A Comprehensive Framework for Big Data Analytics in Education

Ganeshayya Shidaganti[1], Prakash S[2]
Research Scholar, Visvesvaraya Technological University, Belagavi and Assistant Professor, Department of C.S.E[1]
M.S. Ramaiah Institute of Technology (Affiliated to VTU, Belagavi), Bengaluru, Karnataka, India[1]
Professor, Department of C.S.E, East Point College of Engineering and Technology, Bengaluru, India[2]

***Abstract*—With the adoption of cloud services for hosting knowledge delivery system in educational domain, there is a surplus quantity of education data being generated every day by current learning management system. Such data are associated with certain typical complexities that impose significant challenges for existing database management and analytics. Review of existing approaches towards educational data highlights that they do not offer full-fledged solution towards analytics and still there is an open-end problem. Therefore, the proposed system introduces a comprehensive framework which offers integrated operation of transformation, data quality, and predictive analytics. The emphasis is more towards achieving distributed analytical operation towards educational data in cloud. Implemented using analytical research methodology, the proposed system shows better analytical performance with respect to frequently used educational data analytical approaches.**

Keywords*—Big data; data analytics; educational big data; predictive analytics; text mining; machine learning; education technology

I. INTRODUCTION

There has been significantly increase in adoption of technique in the area of education system in recent time. It is because the process of knowledge delivery system is required to be carried out in order to reach number of users. However, this mechanism of online educational knowledge delivery system will demand heavier platform [1]. In this aspect, cloud computing offers a better option for hosting educational delivery system for online learning management [2]. However, such online learning management scheme is not only for delivering the live classes but it is also about understanding the entire process of service delivery from quality perspective [3]. Different information in the form of archives of study materials, online forums, information about students, instructors, payment and service syncing, etc. are required to be stored. With proliferation of mobile internet, adoptions of such educational services are tremendously increasing [4]. This results in generation of massive amount of data in the form of stream. Some data are generated voluntarily while many other data are generated involuntarily. The voluntary data will consist of study material, online notes, and respective information about students /instructor. The involuntary data consists of data that is generated autonomously e.g., positional data, trace data, behavioral specific data, etc.

There is problem associated with such form of educational data viz. i) the data generated is so massive that it cannot be supported by small scale deployment. These large data cannot be stored directly to the storage unit of cloud efficiently in distributed fashion, ii) the next issue is into form of the data that is being subjected to mining. Normally, the data is either semi-structured or unstructured which makes them ineligible for storing it into conventional storage units, iii) another bigger issue is the distributed deployment of educational data. If the cloud-based educational services are running from different geographical regions, than there will be different origination point of such educational big data. All these data are required to be aggregated in distributed manner before deploying them to the analytical process. In case of incomplete or imperfect aggregation, the mining process will be carried out over impartial data resulting in outliers. Hence, mining will be not effective in such way, iv) Finally, the problem is in applying machine learning in order to carry out predictive value of it as the form of knowledge extraction. A data is of no use until and unless it is not predictive. Therefore, it is not simpler mechanism to perform storage, processing, and analyzing existing educational big data.

At present, there has been various works being carried out towards analyzing such education data. Existing studies on big data is found to use sophisticated tools and framework [5-8]. However, such complex adoption cannot be taken into consideration in real cases. Apart from this, existing approaches of educational big data [9] lacks various conceptualization like heterogeneity in data, multiple and large-scale origination of educational data, streaming of educational data, etc. Irrespective of presence of multiple problems, only few problems are addressed symptomatically in existing approach. Therefore, this paper presents discussion of the unified approach where multiple problems associated with the educational big data is addressed.

The organization of the paper is as follows: Section 2 discusses about the existing studies while Section 3 discusses about the research problems, Section 4 discusses about the adopted research methodologies, while Section 5 discusses about the algorithm implementation. Results are discussed in Section 6 while conclusion is discussed in Section 7.

II. Related Work

There have been various forms of work carried out towards educational big data in recent times [10-11].

Uses of big data approach in educational domain were witnessed to explore the popular topic of study [12] Inclusion of deep learning concept has assisted in constructing such framework that could further facilitate in extraction of key-words. The work carried out by [13] have developed a model using structural equation modeling. The framework is designed using conventional technical adoption model toward educational domain targeting to find the comfortability of end user. Existing studies highlights that combined usage of warehouse, business analytics, and enterprise architecture can be used for improving the analytical operation [14]. Apart from this, the researchers have also offered an importance towards clustering process of educational data which consist of three stages viz. pre-processing, standardization, and modeling [15]. Clustering approach toward online learning can be personalized for improving the knowledge delivery process [16]. Along with clustering, classification-based approaches are also used for improving analytical processes over educational data [17]. All such advanced mechanism gives rise to evolution of smart campus; however, there are still issues associated with integrating such devices. Development of smart campus can be carried out over a platform of effective data fusion [18]. Inclusion of ubiquitous approach in the form of framework for facilitating distance learning is carried out by [19]. In order to offer user friendly experience, visualization of data can offer faster access to the knowledge. There are various visualization tools at present that can carry out this task [20-21] have presented another visual analytical scheme for online courses. This scheme is deployed for offering visual representation of the learning groups. Machine learning has been consistently found to be adopted in existing approaches towards incorporating smart features in framework building. Such framework is witnessed to assess the competencies of student [22]. Existing approach has also witnessed the usage of collaborative filtering process using predictive method [23]. This model predicts scores about the courses. The work of [24] has implemented a unique approach where the study contents are emphasized during the knowledge delivery process. Nearly similar predictive approach is presented in the work of [25] where the idea is to perform identification of the students that are in the level of risk. The work carried out by [26] has discussed that adoption of mining approach as well as analytical approach can be mechanized. This work has presented a comprehensive representation scheme of the involuntary regulation of the behavior of student. Apart from this, there are various other schemes towards big data analytics e.g., tensor-based scheme [27], compression based on context [28], pattern analysis [29], deep learning [30], clustering technique [31]. Existing system has also witnessed an extensive usage of Hadoop framework. Some of the existing approaches are discussed by [32-35]. Adoption of machine learning is another frequently used approaches in educational analytics viz. [36-39] finally, text mining is another frequently used approach for educational analytics e.g. [40-44]. Ifenthaler [45-46] have carried out study towards proving that all the upcoming forms of education system will be requiring advanced forms of analytics. The similar forms of proposition towards adoption of learning-based analytics during pandemic is also studied by Beerwinkle [47]. Further, the recent work of Lee et al. [48] have discussed about various innovative practices of using analytics over educational data in order to cater up both technological and pedagogical demands of students. The next section discusses about the research problems.

III. Research Problem

After reviewing the existing approaches of analytical operation associated with educational domain, it has been observed that that present schemes have certain loopholes viz.

Simplified Transformation Scheme: Majority of the transformation scheme is meant for making the data suitable for structurization and mining where complex operation is involved. Moreover enough emphasize is not offered on transformation operation. Educational big data is highly heterogeneous in its form and it requires cost effective transformation scheme. Data that can actually be carried out on educational data.

Presence of Artifacts in Transformed Data: Usually storage and analysis is carried out in explicit manner in two different places. A transformed data is forwarded to special block of operation that is meant to be carrying out analytical operation. Owing to various impediments in communication medium, there are fair chances of inclusion of artifacts in transformed data. Existing system has no scheme to solve this problem.

Cost Ineffective Predictive Schemes: Existing schemes uses various predictive techniques where majority of this techniques are highly iterative and depends upon the trained data. Higher the trained data, higher is accuracy in prediction.

Therefore, all the above points are required to be addressed in order to mechanism a truly distributed analytical modeling of educational data. Until and unless, these research problems are not addressed, it is challenging to evolve up novel solution.

IV. Research Methodology

This part of the proposed study focuses on achieving a high-quality data in order to make it suitable to apply analytics for better value extraction. Following are the details of proposed implementation:

The proposed system addresses an explicit problem of analysis the complex form of educational big data. With the generation of data from multiple sources, there is higher feasibility of inclusion of errors in the form of noise. These errors could be human-based, machine-based, as well as network-based. Hence, the presence of errors will generate significant outliers, which is detrimental to carry out analysis. At the same time, solution to eliminate or reduce errors cannot be carried out locally as it will be not be cost effective and moreover, it cannot offer instantaneous query processing capability. The complete implementation of the proposed study was carried out considering an explicit case study. In order to solve the above-mentioned problem, the proposed study considers a case study. Referring to Fig. 1, the study considers m number of data node (d_n) to represent repository of

educational data in different geographic location with an inclusion of errors α in each data nodes of k types. For simplicity, the study considers k<<m, which will mean that number of error types are considerably lower than number of data nodes. As the computation for error elimination cannot be carried out in local level, so the study considers the presence of a memory stream which can connect to indexes of all the data nodes and an external cluster hosted in cloud is considered as the prime location where the data aggregation is carried out. By data aggregation, it will mean consolidation of all the individual data along with the level of errors maintained within each data nodes. The proposed algorithm for error elimination is then applied over the cluster in order to finally generate an error free data (dbef).

Fig. 1 highlights an adopted solution strategy where the prime agenda of the proposed system is to make the incoming stream of educational big data effectively structured and prepared for high end and cost-effective analytical operation. For this purpose, the first preference is offered towards rectifying the structuredness of the raw data by implicating a simple and novel data transformation scheme. After the data transformation scheme is implemented, the next focus of the proposed system will be towards identifying the presence of artifacts in that transformed data when they are forwarded from various data nodes via memory streams. A superlative indexing scheme is implemented in the proposed system which indexes all forms of data especially when the data is further classified into two forms. One form of data is permanently saved while other form is stored in volatile memory system and the complete algorithm is implemented over the volatile memory system. Thereby, a significant saving of storage units is emphasized in proposed system. The proposed system also assists in identification of the position in the cell over the temporary storage units with respect to error prone data. Such data are not only identified but also substituted by statistically computed value. This mechanism assists in maintaining higher degree of data purity. Once the quality data is obtained, the next process is to carry out predictive analysis using a novel deep learning mechanism. The overall scheme of the proposed system is to offer the complete educational mined data to be used in the form of cloud-based services. So that a new avenue of analytical application can be used for automating the knowledge delivery system over educational domain in various perspective.

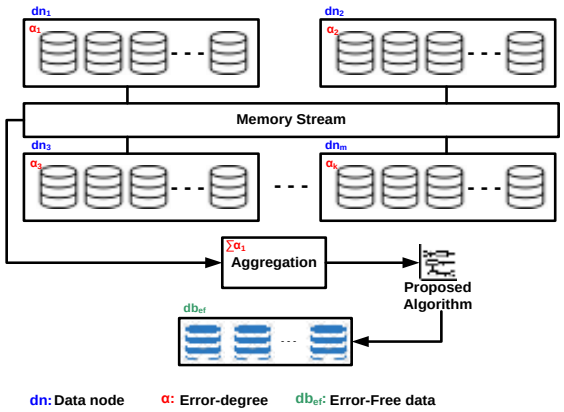


Fig. 1. Adopted Solution Strategy.

V. Algorithm Implementation

This proposed algorithm uses a series of three essential operations in order to facilitate development of a comprehensive analytical framework. The complete implementation is carried out with respect to three sequential phases of operation i.e., i) Data Transformation Phase, ii) Data Quality Incorporation Phase, and iii) Data Predictive Analysis Phase. It should be noted that all the above-mentioned approaches are applied over an educational big data. The discussion of this implementation phases is carried out as follow:

A. Data Transformation Phase

This is the preliminary phase of implementation where the emphasis is offered towards processing followed by an effective transformation of the data. The study considers data transformation as one of the essential operations which makes the incoming stream of data more suitable to be subjected to analytical operation. For this purpose, the study considers one unit of educational data in following form as shown in Fig. 2.

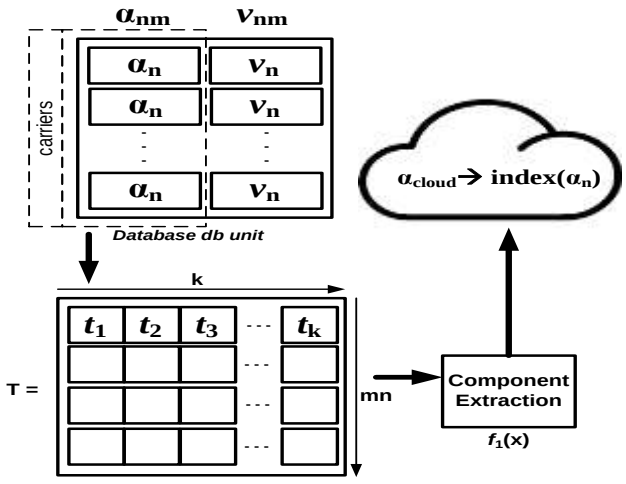


Fig. 2. Structure of Educational Data.

Fig. 2 highlights the structure of one database unit from the massive stream of educational big data. The study consider that each database unit has two essential parts i.e., carriers and value. The carrier is static information type of the educational data it targets to carry different explicit information called as components in each database units. The study considers αnm to represent components where n and m represents number of component and number of carriers in one database unit respectively. In case of educational big data, the constraint in this regards is,

$$m>>n \quad (1)$$

There are various possibilities of the number of component n, but during analysis, the proposed study can select a fixed integer value to obtain measurable outcome. Considering the educational data with respect to text mining approach, the study constructs a matrix T that retains all the textual contents in the complete database which can now be subjected to next phase of operation. The proposed algorithm for the process of data transformation is as follows:

Algorithm for Data Transformation

Input: c (carriers), α (component)

Output: O1 (transformed mined data)

Start

1. For i=1: c c is carriers
2. $\alpha f_1(T)$
3. dataβ {stream(v_{nm}) (α_{cloud})}
4. posg(data)
5. φ (γ)data
6. generates O_1

End

The description of the algorithmic steps are as follows: The algorithm takes the input of c(carriers) and α(component) for all the database units (Line-1). An explicit function f1(x) is applied over the matrix T which retains all the textual contents with a dimension size of (k x mn). Applying this function results in extraction of all the essential components which α which are then retain in cloud storage system permanently (Line-2).

Each individual component is further indexed within the cloud storage which ensures the rejection of storage of similar components as well as it can be called upon for looking for ownership of all individual values during query processing. This results in further storage optimization while values are never forwarded in this stage. The next part of the algorithm is to carry out data organization as follows:

$$\alpha\text{cloudindex}(\alpha n) \quad (2)$$

The expression (2) exhibits that all the incoming streams of educational data value vnm are now assessed for their owner components which now resides in cloud αcloud considered for all database db. (Line-3). The expression (3) exhibits that proposed system applies a function β(x) where web-script tags are applied on both components in cloud αcloud and its respective value vnm. This operation results in well-formatted data that could be supported on any client application over cloud interface in semi-structured manner. The algorithm also extracts the position information of the individual data which directly assists in lowering the search time during the query processing (Line-4). A function g(x) is designed for extracting all the positional information from the data and stored it in pos matrix. Finally, the knowledge extraction is carried out where γ syntactical rule set is used as following:

$$\gamma = \{r1, r2 \ldots rl\} \quad (3)$$

In the above expression, the variables r is l number of rule set which offers semantic information of the chosen text from the value of data. The algorithm constructs a function φ(x) which computes the syntactical correlation between rl and all the incoming data (Line-5). This operation results in the generation of knowledge as the outcome O_1 (Line-6). This outcome can be now stored back in the cloud storage system. The algorithm therefore offers a good balance between storage optimization (by storing only the mined data and non-repeating components) and data transformation operation. The study outcome is now assessed for next level of challenge with solution.

B. Data Quality Incorporation Phase

This module of operation is executed after the first algorithm is successfully executed that results in transformed data O1. It should be noted that execution of first algorithm is accompanied by storage of component information αnm in the indexed cloud storage permanently. However, the evaluated mined data O1 is ready to be stored distributed manner in cloud. In the process of distributed transmission of the aggregated transformed data O1, there are various possibilities of inclusion of further artifacts associated with network-based transmission. This phenomenon could significantly affect the quality of data resulting in inclusion of storage of artifact-incorporated transformed data. Hence, this part of the algorithm is mainly focused on rectifying the artifacts and substitutes the artifact-based transformed data into more quality data O2. The process flow of this part of implementation is as follows:

Fig. 3 highlights the process implementation towards identification and removal of the artifacts in order to retain higher degree of data quality. By data quality, it means that complete structure of the distributed database db should be fulfilled. Presences of any missing value of noisy values are easier to find but difficult to be rectified in distributed manner. For this purpose, the proposed system carries out following steps of operation in the form of algorithm: The algorithm initially takes the input of all the transformed data released from prior algorithm and computes its size (Line-1). The next part of the implementation is to split the complete text-based data T (considering both components α and their respective values v) into smaller components T1, T2, . . . Th, where h is number of splits of the data carried out on the basis of total number of available storage slots H in cloud (Line-2). It will eventually mean that the proposed algorithm offers a better form of elastic cloud usage where the on-demand scaling process of the data splitting operation is carried out. This operation is one significant step towards i) storage optimization as well as ii) faster query processing in distributed manner over cloud environment. By splitting the aggregated data into different segments, the network overhead towards processing the mined data is potentially controlled. Apart from this, it also offers a significant level of mined data availability which is also a part of solution towards dense state of traffic.

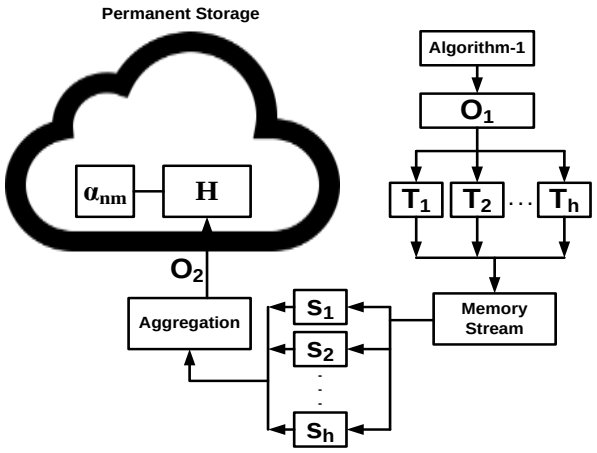


Fig. 3. Process of Data Quality Incorporation.

Algorithm for Data Quality

Input: O_1 (transformed data)

Output: O_2 (quality data)

Start

1. For i=1: size (O_1)
2. construct T= {T_1, T_2…., T_h} h⩽H
3. create s= {s_1, s_2……s_u }|u = h
4. For j=1: num (α_{nm})
5. search (v_{nm})err
6. $(v_{nm})_{corr} f_2((v_{nm})err, (v_{nm}+1))$
7. $O_2(v_{nm})_{corr}$

End

All the split data (T) is now forwarded to the memory stream which is an explicit buffer maintained over the cloud cluster or edge server where adaptive queue management is carried out. The next part of the implementation is to perform allocation of this split data with distinct streams. The algorithm creates a matrix s which has u number of streams where each stream will carry different split of data. In order to perform optimization of the network resources, the algorithm considers the equivalent value of both u and h (Line-3). This consideration prevents the system memory stream in edge server to create unnecessary streams of data. This will eventually mean that outgoing traffic of s doesn't offer any form of data overhead over the cloud interface prior to storage. The algorithm then looks for all the number num of value vnm (Line-4) from this stream of data to find if there is presence of any values with artifact (Line-5). The study considers that there are no artifacts associated with components as all the components are stored permanently over cloud. Hence, lesser chances of error prone err component information and all the errors will be related to values vnm itself.

The prime contribution of this algorithm is that it constructs a function f2(x) which performs statistical calculation in following manner. The columns of all the respective values are considered which have presence of artifact data followed by computation of statistical value (mean) of it. This extracted statistical value is now compared with all the columns of remaining streams of data (Line-6). Only the highest correlated data is extracted and substituted in the target value with prior artifacts (Line-7). It will mean that the final steps of this algorithm result in substitution of statistically computed data in the cell which has priory data with artifacts. Assuming that the proposed system works on defined domain of heterogeneous data, there are no chances of computed data with higher fluctuation or deviation. Therefore, the computation of mean value offers faster and reliable substitution of computed data. This operation ensures that any incoming data should never have any artifacts and in case there are any artifacts than they are going to be searched upon by this algorithm and substituted with accurate value. Hence, the process eliminates any presence of data uncertainty issue and offer inclusion of higher quality of data in cost effective manner.

C. Data Predictive Analysis Phase

This is the final part of execution which applies machine learning mechanism over the quality data O2 obtained from prior algorithm implementation. The part of the implementation considers the similar distributed scenario where the quality data O2 is assumed to be generated in distributed manner. The process flow of the proposed system is shown in Fig. 4.

After the data is considered to be distributed i.e., O21, O22. . . O2H, it is subjected to dual sequential operation of indexing and then sorting. This process is slightly different in contrast to conventional deep learning approach where there are possibilities of various numbers of features. The essential steps of processing of proposed algorithm are as follows:

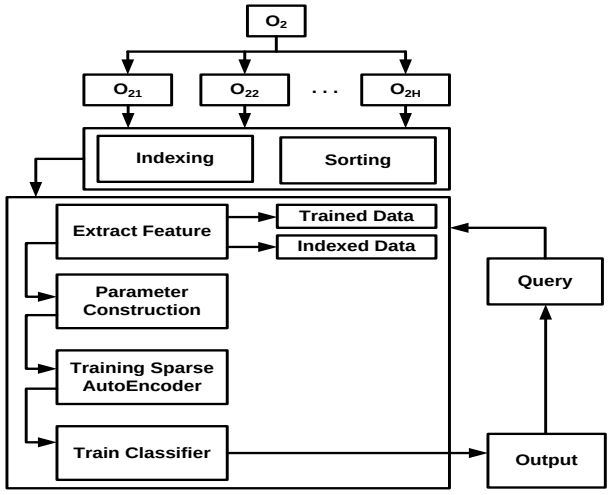


Fig. 4. Process of Predictive Analysis.

The proposed system takes the input of the quality data O2 from the prior algorithm implementation and computes its size (Line-1). The process of data aggregation is further carried out using a function agg applied over quality data O2 (Line-2). The proposed system considers very simple form of feature feat, which are indexed data μ and sorted data τ (Line-3). This step is uniquely carried out in proposed system. The next step is to carry out construction of various fundamental parameters e.g., size of input data, number of classes which are length of uniquely trained data, weight decay parameter and anticipated activation of hidden layers, weight of sparsity penalty term, and maximum number of iterations. The proposed system makes use of the auto-encoding system, which are subjected to training depending upon the fact of replication of input with the outcome data. Upon encoding the data, the transformation is carried out over the obtained set of features which is actually chosen by the auto-encoders itself. This is further used for the decoding purpose in consecutive process. Finally, the error is computed from the difference obtained from decoded data and input data, while this information is further used for minimizing the rate of error resulting in predictive outcome. In this entire process, the proposed system also adopts the usage of the optimization process towards solving non- linear unconstrained problems using iterative process. The essential steps of the proposed system are as follows:

Algorithm for Prediction

Input: O_2 (quality data)

Output: O_3 (predicted value)

Start

1. For i=1: size (O_2)
2. a=agg (O2) = {O_{21}, O_{22}, O_{23}... O_{2H}}
3. [μ τ] [index (a), sort (a)]
4. feat [μ τ]
5. trf_3(feat) ^{sc}
6. O_3tr

End

The next part of the algorithm implementation is about applying a sparse auto encoding mechanism which has many numbers of neuron present in hidden layers in contrast to its respective input later. However, there is a good possibility of allocating of single neuron for one input data. This problem is overcome, the proposed system performs frequent switching of neurons over varied ranges of iteration. This mechanism allows the precise encoding of the features leading to better form of predictive analysis. The proposed system furthers carry out a unique mechanism of training. It trains the initial layer of the sparse auto-encoder with numerical optimization mechanism such that presence of any instances related to the non-smooth optimization can be considered. This optimization technique ensures the presence of zero gradients for all the required condition in order to achieve optimized performance. The next sparse auto encoder is feed forward mechanism considering the trained data, size of input layer, hidden layer. This mechanism is nearly equivalent to the single layer perception where the number of layers for both input and output are same but it can have varied number of hidden layers. The prime agenda in this part of operation is to reduce the significant difference between the input and output. One of the significant contributions of this algorithm is that it can carry out unsupervised form of learning mechanism without any form of dependency towards indexed data in its input layer in order to carry out learning operation. This characteristic of the proposed operation ensures that even if the incoming stream of the data is not indexed appropriately, then also the proposed system is able to perform the encoding operation. The final part of the implementation is followed by performing classification of the trained data that offers the complete probability of all the classes of indexed data. This form of the classification is essentially a binary type of the statistical regression that offers a significant precision in the process of classification. One of the interesting points of proposed algorithm is the generation of elite feature in each cycle of training which significantly improves the accuracy level. It also offers higher scope of utilization of the unstructured educational data and yet maintaining better for of predictive performance. By its unique mechanism of training, the algorithm also reduces the cost of training operation as better results are obtained in reduced number of training cycle. Apart from this, the accuracy doesn't get affected in presence of indexed data; however, it is used for further improving the classification performance.

VI. Result Analysis

This section discusses about the outcome obtained after implementing the proposed logic in the prior section. The proposed study implements a comprehensive mechanism which aggregates, organizes, transforms, processes, and performs predictive operation over educational big data. This section discusses about the strategy adopted for analysis, the dataset considered for the result analysis, test case used, and discussion of the accomplished results.

A. Analysis Strategy

A closer look at the proposed system shows that it carries out three sets of sequential operation in order to carry out comprehensive analytical operation. However, in order to ascertain the effectiveness of the proposed system, there is a need of using certain strategy to carry out analysis. The proposed system makes use of three essential strategies in order to measure the effectiveness of the proposed system. The primary strategy of the analysis is to assess all the performance parameters with respect to size of the data. There are multiple reasons behind this adopted strategy. The first reason is associated with the scalability factor of the proposed system. By scalability, it will mean that proposed system will successfully deliver similar form of optimal services towards data analysis. Normally, the capacity of the cloud servers is finite in spite of using distributed environment, therefore, increasing traffic will have possibility towards overloading the task towards this cloud server that could affect the database management for heavier and uncertain traffic condition. The second reason for assuming size of data is because in educational domain, the size of the data is always exponentially increasing in shortest span as well as in various forms. Therefore, analyzing size of data contributes towards effective review of scalability factor of proposed system. The secondary strategy of the proposed analysis is carried out for comparative analysis of existing approach that is frequently used for classification purpose while boosting the analytical operation. By comparing with the existing approach of analytics, the accomplished outcome can be generalized for the effectiveness towards the distributed data mining operation associated with educational domain. The tertiary strategy of analysis is to perform selection of multiple forms of performance parameter over the same test environment along with other existing approaches. The proposed system uses data transformation time and data transformation accuracy as the performance parameters for assessing first algorithm. In order to assess the second algorithm, the proposed system uses data fusion time and data fusion quality while data prediction time and data prediction accuracy is used for assessing the last algorithm. Therefore, a comprehensive set of performance parameters are used for this purpose. The study outcome is assessed and compared with respect open-source distributed framework Hadoop, machine learning, and conventional text mining approach. All the assessment has been carried out considering similar environmental variable in order to obtain an unbiased outcome of the proposed system. The efficiency of mining operation is assessed using different methods in proposed system considering educational dataset.

B. Dataset Considered

The discussion of dataset is quite important in proposed system. In order to testify the effectiveness of the proposed system, the input data should have certain criteria to fulfill viz. i) the input data should be associated with educational domain, ii) the input data should be stream of educational data originated from distributed systems in institution, iii) the input data should have the characteristics of massive size, inclusion of artifacts, lesser valuable data. However, availability of such form of data is quite difficult. Therefore, the proposed system initially reviews the publically available big data in order to understand the form and specification of big data. It is found that existing big dataset has a finite smaller size and it doesn't possess the 3rd criteria discussed about the suitable input file. Therefore, the proposed system chooses to construct a synthetic data considering the domain of educational system. The dataset considered for the analysis of the proposed system has an overall size of 100 GB with 7,500 plain text files consisting of specific performance information about educational delivery system.

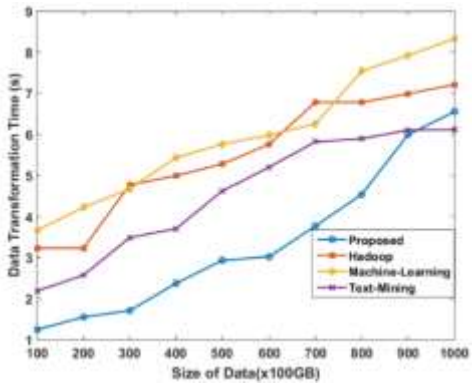
Table I highlights one set of information among many existing within one plain text file. It can be seen that there are 9 categories involved for one course undertaken and a single plain text file can contain many more such course information within it. The study assumes that Course ID is a unique number which is allocated by the system for every course undertaken by the scholar. Hence it is non-repetitive in nature in overall dataset considered for assessment. Course Title represents the name of the course undertaken by the scholar and it is also fixed and non-iterative in nature. It should be noted that every Course Title has uniquely allocated Course ID which is non-iterative. Course Type represents the do- main of the course viz. i) social, ii) political, iii) engineering, iv) medical, v) literature, etc. Date represents the start point of course while Location represents the geographical position of the knowledge delivery point of the scholar. This information can be easily retrieved from IP address of scholar. The category of Course Status is either active or inactive. The flag active will represent ongoing course and inactive will represent already delivered course. The category Total Episode will represent total number of online sessions allocated for specific course. Scholar Name is name of the client who has privilege access to this cloud-based knowledge delivery application and receives training from instructor while the category Scholar Feedback is the response given by the user for either the ongoing courses or the completed course.

Scripted in MATLAB using normal windows machine, the proposed system is analyzed for all the three discrete and sequential algorithms in order to measure their level of effectiveness. The proposed system is basically an analytical model and MATLAB is one of the best tools for this. The prime reason to use MATLAB is that it offers simpler operation for carrying out extensive evaluation where the focus can be purely on the analysis without any concern of tools, writing bigger scripts, or dependency of extensive code. The first algorithm is assessed with respect to data transformation time and accuracy (Fig. 5). The outcome shown in Fig. 5 (a) highlights that although transformation time increases with increase in data, yet proposed system offers reduced data transformation time. The similar trend is also observed in Fig. 5 (b) where the proposed system is found to offer increased data transformation accuracy compared to existing approaches. The prime reason behind this is that proposed data transformation algorithm is carried out in a smaller number of non-iterative steps without offering any dependencies of third parties causing faster processing time. Apart from this, the proposed system makes use of semantics for the purpose of extracting the essential elements of data during mining process causing higher accuracy. Such semantics are further user-defined and can be constructed depending upon the demands making the proposed system free from any lexical database system. On the other hand, adoption of Hadoop has higher dependencies on system requirements and construction of higher number of tall arrays in order to deal with increasing dimension of dataset. Hence, transformation time evidently increases while accuracy decreases in such case. In the case of machine learning approach, the approach is extensively iterative in its operation as well as there is a dependency on training dataset. Finally, all the text mining approaches are assessed to find out that they consume more time to perform transformation of educational data. Apart from this, there is also an increasing dependency on lexical dataset in order to extract the logical meaning of the elements within corpus.

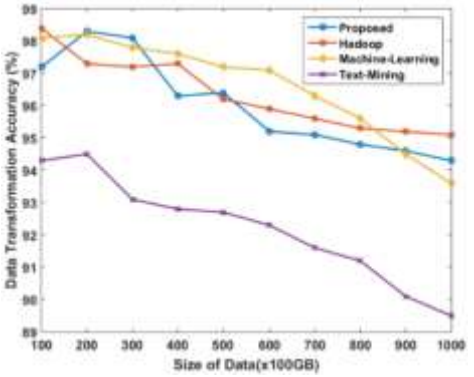
The effectiveness of the second algorithm is assessed and outcomes are shown in Fig. 6. The proposed system offers highly reduced data fusion time in comparison to existing approaches (Fig. 6 (a)). The prime justification for this lowered fusion time is that proposed system performs the queuing of the incoming traffic in its data node which is capable of carrying out distributed stream management. The likability of the data and traffic are highly maintained owing to an effective indexing policy executed in this part of implementation. This causes faster aggregation of data as although the data nodes are distributed but they have a good linkage causing an effective redundancy management too. At the same time, the proposed system also offers increasing data quality without using any third party. Hadoop has increasing dependencies on constructing arrays in order to save increasing data. It uses too many mechanisms for compacting the data which consumes time for performing data fusion. Machine learning has inclusion of training while text mining approaches has too much simplified and often fails to understand the relationship among the data especially if the database is of massive and uncertain scale.

TABLE I. DATA SPECIFICATION

#	Categories	Character Range	Data-Type
1	Course ID	1-5	Number
2	Course Title	1-15	String
3	Course Type	1-15	String
4	Date	10	Number
5	Location	1-15	String
6	Course Status	6/8	String
7	Total Episodes	1-3	Number
8	Scholar Name	1-20	String
9	Scholar Feedback	1-200	String

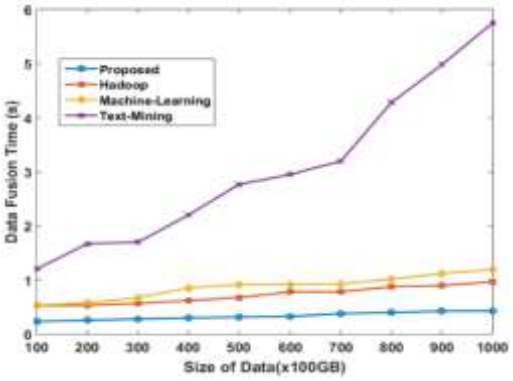


(a) Data Transformation Time.

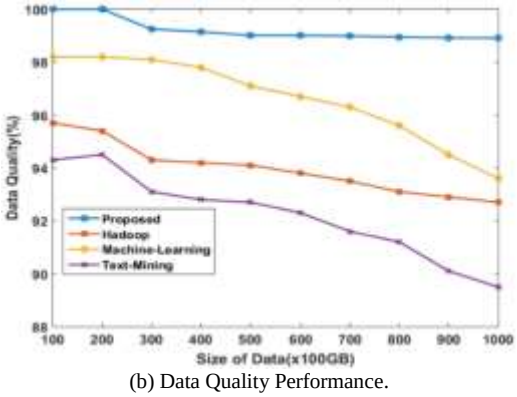


(b) Data Transformation Accuracy.

Fig. 5. Outcome for Algorithm of Data Transformation.



(a) Data Fusion Time.

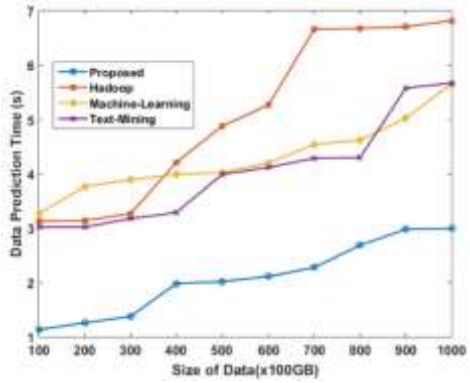


(b) Data Quality Performance.

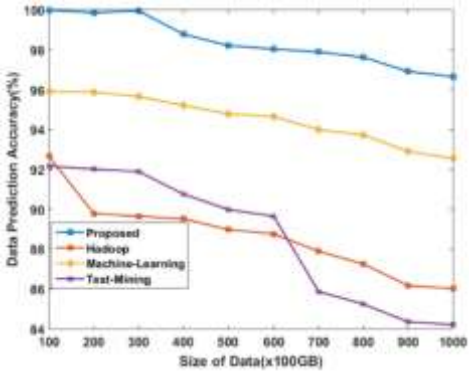
Fig. 6. Outcome for Algorithm of Data Quality.

The proposed system makes use of simplified statistical measures in order to offer an improved data quality in much reduced number of steps, which is not found in any existing approaches (Fig. 6 (b)). Hadoop has the better capability of storing and managing large and distributed data; however, they don't offer any form of identification and substitution of artifact data by itself. This causes slight increase in data fusion time and it takes assistance of zookeeper and other metadata management present in its architecture. This is also the reason of reduced data quality in Hadoop. Machine learning offer nearly equal time of operation for data fusion just like Hadoop, but its performance towards data quality depends upon its epoch level. Machine learning is capable of offering higher accuracy but the accuracy for higher size of data in present state is found to be reduced in it. The reason is simplified as the activation function in it is incapable of offering higher accuracy. On the other hand, text mining approach does not offer any form of substitution operation for the artifact elements in the corpus. Therefore, the proposed system is capable of offering a better form of data quality along with replacement of the error-prone data in faster way.

In order to assess the predictive analysis of the proposed system, the processing time as well as accuracy is the prominent indicator of its performance. The outcome shown in Fig. 7 eventually highlights that proposed system offers better predictive performance in contrast to existing approaches. It should be noted that the outcome of the predictive analysis is also a mined data and this data is essential for the data analyst (or stakeholder of the data). The first prominent reason for data prediction time is an inclusion of lesser number of iterations toward reaching the minimum gradient (Fig. 7 (a)). As the proposed system offers an exclusive extraction of sequential mined data with progression of each algorithms, therefore, a greater number of information is obtained in this process which reduces the decision-making time for the proposed system in order to make prediction. On the other hand, availability of more precise and larger set of filtered information also results in higher accuracy in proposed system (Fig. 7 (b))). On the other hand, Hadoop doesn't have extensive capability to carry out data prediction and hence it offers extensive time processing (Fig. 7 (a)). Hadoop doesn't address the problems associated with data uncertainty although it can offer better data transformation scheme for homogeneous data. This adversely affects the accuracy score of Hadoop. The machine learning scheme is found to offer increased consumption time owing to an inclusion of training however, its accuracy is the next better score after the proposed system. Text mining approach offers simplified mechanism; however, knowledge extraction process is quite length process in it resulting in higher involvement of prediction time. This also results in reduced accuracy.



(a) Data Prediction Time



(b) Data Prediction Accuracy.

Fig. 7. Outcome of Algorithm for Predictive Analysis.

VII. CONCLUSION

This paper has presented a framework that is meant to carry out comprehensive operation that leads towards an effective analytical operation over educational data. The proposed system has emphasized over data transformation, data quality incorporation, and predictive analytics in educational data. Scripted in MATLAB, the study highlights that it is capable of better analytical operation in contrast to text mining approach, machine learning approach, and Hadoop, which are the most used techniques in data analytics over educational domain. Our future work will towards optimizing the performance more by exploring more approaches towards its implication on real-time applications.

REFERENCES

[1] Al-Maroof, R. S., Alfaisal, A. M., & Salloum, S. A. (2021). Google glass adoption in the educational environment: A case study in the Gulf area. Education and Information Technologies, 26(3), 2477-2500.

[2] Sharma, D., & Kumar, V. (2021). A framework for collaborative and convenient learning on cloud computing platforms. In Research Anthology on Architectures, Frameworks, and Integration Strategies for Distributed and Cloud Computing (pp. 629-650). IGI Global.

[3] Ammenwerth, E., Hackl, W. O., Hoerbst, A., & Felderer, M. (2021). Indicators for cooperative, online-based learning and their role in quality management of online learning. In Research Anthology on Developing Effective Online Learning Courses (pp. 1709-1724). IGI Global.

[4] Huang, R., Tlili, A., Chang, T. W., Zhang, X., Nascimbeni, F., & Burgos, D. (2020). Disrupted classes, undisrupted learning during COVID-19 outbreak in China: application of open educational practices and resources. Smart Learning Environments, 7(1), 1-15.

[5] Baig, M. I., Shuib, L., & Yadegaridehkordi, E. (2020). Big data in education: a state of the art, limitations, and future research directions. International Journal of Educational Technology in Higher Education, 17(1), 1-23.

[6] Fischer, C., Pardos, Z. A., Baker, R. S., Williams, J. J., Smyth, P., Yu, R., ... & Warschauer, M. (2020). Mining big data in education: Affordances and challenges. Review of Research in Education, 44(1), 130-160.

[7] Williamson, B. (2017). Big data in education: The digital future of learning, policy and practice. Sage.

[8] Huda, M., Maseleno, A., Atmotiyoso, P., Siregar, M., Ahmad, R., Jasmi, K., & Muhamad, N. (2018). Big data emerging technology: insights into innovative environment for online learning resources. International Journal of Emerging Technologies in Learning (iJET), 13(1), 23-36.

[9] Ogata, H., Oi, M., Mohri, K., Okubo, F., Shimada, A., Yamada, M., ... & Hirokawa, S. (2017). Learning analytics for e-book-based educational big data in higher education. In Smart sensors at the IoT frontier (pp. 327-350). Springer, Cham.

[10] Quadir, B., Chen, N. S., & Isaias, P. (2020). Analyzing the educational goals, problems and techniques used in educational big data research from 2010 to 2018. Interactive Learning Environments, 1-17.

[11] Zhang, W., & Qin, S. (2018, March). A brief analysis of the key technologies and applications of educational data mining on online learning platform. In 2018 IEEE 3rd International Conference on Big Data Analysis (ICBDA) (pp. 83-86). IEEE.

[12] Wang, B., Yang, B., Shan, S., & Chen, H. (2019). Detecting hot topics from academic big data. IEEE Access, 7, 185916-185927.

[13] Al-Rahmi, W. M., Yahaya, N., Aldraiweesh, A. A., Alturki, U., Alamri, M. M., Saud, M. S. B., ... & Alhamed, O. A. (2019). Big data adoption and knowledge management sharing: An empirical investigation on their adoption and sustainability as a purpose of education. IEEE Access, 7, 47245-47258.

[14] Moscoso-Zea, O., Castro, J., Paredes-Gualtor, J., & Luján-Mora, S. (2019). A hybrid infrastructure of enterprise architecture and business intelligence & analytics for knowledge management in education. IEEE Access, 7, 38778-38788.

[15] Dutt, A., Ismail, M. A., & Herawan, T. (2017). A systematic review on educational data mining. Ieee Access, 5, 15991-16005.

[16] Kausar, S., Huahu, X., Hussain, I., Wenhao, Z., & Zahid, M. (2018). Integration of data mining clustering approach in the personalized E-learning system. IEEE Access, 6, 72724-72734.

[17] Yu, L., Wu, X., & Yang, Y. (2019). An online education data classification model based on Tr_MAdaBoost algorithm. Chinese Journal of Electronics, 28(1), 21-28.

[18] Yang, A. M., Li, S. S., Ren, C. H., Liu, H. X., Han, Y., & Liu, L. (2018). Situational awareness system in the smart campus. Ieee Access, 6, 63976-63986.

[19] Mehmood, R., Alam, F., Albogami, N. N., Katib, I., Albeshri, A., & Altowaijri, S. M. (2017). UTiLearn: a personalised ubiquitous teaching and learning system for smart societies. IEEE Access, 5, 2615-2635.

[20] Liu, J., Tang, T., Wang, W., Xu, B., Kong, X., & Xia, F. (2018). A survey of scholarly data visualization. Ieee Access, 6, 19205-19221.

[21] Chen, Q., Yue, X., Plantaz, X., Chen, Y., Shi, C., Pong, T. C., & Qu, H. (2018). Viseq: Visual analytics of learning sequence in massive open online courses. IEEE transactions on visualization and computer graphics, 26(3), 1622-1636.

[22] Chou, C. Y., Tseng, S. F., Chih, W. C., Chen, Z. H., Chao, P. Y., Lai, K. R., ... & Lin, Y. L. (2015). Open student models of core competencies at the curriculum level: Using learning analytics for student reflection. IEEE Transactions on Emerging Topics in Computing, 5(1), 32-44.

[23] Huang, L., Wang, C. D., Chao, H. Y., Lai, J. H., & Philip, S. Y. (2019). A score prediction approach for optional course recommendation via cross-user-domain collaborative filtering. IEEE Access, 7, 19550-19563.

[24] Xie, T., Zheng, Q., Zhang, W., & Qu, H. (2017). Modeling and predicting the active video-viewing time in a large-scale E-learning system. IEEE Access, 5, 11490-11504.

[25] Hung, J. L., Shelton, B. E., Yang, J., & Du, X. (2019). Improving predictive modeling for at-risk student identification: A multistage approach. IEEE Transactions on Learning Technologies, 12(2), 148-157.

[26] Fincham, E., Gašević, D., Jovanović, J., & Pardo, A. (2018). From study tactics to learning strategies: An analytical method for extracting interpretable representations. IEEE Transactions on Learning Technologies, 12(1), 59-72.

[27] Kaur, D., Aujla, G. S., Kumar, N., Zomaya, A. Y., Perera, C., & Ranjan, R. (2018). Tensor-based big data management scheme for dimensionality reduction problem in smart grid systems: SDN perspective. IEEE Transactions on Knowledge and Data Engineering, 30(10), 1985-1998.

[28] Dong, D., & Herbert, J. (2017). Content-aware partial compression for textual big data analysis in hadoop. IEEE Transactions on Big Data, 4(4), 459-472.

[29] Edstrom, J., Chen, D., Gong, Y., Wang, J., & Gong, N. (2017). Data-pattern enabled self-recovery low-power storage system for big video data. IEEE Transactions on Big Data, 5(1), 95-105.

[30] Yang, Y., & Chen, T. (2019). Analysis and visualization implementation of medical big data resource sharing mechanism based on deep learning. IEEE Access, 7, 156077-156088.

[31] Kumar, S., & Singh, M. (2019). A novel clustering technique for efficient clustering of big data in Hadoop Ecosystem. Big Data Mining and Analytics, 2(4), 240-247.

[32] Parsola, J., Gangodkar, D., & Mittal, A. (2019, July). Mobile Application for Storage and Retrieval of e-learning videos Using Hadoop. In 2019 International Conference on Communication and Electronics Systems (ICCES) (pp. 757-762). IEEE.

[33] Jagtap, A., Bodkhe, B., Gaikwad, B., & Kalyana, S. (2016, January). Homogenizing social networking with smart education by means of machine learning and Hadoop: A case study. In 2016 International Conference on Internet of Things and Applications (IOTA) (pp. 85-90). IEEE.

[34] Tian, X., Cui, B., Deng, J., & Yang, J. (2016, July). The Performance Optimization of Hadoop during Mining Online Education Packets for Malware Detection. In 2016 10th International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS) (pp. 305-309). IEEE.

[35] Cholissodin, I., & Supianto, A. A. (2019, September). Enhancement Full Open Source Hadoop Distribution Universal Big Data Up Projects (UBig) From Education To Enterprise. In 2019 International Conference on Sustainable Information Engineering and Technology (SIET) (pp. 90-93). IEEE.

[36] Wu, C. H. (2019, July). A Concept Framework of Using Education Game With Artificial Neural Network Techniques to Identify Learning Styles. In 2019 International Conference on Machine Learning and Cybernetics (ICMLC) (pp. 1-6). IEEE.

[37] Huang, L., & Ma, K. S. (2018, October). Introducing Machine Learning to First-year Undergraduate Engineering Students Through an Authentic and Active Learning Labware. In 2018 IEEE Frontiers in Education Conference (FIE) (pp. 1-4). IEEE.

[38] Gouripeddi, P. S., Gouripeddi, R., & Gouripeddi, S. P. (2019, December). Toward Machine Learning and Big Data Approaches for Learning Analytics. In 2019 IEEE Tenth International Conference on Technology for Education (T4E) (pp. 256-257). IEEE.

[39] Jeon, H., Oh, H., & Lee, J. (2018, October). Machine Learning based Fast Reading Algorithm for Future ICT based Education. In 2018 International Conference on Information and Communication Technology Convergence (ICTC) (pp. 771-775). IEEE.

[40] Sriyanong, W., Moungmingsuk, N., & Khamphakdee, N. (2018, July). A Text Preprocessing Framework for Text Mining on Big Data Infrastructure. In 2018 2nd International Conference on Imaging, Signal Processing and Communication (ICISPC) (pp. 169-173). IEEE.

[41] Wang, H., Wang, Q., & Wang, W. (2018, September). Text mining for educational literature on big data with Hadoop. In 2018 IEEE International Conference on Smart Cloud (SmartCloud) (pp. 166-170). IEEE.

[42] Çakir, M. U., & Güldamlasioğlu, S. (2016, June). Text mining analysis in Turkish language using big data tools. In 2016 IEEE 40th Annual Computer Software and Applications Conference (COMPSAC) (Vol. 1, pp. 614-618). IEEE.

[43] Larson, R. R., Marciano, R., Hou, C. Y., Watry, P., Harrison, J., Aguilar, L., & Fuselier, J. (2014, October). Integrating Data Mining and Data Management Technologies for Scholarly Inquiry. In 2014 IEEE International Conference on Big Data (Big Data) (pp. 67-71). IEEE.

[44] Cavallaro, G., Riedel, M., Richerzhagen, M., Benediktsson, J. A., & Plaza, A. (2015). On understanding big data impacts in remotely sensed image classification using support vector machine methods. IEEE journal of selected topics in applied earth observations and remote sensing, 8(10), 4634-4646.

[45] Ifenthaler, D., & Yau, J. Y. K. (2020). Utilising learning analytics to support study success in higher education: a systematic review. Educational Technology Research and Development, 68(4), 1961-1990.

[46] Gibson, D., & Ifenthaler, D. (2020). Adoption of learning analytics. In Adoption of data analytics in higher education learning and teaching (pp. 3-20). Springer, Cham.

[47] Beerwinkle, A. L. (2021). The use of learning analytics and the potential risk of harm for K-12 students participating in digital learning environments. Educational Technology Research and Development, 69(1), 327-330.

[48] Lee, L. K., & Cheung, S. K. (2020). Learning analytics: Current trends and innovative practices. Journal of Computers in Education, 7(1), 1-6.