and others are rightly concerned about companies' ability to mine large amounts of sensitive student data and act in ways that are not necessarily focused on bettering individual students' futures. Fears have been raised that student data that are inappropriately shared or sold could be used to stereotype or profile children, contribute to tailored marketing campaigns, or lead to identity theft. Data privacy issues are exacerbated in K–12 settings, where students are children and participation in educational activities is mandatory.

Though the risks of sharing student data generate the most publicity, there are also risks to not sharing student data. Colorado has the strictest student data–sharing policies in the United States, according to the Parent Coalition for Student Privacy (2019). Yet data sharing is so strict that, according to the Right to Know (2019) coalition , the public is robbed of the information necessary to evaluate the performance of schools and educational programs in the state and their impact on diverse students.

Finding the right balance between individual privacy and the public interest is very challenging. This is, in part, because the large amount of data available in big data sets makes it very difficult to prevent the "reidentification" of de-identified data, even if all direct identifiers are removed. It is thus impossible to combine maximal privacy with maximal utility. Instead, educational institutions and researchers face a choice between maximizing privacy and limiting the utility of the data set or maximizing utility but leaving the data subject to possible reidentification with sufficient effort.

The challenges of sharing macrolevel data are even greater, since there is an unlimited number of ways in which students can reveal their identity in their writing. Addressing these challenges requires different kinds of strategies for different audiences and purposes. The U.S. Family Education and Privacy Act allows schools and institutions to share data with organizations conducting studies for the purpose of improving instruction.

Organizations such as the Inter-University Consortium for Political and Social Research host data sets with a wide range of restrictions. Data sets that favour utility (but sacrifice maximal privacy) can be made available to other research teams that are governed by institutional review board protocols, while data sets that limit utility but maximize privacy can be shared with the general public. Of course, even groups that are inclined to make data available for research may be hesitant to do so due to the extra steps and expenses required to ensure an appropriate level of de-identification.

ANALYZING BIG DATA

As with accessing big data, analysing big data also poses challenges regarding researchers' skills. As noted above, few education researchers know key programming languages used for data science, such as Python. Education research graduate programs seldom offer instruction in the data clustering, modelling, and prediction techniques used to analyse big data.

Even for researchers with such skills, error rates and noise pose additional challenges. For example, although predictive models can provide systematic improvements in prediction quality on average over base rates, high error rates may indicate the occurrence of significant exogenous factors at play not captured even in large amounts of data. When such predictive results facilitate the decision making of instructors or institutional policymakers, these errors may harm students' short-term learning or long-term success. In addition, large data sets with large numbers of predictor variables may result in models that are quite complex and difficult to interpret and that may not necessarily help stakeholders more than simpler models. This suggests that predicting student outcomes at a macro, "long time scale" level is inherently difficult and relationships between predictors and "downstream outcomes" can be complex, with many different factors affecting student outcomes that may potentially not be measured.

One way to mitigate these challenges is to combine macrolevel data with micro or macrolevel data. For instance, Aguiar et al. (2014) exemplified how non macro data can be useful in predicting student outcomes. The authors investigated different data sources for predicting student dropout of engineering courses at Notre Dame after their first term, treated as a binary classification problem. In terms of institutional (macro) data sources, the authors used predictor variables based on academic performance (i.e., SAT scores, first-term GPA) and demographics (i.e., gender, income group). Microlevel predictor variables included online student engagement during the first college term. The results were strikingly clear: Online engagement variables had significantly more predictive power than academic performance or demographic variables across a variety of classification models. Similarly, Miller found that predictive models constructed to predict learning outcomes for students taking undergraduate computer science courses could benefit significantly from including online student interaction data. These studies indicated that the addition of predictors based on noninstitutional data (e,g., online engagement data) can provide significant additional predictive power beyond that of institutional data alone.

USING BIG DATA

Finally, even if we successfully access and analyse big data, additional issues arise related to how such data are used. As education researchers increasingly turn to data mining, they will have to confront the tension between explanation and prediction. They argue that psychology's focus on explaining the causes of behaviour has led the field to be populated by research programs that provide intricate theories but have little ability to accurately predict future behaviours. They further suggest that increased focus on prediction using data mining and machine learning techniques can ultimately lead to a greater understanding of behaviour.

We also believe that this is true in education research, as seen the example of Assessment 2 Instruction (A2i) professional support system for reading instruction. Literacy research has been marked by the so called

reading wars between advocates of code-focused (e.g., phonics) versus meaning-focused (e.g., comprehension) instruction. Though a consensus has emerged over time on the critical value of the former, how much it should be supplemented by the latter is a continued debate. Connor's team tackled this issue in a highly creative way, adding a less-talked-about but also important question: Are elementary students best served by individualized (child managed) or whole-class (teacher managed) instruction?

The research team collected vast amounts of data on how much time children spent in (a) code- versus meaning-focused and (b) child- versus teacher-managed reading instruction, as well as (c) children's progression in reading proficiency throughout the year. Data mining techniques were used to develop and refine models indicating what combinations of instruction work best for children at different levels of proficiency and at different points in the school year. These models were developed into a software recommender system (A2i) that would assist teachers with grouping students to receive the types of instruction best suited to their needs. Randomized controlled trials were used to compare reading achievement in classrooms using A2i with that in classrooms teaching reading without it, finding strong positive effects for the former. This project thus not only built a valuable predictive tool that can guide teachers and improve literacy outcomes but also added explanatory value as to the differential contributions of code- versus meaning-focused and child- versus teacher-managed instruction.
Finally, in using big data, it is critically important to examine and address potential issues of bias, particularly when algorithms associated with big data lead to predictions and/or policy. For example, much attention has been focused on the potential for racial bias in predictive algorithms used in policing. The European Union Agency for Fundamental Rights (2018) provides a well justified set of recommendations for how to minimize bias in big data–derived algorithms. These include ensuring maximum transparency in the development of algorithms, conducting fundamental rights impact assessments to identify potential biases and abuses in the application of and output from algorithms, checking the quality of data collected and used, and ensuring that the development and operation of the algorithm can be meaningfully explained.

III. OPPORTUNITIES:

Big data provides an opportunity to HEIs to use them information technology resources strategically to improve educational quality and guide students to higher rates of completion, and to compare with student's persistence and outcomes. In the management department of any HEIs, there are large volumes of student data, including enlistment, employee and disciplinary records, and higher education institutions have the informational database expected to benefit from the targeted analytics.

Analyzing and managing such data is important as it can bring accountability and transparency in the education sector, help both learners and organizations to identify their achievements and areas of weaknesses and compare with other such organizations. Besides, the faculty and students can track their faculty members and behavioral progress through big data analytics. Other than that, big data analytics could also create opportunities like improving learning effectiveness through the self-measurement of learners and educators, cost reduction through managing financial performance which could be possible and improving the number of retention and graduation rates.

IV. CONCLUSION :

Big data analytics in advanced education can be transformative by adjusting to current procedures of the organization by adding the approach and practice results which can help in contemporary difficulties that are confronting education institutes. Education industry is no different than any other businesses that need to survive. Behind the student success goals, institutions conform to a system that values students getting good grades and shows continued progress towards finishing their studies to build revenue and to remain in business. Macrolevel data provide a deep window into cognitive processes by examining individuals'

writing, but they are prone to many of the broader challenges of using automated tools for writing measurement. Macrolevel data can be valuable for taking the broadest look at student persistence and achievement, but the smaller size and coarse measurements of macrolevel data sets may make it difficult to identify the fine-grained mechanisms at play. Without continued growth in enrolment, and students persisting on graduation, Education industry will struggle with funding. This is because education industry need to get their product to market: the graduating students.

V. REFERENCES :

[1] Allen, L., Mills, C., Jacovina, M. E., Crossley, S., D'Mello, S., & McNamara, D. S. (2016). Investigating boredom and engagement during writing using multiple sources of information: The essay, the writer, and keystrokes. In Proceedings of the Sixth International Conference on Learning Analytics and Knowledge (pp. 114–123). Association for Computing Machinery. https://doi.org/10.1145/2883851.2883939

[2] Atapattu, T., & Falkner, K. (2018). Impact of lecturer's discourse for student video interactions: Video learning analytics case study of MOOCs. Journal of Learning Analytics, 5(3), 182–197. https://doi.org/10.18608/jla.2018.53.12

[3] Baker, R. S., & Siemens, G. (2014). Educational data mining and learning analytics. In R. K. Sawyer (Ed.), Cambridge handbook of the learning sciences (2nd ed., pp. 253–274). Cambridge University Press.

[4] Baker, R., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. Journal of Educational Data Mining, 1(1), 3–17. https://doi.org/10.5281/ zenodo.3554657

[5] Bauer, A., Flatten, J., & Popović, Z. (2017). Analysis of problem-solving behavior in Openended scientific-discovery game challenges. In X. Hu, T. Barnes, A. Hershkovitz, & L. Paquette (Eds.), Proceedings of the 10th International Conference on Educational Data Mining (pp. 32–39), Wuhan, China. https://pdfs.semanticscholar.org/fa02/2ca5d9b1f53 64d5346ce0a6ee1cff0976840.pdf

[6] Clement, B., Roy, D., Oudeyer, P.-Y., & Lopes, M. (2015). Multi-armed bandits for Intelligent tutoring systems. Journal of Educational Data Mining, 7(2), 20–48. https://doi .org/10.5281/zenodo.3554667

[7] Daniel BK, Butson R. Technology Enhanced Analytics (TEA) in Higher Education. International Association for Development of the Information Society. 2013.

[8] Vatsala, Rutuja J, Athyaraj R. A review of big data in higher education sector. International Journal of Engineering Science. 2017; 7(6):25–32.

[9] Siemens G, Long P. Penetrating the fog: Analytics in learning and education. Educ. Rev. 2011; 46(5):30.

[10] Daniel B. Big data and analytics in higher education: Opportunities and challenges. British Journal of Educational Technology. 2015; 46(5):904–20.

[11] Shitut N. Five skill you need to know to become a big data analyst. Analytics Traininng. 2017. PMid:28392961 PMCid:PMC5374337