Research Review on Big Data Usage for Learning Analytics and Educational Data Mining: A Way Forward to Develop an Intelligent Automation System

Research Review on Big Data Usage for Learning Analytics and Educational Data Mining: A Way Forward to Develop an Intelligent Automation System

A Yunita[1,2], H B Santoso[1*], Z A Hasibuan[3]

[1] Faculty of Computer Science, Universitas Indonesia
[2] Faculty of Science and Computer Science, Universitas Pertamina
[3] Faculty of Computer Science, Universitas Dian Nuswantoro

harrybs@cs.ui.ac.id

Abstract. Digitalization and the development of information technology, especially Artificial Intelligence, have been embraced in all fields. At the same time, data has grown mostly from the digital footprints or any technology information system. The development of technology and big data offers enormous opportunities to conduct big data analytics in any field, including education. This study aims to review current research related to big data analytics in education and explain future research direction. Using Kitchenham's technique, we selected and clustered the literature into the types of data, methods, type of data analytics and learning analytics application used. The results show that research of big data learning analytics generally aims to improve the learning process, analyze learner behaviour for student profiling, improve student retention and evaluate student feedback in the context of MOOCs and Learning Management System. Several future directions for this topic are: 1) building a big open dataset including data pre-processing and addressing the problem of imbalanced dataset, 2) process mining for learning log activity to gain knowledge and insights from online behaviour, not only from the perspective of the learner but also from the activities of the teacher, 3) designing an automated framework which uses big data and allows descriptive, predictive, prescriptive analytical learning to be carried out. To summarize, embracing big data to learning analytics and educational data mining is an open research area that seems very powerful in education.

1. Introduction

In the past, it was thought that those who gained data and information would be the winners. However, today, with the development of technology, acquiring data and information is only the beginning. Data and information should be processed to the next stage, which is knowledge to predict, and the next step is presenting recommendations. To process data in advance, the expansion of technology, which can be seen as computing power, increased dramatically. Consequently, it has enabled the researcher to process data that are characterized as 5V (Volume, Variety, Veracity, Velocity, and Value), commonly known as big data [1].

The development of big data research has grown exponentially and is embraced in every field to support daily life activities until decision-making. Big data analytics has been implemented in every sector due to the ability to gain insights, behaviour trends, and profiling users. Like any other sectors, education also uses big data using data mining and machine learning approach although it is still in its

infancy. Two communities built around a shared interest in how big data can be used in education are Learning Analytics and Educational Data Mining.

Several review studies regarding big data in education have been made [2]–[4]. However, the process of gathering literature has not been explained systematically. This study reviews big data in education using the Systematic Literature Review (SLR) method proposed by Kitchenham et al. [5]. This study aims to present a broader picture of the use of big data for learning analytics and educational data mining.

The paper is organized as follows. Section two will discuss the research method, including the research objectives, keywords for the search, the database used, the search criteria, and the iteration process for review. The next section will provide the results and discussion. Furthermore, it will discuss the challenges, opportunities, and future directions for research on big data in learning analytics and educational data mining. The final section will present the conclusion, limitations, and future research.

2. Literature Review

This study follows the rule from Kitchenham for doing SLR [5]. According to Kitchenham et al. [5], there are three steps in conducting SLR: planning, conducting, and reporting. First, the objectives of doing literature review will be defined. This study's objective is to review previous works related to embracing big data for learning analytics and educational data mining from the perspective of empirical study. The database used in this study is the Scopus database because it covers peer-reviewed publications [6]. To start, we used the keywords "big data" and “education” which resulted in 3,866 papers. Next, we refined the keywords to be more specific in the field of study that is commonly called "learning analytics" and “educational data mining”. The keywords used for this research were "learning analytics" or "educational data mining", and "big data". We used Boolean operators like AND, OR in our search strings.

After defining the research objectives and keywords for the search, several criteria were determined in this study. The first criterion was that only peer-reviewed publications would be reviewed in this study. For the publication year and the type of papers, we only included proceeding papers and journal papers published on 2016 to 2020. Book chapters and lecture notes were rejected. Another criterion was only empirical papers published in English would be included in this study. Theoretical or conceptual papers without any evidence would be disregarded. If any duplications or the same papers were found in the search, one of them was eliminated.

The whole process can be seen in Figure 1. The first iteration was a search based on the selected keywords that resulted in 480 papers. After applying the inclusion-exclusion criteria, 309 papers were identified. The titles and abstracts of those 309 papers were reviewed based on the research objectives. Ninety-four papers were selected from the iteration. Papers that did not contain an explanation about related works were disregarded. Finally, 42 final documents were selected for this study.

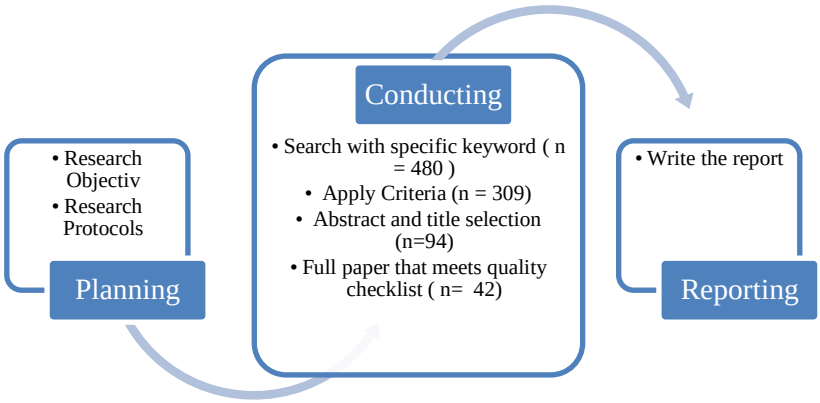

Figure 1. Research Phases [5]

3. Results and Discussion

This section shows SLR results from the 42 final papers whose general themes were classified based on the research context, the technique used, type of analytics, type of data, and research objectives. As shown in Table 1 and Figure 2, each theme was narrowed down into a sub-theme, and several papers reviewed were categorized into each sub-theme.

3.1. Research Context

First, related to the research context, as can be seen in Table 1, most of the literature of big data learning analytics researched was in the context of Massive Open Online Courses (MOOCs) and Learning Management System (LMS). For example, Elia et al. [7] evaluated student feedback from questionnaires using sentiment analysis in online collaborative learning, and Al-Shabandar et al. [8] has analyzed learner behaviours using datasets from online courses provided by Harvard University and Massachusetts Institute of Technology through the edX MOOCs platform. In line with that, Shi et al. [9] also analyzed learner behaviour in MOOC, but they used different sources, a MOOC offered from FutureLearn. Another study used a dataset derived from MOOCs and the e-book reading system [10].

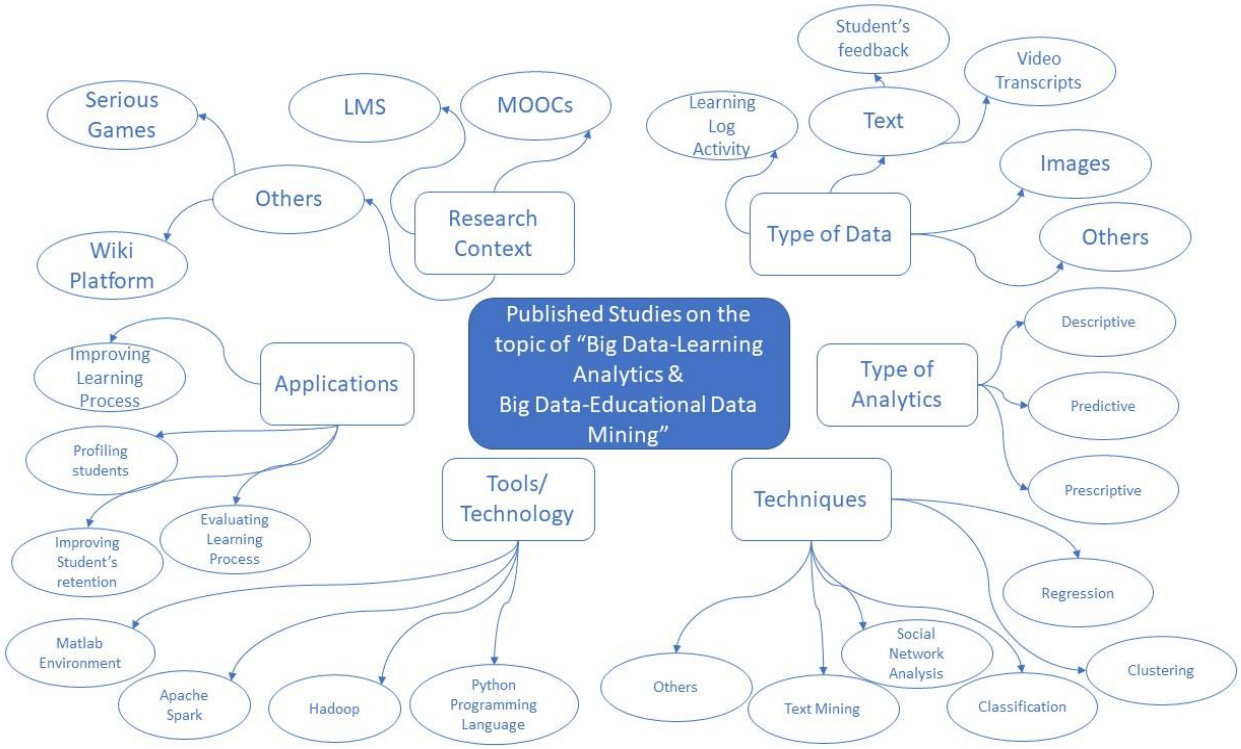

Figure 2. Components of the Literature

Table 1. List of Study Clustered by Theme

Theme	Sub-Theme	Reference
Research Context	MOOCs	[7], [8], [17]–[21], [9]–[16]
	LMS/e-Learning/VLE	[7], [22], [31], [32], [23]–[30].
	Others	[33], [34], [43]–[47], [35]–[42]
Type of Data	Learning log Data	[8], [9], [18], [19], [21], [23]–[29], [10], [31]–[33], [36], [37], [40], [41], [43], [47], [11]–[17]
	Text	[7], [20], [22], [38], [46]
	Images	[42]
	Others	[7], [19], [30], [33]–[35], [39], [44]–[46]
Type of Analytics	Descriptive	[7], [9], [41], [44]–[47], [19], [21], [22], [25]–[27], [29], [36]
	Predictive	[8], [10], [26], [28], [31]–[36], [38], [39], [11], [40], [42], [43], [46], [12], [15]–[18], [22], [23]
	Prescriptive / Recommendation	[39], [44]
Techniques	Regression	[9], [14], [18], [21], [41], [43], [45], [47]
	Association Rules	[26], [37]
	Clustering	[7], [9], [13], [27], [30], [35], [46]
	Classification	[7], [8], [31]–[33], [38], [40], [46], [10]–[12], [15]–[17], [23], [28]
	Text Mining	[7], [20], [22], [38]
	Social Network Analysis	[25], [26], [29], [35], [46]
	Deep Learning	[34], [42]

Theme	Sub-Theme	Reference
Applications	Speech-to-text	[12]
	Recommender Algorithm	[39]
	Improving Learning process	[12], [19]–[21], [37], [42], [44], [46], [47]
	Profiling students	[8], [9], [40], [41], [13], [24], [25], [27], [29], [30], [33], [36]
	Improving Student's retention	[10], [11], [32]–[34], [43], [15]–[18], [23], [26], [28], [31]
	Evaluating Learning Process / Feedback	[7], [22], [38], [45]
	Others	[35], [39]

3.2. Type of Data Used in Literature

The type of data used in the literature varies, as can be seen in Figure 3a. The two most common types of data used in the literature are learning log data and texts, with a percentage of 65% and 22%, respectively. Most of literature in this study used log data from online learning environment, as can be seen in Table 1, to analyze learner behaviour and predict student academic performance. Learning log activities might represent two types of data: behavioural and temporal data [8]. Behavioural data interpret how a learner interacts with the learning system, such as how many times the learner views the course module in a week, the number of posts in a discussion, the number of videos viewed, and the number of assignments submitted. Learning log data also can represent the temporal data from learners, such as how many unique days learners access the course, the first date, and the last date of accessing the course. Besides log data, several studies also combined behaviour data with demographic and academic transcripts [23], [26], [31].

Another type of data used in literature is text, mainly from video transcripts [12], [20], and opinions from students [22], [38]. A study conducted sentiment analysis to mine student views from social media and LMS [22], but the dataset is still unclear for readers due to the disclosure agreement. Another study also conducted opinion mining using Sentiment Analysis, but the data was retrieved from a student evaluation online survey of the Middle East College in Oman [38]. On the other hand, a study proposed the use of video transcripts combined with the video-based learner behaviour [20]. When the learner plays, pauses, stops the video, probably they do not understand the topic explained in the video, so they could clarify the specific content based on the results of text mining.

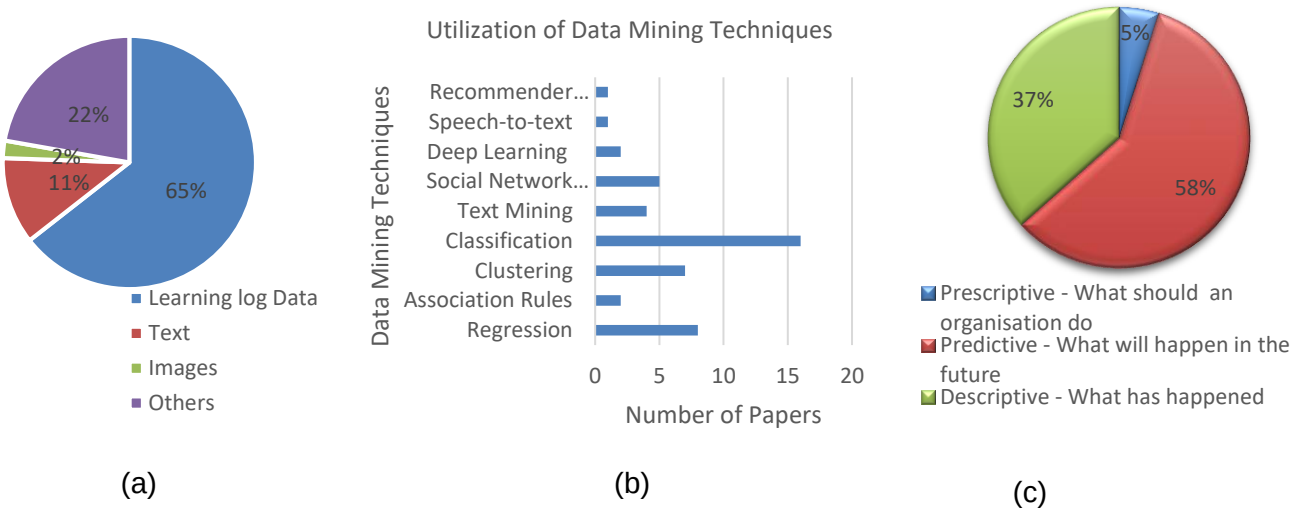

Figure 3: (a) Type of Data Used, (b) Utilization of Data Mining Techniques, (c) Type of Analytics Used in Percentage

Other types of data used in the literature are images [42], student profiles such as marks or grades [30], [35], [39], and student attendance [30]. Furthermore, a study portrayed student personal behaviour related to their internet access, meal habits, and previous academic records to predict academic performance [34]. Internet access includes how much internet flow charged, the length of time they were logged into the internet, and how long they spent accessing the internet during a class. Meanwhile, student meal consumption includes what time they eat breakfast when having a lecture and the average breakfast time [34].

The type of data for big data learning analytics varies, and is highly related to the research context, what type of analytics and algorithms were used in the research. While the research context was explained in the previous section, the next section will discuss whether descriptive, predictive, or prescriptive analysis were used and the type of data mining techniques used.

3.3. Type of Analytics and Data Mining Algorithms

There are at least three types of data analysis to process raw data, gain knowledge and insight, up to decision support. As can be seen in Figure 3c, more than half of the papers reviewed in this study implemented predictive analytics, 37% used descriptive analytics and only 5% implemented prescriptive analytics. Descriptive analysis is the basic type of data analysis that mainly describes what happened in the past and can be analyzed using data statistics, including mean, modus, and median. Several examples of descriptive analytics use inferential statistics to describe how students interact with learning objects [15] and use descriptive statistics and data visualization to describe online learners [29].

Furthermore, after descriptive analysis, predictive analysis can be conducted using machine learning techniques to train the model and predict what will happen in the future. For example, predicting academic performance [10], [18], [26], [31], [33], [34] , and student level of motivation [8]. The more advanced analysis is prescriptive to support strategic management for decision making and recommend what a stakeholder should do after acquiring knowledge about the past. We categorized [39], [44] into the category of studies that implement prescriptive analysis. Shabaninejad in [44] built an automated recommendation tool in Learning Analytics Dashboard, and Dwivedi and Roshni [39] built a recommendation model for students to choose elective courses. Both studies developed recommendation tools towards the prescriptive analysis, even though the prescriptions were still unclear.

Various data mining and machine learning techniques were employed to conduct descriptive, predictive and prescriptive analytics, such as regression [9], [14], [18], [21], [41], [43], [45], and

association rules [26], [37]. The frequency of each technique used by the literature is described in Figure 3b. The figure shows that classification is the most commonly used as classifier algorithms that purpose to predict.

For each data mining technique, existing algorithms are used in the literature. For example, Apriori algorithm is a classical algorithm [48] in data mining for mining frequent item sets and relevant association rules to find a dataset pattern. Mouri and Yin [37] employed the Apriori algorithm to detect the patterns from e-book logs to improve learning materials, such as finding small fonts and small pictures that instructors should improve. Another example is Multiple Linear Regression (MLR), which are a statistical technique to model the relationship between two or more explanatory variables. Yang et al. [18] proposed MLR for predicting academic performance combined with Principal Component Analysis to improve the accuracy using data from student behaviour and student grades.

Other data mining and machine learning techniques used in the articles reviewed are clustering and classification. Clustering is a self-organized learning algorithm that groups similar classes of objects. Among the reviewed articles, five out of seven articles used K-means for clustering [7], [9], [13], [27], [46]. In the meantime, classification is a supervised training algorithm in which there are pairs of inputs and outputs so that there is output knowledge of what will be produced. Several existing classifier algorithms were used in the literature, such as Naïve Bayes, KNN, Decision Tree, and Artificial Neural Network. Clustering and classification are often used all together, which means that in the beginning, the class is unknown, and clustering will be first conducted. This is commonly known as semi-supervised learning. Afterward, the clustering results will be the set of classes to be classified, such as in reference [7], [46]. An illustration of how clustering is used before classification can be seen in Figure 4.

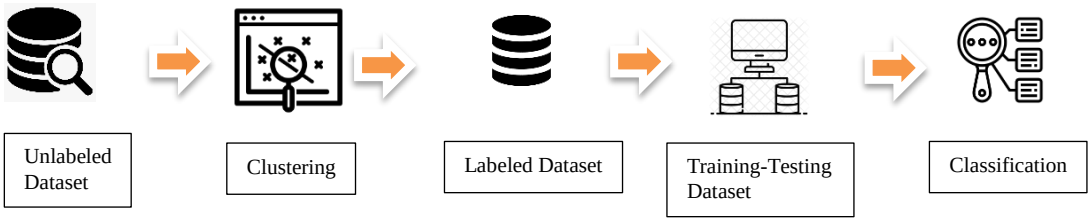

Figure 4. An illustration of semi-supervised learning using clustering and classification in a research

Besides those data mining algorithms, other Artificial Intelligence techniques were employed by several reviewed literature, such as sentiment analysis, speech to text, and Social Network Analysis (SNA). SNA is represented as a graph to show the relationship between nodes as can be found in several studies [25], [26], [29], [35], [46]. Sentiment analysis used for text mining of student feedback as conducted by Jena [22] and Dhanalakshmi, Bino and Saravanan [38].

Tools for analysing data mining vary. Python programming language has been used widely, such as [12], [17], [22], [36], [46]. Khousa and Atif [35] and Kausar et al. [30] used Matlab environment for their study, while [22], [24] used Hadoop. Another tool, Apache, was used by Elia et al. [7], [12].

4. Research on Big Data Learning Analytics

This section reviews the literature by comparing, contrasting, criticizing, summarizing, and synthesizing. A search of the literature published over a period of five years revealed 42 studies related to big data learning analytics and educational data mining, of which 19 were journaled papers. This section will review the published studies, especially the journaled papers grouped by the general research objectives.

4.1. Improving the Learning Process

Several published studies contributed to improving the learning process. Martin-Monje [15] attempted to investigate which learning objects engaged students most. This study found that video is the most

suitable for learners, both for male and female MOOC learners. Using inferential statistics, this study also determined what type of learners are the most prominent in MOOC. This study shows that 'viewers', who view the videos without doing any tasks or interaction are the most common. The insight gained from this study is that instructors should design the courses to be more interactive so that the learning process can be improved. This study has not embraced big data in the research, but they plan to use big data to improve learning and teaching, especially in Language MOOCs.

Similarly, Rienties [47] determined whether the learning design approach influences student engagement in Virtual Learning Engagement (VLE). Their findings have shown that 55% of the variance in weekly online participation in these four modules were explained by the manner in which language teachers planned the weekly learning design activities. This study attempted to reveal whether the instructor's design can predict student behaviour. Unfortunately, predictive modelling has not been included in this study.

Another study by Dessi [17], classified videos for learning by converting the audio to a transcript (text format) then classifying them into specific categories. This study proposes enhancing the learning process to understand pieces of knowledge through short videos. The result of the study shows that the accuracy is still under 75%. Nevertheless, it cannot be claimed that the study has poor performance because other studies should be analysed for comparison. Another study that improved the learning process but used different approaches from the other studies mentioned in the previous paragraph is Jin [34] which attempted to help a teacher teach how to draw using Generative Adversarial Network.

Synthesizing several published studies in improving the learning process, it was found that the published studies used different research methods and various aspects to help teachers to teach, to find the correlation between students and learning objects [21], [47], or to classify the learning materials [12].

4.2. Analyzing Learners' behaviour

For profiling learners, the literature used different approaches to cluster learners. An example of literature that analyzed student behaviour is a study conducted by Al-Shabandar et al. [8]. This study categorized MOOC learners based on the Incentive Motivation theory: amotivation, extrinsic and intrinsic. Thus, the educators might flag learners who are deficient in motivation and give them some intervention. Similarly, Shi et al. [9] revealed online learner patterns based on three parameters: the number of times they accessed the online course, the number of attempts to do the quiz, and the number of comments. They found that there are seven clusters based on behaviour. They also described which groups needed support and intervention.

Another study also discussed about clustered learner behaviour. Filvà et al. [36] categorized student clickstream behaviour but in a different context, i.e. in a virtual learning environment, then analyzed the correlation between behaviour and student grades. Other studies that also analyzed the behaviour of learners are [24], [29], [30], [33], [40], [41]. In contrast, Owen and Baker [45] and Liu, Li, Pan and Pan [46] analyzed learning behaviour in the context of games. Unlike others studies, Kausar et al. [30] focused on the algorithm for personalized learning and proposed a novel clustering approach. To synthesize, most studies aimed to detect groups that needed support and intervention using existing algorithms, such a K-means and Agglomerative Hierarchical Clustering algorithms [8], [9], [29], and only one study proposed a novel algorithm [30].

4.3. Improving Student Retention

The third benefit of big data learning analytics are improving student retention. All attempts to predict at-risk students [32], student grades [17], [26], academic performances [10], [18], [33], [43], drop-out [16], academic achievement [34] were classified into an attempt to improve student retention. Yang et al. [18] predicted student academic performance from learning activity datasets built from open EdX and Maple TA. Due to the strong correlation among the dataset features, Principal Component Analysis (PCA) was applied in this study before doing Multiple Linear Regression. However, the proposed model does not seem to be applicable to other courses with different learning activities.

Similarly, Huang et al. [10] uses big data in learning analytics to increase student retention using three different university datasets. However, several issues might arise when using big data from various sources, such as the possibility of a different grading policy in each university which might lead to bias. Another study used student learning logs from the Hellenic Open University to compute student drop-out probability [23]. Unfortunately, the system used was still not automated. In contrast, Qu, Li, Zhang and Wang [34] captured an entirely different type of data: student consumption time and web login activity using Layer-Supervised Multi-Layer Perceptron (LSMLP). To summarize, all studies attempted to predict student academic achievement by using big data in learning analytics with a different approach in terms of data, tools, and algorithms.

4.4. Evaluating the Learning Process

The fourth application of big data learning analytics evaluated the learning process that might be conducted by analyzing student feedback. A study examined forum discussions and LMS questionnaires using sentiment analysis to find out learner satisfaction [7]. Jena [22] and Dhanalakshmi, Bino and Saravanan [43] correspondingly conducted text mining of student opinions to evaluate the learning process. One study proposed a big data approach to do sentiment analysis [22], while another, [38], did not use big data to conduct opinion mining.

To synthesize, two studies [7], [22] proposed the real-time model that embraces big data to do sentiment analysis for analyzing student feedback. Other studies were not categorized into the four categories of applications, such as a study that developed a recommender system of elective courses for students using Collaborative Filtering [39] and another study that predicted careers for students to meet the industrial needs [35].

5. Future Directions for Research on Using Big Data Learning Analytics

After conducting a comprehensive literature review, we analyzed challenges and opportunities for shaping research direction in using big data for learning analytics and educational data mining. This section recommends several research directions, especially those related to building big datasets, learning processes, and automated systems for big data learning analytics and educational data mining.

First, with regard to the data or dataset, recent studies tend to capture all possible data from students, such as in [10], [22], [34], [47]. Another study attempted to collect several embedded learning sources to develop multidimensional data-learning behaviour from online and offline learning [47]. On the other hand, another study went beyond ordinary data, not learning behaviour, but meal habits and internet access [34]. It seems that there is an opportunity to collect all possible data from students and analyze learner behaviour using all possible student profiles.

However, building a large dataset from different sources might be challenging, for example, collecting data from different courses. The challenge is because each course is unique, and learners in one course have different activities from learners in other courses [10]. Therefore, it is suggested that the instructor optimizes the activities in each class, so that the dataset from each class will be equal. Another challenge is the fact that academic achievement mainly depends on teacher grading policy [10], and as a consequence, it might be challenging to build a dataset.

Another challenge of big datasets is imbalanced datasets; for example, predicting student academic achievement in which data from failing students is not equal to that of successful students. The problem of the imbalanced datasets will be faced when detecting at-risk students. Gkontzis et al. [23] suggested using a specific algorithm for this problem. They stated that MetacostSMO produced highest accuracy for an imbalanced dataset, while for a balanced dataset, Decision Tree is the best algorithm amongst KNN, Neural Network, and Naïve Bayes.

To summarize, the future direction for conducting research in big data learning analytics related to dataset is building open datasets, including data pre-processing and tackling the imbalanced dataset problem. Since existing datasets for learning analytics are still limited, this might be a future direction for researchers to contribute to building datasets and making them open and public. It is interesting to

note here that several well-known available datasets for learning analytics are Open University Learning Analytics (OULA) dataset [49], and MOOC datasets, such as Coursera and edX.

The second issue is regarding the process of mining log activity to portray the learning process. As explained in the previous section, learning log activity is the most used in the literature. Since learning process is complex, instead of analyzing learner's behaviour, one study suggests analyzing from another point of view to analyze the lecturer or instructor behaviour [24] to gain a broader picture of the learning process.

Besides the perspective of the dataset, several challenges about the learner behaviour are related to how to face student uniqueness, representing the unpredictability of human behaviour. Strang [50] stated that learning analytics failed to predict student academic achievement since empirical evidence shows that higher achiever students seem less active in viewing courses and assignments. However, it is important to note that this research in Learning Management System is not fully online learning. It might be different when predicting academic achievement, especially student retention in the context of MOOCs. Future direction to tackle this issue is by using micro-level fine-grained data [47], or in other words, to conduct process mining in learning log activities, such as mining student engagement, self-regulated learning strategies, learning styles, and other aspects based on the learning log activities.

Other future research directions are to develop an automated system for learning analytics. One study recommended future research to capture, analyze, and aggregate data in their system [35], thus, making it possible to implement this system for other purposes, such as developing an automated system to use big data in mining student opinions, predicting student academic achievement, and predicting student career paths.

Another recommendation is, as explained in the previous section, related to the limited literature on the application of prescriptive analytics. Therefore, another opportunity is to build a prescriptive model for students. The prescription should also apply to the lecturers enabling them to improve the learning process and strategic management that develops the policy. However, prescriptive analytics needs descriptive and predictive analytics. To summarize, future research directions lead to building automated systems that can automatically capture, analyze and aggregate big data, then produce descriptive, predictive, and prescriptive analysis for learning analytics objectives.

6. Conclusion and Future Research

The exponential growth of data and the increase of computation power has led to the penetration of big data into education, explicitly learning analytics and educational data mining research. Several research contexts, types of data, types of analytics, algorithms, and applications that were used in the previous research have been reviewed. Most of the studies related to MOOCs and LMS improve student retention using various data types, but the most common type of data is log data generated by the learning system. In general, the research objectives of the literature are to improve the learning process, analyze learner behaviour for student profiling, improve student retention and evaluate student feedback in the context of MOOCs and Learning Management System. Furthermore, various machine learning techniques have been employed for descriptive, predictive, and prescriptive analysis to analyze data. A significant number of studies used descriptive and predictive analytics, but only few used prescriptive analytics. For the more advanced SLR, it is recommended to include cited literature or cited by other related papers that might provide additional insights and a broader perspective. We recommend several research directions for this topic: 1) building big public datasets including data pre-processing and tackling the problem of imbalanced datasets, 2) process mining for learning log activities to gain knowledge and insights from the online activity, not only from the learner’s perspective but also the teacher’s activities, 3) develop an automated system that uses big data and enables conducting descriptive, predictive, prescriptive learning analytics. Our future research will explore the possibility of collecting all data from students to build a descriptive, predictive, and prescriptive system for improving student retention.

Acknowledgement

This research was supported by Hibah Publikasi Terindeks Internasional (PUTI) Prosiding 2020 at Universitas Indonesia (Number: NKB-846/UN2.RST/HKP.05.00/2020).

References

[1] Ishwarappa and Anuradha J, 2015 A brief introduction on big data 5Vs characteristics and hadoop technology *Procedia Comput. Sci.* 48, C p. 319–324.
[2] Sin K and Muthu L, 2015 Application of Big Data in Education Data Mining and Learning Analytics – a Literature Review *J. Soft Comput.* 5 4 p. 1035–1049.
[3] Daniel B K, 2019 Big Data and data science: A critical review of issues for educational research *Br. J. Educ. Technol.* 50, 1 p. 101–113.
[4] Ang K L-M Ge F L and Seng K P, 2020 Big Educational Data Analytics: Survey, Architecture and Challenges *IEEE Access* 8 p. 116392–116414.
[5] Kitchenham B Brereton O P Budgen D Turner M Bailey J and Linkman S, 2009 Systematic literature reviews in software engineering – A systematic literature review *Inf. Softw. Technol.* 51, 1 p. 7–15.
[6] Martín-Martín A Orduna-Malea E Thelwall M and Delgado López-Cózar E, 2018 Google Scholar, Web of Science, and Scopus: A systematic comparison of citations in 252 subject categories *J. Informetr.* 12 4 p. 1160–1177.
[7] Elia G Solazzo G Lorenzo G and Passiante G, 2019 Assessing learners' satisfaction in collaborative online courses through a big data approach *Comput. Human Behav.* 92 p. 589–599.
[8] Al-Shabandar R Hussain A J Liatsis P and Keight R, 2018 Analyzing Learners Behaviour in MOOCs: An Examination of Performance and Motivation Using a Data-Driven Approach *IEEE Access* 6 p. 73669–73685.
[9] Shi L Cristea A I Toda A M and Oliveira W, 2019 Revealing the Hidden patterns: A comparative study on profiling subpopulations of MOOC students in *Proceedings of the 28th International Conference on Information Systems Development: Information Systems Beyond 2020*
[10] Huang A Y Q Lu O H T Huang J C H Yin C J and Yang S J H, 2020 Predicting students' academic performance by using educational big data and learning analytics: evaluation of classification methods and learning logs *Interact. Learn. Environ.* 28 2 p. 206–230.
[11] Xi J Chen Y and Wang G, 2018 Design of a personalized massive open online course platform *Int. J. Emerg. Technol. Learn.* 13 4 p. 58–70.
[12] Dessì D Fenu G Marras M and Reforgiato Recupero D, 2019 Bridging learning analytics and Cognitive Computing for Big Data classification in micro-learning video collections *Comput. Human Behav.* 92 p. 468–477.
[13] Ocaña M Khosravi H and Bakharia A, 2019 Profiling language learners in the big data era in *Conference Proceedings - 36th International Conference of Innovation, Practice and Research in the Use of Educational Technologies in Tertiary Education: Personalised Learning. Diverse Goals. One Heart* p. 237–245.
[14] Kidziński Ł Sharma K Boroujeni M S and Dillenbourg P, 2016 On generalizability of MOOC models in *Proceedings of the 9th International Conference on Educational Data Mining* p. 406–411.
[15] Laveti R N Kuppili S Ch J Pal S N and Babu N S C, 2017 Implementation of learning analytics framework for MOOCs using state-of-The-Art in-memory computing in *Proceedings -5th National Conference on E-Learning and E-Learning Technologies,*.
[16] Liang J Yang J Wu Y Li C and Zheng L, 2016 Big data application in education: Dropout prediction in edx MOOCs in *Proceedings - 2nd International Conference on Multimedia Big Data, BigMM* p. 440–443.

[17] Lemay D J and Doleck T, 2020 Grade prediction of weekly assignments in MOOCS: mining video-viewing behaviour *Educ. Inf. Technol.* 25 2 p. 1333–1342.
[18] Yang S J H Lu O H T Huang A Y-Q Huang J C-H Ogata H and Lin A J Q, 2018 Predicting students' academic performance using multiple linear regression and principal component analysis *J. Inf. Process.* 26 p. 170–176.
[19] Shi L and Cristea A I, 2018 Demographic indicators influencing learning activities in MOOCs: Learning analytics of futurelearn courses in *Proceedings of the 27th International Conference on Information Systems Development: Designing Digitalization, ISD 2018.*
[20] Huang N-F *et al.*, 2017 VideoMark: A video-based learning analytic technique for MOOCs in *2017 IEEE 2nd International Conference on Big Data Analysis, ICBDA 2017* p. 753–757.
[21] Martín-Monje E Castrillo M D and Mañana-Rodríguez J, 2018 Understanding online interaction in language MOOCs through learning analytics *Comput. Assist. Lang. Learn.* 31 3 p. 251–272.
[22] Jena R K, 2019 Sentiment mining in a collaborative learning environment: capitalising on big data *Behav. Inf. Technol.* 38 9 p. 986–1001.
[23] Gkontzis A Kotsiantis S Panagiotakopoulos C and Verykios V, 2019 A predictive analytics framework as a countermeasure for attrition of students *Interact. Learn. Environ.*
[24] Cantabella M Martínez-España R Ayuso B Yáñez J A and Muñoz A, 2019 Analysis of student behaviour in learning management systems through a Big Data framework *Futur. Gener. Comput. Syst.* 90 p. 262–272.
[25] Zhang J-H and Zou Q, 2016 Group learning analysis and individual learning diagnosis from the perspective of Big Data in *Proceedings of 2016 IEEE International Conference on Cloud Computing and Big Data Analysis, ICCCBDA 2016* p. 15–21.
[26] Mouri K Okubo F Shimada A and Ogata H, 2016 Bayesian network for predicting students' final grade using e-book logs in university education in *Proceedings - IEEE 16th International Conference on Advanced Learning Technologies, ICALT 2016* p. 85–89.
[27] Islam O Siddiqui M and Aljohani N R, 2019 Identifying online profiles of distance learning students using data mining techniques in *ACM International Conference Proceeding Series* p. 115–120.
[28] Farokhmehr M and Fatemi S O, 2016 Implementing machine learning on a big data engine for e-learning in *Proceedings of the European Conference on e-Learning, ECEL* p. 188–193.
[29] Zhang J-H Zhang Y-X Zou Q and Huang S, 2018 What learning analytics tells us: Group behaviour analysis and individual learning diagnosis based on long-term and large-scale data *Educ. Technol. Soc.* 21 2 p. 245–258.
[30] Kausar S Huahu X Hussain I Wenhao Z and Zahid M, 2018 Integration of Data Mining Clustering Approach in the Personalized E-Learning System *IEEE Access* 6 p. 72724–72734.
[31] Waheed H Hassan S U Aljohani N R Hardman J Alelyani S and Nawaz R, 2020 Predicting academic performance of students from VLE big data using deep learning models *Comput. Human Behav.* 104 p. 106189.
[32] Kondo N Okubo M and Hatanaka T, 2017 Early Detection of At-Risk Students Using Machine Learning Based on LMS Log Data in *Proceedings - 6th IIAI International Congress on Advanced Applied Informatics,* p. 198–201.
[33] Buenaño-Fernández D Gil D and Luján-Mora S, 2019 Application of machine learning in predicting performance for computer engineering students: A case study *Sustain.* 11, 10.
[34] Qu S Li K Zhang S and Wang Y, 2018 Predicting Achievement of Students in Smart Campus *IEEE Access* 6 p. 60264–60273.
[35] Khousa E A and Atif Y, 2018 Social network analysis to influence career development *J. Ambient Intell. Humaniz. Comput.* 9 3 p. 601–616.
[36] Filvà D A Forment M A García-Peñalvo F J Escudero D F and Casañ M J, 2019 Clickstream for learning analytics to assess students' behaviour with Scratch *Futur. Gener. Comput. Syst.* 93 p. 673–686.

[37] Mouri K and Yin C, 2017 E-Book-Based Learning Analytics for Improving Learning Materials in *Proceedings - 2017 6th IIAI International Congress on Advanced Applied Informatics, IIAI-AAI 2017* p. 493–497.
[38] Dhanalakshmi V Bino D and Saravanan A M 2016 Opinion mining from student feedback data using supervised learning algorithms in *2016 3rd MEC International Conference on Big Data and Smart City* p. 332–336.
[39] Dwivedi S and Roshni V S K, 2017 Recommender system for big data in education in *Proceedings - 2017 5th National Conference on E-Learning and E-Learning Technologies, ELELTECH 2017*.
[40] Owen V E and Baker R S 2020 Fueling Prediction of Player Decisions: Foundations of Feature Engineering for Optimized Behaviour Modeling in Serious Games *Technol. Knowl. Learn.* 25, 2 p. 225–250.
[41] Liu M Li C Pan Z and Pan X 2019 Mining big data to help make informed decisions for designing effective digital educational games *Interact. Learn. Environ.*
[42] Jin Y Li P Wang W Zhang S Lin D and Yin C, 2019 GAN-based pencil drawing learning system for art education on large-scale image datasets with learning analytics *Interact. Learn. Environ.*
[43] Lu O H T Huang A Y Q Huang J C H Lin A J Q Ogata H and Yang S J H, 2018 Applying learning analytics for the early prediction of students' academic performance in blended learning *Educ. Technol. Soc.* 21 2 p. 220–232.
[44] Shabaninejad S Khosravi H Indulska M Bakharia A and Isaias P, 2020 Automated insightful drill-down recommendations for learning analytics dashboards in *ACM International Conference Proceeding Series* p. 41–46.
[45] Menon A Gaglani S Haynes M R and Tackett S, 2017 Using "big data" to guide implementation of a web and mobile adaptive learning platform for medical students *Med. Teach.* 39 9 p. 975–980.
[46] Sancho J, 2016 Learning Opportunities for Mass Collaboration Projects Through Learning Analytics: A Case Study *Rev. Iberoam. Tecnol. del Aprendiz.* 11, 3 p. 148–158.
[47] Rienties B Lewis T McFarlane R Nguyen Q and Toetenel L, 2018 Analytics in online and offline language learning environments: the role of learning design to understand student online engagement *Comput. Assist. Lang. Learn.* 31 3 p. 273–293.
[48] Agarwal R Srikant R and others, 1994 Fast algorithms for mining association rules in *Proc. of the 20th VLDB Conference* p. 487–499.
[49] Kuzilek J Hlosta M and Zdrahal Z, 2017 Data Descriptor: Open University Learning Analytics dataset *Sci. Data* 4 p. 1–8.
[50] Strang K D, 2017 Predicting student satisfaction and outcomes in online courses using learning activity indicators *Int. J. Web-Based Learn. Teach. Technol.* 12 1 p. 32–50.