A Preliminary Study on Student Classroom Reading Vs Digital Reading Pattern Behavior Analysis during Pandemic

Divanshi Priyadarshni Wangoo *1 and* S.R.N Reddy *2*

12 Indira Gandhi Delhi Technical University for Women, New Delhi, India

Abstract

This study presented in the paper aims at analyzing reading pattern and behavior analysis of the students in classroom and digital learning environment from the reading behavior log data activities analyzed on the dataset provided for the year 2019 and 2020 courses. The analysis is able to detect reading behavior changes in student reading pattern from the onsite classes in 2019 to the online classes in 2020 in the same education institution based on the overall operation counts, aggregate page transition analysis and time aggregation analysis for the operations and pages ready by the students across two courses A and B for the year before and the year after the pandemic of 2019 and 2020 respectively. The shift to the online learning system and its impact on the student learning is based on various features. The number of page navigation operation counts, page transitions, time range aggregations for page access done by the student for a specific time interval are useful factors for assessing his or her reading pattern. This paper emphasizes on the importance of relevant features in analyzing a student's reading pattern behavior with page transition and aggregation time intervals of students in the offline and online reading activities. Learning Analytics intervention and its integration in the students reading pattern analysis is very important for giving feedback in the way of learner's reading and learning journey. This paper would possibly contribute to the reading learning behavior pattern interpretation of the learner in the world of digital learning which holds the future of smart education. The results are preliminary in nature and with further depth analysis will lead to automated student reading behavior pattern prediction systems using the machine learning techniques. The paper has a reasonable insight into the reading strategies of students during the two different years of onsite 2019 classroom reading activities and 2020 digital reading activities respectively. The results of this study indicate that reading time of a user in digital reading activities is more as with the onsite reading activities whereas the navigation operation counts are more with onsite reading as compared to the digital reading. It will motivate the researchers for further study and analysis for predicting the various student related reading behaviors using Learning Analytics and Machine Learning.

Keywords

Learning Analytics, Reading Behavior, Reading Pattern, Machine Learning

1. Introduction

Reading is an ongoing educational activity that humans embrace since their childhood. It is a constant mechanism for all the stages of human cognitive development from childhood till old age. The Education system has moved from traditional classroom-based systems to the new online based learning. The flipped based learning combines the best of both the worlds. The rise and need for digital education and learning seems to flourish in the coming years of human existence. With the advent of educational technologies and smart education there has been a rising need for assessing the impact of

the technologies on student learning. Intelligent textbooks and reading devices have a plethora of data for analyzing the student's reading behavior and pattern. The impact of digital education on students reading is a matter of concern for all student age groups. The benefits of digital learning are numerous as compared to the in classroom reading but an empirical study is needed to carry out the seed indicators in order for supporting the advantages [9]. But there remains a need to understand the variation of the change in reading patterns of the students in classroom and digital learning scenarios. This paper is a work in progress and aims at empirically validating the difference if any exists between the two learning reading scenarios given in the provided dataset files [6,8] and at the same time find out the reading patterns of the learning. This would give us an insight into the trends associated with digital learning and its impact on the students' performance. The preliminary machine learning classification analysis of the reading patterns classifications for the users of the A-2020 course is presented in the paper and lays fertile ground for further analysis by the researchers in the future work of the study. The present study aims at analyzing the reading pattern changes and reading behavior detection with the analysis done using the Open LA Python library provided by the workshop organizers [6.8]. With the current analysis the following research questions addressed

Research Question 1: What are the reading patterns behavior of the users for a specific lecture during the onsite class reading and comparing it to the online class readings. Research Question 2: What is the reading time patterns of the users based on the during the onsite class reading as compared to the online class readings. Research Question 3: What are the significant features that contribute to the reading pattern behavior analysis of the student. The paper organization follows with the further sections presenting the literature review followed by the reading time behavior pattern evaluations, methods and analysis to answer the above questions with results and discussions presented along with the conclusions. This study fosters There is a need for deriving the reading pattern scores that would be calculating the reading pattern behavior of the students which would be fed into strong machine learning classifiers for building accurate reading behavior detection models. Future work will be motivated by building and deploying of such models for real time analysis.

2. Background

The analysis of the reading behaviors of the learners have been studied in the past educational theories and with the advent of educational technologies have laid a fertile ground to validate their results for a useful scientific study. The results of the study would be helpful in understanding the reading difficulties faced by the students with their present reading behavior patterns and would also help in designing the personalized material by the teachers based on the results and feedback received. Reading behavior analysis holds significant importance in predicting the students' performance. Various researches have been done in the reading behavior analysis of the student such by analyzing learners reading logs and based on the clickstream data various reading profiles have been calculated as demonstrated by Majumdar et al [1]. Learning Analytics for estimating the learner engagement and reading styles in e-books have been studied by Boticki et al., for estimating the higher-level estimation of the skills developed by the learner with reading log analysis [2]. Twilhaar et al., have studied the neurocognitive phenomena underlying with the academic difficulties faced by the children who have preterm born history that could be helpful for understanding the development of corrective behavioral techniques for these children to excel in academics [3]. The integration and significance of Learning Analytics in Learning Management Systems for Educations Institutions have been discussed in length and breadth by Flanagan et al. with the explanation of the working of two platform of Learning Analytics based LMS and Learning Analytics without LMS with their respective outcomes in context to the Educational Institutions practices [4]. Hsu et al. have performed a study for developing monitoring reading concentration with sensor technologies in e-book systems with Artificial Intelligence based optimization approaches and algorithms [5]. Flanangan et al., have focused on the privacy issues in the Learning Analytics and have proposed an authentication-based system approach for the same which holds importance as privacy is a major concern for taking the reading logs of the users with their informed consents [6]. Learning Analytics and its impact in the education systems has seen a rapid growth since its inception and continues to grow in huge successes for the years to come. Ogata et al., have discussed about the importance of educational data and Learning Analytics based E-

book systems for higher education and its active usage in university education [7]. Ogata et al. have studied the E-book for Learning Analytics for Higher Education System practices and its impact on incorporating the educational big data research into current educations practices [8]. The importance of the digital reading and its comparison with the offline reading has been variedly studies which can be seen from the study of Coiro et al., emphasizing on the advantages of digital learning [9].The reading style indicators are necessary to derive based on the log data as done by Boticki et al., which can be further validated with the help of large datasets [10] Which Parameters are important among a list of parameters requires a further analysis for its relevance into the using of the machine learning techniques in reading behavior analysis.

3. Reading Time Behavior Pattern Evaluation

The Reading time of the students in the pandemic has more shifted towards digital learning. There comes a need to access the impact of the digital book reading on the normal reading behavior of the student and at the same time measure its change with the traditional reading systems of the classroom. The change in variation can be seen more importantly from the variation in the reading time over a range of years of the student reading activities. This section discussed a preliminary analysis of the students reading pattern with the study of the reading pattern of the student. The page transition of students with time is being analyzed and studied for the courses during the year 2019 and 2020 as provided in the dataset [6,8].

### 3.1. Reading Behavior Change Analysis

The change in the reading behavior of the learners can be seen from the operation counts, page transition and page aggregation functions as shown in Figure1-4. The Figure 1 below gives the average operation count of all the users in all the contents across both the courses of 2019 and 2020. The X-axis label defines the operation count with values open, close, add marker, add memo and add bookmark. The Y-axis represents the content id and depends on the number of contents in the particular course and year. It can be clearly observed that the number of open and close navigation operations are higher in the Year of 2020 for online classes as compared to Year 2019 for onsite classes. The A- 2019 course open and close navigation operations are similar to A-2020 and B-2020 courses and differ in the add marker operation. The teacher and the contents uploaded for the particular course can be a reason for the similar open and close operations counts. Also, it can be seen that the add bookmark has the highest count in the B-2019 course as compared to the other courses. Therefore, it can be inferred that the reading pattern behavior difference of the users in for the A-2020 and A-2019 courses in terms of open operation and varies very significantly for the close operation as the students may have the same open navigation pattern across the two years. Also add bookmark operation is present in the A-2020 course whereas add marker operation is present in the A-2019 course. Similarly for the B-2020 course there can be seen a significant difference between the number of operations counts for the open and close operation which is highest in B-2020 course as compared to the add marker operation highest for B-2019 course followed by all other bookmarks.

X-axis = operation count
Y-axis = content id

B-2020 Course

B-2019 Course
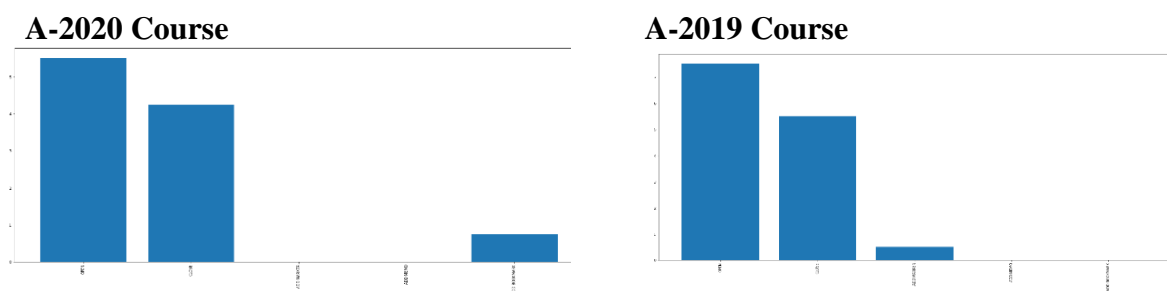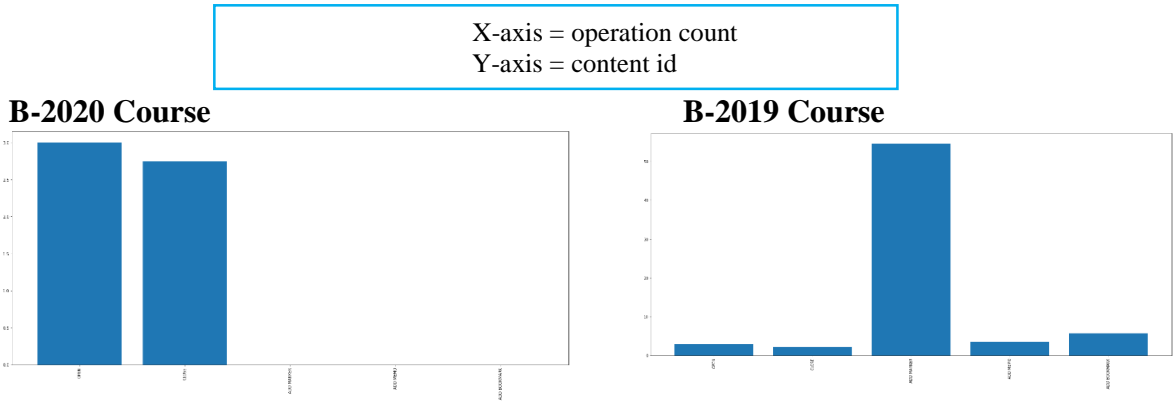
Figure 1: Reading Pattern Analysis based on the operation count taking the average operation count of all users in all contents for the A-2020 and B-2020 online classes comparing with A-2019 and B-2019 online classes where X-axis represents the operation type and Y-axis represent the Contents ID

From the Figure 2 below the average reading time using the page-wise aggregation module. The average operation count of all the users in each page are plotted with page number on the X-axis, average score of the operation count on the Y-axis and reading minutes in the page for all the users accessing a particular content which is around the same lecture time across all the lecture weeks of 1-8 in 2019 and 2020 courses. As can be seen from the plots the average reading time in the 2019 courses are higher as compared to the 2020 courses. Also, the number of close operations is higher in A-2020 as in A-2019 courses going along with the reading minutes in the page as students might be reading quickly and closing the page to move on to the next page. The number of open operation count is higher in A-2020 course. Also add marker and add bookmark operation has been used by the more by the students of the 2019 courses as compared to the 2020 courses. Thus, it can be inferred that the onsite classroom reading activities in terms of the operations counts for a particular week content is read more often by the 2019 courses as compared to the 2020 courses. From the Figure 2 graph it has been observed that there are a smaller number of navigation operation activity in 2020 courses which can be a probable measure of good flow and reading concentration in the online reading as compared to the onsite reading activity.

A-2020 Course

A-2019 Course

X-axis = page
Y-axis = Operation count

B-2020 Course

B-2019 Course
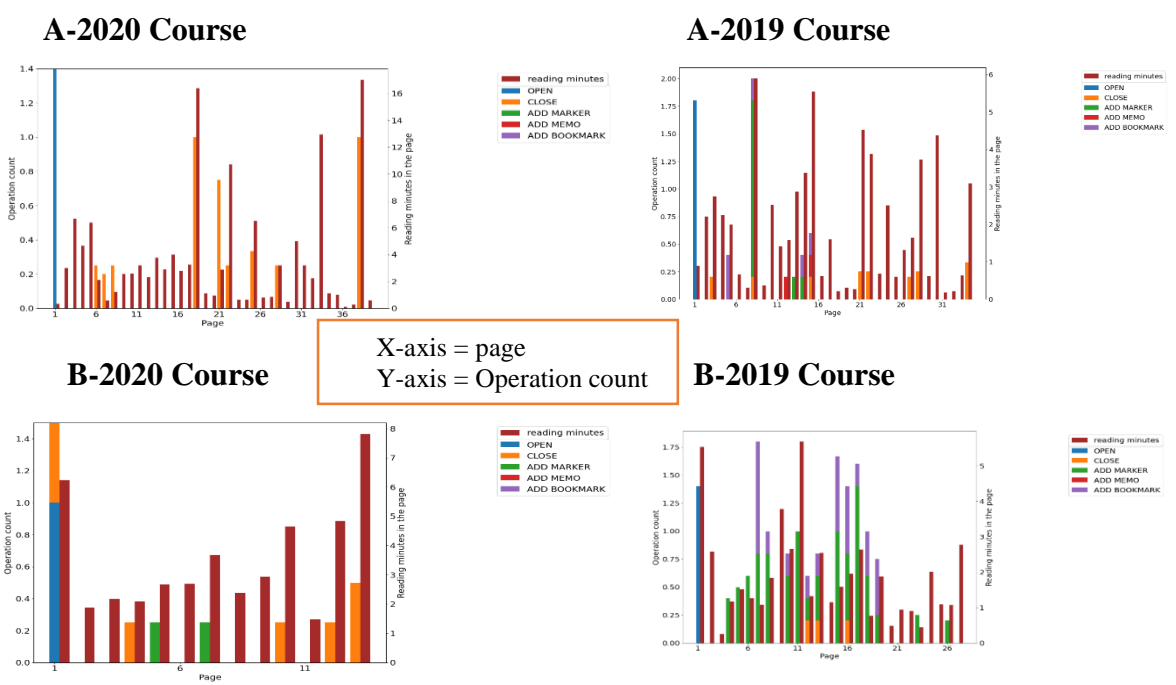
Figure 2: Reading Pattern Analysis based on the Page -Wise Transition Aggregation calculated by the average operation count of all users in each page for the A-2020 and B-2020 online classes comparing with A-2019 and B-2019 online classes

As can be seen from the Figure 3 and with the help of the time range aggregation module of the Open LA library by calculating the average operation count of all the users in a particular lecture which is

taken across the same week for all the two courses of the two years, the next and previous operation counts are higher in the initial minutes of 0-50 for A-2020 as compared to 50 to 100 for A-2019 and higher in later minutes of 50 to 90 for B-2020 courses as compared to 70 to 90 for B-2019 courses. Therefore, it can be inferred that the students are initially active during the initial time periods in digital reading whereas in off line reading the students become active in the middle of the reading intervals. Also, as can be seen from the Figure 4 graphs, the pages the user read in each time for a particular lecture is higher in the 50–150-minute time interval for A-2020 and A-2019 courses, 70 -90 B-2020 and 50 to 90 for B-2019 courses. It can be inferred that the users read more pages in the mid time interval after gaining momentum in the initial elapsed minutes.
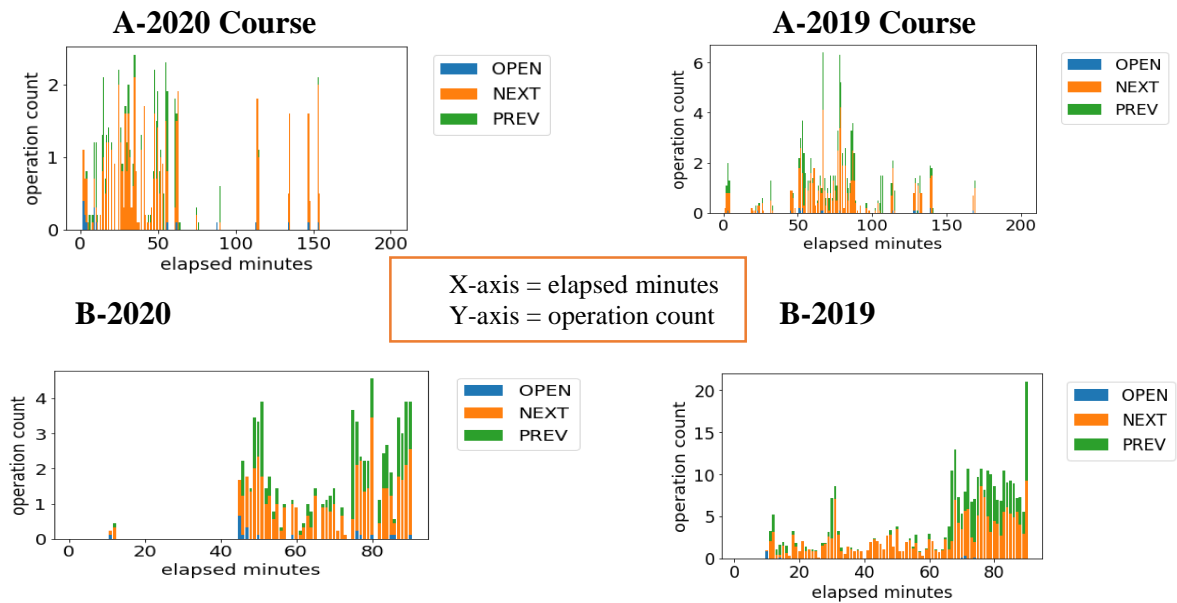
Figure 3: Reading Pattern Analysis based on the Time range aggregation calculated by the Average Operation count of all the users in Lecture 1 for the A-2020 and B-2020 online classes comparing with A-2019 and B-2019 online classes
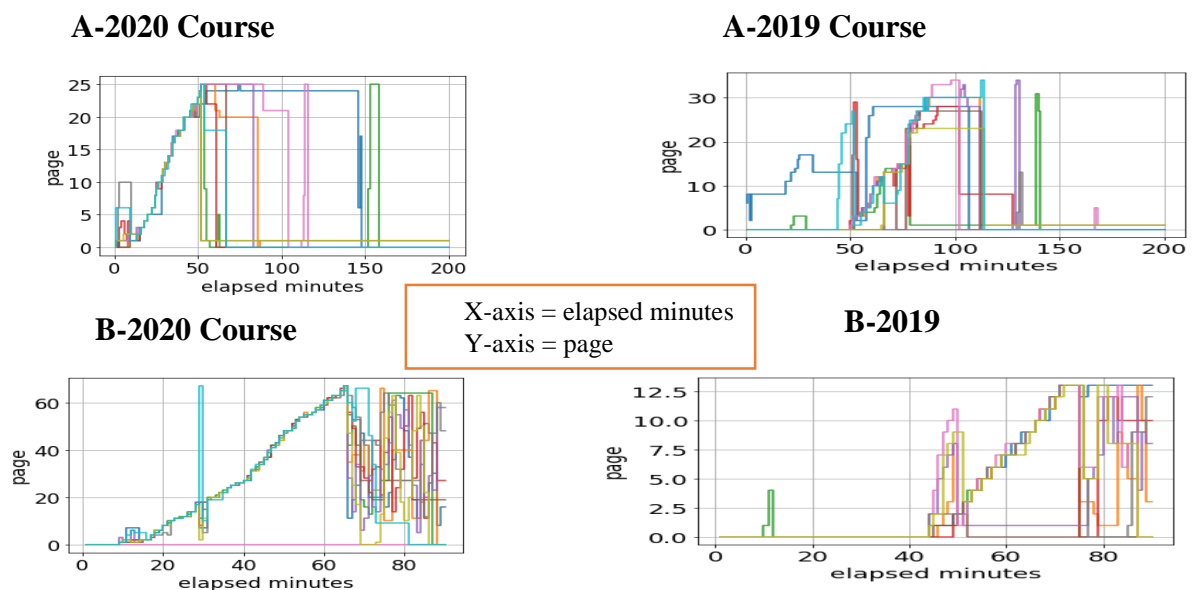
Figure 3: Reading Pattern Analysis based on the Time range aggregation calculated by the pages users read in each time in Lecture 1 for the A-2020 and B-2020 online classes comparing with A-2019 and B-2019 online classes.

3.2 Variation in Students Reading Pattern

The variation in the student reading patterns has been observed through the data aggregation module methods of the Open LA library which is described in detailed in the further subsections. Although more data analysis and mathematical derivations are needed to infer the claims, it can be partially inferred that the reading pattern between the classroom and digital learning activities varies significantly in terms of the number of pages viewed, average reading time, navigation pattern of operation counts, page wise aggregation and time range aggregation of the reading activities. A further work using machine learning, feature engineering and prediction techniques will be able to provide us with a precise impact of the above on the students' performance based on reading behavior.

4. Methods and Analysis

This section presents the methods used for the dataset analysis. The basic methods of the Open LA library have been used in carrying out all the processing related to the operation counts, page wise transition aggregation and time-range aggregation of the data modules of the library. The Table 1 and Table 2 below describe the dataset files and the features used which is same for all the dataset files of Course A and Course B for 2019 and 2020 years. The comparison thus holds valid as the same variables are being compared with around the same week contents for all the obtained results.

Table 1 Dataset Files Types

Dataset File Type (.csv files)	Years (Online/Onsite Class Years)	Courses	Features
EventStream	2020 / 2019	A, B	userid, contentsid, devicecode, marker, eventtime, operationname, pageno
LectureMaterial	2020 / 2019	A, B	lecture, contentsid, pages
LectureTime	2020 / 2019	A, B	lecture, starttime, endtime
GradePoint	2020 / 2019	A, B	userid, grade

With the increase in number of navigations for the 2019 courses it can be inferred that the users have high navigation pattern in the on-site classes as compared to the online classes. The smaller number of navigations in 2020 courses can probably amount to the increase concentration and flow with less distractions in online reading activity.

Table 2 Dataset Description

File Type	Features	Feature Type
EventStream	userid	U1-163 (Overall)
	contentsid	C1 – C20(A-2020 Courses), C1-C20 (A-2019 Courses), C1-C9(B-2020 Courses) and C1-C8(B-2019 courses)
	devicecode	'pc', 'mobile', 'tablet'
	marker	'rgb (255,0,0)', 'rgb (255,255,0)', 'rgb (0,255,255)'
	eventtime	Timestamps
	operationname	'OPEN','NEXT','PREV','GETIT','ADD BOOKMARK','CLOSE','PAGE_JUMP','NOTGETIT','ADD MARKER','DELETE, BOOKMARK','BOOKMARK_JUMP','DELETE MARKER','SEARCH','SEARCH_JUMP','ADD MEMO','CHANGE MEMO','DELETE_MEMO'
	pageno	Varies
LectureMaterial	lecture	1-6(2020 Course_A and 2019 Course_B), 1-7 (2019 Course_A) and 1-8 (2019 Course_B)
	Contentsid	C1 – C12(2020 Courses) and C1-C21 (2019 Courses)
	pages	Varies
GradePoint	grade	U1-U50
	userid	A-F
LectureTime	lecture	1-7(2020 Courses) and 1-8 (2019 Courses)
	starttime	Timestamps
	endtime	Timestamps

Table 3

Reading Behavior Analysis in terms of total reading time in hours, minutes and seconds along with total reading seconds in the data and number of unique pages for the mid-week lecture contents of a Particular User U_12 Across all the Courses in 2019 and 2020

Courses	User	Content	Total Reading Time in the Data (in hours)	Total Reading Time in the Data (in minutes)	Total Reading Time in the Data (in seconds)	Total Reading Seconds in the Data	No of Unique Pages in the Data
2020 Courses_A	U12	C6	1.8258	109.55	6573	6573	33
2019 Courses_A	U12	C3	0.703	42.233	2534	2534	33
2020 Courses_B	U12	C6	1.7355	104.133	6248	6248	13
2019 Courses_B	U12	C3	1.549	92.9666	5578	5578	54

Table 4 Analysis of the Operation Count in Order of Increasing Sequence Access of the Operation Mentioned in the Table Legends of Figure 1-3

S. No	Data Function Type	A-2020	A-2019	B-2020	B-2019
1.	Total Operation Count	O,C,AB	O, C,AM	O,C	AM, AB, AMe,O,C
2.	Total Page Transition Page-wise Aggregation	O,C	O, AM, AB, C	O, AM, C	AB, AM, O, C
3.	Total time-range Aggregation- Average Operation Count of all the Users in Lecture 1	N,P,O	P, N, O	N,P,O	P,N,O

Table Legends* O=Open, C=Close, N=Next, P=Previous, AM=Add Marker, AMe=Add Memo, AB=Add Bookmark

The reading minutes of the same user in the middle of the course around week 4 can talk about the motivation of the student for continuing reading. For 2020 courses the reading minutes across a particular content in the same lecture week are higher as compared to th2 2019 courses. With A-2020_C6, A-2020_C7, B-2020_C6, B-2020_C7 having higher average reading minutes as 54.2, 0.78, 4.05 and 3.13 respectively whereas with A-2019_C3, B-2019_C3, B-2019_C4 courses the reading minutes are 3.433, 0.366 and 25.566 respectively. Therefore, it can be inferred that the reading minutes of the user are higher in 2020 as compared with 2019 courses.

Table 5 Variation in Reading Times between Lecture 4 Around Middle of the Lecture Weeks of the Courses of a Particular User U-11 for two Contents C6 and C7 for 2020 Courses and Courses 2019

userid	contentsid	reading_minutes
A-2020_U12	A-2020_C6	54.2
A-2020_U12	A-2020_C7	0.7833333333333333
A-2019_U12	A-2019_C3	3.433333333333333
B-2020_U12	B-2020_C6	4.05
B-2020_U12	B-2020_C7	3.1333333333333333
B-2019_U12	B-2019_C3	0.36666666666666664
B-2019_U12	B-2019_C4	25.566666666666666

5. Results and Discussions

As discussed in the previous sections the reading behavior analysis done is discussed here. The change in the reading behavior of the learners can be seen from the operation counts, page transition and page aggregation functions as shown in Figure1-4. It can be clearly observed that the number of open and close navigation operations are higher in the Year of 2020 for online classes as compared to Year 2019 for onsite classes. Similarly, the average reading minute per page is higher in 2020. The average operation count of all the users and the pages the user read in each time in Lecture 1 are most accessed for the online reading. Also, as seen from the results in Table 3 the number of reading hours, minutes and seconds are higher for courses taken in 2020. The user navigation with the operations counts are more for the open, add bookmark, next and previous operations that clearly signifies increase in navigation activities by the learners in the online reading systems. To further analyze the importance of the study for educational reading-based application an initial machine learning classification has been done for the 2020 Year page transition dataset processed through the OpenLA library and derived from the operations to classify the users based on their contents navigated, their average reading time, the number of access operations and page numbers. The initial results and accuracy achieved by the classifiers are described in Table 6 along with their ROC curves in Figure 6 below. The research questions that were presented for the purpose of this study can be answered with this study analysis.

Firstly, for answering the Research Question 1 related to "what are the reading patterns behavior of the users for a specific lecture during the onsite class reading and comparing it to the online class readings" can be seen from the reading patterns behavior of the users based on the total operation counts, total page transitions aggregation operation counts and total time-range aggregation operation counts and pages users read in each time interval of one minute for a specific lecture during the onsite class reading and comparing it to the online class readings. Secondly, for the Research Question 2 related to "What is the reading time patterns of the users based on the during the onsite class reading as compared to the online class readings" can be seen from the reading time patterns of the users based on the average reading score during the onsite class reading as compared to the online class readings of a specific user. Lastly, for answering the Research Question 3 related to "what are the significant features that contribute to the reading pattern behavior analysis of the student" the reading minutes, average reading time and average operation count are relatively important features for reading pattern behavior analysis. For validating the above claim further machine learning based feature importance analysis using Random Forest classifiers would be required as a suitable solution.

As the log dataset files for the event stream files correspond to 13, 937 instances and also for model building for various types of classification and prediction models there arises a need for the machine learning based analysis for the same. Therefore, this paper also presents an initial work on the implementation of machine learning models for classifying the users based with their unique userid on the features such as contentsid, page number, number of page visits, average reading seconds, reading seconds and type of navigation operations such as open, close, add marker, add memo and add bookmark. This is an initial analysis work for only one course of the year 2020 using the page wise

aggregation module of the data conversion module of the Open LA library. The page wise aggregation stores the page navigation behavior by calculating the total staying seconds and operation count per page of the users in relation to the contents accessed by the user. For modeling various types of machine learning algorithms were used to train the model ranging from Neural Networks, Support Vector Machines (SVM), Decision Trees to Ada Boost that were implemented using MATLAB software. The initial analysis was done using the training data which resulted in a validation accuracy of 98.0% with Neural Network classifiers with one fully connected layer of size 100, 84.1% with Fine Gaussian SVM, 81.5% with Decision Trees and 71% with Ada Boost classifiers. The Neural Network Classifier gives the highest accuracy among all the classifiers. But with the 10 Cross-Fold Validation the AdaBoost algorithm achieved the highest accuracy of 64.4% among all the classifiers. For different educational datasets classifications and prediction-based applications, a high accuracy model is the need of the hour. Thus, the future work lies in the choosing of various other models and increasing the validation accuracy as well as testing our trained model on the unseen test data. These kind of behavior-based prediction models will help in providing valuable feedback to the students in knowing where they are spending most of their time in reading and at the same time would help teachers keep track of their students reading behavior pattern to prepare schedule according to the specific user needs. The Table 6 below lists the accuracy results followed by the ROC curves of the classifiers processed in MATLAB.

Table 6 Users Classifications based on the Page Transition Behavior of Users (Four Unique Users) of A-2020 course with Machine Learning Classifiers in MATLAB

Machine Learning Classifier	Accuracy Validation	Accuracy 10 Cross-Fold Validation
Neural Network Classifier with one fully connected layer of size 100 With ReLU Activation	98.0%	59.5%
Fine Gaussian SVM	84.1%	59.4%
Decision Tree	81.5%	62.5%
Ada Boost (Ensemble Boosted Trees)	71%	64.4%

X-axis = False Positive Rate
Y-axis = True Positive Rate
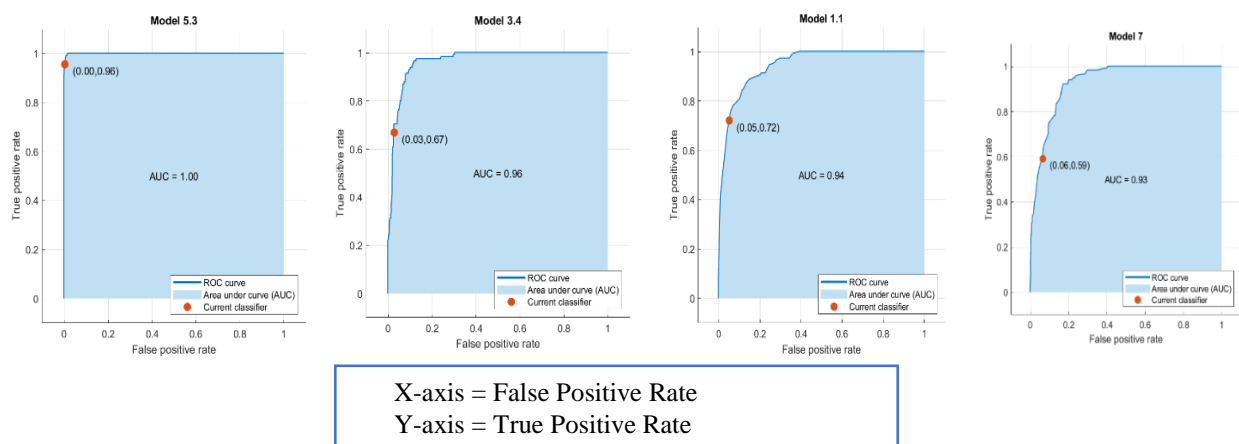
Figure 4: ROC Curves of the Classifiers in MATLAB (a) Neural Network (b) SVM **(c)** Decision Trees (d) AdaBoost

6. Conclusion and Future Work

This aim of this paper is to do a preliminary analysis study for emphasizing on the importance of change in reading patterns and behavior of the students in the classroom and digital reading activities.

The number of operation counts, page transition, page access done by a student is an important factor for assessing his or her reading pattern. This paper emphasizes on the importance of relevant features in analyzing a student's reading pattern behavior with page transition and aggregation time intervals of students in the offline and online reading activities. The results are preliminary in nature and with the completion of the work with lead to the final results using machine learning techniques. There is a need for deriving the reading pattern scores that would be calculating the reading pattern behavior of the students which would be fed into strong machine learning classifiers for building accurate reading behavior detection models. Future work will be motivated by building and deploying of such models for real time analysis.

7. References

[1] Majumdar, R., Bakilapadavu, G., Majumder, R., Chen, M. R. A., Flanagan, B., & Ogata, H. (2021). Learning analytics of humanities course: reader profiles in critical reading activity. Research and Practice in Technology Enhanced Learning, 16(1), 1-18.

[2] Boticki, I., Akçapınar, G., & Ogata, H. (2019). E-book user modelling through learning analytics: the case of learner engagement and reading styles. Interactive Learning Environments, 27(5-6), 754-765.

[3] Twilhaar, E. S., De Kieviet, J. F., Van Elburg, R. M., & Oosterlaan, J. (2020). Neurocognitive processes underlying academic difficulties in very preterm born adolescents. Child Neuropsychology, 26(2), 274-287.

[4] Flanagan, B., & Ogata, H. (2018). Learning analytics platform in higher education in Japan. Knowledge Management & E-Learning: An International Journal, 10(4), 469-484.

[5] Hsu, C. C., Chen, H. C., Su, Y. N., Huang, K. K., & Huang, Y. M. (2012). Developing a reading concentration monitoring system by applying an artificial bee colony algorithm to e-books in an intelligent classroom. Sensors, 12(10), 14158-14178.

[6] Flanagan, B., & Ogata, H. (2017, November). Integration of learning analytics research and production systems while protecting privacy. In The 25th International Conference on Computers in Education, Christchurch, New Zealand (pp. 333-338).

[7] Ogata, H., Yin, C., Oi, M., Okubo, F., Shimada, A., Kojima, K., & Yamada, M. (2015, November). E-Book-based learning analytics in university education. In International conference on computer in education (ICCE 2015) (pp. 401-406).

[8] Hiroaki Ogata, Misato Oi, Kousuke Mohri, Fumiya Okubo, Atsushi Shimada, Masanori Yamada, Jingyun Wang, and Sachio Hirokawa, Learning Analytics for E-Book-Based Educational Big Data in Higher Education, In Smart Sensors at the IoT Frontier, pp.327-350, Springer, Cham, 2017.

[9] Coiro, J. (2021). Toward a multifaceted heuristic of digital reading to inform assessment, research, practice, and policy. Reading Research Quarterly, 56(1), 9-31.

[10] Boticki, I., Ogata, H., Tomiek, K., AKCAPINAR, G., Flanagan, B., MAJUMDAR, R., & HASNINE, N. Identifying Reading Styles from E-book Log Data.