Lappeenranta-Lahti University of Technology LUT
LUT School of Engineering Science
Industrial Engineering and Management
Business Analytics

Arnob Islam Khan

DATA DRIVEN DECISION MAKING IN DIGITAL EDUCATION: A CASE STUDY FROM FINLAND AND RUSSIA

Master's Thesis

Updated 19.07.2019

Examiner(s): Professor Leonid Chechurin

ABSTRACT

Lappeenranta-Lahti University of Technology LUT
Industrial Engineering and Management
Business Analytics

Arnob Islam Khan

DATA DRIVEN DECISION MAKING IN DIGITAL EDUCATION: A CASE STUDY FROM FINLAND AND RUSSIA

Master's thesis

2019

81 pages, 32 figures and 14 table

Examiners: Professor Leonid Chechurin

This Master's thesis is contemplated as a part of CEPHEI project. One of the major goals of the project is to increase the digitalization the course contents of Industrial Innovation by standardization of its e-learning elements. The assessment of course design and students' performance is one of the integral part of e-learning/digital course. However, the standards in learning assessment for digital courses remains vague. For teachers to find the proper evaluation method for online course has become a big challenge. It is crucial to determine the learning outcomes and check, whether the e-learning deliverable serve its purpose or not. Evaluation of eLearning course makes it possible to assess its quality and efficiency and, most importantly, to comprehend what modifications and improvements are needed.
The objective of the thesis is to evaluate a supervised learning based assessment methods for digital education. The method is evaluated based on the available data of "Systematic Creativity and TRIZ basics" course at Lappeenranta University of Technology, Finland and "Computer Science" course at Tomsk State University of Control Systems and Radioelectronics, Russia. This work is an attempt to illustrate the obtained results by the described evaluation methods and provide an initial speculation on the usability of learning analytics.

Keywords: Social-network analysis; Online-discussion forums; Online learning; Sentiment analysis; NLP; Moodle; Decision Tree; KNN; Regression

ACKNOWLEDGEMENTS

I would like to start by expressing my gratitude to the almighty for his blessings. Firstly, I believe no word is enough to express my deep gratitude to Professor Leonid Chechurin, my supervisor and mentor who always has been a great source of inspiration, motivation, knowledge and guidance since my first days of Master's degree at LUT University. I would like to say special thanks to my colleagues Iuliia and Vasilii, whose help and support made this work possible. This work was partially supported by CEPHEI project of ERASMUS+ EU framework. My grateful thanks to all the project members of CEPHEI for their cordial support and sharing knowledge throughout the work. It was my privilege and honor to get this excellent chance to study at LUT University. My thanks and appreciation to all the faculty members, staff and peer students at LUT who were excellent source of knowledge and support throughout my study period.

I would like to give special thanks to my parents and younger brother for their endless encouragement and blessings. Lastly, my warmest thanks to my wife for her continuous support. Finally, thanks to my friends at LUT who made me feel like home.

Arnob Islam Khan
Lappeenranta 06.07.2019

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS AND ABBREVIATIONS

AR	Audience Retention
CNN	Convolution Neural Network
EDM	Educational Data Mining
GDPR	General Data Protection Regulation
KNN	K-nearest Neighbor
LA	Learning Analytics
LDA	Latent Dirichlet Allocation
LMS	Learning Management System
ROMA	Rapid Outcome Mapping Approach
SNA	Social Network Analysis
SVM	Support Vector Machine
VBL	Video Based Lecture Series

1 INTRODUCTION

In the new era of digitalization, education sector is experiencing changes in terms of learning design, teaching methods, engagement of the learners and integration of technology. Due to worldwide digital shifts and the need of constant rising "YouTube learners", the ways of teaching and education experience have undergone through rapid reformation. New pedagogic standards, innovative practices and methodologies in education have emerged as traditional classroom learning has transited to blended or e-learning. Online learning provides opportunities for students and teachers from around the world to increase learning efficiency(Beichner et al., 2007). In traditional classroom courses, available teaching/participation hour of the teachers and adopting more number of students is always a contradiction. Transition from traditional course to online or flipped format, allow teachers to scale their courses for thousands of students with same teaching hour. At the same time, students also enjoy flexibility in learning with respect to time and geographic location. It facilitates more participatory and interacting learning materials for the students. It allows students to learn at their own pace as they can go multiple times through the various pieces of the material and are not bound to time and place (Prober and Khan, 2013; Antonova, Shnai and Kozlova, 2017). Converting traditional classroom content to e-learning contents is not a new concept anymore. This new form of learning has been adapted and accepted globally by the learners' community. Different pedagogic models (e.g. ADDIE Model) have also been introduced by the educational researchers and instructional technologists to develop the e-learning courses. Due to advantages in teaching and learning process, the format of online learning is gaining increasing popularity(Bergmann and Sams, 2009) which opens up new research problems, such as appropriate ways of evaluating students' performance in a flipped or online learning(Ferguson, 2013).

Unfortunately, the proper assessment method to evaluate the learning of these e-learners has become a big challenge in this new form of learning. Due to lack of proper evaluation method, the e-learning courses may fail to serve its main purpose, which is providing effective learning to the learners through a powerful and memorable digital learning experience. Since course teacher is not able to engage in one to one personal interaction, he/she does not have many options for measuring students' activity during the course.

Obviously, one teacher cannot answer all the questions from thousands of students and assess their work. Therefore, the courses are designed in such way that teachers do not need to participate in grading. There are automatic ways of assessment, such as tests and quizzes. However, these methods of evaluation provide insufficient information about students' learning process and relate much on students self-grading, and self-accountability. It also restricts the teacher to gain valuable insights whether the delivered teaching materials provide desired learning outcome or not. As a result, the teachers conduct continuous development and upgrade of course contents based on intuitions and experiences rather than following systematic process.

On the contrary, the potential of the data analysis in the virtual environment is enormous. Practically each click can be traced and described. With the analytical tools and systems, embedded in each digital medium, huge amount of data can be gathered, systematized and visualized much easier. As discussed earlier, the traditional evaluation approaches like surveys or scores tracking do not assess even approximate half of the available data for the virtual course. Whereas, learning analytics sources like learning management systems (lms), video host and discussion forum provide quantitative, precise information about student experience. Since there is no standards in learning assessment, it has become a challenge for the teachers to utilize the proper evaluation method in order to capture all these information and make sense out of it. To determine whether the learning content/the course properly satisfies the learning outcomes or not, evaluating is crucial. It makes possible to assess quality and efficiency of the learning contents and, most importantly, to comprehend what modifications and improvements are needed.

This work is conducted under EU ERASMUS+ project: Cooperative eLearning Platform for Higher Education in Industrial Innovation (CEPHEI). It is a consortium of 9 (nine) universities: "Lappeenranta University of Technology (LUT), Peter the Great St. Petersburg Polytechnic University (SPbSPU), University of Twente (UoT), Tomsk State University of Control Systems and Radioelectronics (TUSUR), Royal Institute of Technology (KTH), MEF University (MEF),Tianjin University (TJU), Gubkin Russian State University of Oil and Gas (GUBKIN), Hebei University of Technology (HEBUT)". One of the major goal of the project is to digitalize the educational contents related to industrial innovation in Partner Universities and EU. The project also focuses on developing standards covering both

technological and course content related aspects. These standards can be used as a guideline to create digital courses. The assessment/evaluation method in an online course is significant in terms of assessing the quality of course contents and measuring learners' performance. This work is an initial attempt to develop suitable evaluation methods of an online course that can be beneficial for the teachers to continuously improve the course contents and assess the students' performance. This report describes analysis results based on the proposed evaluation methods of two courses in Finland and Russia. One of the course is of "Systematic Creativity and TRIZ basics" at LUT University, Finland. From 2011 to 2015, the course was conducted in traditional class settings. In 2016, it was gradually converted to flipped classroom course then to online course. The course designers and teachers introduced different evaluation method, constantly improving student experience by the formula: Data Collection + Editing = eLearning Course Improvement. Another is "Computer Science" course at Tomsk State University of Control Systems and Radioelectronics, Russia. The data was gathered from different sources. The authors collected the data from surveys, general statistics, observations, learning management systems, specific experiments and scores tracking.

1.1 Objective of the Thesis

The objective of the thesis is to evaluate a supervised and unsupervised learning based assessment method for digital education. The method is evaluated based on the available data of "Systematic Creativity and TRIZ basics" course at Lappeenranta University of Technology, Finland and "Computer Science" course at Tomsk State University of Control Systems and Radioelectronics, Russia.

1.2 Research Question and Framework

The primary research question of this work is, "How learning analytics can lead to systematic improvement of the course contents and assessment of students' performance." Based on the primary research question, there are three sub-questions illustrated in Table 1,

Table 1. Research questions

Research Questions (RQ)	Goal	Hypothesis
RQ1: How to engage/hook up audience in the video?	Identifying the parameters, which illustrates audience interaction with the video contents and improve video contents accordingly.	Video length corresponds with audience retention.
RQ2: How to assess the students' online discussion engagement?	Providing a score/set of score and ranking for each of the student based on their participation in the online discussion forum. This will help the teachers' to assess students' performance in the discussion more numerically based on certain criteria. It will also indicate the interactivity of the course.	High frequency of meaningful words leads to high degree of students' engagement in an online discussion forum.
RQ3: How to predict learners' success based on the LMS activity?	Understanding the activities of the learners' in the LMS and modeling of a learning path of the students to predict their probability of success. Another goal is to speculate the frequency of usage of the LMS to complete the course.	Frequency of activity in LMS predicts has higher degree of correlation with learners' final score.

Figure 1 illustrates the proposed framework of assessing the research questions.

Figure 1. E-learning/flipped learning analytics and assessment framework

1.3 Structure of the thesis

The thesis report is divided into six chapters focusing on their output. The breakdown of the chapter is given in below,

Table 2. Structure of the thesis

Chapter	Aim
1- Introduction	Provides the background of the research. It also covers the goal and delimitations of the research work.
2- Background and Related Work	Mainly explains the literature related to this research work of learning analytics in education. This chapter also explains the common methodology and approach related to learning analytics explained in these research works

Table 3 continues. Structure of the thesis

Chapter	Aim
3- Proposed method and Framework	Illustrates the proposed methods and framework. This chapter also explains the goal of each methods aligning with the specific research questions
4- Data Ethics	Depicts the significance and effect of data ethics and privacy in adoption of learning analytics
5- Results and Discussion	Explains analysis results and implication in answering the research questions
6- Conclusion and Future Work	Summaries the whole work and possible future work

2 BAKGROUND AND RELATED WORK

Due to its increasing popularity, digital learning has become an essential part of the educational process around the world. Traditional lecture based teaching also referred as Teacher-centered approach lacks in keeping the students' attention and intriguing their involvement (Sams and Bergmann, 2012). It allows passive participation of students that limits the learners' learning and understanding since the knowledge transfer in this process primarily based on pure listening(Bergmann and Sams, 2009). According to "Dale's Cone of Learning", students' only remember 20% of the lecture without involvement of any active participation(Beichner *et al.*, 2007). Literature suggests that students' active engagement is significant for learning and one of the principal component of effective teaching(Bryson and Hand, 2007; Early, 2011). Another challenge of traditional classroom is lack of resource (classroom, technology, teacher's time) for large scale of students and opportunity to provide tailored lessons for each individuals' as per their need.

2.1 Different Types of Digital Learning

Educational institutions worldwide are constantly trying to work around the challenges of traditional learning thus facilitate effective teaching. In order to address these challenges, educational sector has experienced shift in its learning design in many formats from Traditional learning design to online/flipped/blended learning design based on the context and target group of students.

2.1.1 Online Learning/ E-learning

In this type of digital learning, the learning process is facilitated through distance by the use of internet. The term e-learning coined up in the 90s and gained its popularity because of its flexibility in geographic locations and time. The learning environment can be both synchronous and asynchronous. All the learners connect at the same time in synchronous learning. On the contrary, participants are engaged in learning at different times in asynchronous environment. It also allows students' to be involved in active learning rather than passive learning in traditional classroom(Mason, 2014). Cisco CEO John Chambers

stated that E-learning removes time and distance barriers, creating universal, on-demand learning opportunities for individuals, businesses and countries(Galagan, 2001).

MOOC(Massive Online Open Course) is the new evolution of E-learning, which provides open access to the platforms and allows interactive engagement of the participants through online(Adamopoulos, 2013). MIT first initiated the movement towards open education by uploading 50 courses at their open courseware (OCW) in 2002. After a year, MIT uploaded more 500 courses and other universities around the world started to follow the new business model of e-learning. (Pantò and Comas-Quinn, 2013) Now a days, MOOCs focuses on three aspects: sharing knowledge and learning materials around the world, cooperative programs and process standardization (Baron, Willis and Lee, 2010). There has been extensive number of researches to explain the different dimensions of MOOCS. Cormier and Siemens discussed how the role of professors have changed in online education and course design for positive student experience(Cormier and Siemens, 2010). Xu and Jaggars analyzed the students' performance in online contrast to face-to-face course(Xu and Jaggars, 2014). In another study, (Mak, Williams and Mackness, 2010) studied the significance of blogs as a communication and interaction tool in MOOC.

2.1.2 Flipped Learning

Flipped Classroom is one of the most significant innovations of this decade in learning pedagogic design. This concept emerged from its predecessor blended learning following the ongoing trends of distance learning and MOOCS (Massive Open Online Courses). "Flipped Classroom" was first introduced by Bergman and Sams in 2012 (Sams and Bergmann, 2012). Others termed it as inverted classroom, Post-lecture classroom(Lage and Platt, 2000; Plasencia and Navas, 2014). It can be also regarded as blended learning. The core principle of "Flipped Classroom" is inverting the common instructional activities (lectures) of class in advance via video lectures and interactive homework. As a result, the students can participate in active engagement, discussion, peer-to-peer collaboration and problem solving which deepens their knowledge and practical skills (Tucker, 2012; Jensen, Kummer and Godoy, 2015). The number of research publication on "Flipped Classroom" has been increasing exponentially each year and Scopus has already reached 2339 publication in 2019. United States is the leading country in research and implementation on flipped concept, it has published one third of the total published paper. The literature also

suggests that “Flipped Classroom” has primarily been popularized widely at K-12 education systems in the United States(Bergmann and Sams, 2009; Ash, 2012). Researcher suggests that it allows the teacher to invest more time in class focusing on student-centered activities, deliver tailored lessons for each individuals’ need and individualized advocacy in project based education(Prober and Khan, 2013). There has been numerous attempts and research to employ “Flipped Classroom”. North Carolina University and Dozon of schools collaborated under SCALE-UP Project to prepare and disseminate alternative educational curriculum in the field of physics, chemistry, and biology. In their research, they have observed significant increase in students’ conceptual knowledge and performance especially for females and minorities in “Calculus-based Introductory physics” course at North Carolina State University(Beichner *et al.*, 2007). Researchers at Pennsylvania State University implemented “Flipped Classroom” in undergraduate architectural engineering class and students’ feedback illustrates that additional time for project activity in class improved their understanding(Zappe *et al.*, 2009). In undergraduate and postgraduate education in medical centers of USA, “Flipped Classroom” principal was implemented to enhance learning and overcome the constrains of residency duty hours(Martin, Farnan and Arora, 2013). Prober and Khan proposed a “Flipped Classroom” model for medical education which was highly favored by the students’ compared to traditional lectures(Prober and Khan, 2013). Critz and Wright introduced first approach of FC in nursing in education with high satisfaction rate from students(Critz and Knight, 2013). In order to test the hypotheses: “Flipped Classroom” would result in higher grade, implementation of FC among 589 in U/G 1st and 2nd year nursing students resulted in 47 students passed more than previous year(Missildine *et al.*, 2013). Other attempts in this domain has also been reported with positive outcomes(McDonald and Smith, 2013; Schlairet, Green and Benton, 2014). Redesigning a second year U/G pharmacy course resulted in more time for the students to conduct project work in small group and improved grades though significant number of negative comments highlighted the need of larger team to assist in implementation(Ferreri and O’Connor, 2013). The same study was extended by offering the course for two different distance campuses depicted that students’ considered FC approach led to more engaging class(McLaughlin *et al.*, 2013). In another attempt of flipping a pharmacy course reported 80% of students’ satisfaction rate(Pierce and Fox, 2012). Department of Mechanical Engineering at Seattle University, Seattle, WA, US revealed from their attempt of flipping “Control Systems” course that it allowed the instructors’ to cover more material and

students' performance increased compare to the traditional course(Mason, Shuman and Cook, 2013). The attempt of "Flipped Classroom" with satisfactory indicators has also been reported in different domain ranging from statistics, chemistry, management, economics, philosophy and software engineering(Albert and Beatty, 2014; McLaughlin *et al.*, 2014). Many universities are now preparing MOOCS courses and using them to flip their classes so that both physically or virtually attending students watch the same videos.

2.2 Emergence of Learning Analytics

In 2008, a group of researchers from Taiwan projected 35% growth in e-learning market. They have also discussed about significant failure rate and importance of users' satisfaction to the success of online learning. They have identified six categories that contributes to users' satisfaction.(Sun *et al.*, 2008) Figure 2 below, illustrates the categories.

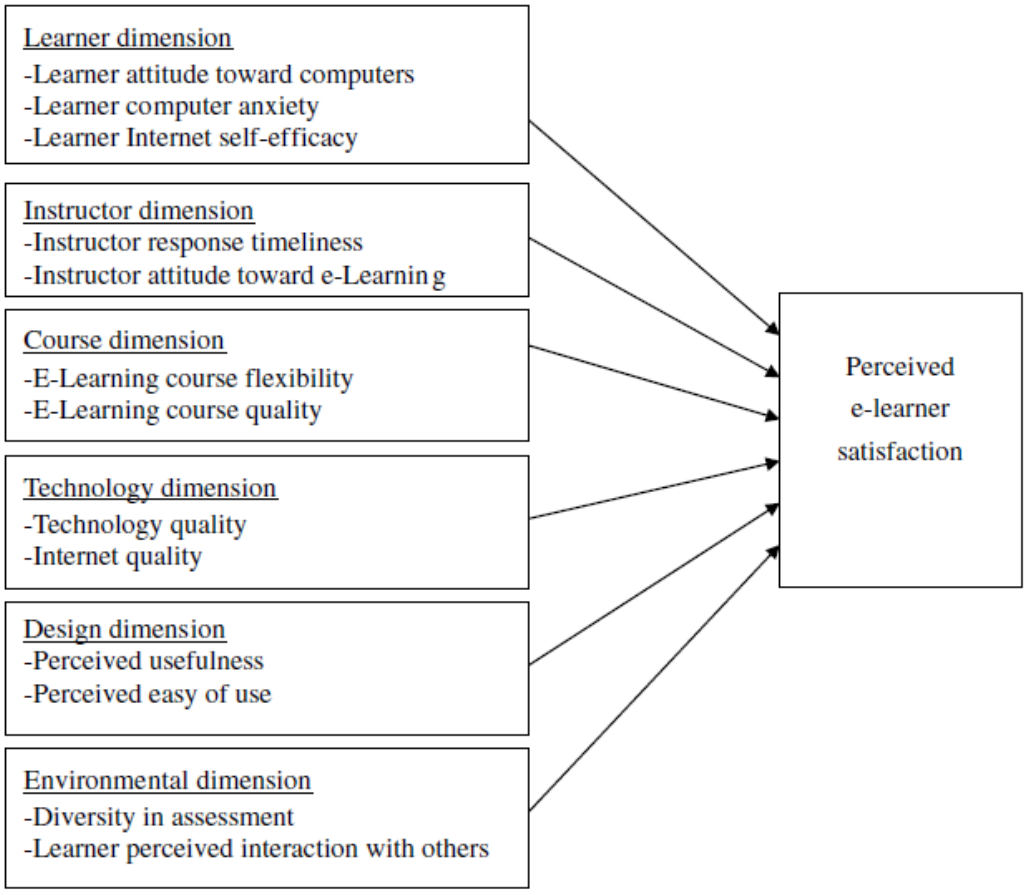


Figure 2. Characteristics to perceive online learners' satisfaction(Sun *et al.*, 2008)

Despite the increasing popularity of MOOCS, the student retention rate is increasing exponentially. It has been found that completion rate of MOOC course is 13% out of every

1000 enrolled students (Onah, Sinclair and Boyatt, 2014). Moreover, it is becoming common phenomenon that MOOCs that consists of millions of learners, have relatively low number of certificate of completion. In a study, it is claimed that CourseEra has 45% of completion rate of the students. (Kolowich, 2013)

Figure 3 illustrates the results of another study analyzing course completion versus course length in CourseEra. It illustrates that many students drop the course in the middle by losing their interest due to long length. (Jordan, 2014)

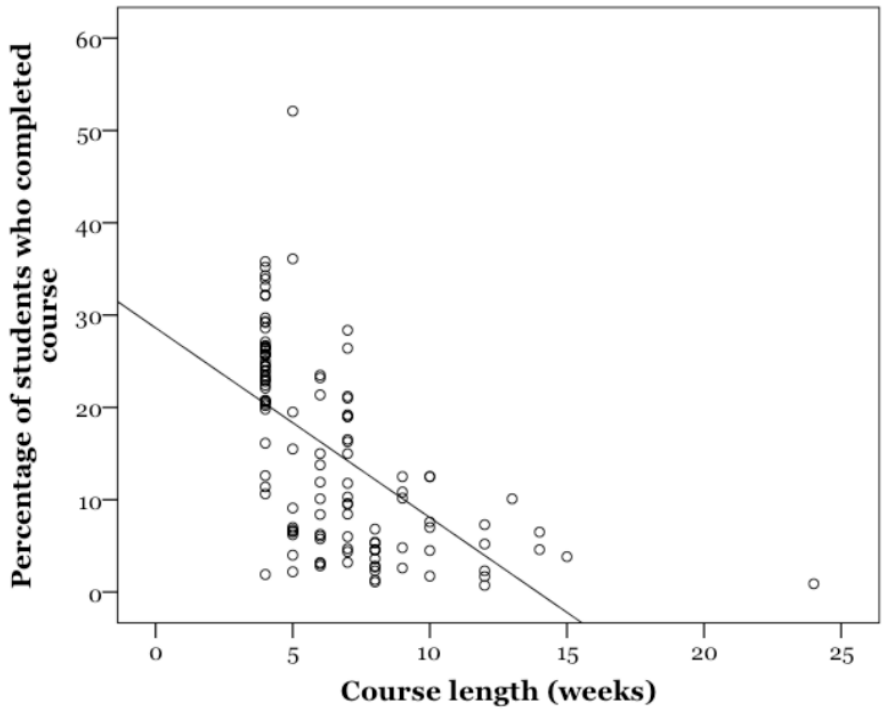


Figure 3. MOOCs Participants' dropout rate(Jordan, 2014)

In another study, the results from analysis of 221 MOOCs course depicts that with a median value of 12.6%, course completion rate varies from .7% to 52.1%. The dropout rate is high in the starting of the course. The completion rate varies based on "course length (longer courses having lower completion rates), start date (more recent courses having higher percentage completion) and assessment type (courses using auto grading only having higher completion rates)".(Jordan, 2015) Researchers also found positive correlation among dropout rates and difficulty of MOOCs course (Vihavainen, Luukkainen and Kurhila, 2012). In an experiment at the university in Cairo, 32.2% out of 122 participants successfully completed the MOOC courses(Hone and El Said, 2016). In another experiment on 34 learners identified learners' perception on the course content and course design one of the

major reasons behind dropout. All of these learners at least completed two MOOCs and participated in a qualitative interview. (Eriksson, Adawi and Stöhr, 2017)

From the results of these studies, it becomes evident that learners' retention in the online course is associated with categories illustrated in **Figure 2**. In order to engage the students on the online course, the instructors need to update the course contents after each cohort. The update process should be data driven and systematic rather than intuitive. For an example, the instructors can update the course videos if they know the points where the number of learners dropped significantly. On this regard, learning analytics allows the instructors to find relevant information/indicators regarding the course design and learners' perception.

2.3 Learning Analytics Overview

From the last decade, researchers and educational community developers commenced exploring the possibility of adopting similar techniques to gain insight into the behavior of online learners. With an increase of global hype on Big Data, Learning analytics emerge as a new branch of science from Educational data mining (EDM). Siemens provide the definition of learning analytics as, "Learning analytics is the use of smart data, data produced by learners and analytical models to discover information and social connections, and to predict and advise on learning."(Siemens, 2013)

Learning analytics is a research area, combination of artificial intelligence, web analytics, action analytics and predictive analytics. It has different objective, goal and outcome based on the context of application. Learning analytics enables to improve teaching or curriculum design. For an example, teachers can identify how the students are taking each module and they can improve the subsequent module based on students' interaction. The teacher can also improve the overall course design and content for each cohort based on analytics from feedback assessment(Chen and Chen, 2009). Learning analytics also can be used for prediction of students' retention and modeling the behavior/learning path. In different studies, it has been observed that number of quizzes passed act as a main determinant of obtaining higher grade. By applying machine learning techniques on the LMS log data, the students' at high risk of dropping out can be identified.(Dekker, Pechenizkiy and Vleeshouwers, 2009; Li *et al.*, 2012)

Recommendation of suitable contents for each learner is another important application of learning analytics. Several studies report different approach of recommendation of contents by utilizing students' navigation history, personal trait, learner attributes with content based collaborative filtering and sequential pattern mining(Khribi, Jemni and Nasraoui, 2009; Khribi, 2013). Students' motivation can also be boosted by the adoption of learning analytics. When students' can view how they are performing, they can change their learning style and be more aware to adapt the contents properly. In an experiment, researchers explored self-awareness and self-reflection implications using dashboard applications based on "Social Network Analysis" (Clow and Makriyannis, 2011).

In order to get an over view of Learning analytics, extensive search was conducted in different databases including Scopus, ACM digital library, Science Direct and Google Scholar to find relevant papers. The literature review is conducted in several stages:

Stage 1: Conduct the literature search with appropriate keywords

Stage 2: Assessing the results

Stage 3: Reporting the results

In the first stages, ***learning analytics, learning analytics tools, eLearning, education, students*** is used as keywords. With this search, approximate good amount of paper is retrieved. **Figure 4** illustrates the results obtained in Scopus database from 2008 to 2017. The increasing trend line may not always mean that the quality of literatures have increased in this domain. We also experienced increasing trend line for many other subject areas in Scopus database. In general, the increasing trend depicts that the field has emerging focus and interests among researchers.

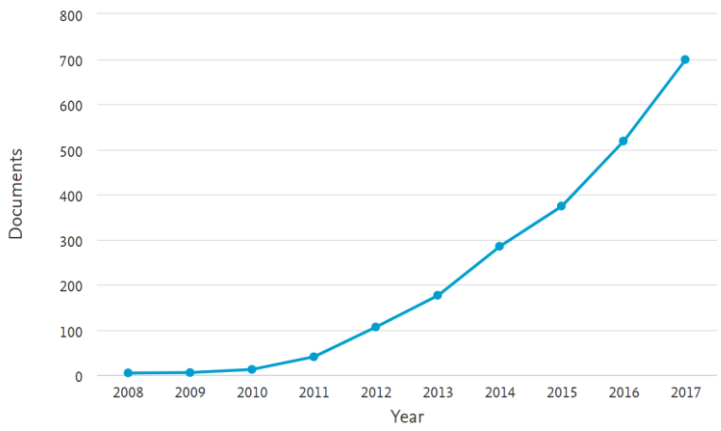


Figure 4. Paper published per year in Scopus

Figure 5 highlights documents per source in Scopus database,

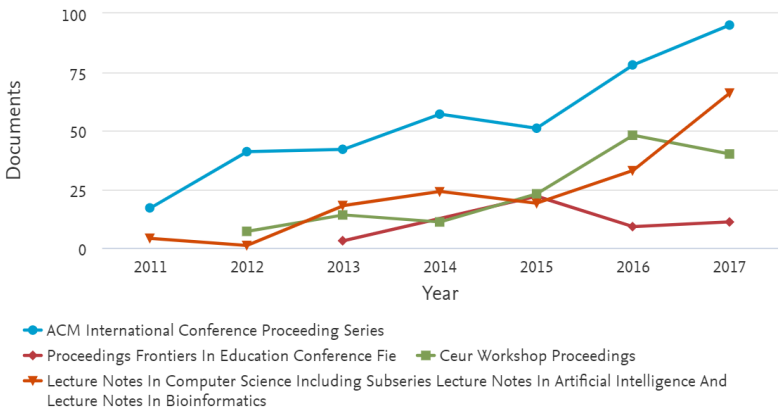


Figure 5. Documents per source

In the next steps, the search results were assessed based on number of citations (citation>50), year of publication (2008-2017) and other criteria. Table 4 below illustrates the selection criteria of papers.

Table 4. Selection Criteria of Papers

Selection standards for inclusion of research papers	
Date	2008-2017
Language	English
Published In	Conference Proceedings, Journal Articles, Book Chapter
Citations	>50
Demographics	International

After removing the duplicate results and based on the selection criteria, 37 papers have been reviewed for the scope of this work. Table 5 highlights the different methods to analyze digital learning data.

Table 5. Different Methods based on selected literature

Method	Related Work
Regression	(Romero-Zaldivar *et al.*, 2012); (Abdous, He and Yen, 2012); (Macfadyen and Dawson, 2010)
Classification	(Barla *et al.*, 2010); (Dejaeger *et al.*, 2012);(Khribi, Jemni and Nasraoui, 2009); (Romero *et al.*, 2008);(Chen and Chen, 2009);(Dekker, Pechenizkiy and Vleeshouwers, 2009);(Huang and Fang, 2013); (Pardos *et al.*, 2016); (Zhuoxuan, Yan and Xiaoming, 2015); (Lin *et al.*, 2013)
Clustering	(Chen and Chen, 2009); (Abdous, He and Yen, 2012); (Khribi, Jemni and Nasraoui, 2009); (Kizilcec, Piech and Schneider, 2013);(Romero *et al.*, 2009)
Text Mining	(Taboada *et al.*, 2011);(Leong, Lee and Mak, 2012);(Wen, Yang and Rosé, 2014); (Chaplot, Rhim and Kim, 2015); (Lin, Hsieh and Chuang, 2009)
Social Network Analysis	(Macfadyen and Dawson, 2010);(Fournier, Kop and Sitlia, 2011);(Scott *et al.*, 2015); (Rabbany *et al.*, 2014);(Stewart and Abidi, 2012);(Wise, Zhao and Hausknecht, 2013)

In the below, the Table 6 illustrates number of studies based on different research goals out of learning analytics.

Table 6. Research goals based on selected literatures

Goals	Related Work
Performance Prediction	(Abdous, He and Yen, 2012); (Pardos *et al.*, 2016); (Romero-Zaldivar *et al.*, 2012);(Huang and Fang, 2013);(Romero *et al.*, 2008);(Macfadyen and Dawson, 2010);
Dropout/Retention prediction	(Giesbers *et al.*, 2013);(Dekker, Pechenizkiy and Vleeshouwers, 2009); (Dejaeger *et al.*, 2012); (Kizilcec, Piech and Schneider, 2013); (Chaplot, Rhim and Kim, 2015)

Table 7 continues. Research goals based on selected literatures

Goals	Related Work
Increase Motivation	(Fournier, Kop and Sitlia, 2011); (Clow and Makriyannis, 2011); (Ali *et al.*, 2012); (Santos and Boticario, 2012); (Macfadyen and Dawson, 2010)
Tailored Content Recommendation	(Romero *et al.*, 2009); (Khribi, Jemni and Nasraoui, 2009); (Thai-Nghe, Horváth and Schmidt-Thieme, 2011); (Verbert *et al.*, 2012)
Feedback improvement	(Chen and Chen, 2009); (Barla *et al.*, 2010);(Ali *et al.*, 2012);(Leong, Lee and Mak, 2012)

However, there are some limitations in literature review process as the whole process is conducted by human reading and research. These limitations can be improved by conducting another research work using Latent Dirichlet Allocation (LDA) algorithm with data analysis software via topic modeling(Blei *et al.*, 2014).

2.4 Learning Analytics Framework

There has been numerous approach to establish a framework on learning analytics. At the first iteration, the methods focus on data collection and preprocessing for further analysis on learning activities. The next step consists of data modeling, visualization of the results and interpretation to inform the instructors and students for performance and goal achievement. However, these approaches diverges with respect to their origins, techniques, fields of emphasis and types of discovery(Mohamed Chatti., *et al.*, 2012; Siemens and Baker, 2012; Romero and Ventura, 2013). Other attempts were presented by (Ferguson, 2013), (Prinsloo and Slade, 2013), (Ferguson *et al.*, 2016), and (Bienkowski, Feng and Means, 2012).

In recent times, a group of researchers from different European universities formed a consortium and launched an EU funded project named “SHEILA” focusing on development of a learning analytics framework. The main objective of the policy framework is to promote formative assessment and personalized learning via learning analytics in institutions across Europe and around the world. In order to setup the policy, they used “participatory action research and the Rapid Outcome Mapping Approach (ROMA)”.(Tsai *et al.*, 2018)

Among of all of these efforts, the framework introduced by Researchers' from RWTH Aachen University, Germany has been recognized widely by other researchers and have highest number of citation among other work related to LA framework (Mohamed Chatti., *et al.*, 2012). They described learning analytics framework based on four dimension:

- Gathered data for analysis (What?)
- Target group (Who?)
- Goal and objective of the analysis (Why?)
- Methodology (How?)

Figure 6 below illustrates the framework proposed by the mentioned work,

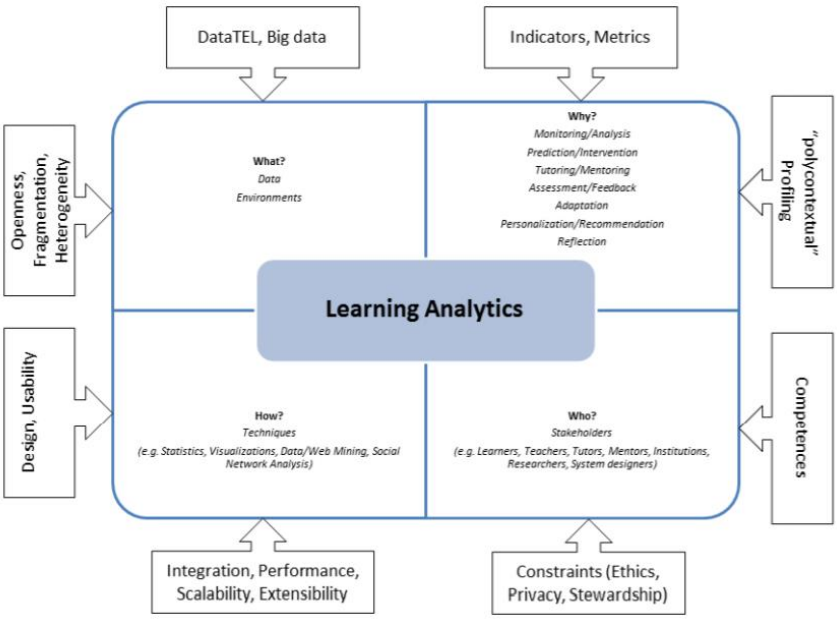


Figure 6. Learning Analytics Framework(Mohamed Chatti., *et al.*, 2012)

2.4.1 Data Source

Learning analytics utilizes different sources and forms of educational data. According to sources, the available educational data for learning analytics can be segregated into two categories: centralized system-generated data and user-generated data. Centralize systems are mainly learning management systems (LMS) such as MOODLE, Blackboard, Canvas, Sakai etc. These data are mainly log generated from the systems consist students interacting and activity in the LMS such as taking rest, reading lessons and uploading assignments (Romero, Ventura and García, 2008). Nowadays, These LMS are often used as an extra tool in face-to-face classroom to enhance learners' learning performance.

User generated contents varies over time and in different learning environment such as blog data. These type of data is mostly unstructured and requires rigorous data preprocessing for meaningful analysis. Generally, researchers track different type of data such as login frequency, resource accessed, response time to reply, duration of taking quizzes, final grades, discussion forum posts to analyze and get desired goal related to learners' performance and effectiveness of course contents.

2.4.2 Target group/Stakeholders

Stakeholders of LA mainly differs based on objectives, perspectives and expectations. The primary stakeholders of LA are teachers and students. Students may be concerned in how analytics can enhance their grades or assist them construct their own learning environments. Teachers may be interested in how analytics can improve their teaching practices according to the needs of learners. Moreover, education institution can use LA for identifying the reasons behind students' retention and decision support system(Campbell, DeBlois and Oblinger, 2007).

2.4.3 Goals

There are different goals of learning analytics based on the stakeholder's preference and context of application. In Table 6**,** different goals: performance prediction, Students' retention and dropout prediction, Motivation increase, content recommendation and feedback improvement have been identified based on the review of the selected papers.

Performance Prediction

The main goal is to build a model to predict the learners' knowledge level and performance based on current activities. This model can be considered as a training set in the language of Artificial Intelligence. This model can be used to predict the performance of the future learners'. The interpretation of the analysis can also result in proactive measures for the students who may require assistance in future in order to improve their learning performance. The common factors contribute to building the predictive model are demographic traits, grades (pre-required courses, evaluation quizzes and final scores), portfolios of learners, multimodal abilities, involvement of learners, registration and activity involvement (Macfadyen and Dawson, 2010; Abdous, He and Yen, 2012; Romero-Zaldivar *et al.*, 2012; Huang and Fang, 2013). For an example: Romero-Zaldivar et al. (2012) analyzed tracked

events (login time, worked time, frequency of post, time spent etc.) to estimate the final performance using multiple regression. They have observed correlation between the tracked events and final score(Romero-Zaldivar *et al.*, 2012). Macfadyen and Dawson (2010) investigated effect of different tracked variables (e.g. number of messages, total online time duration, number of links visited) on the final grade in LMS supported courses. It is necessary to choose the predictors variables properly. A study conducted by Huang and Fang (2013) shows that adding more prediction variables does not improve prediction accuracy to assess learners' performance(Huang and Fang, 2013).

Prediction of Drop out and Retention

The object is to identify the students' at risk or find the probable reasons or pattern of students' dropout. It is associated with instructional design of course contents and by examining student behavior teacher can act on the improvements and future designs of the learning activity. Dekker et al. (2009) tried to predict students' drop out using classification algorithm while Kizilcec et al. (2013) classified MOOCS learners' based on their interaction (video lectures) with course contents. They also identified personal commitments, work conflict and course overload as main reasons of student retention(Kizilcec, Piech and Schneider, 2013). Dejaeger et al. (2012) explored the association between motivation of students and retention. They observed negative correlation between students' satisfaction and course dropout rate. It means dropout rate decreases with the increase of students' satisfaction. They also explored the usage of synchronous tools and students' satisfaction. (Dekker, Pechenizkiy and Vleeshouwers, 2009; Dejaeger *et al.*, 2012; Kizilcec, Piech and Schneider, 2013)

Motivation Increase and Self-Reflection

The objective is to compare data among same courses, classes, contents and interpret the results to measure the effectiveness of learning. In order to receive the expected goal, a set of tailored variables and indicators are needed. Researchers used dashboard like applications to explore self-awareness opportunities in contrast with motivation (Clow and Makriyannis, 2011). In order to increase motivation, researchers used multiple widget technology and embedded feedback to facilitate personalized learning(Ali *et al.*, 2012; Santos and Boticario, 2012). Based on students' interaction and participation in MOOC, triggering moment is identified and different activities have been introduced to engage students in those moments (Fournier, Kop and Sitlia, 2011). Another study revealed that students' engagement increase

with the knowledge of peer activity and they do not consider to be tracked outside the course environment due to privacy concern(Santos and Boticario, 2012).

Content Recommendation

Personalization of learning is one of the major advantages of learning analytics. The objective is to assist learner to decide what to learn next based on learning pattern of other learners' who had identical activities and preferences in similar context. Tailored learning contents leads to higher degree of learning efficiency. In this regard, LA plays a significant role to provide customized learning path for each learner. In a study, researchers evaluated item based and user based filtering to suggest personalized contents for learning (Verbert *et al.*, 2012). Other studies focus on collaborative filtering, content modeling for hybrid filtering, similarity based approaches(Khribi, Jemni and Nasraoui, 2009; Santos and Boticario, 2012).

Researchers also used web mining and user profile information for building recommender system (Romero *et al.*, 2009).

Feedback

The main concept is to increase the learning efficiency by having information on learners' interest and learning context via intelligent feedback. Researchers discuss the significance of using appropriate type of feedback to collect useful information (Macfadyen and Dawson, 2010; Clow and Makriyannis, 2011; Ali *et al.*, 2012). In another study, Leong et al. (2012) investigated application and impact of SMS based feedback to the teachers after each class. The main idea was to perceive positive and negative sentiment after each lecture and improve the contents for next lecture accordingly (Leong, Lee and Mak, 2012). Barla et al. (2010) utilized different classification schemes to automatically selection of relevant text as feedback during adaptive testing(Barla *et al.*, 2010). In another study, researchers developed a tool to measure the satisfaction of students during the mobile formative evaluation(Chen and Chen, 2009).

2.4.4 Methods

In order to identify hidden pattern in educational data set, LA applies different analysis and data mining techniques. The widely used techniques are *"Classification, Clustering, Regression, Text Mining, Social Network Analysis, Descriptive statistics and visualization"*. **Table 5** illustrates the adoption of different techniques based on the selected literatures. Classification is the most popular method, followed by clustering and regression (logistic /

multiple). Furthermore, algorithm based performance comparison techniques like sensitivity, specificity, precision and similarity weights is employed. In the below **Figure 7** depicts the frequency of different techniques within the scope of literature review.

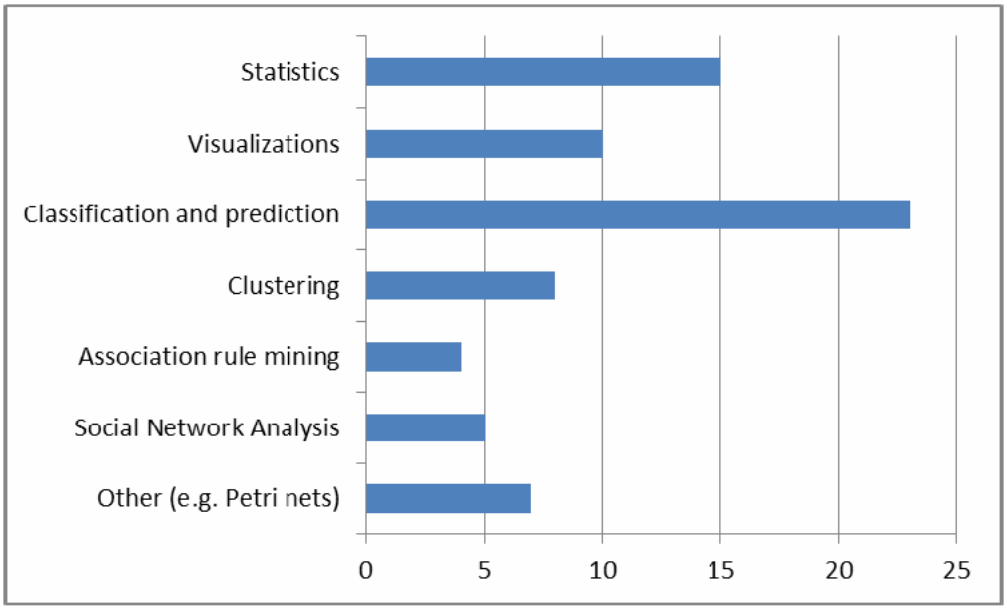


Figure 7. Different Methods Applied in LA (Mohamed Chatti., *et al.*, 2012)

The figure illustrates that classification and prediction remains most used paper followed by clustering techniques (i.e. “K-means clustering and Bayesian networks”). All other studies focused on visualizations, followed by statistics and SNA (22%).

3 PROPOSED METHOD

According to the discussion of related work and literatures, it is evident that there are different approaches of Learning Analytics focusing on the objective, context, segment, applicability and source of data. In this scope of work, we employed the methodologies to answer our research questions based on the proposed framework illustrated in **Figure *1*.** E-learning/flipped learning analytics and assessment framework ***Figure 1***. In Table 8 below outlines the methodology of this work.

Table 8. Proposed Methods

Method	Research Questions (RQ)	Hypothesis	Method	Data Source
Method 1	**RQ1: How to engage/hook up audience in the video?**	Video length corresponds with audience retention.	Regression	YouTube (Video Analytics Data from YouTube studio)
			Cohort Analysis (Nose, Body and Tail approach)	
Method 2	**RQ2: How to assess the students' online discussion engagement?**	High frequency of meaningful words leads to high degree of Students' engagement in an online discussion forum.	Social Network Analysis, Natural Language Processing, Sentiment Analysis	Disquss discussion forum

Table 9 continues. Proposed Methods

Method	Research Questions (RQ)	Hypothesis	Method	Data Source
Method 3	**RQ3: How to predict learners' success based on the LMS activity?**	Frequency of activity in LMS predicts learners' final score.	Decision Tree Classification, K-nearest Neighbor classification, Kmeans clustering	LMS Data (MOODLE)

3.1 Method 1 (Video Analytics)

In the era of YouTube learners, Video lectures has emerged as the elementary component of digital learning. Video-based lecture (VBL) series has received widespread acceptability among students as it provides control to the students in learning at their own pace. Several studies have reported that most of the cases, students take advantage of video lectures if it is available. At least 95-97% of the students view the video lecture once. (Heilesen, 2010; Giannakos *et al.*, 2013) There has been several studies and research conducted focusing on video analytics. Studies using subjective learner survey, reported higher satisfaction of students with the use of video lectures(Galway *et al.*, 2014; Young *et al.*, 2014). Researchers also portray pros and cons of using VBL in flipped/online format of learning in different studies(Traphagan, Kucsera and Kishi, 2010). However, only few of the studies focuses on quantitative measures to evaluate the quality of VBL in accordance with the learning outcome and establishing framework for continuous development of course contents. In a study using 862 video lectures from edX MOOC, researchers have identified five activity pattern of learners to explain their learning behavior with the video lectures (Kim *et al.*, 2014). In another study, researchers found relationship between cognitive load required for each video segment and repeated views(Giannakos, Chorianopoulos and Chrisochoides, 2015). Researchers also conducted empirical study on 6.9 million viewing session on four edX courses and interviews from six edX-stuff to explain the students' engagement with MOOC videos. They came with a set of recommendations for the instructional designer to

design the video lectures and few of the recommendations have been put into practice into edX courses.(Guo, Kim and Rubin, 2014)

All of these studies considered video dropout (e.g. navigating away or stop watching the video before completion) as one of the measurements for engagement and have depicted similar results that shorter video lengths associated with the engagement. In this work, we used audience retention as measurement for engagement along with easy to use video segmentation principal. "Audience retention refers to average percentage of a video people watch". We followed the work of popular studies and tried to come up with easy to understand and applicable method rather than a complicated one. Instructors or anyone with interest in video analytics can adopt this method and get an initial speculation on the quality of their contents.

The main objective of the work is to assess the usability of a simplistic method that can facilitate data driven decisions to the instructors for video lecture content improvement and increase learners' engagement, illustrated in **Figure *1***.

The hypothesis is:

"Video length corresponds with audience retention."

The data for the video lectures is collected by extracting audience retention report from YouTube studio. The first step of the method is to identify the acceptable video length for which the learners' average engagement rate is high. In order to do so, total percentage of video viewed and video length is characterized by using regression analysis.

It models the association of a dependent variable with one or more independent variable using linear equation. The variability of independent variables are derived from the equation of a straight line and considered as the "cause" of results observed in the past.

The equation of a simple linear regression model:

$$Y = a + bX$$

Where,

$Y =$ Dependent variable

$X =$ Repressor/predictor variable

$a =$Y-intercept of the line

$b =$Line slope

In the below, Figure 8 explains different components and architecture of simple linear regression,

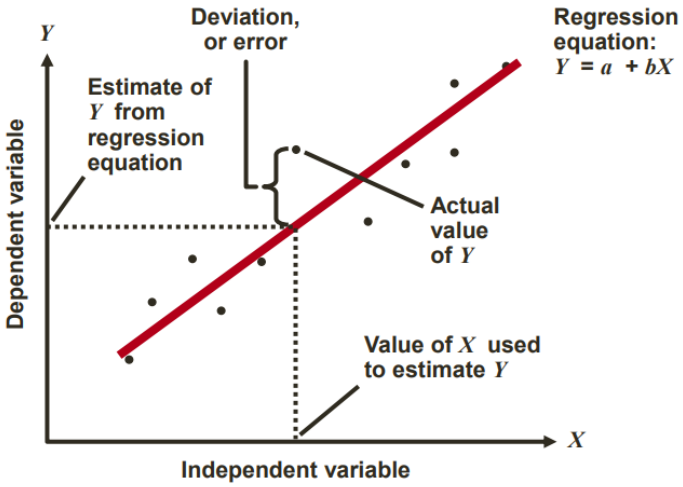


Figure 8. Simple Regression Analysis

In this work, the average percentage of views is considered as dependent variable Y and video length as independent variable X.

The next step consists of applying a systematic approach to improve the contents of the video lectures in each iteration of the course life cycle. The video lectures are segmented into three parts. The segmentation follows the principle of Nose, Body and Tail framework, introduced by Wistia, one of the popular video hosting, creating, managing, and sharing service for business. They split their video into three smaller parts (Nose, Body, and Tail) in order to analyze the performance of each video. (Currier and Fishman, 2015)

- **Nose:** Starting 2% of the video timeline
- **Body:** 96% of the video timeline
- **Tail:** Remaining 2% of the video timeline

The assumptions on the engagement decline of different segments discussed in Table 10.

Table 10. Nose, body, Tail approach by Wistia (Currier and Fishman, 2015)

Nose	Body	Tail
Some people were immediately disinterested in the content and the video failed to keep the interest span after clicking play button. The "average" engagement loss may correlate with video length	A viewer who leaves during the body reflects that they were interested at first and with time, they lost interest because of the video contents. These viewers most likely will not be watching any future videos as well. It may also indicate that information is not uniformly distributed within the body part.	The segment may not be important unless there's some call to action integrated e.g. quiz, survey link, suggested video link.

In summary, the video lectures content are continuously improved by:

Video Content Improvement = Optimal video Length + Video segment Analysis

This allows the instructors implementing data driven decision to improve the pedagogic design of the course. In the below **Figure *9*** illustrates the methodology of the course video analysis,

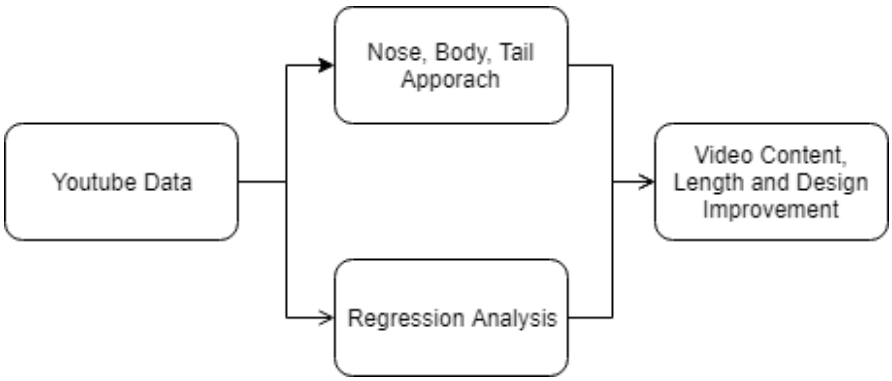


Figure 9. Methodology for course video Analysis

3.2 Method 2 (Discussion Forum Analysis)

The scope of the work focuses on analyzing discussion forum of online course as an indicator to speculate on the overall interactivity of the course. According to Table 1 discussed above, the main goal of analysis is to provide a score/set of score and ranking of students based on their participation in the online discussion forum. This will help the teachers' to assess

students' performance in the discussion numerically based on systematic approach and criteria. The scores from this method can also be used as an empirical input in the proposed framework illustrated in **Figure *1*** to improve the course pedagogic and content over the life cycle of the course in different years.

The hypothesis of the proposed method:

"High frequency of meaningful words leads to high degree of students' engagement in an online discussion forum."

In the first step, the discussion activity of students' in the online discussion forum is illustrated as social network diagram. Then different centrality measures of the Network are calculated to explain the students' activity pattern. In this work, "***Degree Centrality, Betweenness Centrality, Closeness Centrality and Eigenvector Centrality"*** measures have been used for assessment of the students' engagement. In the next step, text analysis and sentiment analysis approach is utilized to analyze the post content of each student.

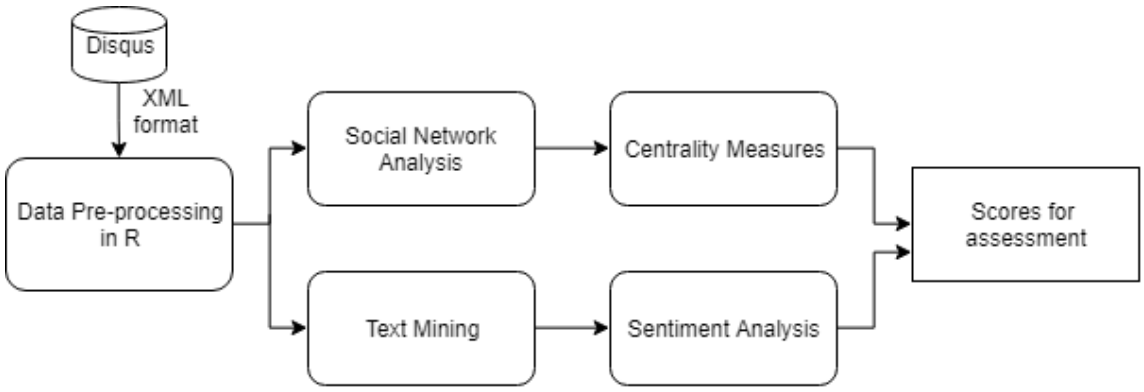


Figure 10. Architecture Diagram for Discussion Forum Analysis

A brief literature review has been conducted to assess the usability and applicability of social network analysis and natural language processing in different domain. The method of analyzing social networks provides an explicit mathematical definition to describe the characteristics of community members and the underlying network (Bonacich, 2002). Based on the relationship between actors (members), these characteristics are evaluated. The introductory analysis of Social Network Analysis was discussed by Scott (Scott *et al.*, 2015), Hanneman (Hanneman and Riddle, 1998)**,** Wasserman and Faust(Faust and Wasserman, 1995). They elaborated the network structures based on different descriptive parameters of the network. Researchers from diverse fields use network analysis to represent community

interests including citation networks (Rosvall and Bergstrom, 2008), World Wide Web (Ferrara, 2018), food websites, biochemical networks. SNA was also used to explore family relations (Widmer, 1999), military C4ISR network (Dekker, 2002), investigate terrorist network. Literature suggests that active engagement is one of the key components of effective learning(Bryson and Hand, 2007; Early, 2011). As per Dale's cone of learning, students learn more by active engagement. In the context of online learning, discussion forum plays as a key tool to facilitate learners ' active involvement. Researchers at the University of Alberta, Canada discuss the application of analysis of social networks to measure the interaction of learners (Rabbany *et al.*, 2014). NICHE Research Group, Dalhousie University, Canada (Stewart and Abidi, 2012) also studied SNA application in a forum on clinical online discussion. They analyzed the communication pattern using SNA in this research to understand how empirical knowledge is exchanged by the community member. In another research, the contribution and response of students to asynchronous online discussion forum was analyzed using SNA (Wise, Zhao and Hausknecht, 2013). There was not much work on the use of student sentiments in predicting attrition. In a study in New York University, it was concluded that students' attitude towards assignment and course material has positive effects on the successful completion of the course(Adamopoulos, 2013). In another study, researchers from Carnegie University identify correlation between student retention rate and the sentiment of discussion forum post (Wen, Yang and Rosé, 2014).

3.2.1 Centrality Measures

Centrality measures (Freeman, 1978; Rajaraman and Ullman, 2011) are based on mathematical calculations from graph theory to classify important vertices of a graph. For an example, identifying influential person(s) in a social network, principal nodes of Internet or transportation networks, and dispersion of infectious diseases. Degree Centrality, Betweenness Centrality, Closeness Centrality and Eigenvector Centrality measures has been used for centrality measures in this scope of work.

Degree Centrality: Degree centrality calculates number of edges of a node, which represents as the degree of the node. Higher value in degree means more centrality.

$$Degree_{ij} = \sum_{ij} N_{ij}$$

In the equation above, we can think of N_{ij} as the value of the cell with the row index i and column index j in a network matrix N. The fundamental intuition is that nodes with more links in a network are more influential and significant. In this scope of work, the higher degree of node reflects higher interactivity the students.(Suraj and Roshni, 2016)

Closeness Centrality:

The second centrality measure is Closeness Centrality. Closeness centrality is the inverse of farness, which is the sum of its distances to all the other nodes. Let us consider,

d_{ij} =The length of the shortest path between nodes i and j,

Then, the average distance l_i is denoted as:

$$l_i = \frac{1}{n} \sum_i d_{ij}$$

We know that the relationship between Closeness Centrality C_i and average lengthl_i is inverse proportional so:

$$C_i = \frac{1}{l_i}$$

It is a measure to calculate information spreading time from one node to another in a sequence manner(Stewart and Abidi, 2012).

Betweenness Centrality:

This centrality method was introduced by Linton Freeman as a quantitative measure to calculate the interaction in human communication network (Freeman, 2006). It counts the occurrence of a node being the bridge between two other nodes in the shortest path. Let us consider, s = Starting node, t = destination node where the input node is i that equals to $s \neq t \neq i$,

n_{st}^i =1, if node i is on the shortest path between s and t; otherwise it is 0. So, Betweenness centrality is denoted as: $x_i = \sum_{st} n_{st}^i$

However, more than one shortest path can exist between s and t. Therefore, the general equation is:

$$x_i = \sum_{st} \frac{n_{st}^i}{g_{st}}$$

In this scope of work, frequent number of post by a learner will lead to higher value in Betweenness centrality.

Eigenvector Centrality:

Eigenvector centrality is a measure of a node's impact in a network. It assigns comparative ratings to all nodes in the network based on the notion that high-scoring node connections contribute more to the node score than equal links to low-scoring nodes. Eigenvector centrality is an extension to degree centrality, centrality of a node is defined proportional to its neighbors' importance. In the first, the centrality of the graph is considered as $x_i = 1$. In the next iteration, new centrality value x_i' is denoted as:

$$x_i' = \sum_j A_{ij} x_j$$

Where A_{ij} = an element of the adjacency matrix, having a value 1 or 0 based on existence of edge between nodes i and j. In matrix notation, it is denoted as $x' = Ax$. In our case, students having higher value in eigenvector centrality represents higher influence in the course discussion forum. (Suraj and Roshni, 2016) In the below **Table 8** illustrates the interpretation and applicability of the discussed centrality measures in this scope of work.

Table 11. Interpretation of each centrality measures in this scope of work

Centrality Measures			
Degree Centrality: The fundamental intuition is that nodes with more links in a network are more influential and significant. In our case, higher the degree of node more interactive the student is. (Suraj and Roshni, 2016)	**Closeness Centrality:** It is a measure to calculate information-spreading time from one node to another in a sequence manner. In our case, it indicates the responsiveness of the students.(Stewart and Abidi, 2012)	**Betweenness Centrality:** In this work, High Betweenness represents that learner has posted in more frequently thus creating more nodes meaning creating more opportunity for discussion. (Barthélemy, 2004)	**Eigenvector Centrality:** In this case, students having higher value in eigenvector centrality represents higher influence in the course discussion forum. (Bonacich, 2007; Suraj and Roshni, 2016)

3.2.2 Natural Language Processing

In order to identify meaningful words, the text of each user's discussion was analyzed using natural language processing. The first step was data cleaning by:

- changing everything to lowercase
- remove punctuations
- remove numbers
- remove white spaces
- remove text within brackets
- replace number with the textual form
- replace abbreviation
- replace contractions
- replace symbols

In the next step, the number of meaningful words is counted. Lexicon-based sentiment analysis is conducted for each user, and each user has been assigned a positivity-based sentiment score based on meaningful words.(Taboada *et al.*, 2011; Wen, Yang and Rosé, 2014; Chaplot, Rhim and Kim, 2015)

3.3 Method 3 (LMS Data Analysis)

Higher educational institutions (HEIs) are evolving constantly to adapt the changes due to widespread acceptability of digital education/e-learning among the students. The traditional learning environment has changed in different dimensions e.g. class size, enrollment number of students, demands from workforce, learners' behavior, expectation and pedagogic design. Learning management systems (LMS) allows educational institutions to tackle the issues of rapid changes in pedagogic design and scalability of learning resources. It is a web based interface which empowers the instructor to shift the classroom activity in online environment using different tools like chat, forum, file storage, assignment, quiz, newsroom, lessons etc. Nowadays, large number of LMS are in use, which falls under two categories commercial and open source LMS such as Blackboard, Canvas, Sakai, Thinkific, Docebo, Absorb, Talent, ispring, Moodle etc. Among them, Moodle is most widely used because it is free, easy to use and flexible to create engaging online courses(Dougiamas and Taylor C, 2003; Cole and Foster, 2007; Cole *et al.*, 2014). According to Moodle statistics page, there are 106,948 registered sites, 19,043,417 courses, 751,605,503 enrolments and 161,747,310 active users in 227 countries till the date of writing this chapter(*Moodle Statistics Page*,

2019). Studies reported that data accumulated from these LMSs is significant to explain the students' learning pattern and pedagogic improvement(Mostow *et al.*, 2005; Lin *et al.*, 2013). These information includes log data on number of content page views, quiz attempted, LMS accessed and message posts. In a study, researchers developed a warning system using LMS data to identify the disconnected students. They have claimed that the system can classify 81% of time correctly the failing students(Macfadyen and Dawson, 2010). In another study, researchers from Spain discussed the applicability of different data mining algorithms and techniques in Moodle data(Romero, Ventura and García, 2008). A study in at Central Queensland University on 92,799 students reported positive correlation between course page views and final grades(Hollander, Saltmarsh and Zlotkowski, 2011).
In this scope of work, we used log data from Moodle to derive a predicted learning path of the students. The work also focuses on identifying the characteristics/pattern of the students using clustering techniques.

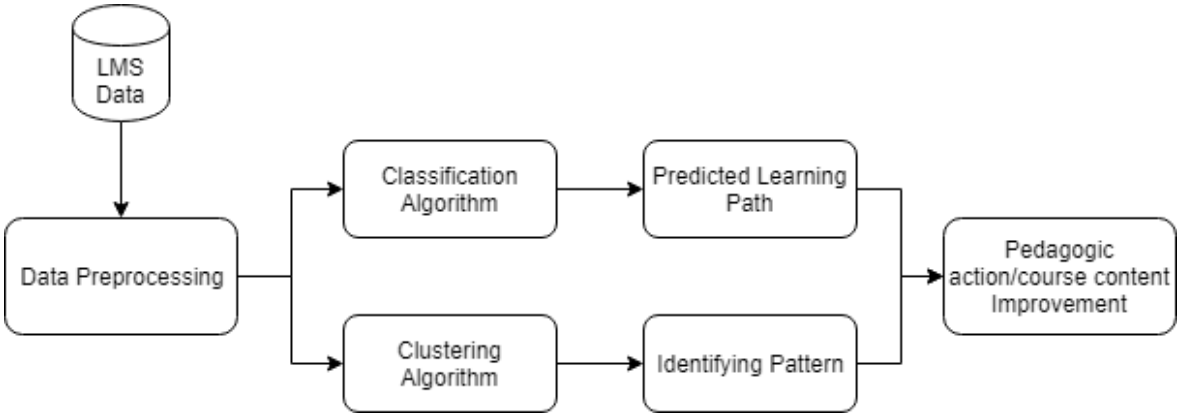


Figure 11. Architecture of LMS data Analysis

We used classification Tree Model to predict the learning path and verify the performance by applying and comparing the results from of the classifier using K-nearest neighbors (KNN) algorithm. In order to identify the characteristics of the students, K-means clustering algorithm is employed.

3.3.1 Classification Tree Model

Classification tree is a popular decision making method in data mining to predict the outcome, after effect, results. It uses tree like structure and learn from a classification function to predict the value of dependent variable based on the given values of independent values (input). It falls under the domain of supervised learning as the parameters are given

or known. It consists of two types of nodes: decision nodes and leaf nodes. The steps of decision tree are:

- **Splitting:** portioning of the data into subsets.
- **Pruning:** shortening the data brunch.
- **Tree Selection:** identifying the smallest tree that fits the data well.

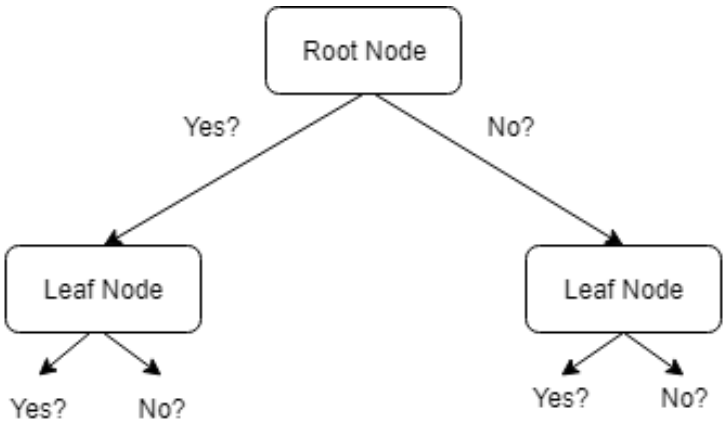


Figure 12. Simple Decision Tree

The calculation follows the principle of information entropy calculation. The main advantage of decision tree is that it is easy to understand and interpretable. It can also deal with categorical and numerical data.

3.3.2 K-means Clustering

A popular clustering algorithm falls under the division of unsupervised learning. K-means clustering is used to classify a given data set (unlabeled data) based on predefined number of clusters (considered as K number of cluster) fixed priori. The main objective is to cluster the data into the same number of groups as the value of K.

Algorithm

- Cluster the unlabeled data into number of groups based on the predefined value of K. For an example, K= 3 means that the data will be partitioned into 3 groups.
- Randomly identify K points as cluster centers.
- By using Euclidean Distance, assign objects to the nearest cluster center.
- Calculate the value of centroids for objects in each cluster
- Label new data based on the centroids
- Iterate the steps in consecutive rounds

Objective function can be denoted as:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^j - c_j \right\|^2$$

Here, $\left\| x_i^j - c_j \right\|^2$ = distance measure a data point and cluster center

c_j = centroid for cluster j; k= number of clusters; x= case i

Apart from Euclidean distance, several other distance functions: Cityblock, Cosine and Correlation distance can also be used. **Figure *13*** illustrates explanation of distance functions from Matlab Help page *(Matlab, 2019)* :

Distance Metric	Description	Formula
'sqeuclidean'	Squared Euclidean distance (default). Each centroid is the mean of the points in that cluster.	$d(x,c) = (x-c)(x-c)'$
'cityblock'	Sum of absolute differences, i.e., the $L1$ distance. Each centroid is the component-wise median of the points in that cluster.	$d(x,c) = \sum_{j=1}^{p} \lvert x_j - c_j \rvert$
'cosine'	One minus the cosine of the included angle between points (treated as vectors). Each centroid is the mean of the points in that cluster, after normalizing those points to unit Euclidean length.	$d(x,c) = 1 - \frac{xc'}{\sqrt{(xx')(cc')}}$
'correlation'	One minus the sample correlation between points (treated as sequences of values). Each centroid is the component-wise mean of the points in that cluster, after centering and normalizing those points to zero mean and unit standard deviation.	$d(x,c) = 1 - \frac{\left(x-\vec{\bar{x}}\right)\left(c-\vec{\bar{c}}\right)'}{\sqrt{\left(x-\vec{\bar{x}}\right)\left(x-\vec{\bar{x}}\right)'}\sqrt{\left(c-\vec{\bar{c}}\right)\left(c-\vec{\bar{c}}\right)'}}$, where • $\vec{\bar{x}} = \frac{1}{p}\left(\sum_{j=1}^{p} x_j\right)\vec{1}_p$ • $\vec{\bar{c}} = \frac{1}{p}\left(\sum_{j=1}^{p} c_j\right)\vec{1}_p$ • $\vec{1}_p$ is a row vector of p ones.

Figure 13. Distance metrics of K-means Clustering(Matlab, 2019)

- **Euclidean distance:** The root of the sum of the squared difference between the coordinates of the pair of objects.
- **Cityblock distance:** The absolute differences between the coordinates of the pair of objects.
- **Cosine distance:** utilizes cosine similarity method.
- **Correlation distance:** use linear relationship between the coordinates of the pair of objects.

(Thakare and Bagal, 2015; Kapil and Chawla, 2017)

In this scope of work, we used Euclidean distance as distance metric to employ k-means clustering.

3.3.3 Confusion Matrix and Performance Measure

It is a performance measure, which utilizes a table to evaluate the performance of classification model. The data set is divided into train and test data set. The performance is evaluated against test dataset (True values are known. In the below, Table 12 illustrates confusion matrix.

Table 12. Confusion Matrix

	Classified Positive	**Classified Negative**
Actual Positive	TP	FN
Actual Negative	FP	TN

- **True Positives (TP):** The cases where the prediction was yes (students' completed the course), and in reality students' completed the course.
- **True Negatives (TN):** predicted was no (did not complete the course), and in reality students' also didn't complete the course
- **False positives (FP):** We predicted yes, but they actually did not complete the course. (Also known as a "Type I error.")
- **False negatives (FN):** We predicted no, but they actually completed the course. (Also known as a "Type II error.")

Performance Measure

- **Accuracy:** It is the proportion of correct predictions and the observations. Accuracy ranges in between value of 1.0 (best accuracy) to 0.0 (worst accuracy).

$$ACC = \frac{TP + TN}{TP + TN + FN + FP} = \frac{(TP + TN)}{P + N}$$

- **Sensitivity (Recall or True positive rate):** Sensitivity (SN) is calculated as proportion of correct positive predictions and the total number of positives observation. It is also called true positive rate (TPR). The value ranges in between 1.0 (best) to 0.0 (worst).

$$SN = \frac{TP}{TP + FN} = \frac{TP}{P}$$

- **Specificity (True negative rate):** Specificity (SP) is calculated as the number of correct negative predictions divided by the total number of negatives. The best specificity is 1.0, whereas the worst is 0.0.

$$TN = \frac{TN}{TN + FP} = \frac{TN}{N}$$

4 DATA ETHICS

Learning analytics has enable the educational institutions worldwide to have greater understanding on the learning behavior and pattern. Analysis of the students' data is the main foundation of learning analytics. Data ethics and privacy is the biggest concern and challenge in the adoption of learning analytics. Though educational data mining has its advantage, misuse of such data can lead to potential disaster scenario like security threat, personal exploitation and discrimination. Now a days, educational institutions and organizations offer digital education via LMS. These LMS are hosted/maintained by third party vendors/support organization. In many cases, the educational institution does not have proper knowledge how the data from the LMS is handled or stored. Most importantly, the educational data may contain personal data and breaching of such data may cause breach of security. In the USA, several incidents reported that students' personal data from the LMS has been sold and used for marketing purposes. An executive from California University expressed her concern regarding the misuse of educational data by the vendors in an article(David, 2018) published in campustechnology.com using the following exact words,

"We started to recognize that vendors were starting to upsell their own products and services to students because they know who the students are and how to get in touch with them."

In order to ensure proper handling data by the vendors, the educational institutions should choose vendors whose processes are compliant with ISO 27001 standards for information security. The 114 control objectives of this information security specifically follow the CIA triad (confidentiality, Integrity and Availability) to deal with sensitive information. In a study related to learning analytics, researchers recommended a set of principles regarding data privacy and security for practitioners and educational designers while employing learning analytics(Pardo and Siemens, 2014). In another study, researchers identified data ethics as external limitation. They suggest that data collection and analysis should follow regional and local standards regarding data privacy(Greller and Drachsler, 2012). In a study by Andrew Cormack, researcher from JISC which is a popular digital learning consultancy organization in UK, proposed a data analysis framework following the approach from data protection law(Cormack, 2016). European Union's General Data Protection Regulation (GDPR) can be a standard guideline to follow as the educational institutions of Europe need

to comply with this regulation by default. Niall Sclater, another researcher from JISC highlighted code of practice and adoption of GDPR in learning analytics (Sclater, 2016). He highlighted the following aspects,

- The objective and purpose of collecting data should be clear and transparent.
- The collected data should be used only for learning analytics. The performed analysis and their objective also should be transparent and disclosed to the students.
- In general, students' consent is not required if the data collection is only for sole purpose in the legitimate interests of the institution. However, students' consent is necessary for collecting special category data such as ethnic origin.
- Students consent is necessary for intervention/action based on the learning analytics. The intervention should be only for the benefits of the students. For an example, learning analytics identify the students' at academic risk of failure. The institution cannot introduce extra lesson/class to these students without their consent. The students also should have a choice to opt out from the intervention.
- Students should have access to the learning analytics platform to see the results of the analysis. The dataset used for learning analytics should be consistent, meaningful and free from spurious correlation.
- The data collection and analysis should met the legal obligation and requirements specified by GDPR. Any collected data or learning analytics using personal data (except specific purpose data e.g. grades) should be destroyed or anonymized based upon request of the students.
- There can other regional requirements in place, which needs to be taken into consideration. For an example, while working for CEPHEI project, the author encountered the issue that a law of Russian Federation states that all user data should be stored in a server physically located in Russia.

There have been several other studies related to data ethical issues in learning analytics. Researchers from all these studies have suggested that personal information of the students should be handled carefully and all the personal information fields/metadata should be encoded/anonymized. In this scope work, there were no personal data in video analytics. Since the data for the social data analysis and LMS data analysis was collected under institutional capacity (academic course) and no intervention was introduced to the students, the consent of the students were not necessary. If the results from the work is used to develop

any kind of student centered action (e.g. extra class), it will be necessary to take the consents of the students. Moreover, the student's personal identity field in discussion forum data has been anonymized so that it complies with data privacy regulation.

5 RESULTS AND DISCUSSION

The objective of the thesis is to evaluate a supervised and unsupervised learning based assessment method, which will assist the teacher/instructors to take data driven decision regarding improvement of the course pedagogic design, content and organizations. In order to do so, Chapter 1 highlights the research questions, framework and structure of the thesis. In Chapter 2, the background, problem statement and related work is portrayed. Chapter 3 explains the used methodologies, parameters and architectures to conduct the analysis of this work. This chapter depicts the analysis results of each block (illustrated in ***Figure 1.*** *E-learning/flipped learning analytics and assessment framework*). The interpretation of results and possible data driven decision measures are also discussed by answering the research questions stated in Table 1.

5.1 Video Analytics

A series of 18 Video lectures on "Systematic Creativity and TRIZ" were developed by Creativity Lab at LUT University, Finland and uploaded to YouTube. The first video was uploaded in June 2015. These videos are being used as online lecture contents in "Systematic Creativity and TRIZ Basic Online" course offered at Summer School and Winter School in LUT University, Finland from 2017. The video lectures are also used as flipped lecture content for "Systematic Creativity and TRIZ Basic" course in the academic M.Sc. program: "Global Management Innovation Technology (GMIT)" at LUT University from 2016. The target audience of the lectures were academic students and young engineers around the world with a view to introduce them with systematic steps of problem solving, innovative thinking and basics of "Theory of Inventing Problem Solving (TRIZ)" methodology.

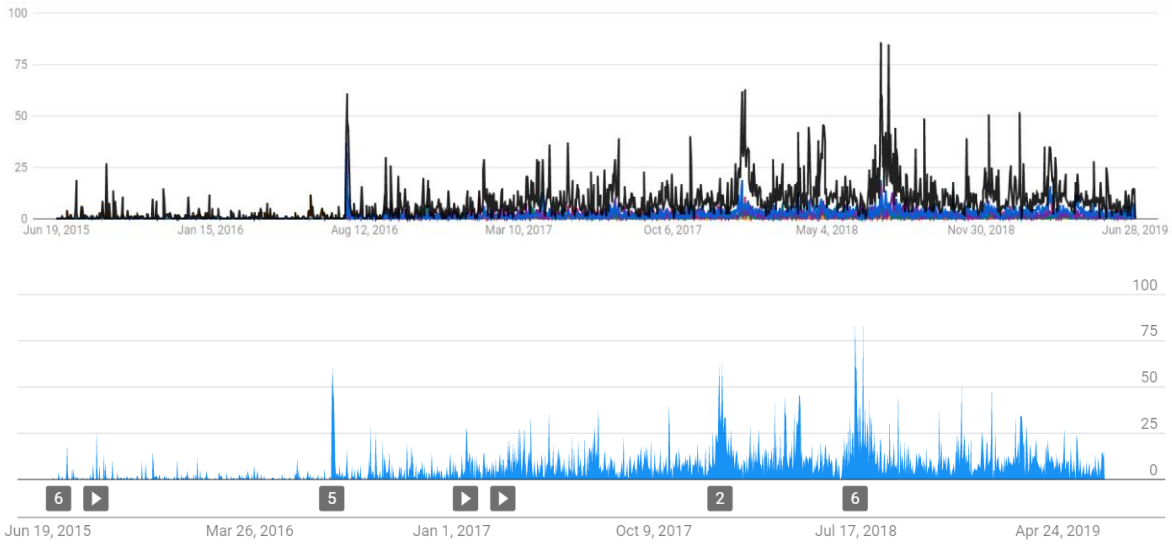


Figure 14. Life Time video views of TRIZ YouTube Channel (Creavity LAB, LUT, 2015)

In the above **Figure *14*** illustrates the lifetime video views of the YouTube channel. The spike in video views timeline represents the time/days when new videos were uploaded and the course was running in LUT University. The authors updated the video contents and design over the years to improve learning and engage more students. In the below Table 13 shows the title, video length, topics and level (type) of each video. Based on the video contents, the videos are mostly categorized into types: concept related videos and example related videos.

Table 13. Details of the video lectures

Title	Video Length (Minutes)	Topics	Level
TRIZ. Contradiction	16.77	Explain the principle and structure of contraction	Concept
TRIZ. Function Modelling	12.20	Explain the concepts of Function Modeling	Concept
TRIZ. Trimming	12.07	Explain the concepts and steps of Trimming	Concept
TRIZ. Ideal Final Result	11.25	Concept, Formulation and importance of Ideal Final Result	Concept

Table 14 continues. Details of the video lectures

Title	Video Length	Topics	Level
Ideal Final Result Basics	12.72	Concept, Formulation and importance of Ideal Final Result	Concept
TRIZ. TESE	12.77	Principal of Trend for Systematic Engineering Evolution	Concept
Example. Ice maker	5.43	Example	Example
Example. Spacecraft rocket launch system	6.83	Example	Example
Role of physics in TRIZ	7.07	Example	Example
Example. Rotation at the required rate	4.65	Example	Example
Case Study on FM, Trimming, IFR	11.53	Example	Example
Example. Diameter of a thin wire	1.92	Example	Example
Short intro to patents	9.45	Basic concepts of Patent Landscaping	Concept
Description of the course	2.6	Course description	Administration
What is the difference between optimization and invention?	8.08	Concept	Concept
Why it is hard to generate new ideas?	2.83	Concept	Concept
Example: Book and Worm	2.02	Example	Example
Example. Thermometer	2.45	Example	Example

In this scope of work, we tried to assess the usability of a simplistic method that can facilitate the instructors to take data driven decisions for video lecture content improvement and increase learners' engagement.

The hypothesis is:

"***Video length corresponds with audience retention.***"

The data for the video lectures is collected by extracting audience retention report from YouTube studio. According to the mentioned methodology discussed at Chapter 3 in **Figure *9.*** Methodology for course video Analysis, the first step was to develop a regression model. In order to do so, total percentage of video viewed and video length is characterized by using regression analysis. The resulted regression equation is:

$$y = -1.67x + 55.42$$

The regression analysis suggests that for TRIZ you-tube channel for 1 unit change in video length, audience retention changes by -1.6777 unit.

From the ANNOVA table illustrated in **Figure *15***, it shows that the relationship is also statistically significant (p value<.05) for 5% confidence interval. As a result, we cannot reject our null hypothesis that means that learners' engagement depends on the video length. Though we could not identify any optimal video length, our finding correlates with other studies that learners' engagement decreases for longer educational videos regardless of the topic, level and subject matter. This portrays the fact that longer educational video creates high cognitive load to the students as a consequence they lose interest to engage with the video for long time. The R^2 value of the model also suggests that model is a better fit.

Regression Statistics	
Multiple R	0.80682206
R Square	0.650961836
Adjusted R Square	0.627692625
Standard Error	5.850445308
Observations	17

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	957.5285	957.5285	27.97524	9.09026E-05
Residual	15	513.4157	34.22771		
Total	16	1470.944			

	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	55.46224188	2.917549	19.00988	6.58E-12	49.24363433	61.6808494	49.2436343	61.68084944
X Variable 1	-1.677655801	0.317187	-5.28916	9.09E-05	-2.35372474	-1.0015869	-2.35372474	-1.001586861

Figure 15. ANOVA table of the regression

In the below **Figure *16*** exhibits the variation of audience retention with respect to video length. We observed a sharp constant decline of AR for all videos. The length of each video is marked against each corresponding audience retention. In this scope work, if we aim for AR value of 50%, the video length should not be more than 6 minutes.

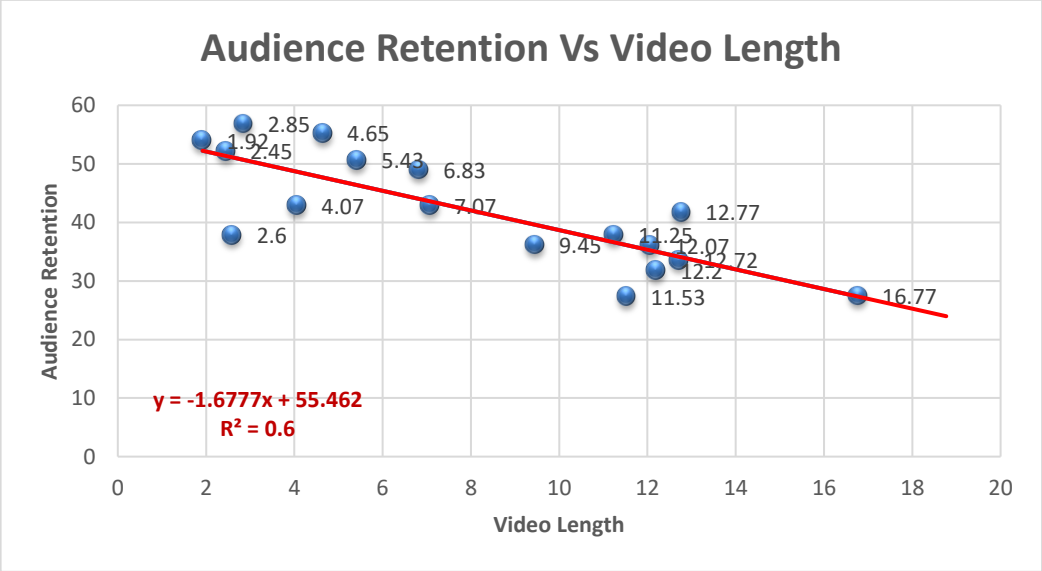


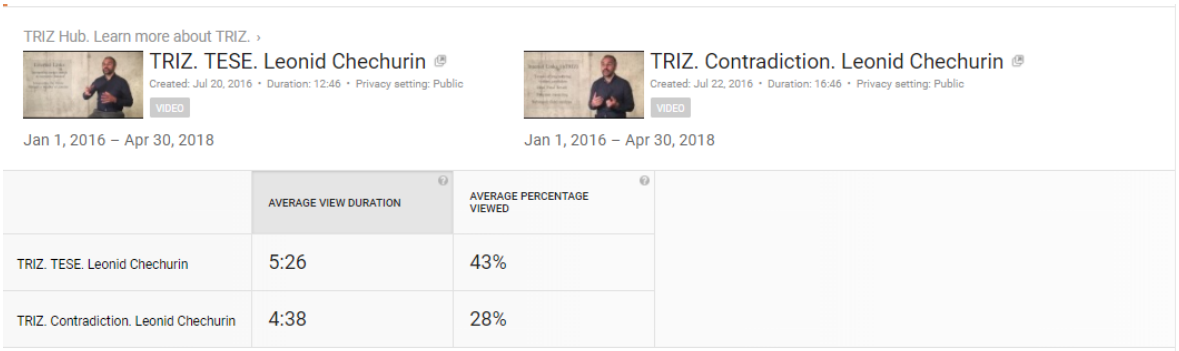
Figure 16. Longer video length exhibits lower audience retention (average percentage viewed)

However, video length may not be the only indicator/ parameter to measure learners' engagement. It also depends on the content of the video. We can see in **Figure *16*** for the video having length 12.77 minutes, exhibits higher AR value (42%) compared to other videos with similar video length. From Table 13, it can be seen that the video lecture is related to TRIZ TESE and covers the core concepts. However, increased AR needs to be interpreted with caution. This may mean that this video explains the subject matter well compared to other videos or the specific subject matter may have interest among learners. In Table 15, we observed that almost all the videos related to concept (according to Table 13) have higher average percentage views despite of having long video length compared to other videos.

Table 15. Comparison of average view

	Average Video Length	**Average view**
Concept Related Video	10.6 minutes	3.9 minutes
Other video	4.94 minutes	2.1 minutes

We may consider shortening the length of concept related videos to 6 minutes, which may be optimal length for this scope of work. In order to analyse the contents of videos, we conducted segment analysis following Nose, Body and Tail (Currier and Fishman, 2015) approach on the most two watched videos having higher average view duration. We used the audience retention graph of these videos directly from YouTube studio in order to keep the analysis simple and easy to adopt for everyone. We noticed most two performing videos based on average view duration are TRIZ contradiction and TRIZ TESE. Both of these videos are concept related videos which supports our previous interpretation that concept related videos on this subject matter has higher learners' interest and it may be reasonable to shorten these videos to an acceptable length as suggested by the regression analysis result. In **Figure *17***, **Figure *18*** and **Figure *19***, the summary and audience retention graph of top two performed videos is illustrated. We segmented the video into 3 parts: Nose (starting 2%), Body (Middle 96%), Tail (end 2%). It is important to note that both videos have similar pedagogic design pattern. We also observed similar audience retention pattern in both of the videos. There is sharp decline in the nose section, steady AR in the body with few spikes and decline in Tail part.



	AVERAGE VIEW DURATION	AVERAGE PERCENTAGE VIEWED
TRIZ. TESE. Leonid Chechurin	5:26	43%
TRIZ. Contradiction. Leonid Chechurin	4:38	28%

Figure 17. Comparison summary of top two performing videos (Creativity Lab, 2016)

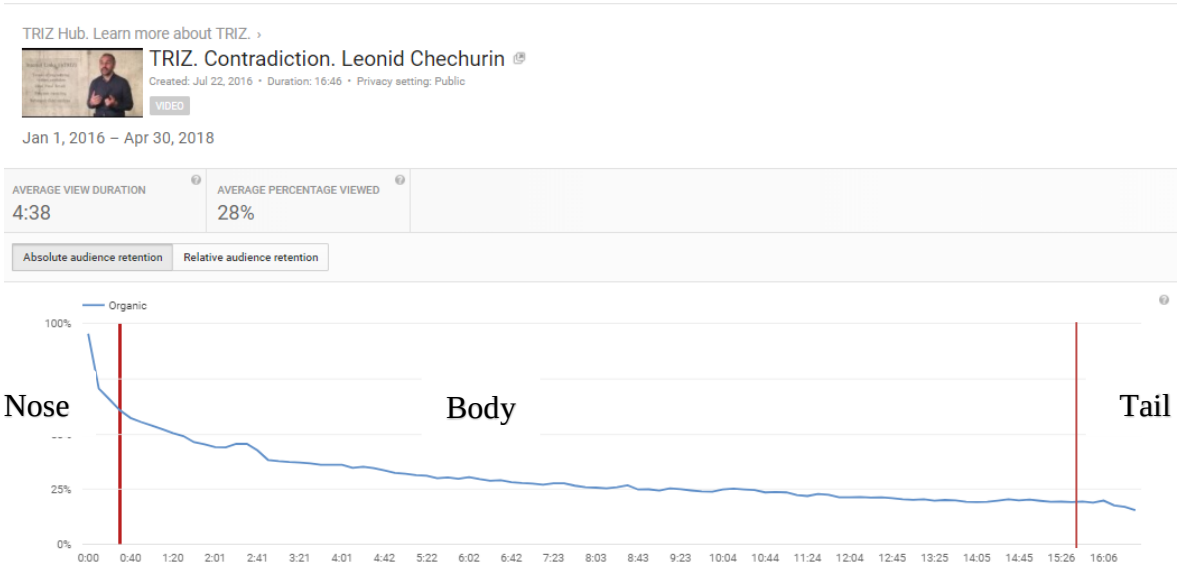


Figure 18. Audience Retention of TRIZ Contradiction video (Creativity Lab, 2016)

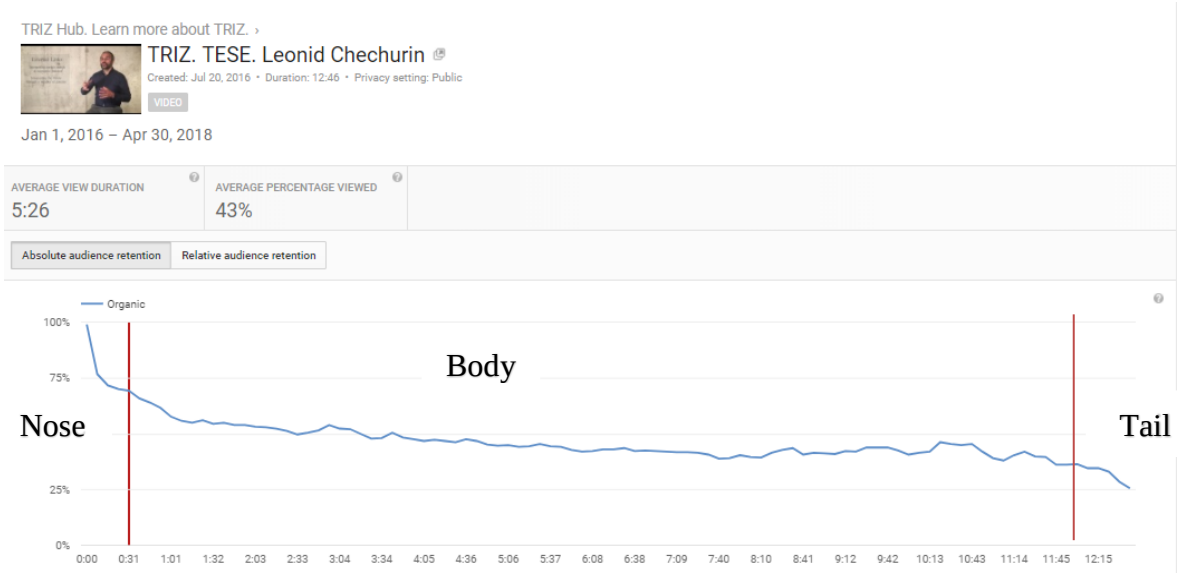


Figure 19. Audience Retention of TRIZ TESE video (Creativity Lab, 2016)

Nose Analysis

There is sharp decline in AR for both of the videos in nose section. After reviewing the nose segment, we found out that content of these segments are mostly intro graphics and logo sequence. In both videos, the professor appears on the screen and start speaking after 17 seconds from the start. As illustrated in (**Figure *18*** and **Figure *19***), this period is the "sharp decline in AR duration" in both videos. This may represent the fact that learners' landed on

the video but lost their interest due to waiting time for human interaction or chunk of information.

Proposed steps:

- Experiment with removing intro graphics and logo sequences
- Starting with a question or trick which triggers dopamine cycle of learners

Body Analysis

In the body segment, we observed steadier AR for both of the videos. However, the body section of TRIZ contradiction video (**Figure *18***) illustrates decline in AR compare to the TRIZ TESE video (**Figure *19***). This may be associated with the longer length of the first video and supports our hypothesis as well as the regression analysis result. We observed a spike in AR for contradiction video from 2.11 minutes to 2.31 minutes and for TESE video from 2.41 to 3.04 minutes. After reviewing the contents, we found out that the peak point of this small spike duration for both of these videos consist the slide “You will learn” which highlights the learning outcome. It seems that many learners were interested to find out the learning outcome of these videos and it may be better to introduce the learning outcome section earlier in the video timeline. In general, the body segments of the videos consists the main information chunk and steady AR suggests that the learning material satisfies learners’ quest for relevant information. These students can also be the students who are genuinely interested in the subject and the students redirected from the course page.

Proposed steps:

- Consider splitting the body into multiple videos.
- Use different shooting locations to break up the monotony.
- Try adding another person to the script. Sometimes the flow of a video is better with two people speaking lines back and forth.
- It is better not to use phrases like "in summary," "that’s about it," or "to wrap things up." These words may indicate to the learners’ that key information is over.
- Avoid recapping the whole thing. The viewer can re-watch the video if they want to hear any information again

Tail Analysis

The Tail segment in both of the videos consist ending and credit slides, which provide signals to the learners’ that the information has ended. As a result, we observed decline in AR for

tail segment. In general, it is common to observe decline at AR in tail section for most of the videos. The instructors should not be concerned about the AR decline unless there is any "Call to Action" element (such as quiz, survey, link to other videos) in the end.
Proposed steps:

- Introduce "Call to Action" elements

Limitation of the results

The significant limitation of this work is the classification of learners. It is not possible to identify which learners come from the course and which learners come organically since the VBL are uploaded publicly in YouTube. However, uploading the VBL publicly allowed us to generate sufficiently large sample size to derive meaningful metedata for developing a LA model. If we had only considered the learners' from the course, the sample size would be reasonably low to develop any interpretable LA. It may be possible to consider the current model as a base model and update it when population size from the course become large after couple of years. Another limitation is considering AR as the measurement for learners' engagement. Continued playing of video may not always reflect the true state of learners' engagement or effective learning outcome. Nevertheless, it can be the best possible measure to conduct such analysis considering the available metadata collection. In addition, the correlation of the results from video segment and content analysis provide evidence in favor of the validity of our assumption.

In summary, this work speculate a simplistic and easy way to adapt method in taking data driven decision for continuous improvement the video lectures over the years. The main advantage of this work is that it is easy to replicate and can be applied for any VBL series.

5.2 Discussion Forum Analysis

The dataset consists of blog discussion from online course of "Systematic Creativity and TRIZ Basic Online" course for two cohort of students (Winter School-2019, Summer School-2018) at LUT University Finland and one cohort of students at Peter the Great St. Petersburg Polytechnic University, Russia. The course was hosted in Thinkific online platform and online discussion was facilitated through disqus blog. The data was extracted as xml format. The xml file was preprocessed and the dataset was transformed as edge matrix for employing Social Network Analysis. The overall ETL (extract, transform, and load) and

analysis as illustrated in **Figure *10*** was conducted using R language. The final dataset contains the following columns:

- **Source:** The person who initiated a discussion
- **Target:** The person for whom the communication was intended for
- **author_messages_list:** all the messages from each individual users

The Hypothesis is:

"High frequency of meaningful words leads to high degree of students' engagement in an online discussion forum."

Network Visualization

To maintain data integrity and privacy, the original names of the participants are coded. In the first step of the analysis, the data for each cohort is illustrated as social network graph using igraph () package of R.

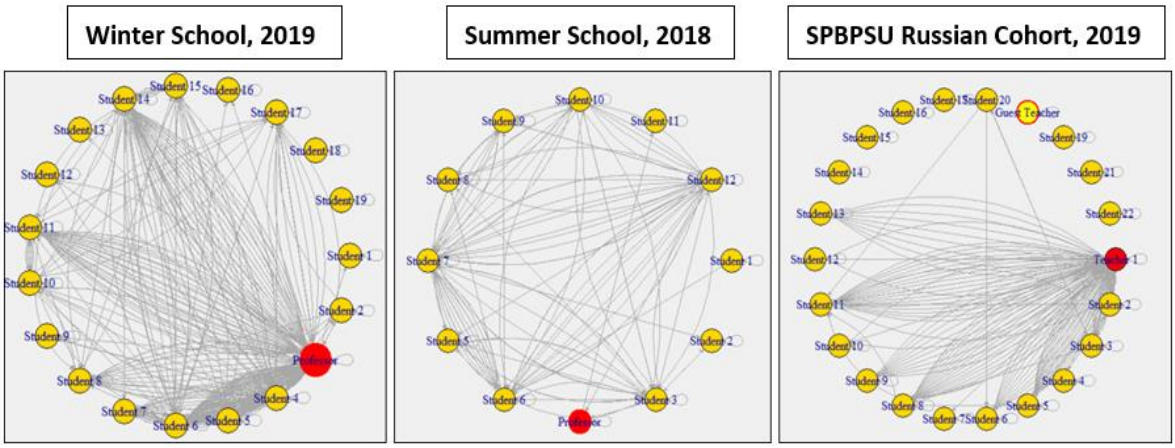


Figure 20. Social Network Graph of each cohort

Figure 20 illustrates the social network graph for each cohort: (Winter School-2019, Summer School-2018) at LUT University Finland and one cohort of students at Peter the Great St.Petersburg Polytechnic University, Russia. Each node (vertex) represents a student and each vertices represents communication between two persons. From the figure, it can be seen that Winter School, 2019 had higher degree of engagement than other two cohort did. In the cohort of Summer school, 2018 students had more interaction among each other rather than with the teacher. The figure also illustrates that most of the discussion was facilitator led in the Russian cohort. This type of visualization can help the instructors' to understand the degree of interactivity of their taught course after each cohort.

Ranking/Scores

In the second step, network centrality measures of every student for each cohort is calculated. Table 11 in Chapter 3, explains the interpretation of each centrality measures "***Degree Centrality, Betweenness Centrality, Closeness Centrality and Eigenvector Centrality***" for this scope of work based on literature review. Degree of centrality illustrates how interactive (central) a student is in the forum. Betweenness centrality indicates frequency (number) of post by each student and Eigenvector centrality depicts the degree of influence of each student.

Natural Language Processing:

In the 3rd step, the text of each student is analyzed and only meaningful number of words are counted. The meaningful word has been scored based on positive sentiment based on lexicon based sentiment analysis method(Taboada *et al.*, 2011; Adamopoulos, 2013; Wen, Yang and Rosé, 2014). Based on these centrality measures, word count and sentiment score; each student is ranked and given score. These scores and ranking can be used as a single assessment tool for the performance measurement of students in online discussion forum. **Figure *21***, **Figure *22*** and **Figure *23*** illustrate the scores of each participant in the discussion forum based on Degree Centrality, Closeness Centrality, Betweenness Centrality, Eigenvector Centrality, meaningful words and positive sentiment for Cohort 1, 2 and 3. The pictures were directly exported from R studio. Here, degC represents Degree centrality; sw_closeness represents Closeness centrality; star_betweenness represents Betweenness centrality; star_eigen represents Eigenvector centrality; Words_count represents number of meaningful words and Sentiment_positive represents Positive sentiment of the meaningful words.

	degC	sw_closeness	star_betweenness	star_eigen	Words_count	Sentiment_positive
Professor	1.00000	1.0000000	1.0000000000	1.000000000	2029	59.53
Student 6	0.58750	0.9391964	0.5377442580	0.939401788	2178	54.19
Student 4	0.51250	0.9391964	0.2488763705	0.776956620	1453	50.00
Student 14	0.43750	0.8842770	0.0845156539	0.677584813	2051	55.00
Student 11	0.26875	0.8505245	0.0502762800	0.401931330	694	61.36
Student 5	0.19375	0.8344271	0.0221146992	0.265065442	577	68.09
Student 10	0.17500	0.8344271	0.0106620074	0.208930810	696	67.92
Student 7	0.15625	0.8344271	0.0104004717	0.189535256	284	66.67
Student 8	0.15000	0.8188175	0.0041247495	0.159537932	529	77.55
Student 15	0.15000	0.8188175	0.0029909095	0.134835042	112	90.91
Student 17	0.11250	0.8036739	0.0022328644	0.123044377	296	61.54
Student 2	0.05625	0.7747036	0.0009918042	0.045121413	128	83.33
Student 9	0.01875	0.7091287	0.0000000000	0.027040493	77	71.43
Student 12	0.01875	0.0000000	0.0000000000	0.022229552	247	42.86
Student 16	0.01250	0.0000000	0.0000000000	0.018431582	23	100.00
Student 1	0.00625	0.0000000	0.0000000000	0.008843787	3	100.00
Student 13	0.00625	0.0000000	0.0000000000	0.008843787	9	0.00
Student 18	0.00000	0.0000000	0.0000000000	0.000000000	32	0.00
Student 19	0.00000	0.0000000	0.0000000000	0.000000000	82	50.00

Figure 21. Scoring of students based on centrality measures for Cohort 1

	degC	sw_closeness	star_betweenness	star_eigen	Words_count	Sentiment_positive
Student 7	1.00000000	1.00000000	1.00000000	1.00000000	268	65.22
Student 12	0.75555556	1.00000000	0.49466526	0.81736362	45	33.33
Student 6	0.73333333	0.90598291	0.48392247	0.77845230	706	41.38
Student 3	0.55555556	0.90598291	0.46753388	0.48339147	389	34.29
Student 10	0.44444444	0.86419753	0.40347213	0.43007513	102	54.55
Student 9	0.22222222	0.86419753	0.01954139	0.33122686	635	54.84
Student 5	0.15555556	0.72401434	0.00000000	0.14015486	191	61.54
Student 8	0.13333333	0.66666667	0.00000000	0.13842912	388	56.25
Student 2	0.04444444	0.66666667	0.00000000	0.04992032	67	33.33
Professor	0.04444444	0.02020202	0.00000000	0.02893202	440	53.85
Student 1	0.00000000	0.00000000	0.00000000	0.02147421	167	56.25
Student 11	0.00000000	0.00000000	0.00000000	0.00000000	479	46.51

Figure 22. Scoring of students based on centrality measures for Cohort 2

	degC	sw_closeness	star_betweenness	star_eigen	Words_count
Teacher 1	1.000000000	1.0000000	1.000000000	1.000000e+00	1614
Student 11	0.222222222	0.9313725	0.198468717	7.173167e-01	493
Student 8	0.212962963	0.9313725	0.085148799	6.878701e-01	1330
Student 5	0.175925926	0.9313725	0.015252608	5.662871e-01	145
Student 2	0.157407407	0.9070457	0.003355781	4.150698e-01	1045
Student 3	0.157407407	0.8990973	0.002541755	3.801739e-01	458
Student 6	0.101851852	0.8680754	0.000000000	3.139156e-01	116
Student 9	0.092592593	0.0000000	0.000000000	2.383327e-01	686
Student 13	0.083333333	0.0000000	0.000000000	2.197365e-01	273
Student 20	0.037037037	0.0000000	0.000000000	5.773877e-02	154
Student 10	0.027777778	0.0000000	0.000000000	5.664850e-02	794
Student 12	0.027777778	0.0000000	0.000000000	3.216265e-02	67
Student 4	0.009259259	0.0000000	0.000000000	1.924626e-02	475
Student 7	0.009259259	0.0000000	0.000000000	1.819566e-02	116
Student 14	0.000000000	0.0000000	0.000000000	3.040169e-18	133
Student 15	0.000000000	0.0000000	0.000000000	3.040169e-18	376
Student 16	0.000000000	0.0000000	0.000000000	2.885609e-18	389
Student 17	0.000000000	0.0000000	0.000000000	2.818244e-18	14
Guest Teacher	0.000000000	0.0000000	0.000000000	2.312523e-18	75
Student 19	0.000000000	0.0000000	0.000000000	2.196485e-18	11
Student 21	0.000000000	0.0000000	0.000000000	1.991940e-18	70
Student 22	0.000000000	0.0000000	0.000000000	0.000000e+00	7

Figure 23. Scoring of students based on centrality measures for Cohort

Centrality degree illustrates how interactive (central) a student is in the forum in this work. Betweenness centrality shows the frequency (number) of each student's post and Eigenvector centrality shows each student's degree of influence. Since the aim of this work is to provide the instructors with specific quantitative measures for evaluating the activity of the students

in the online discussion forum, these measures are further analyzed utilizing Principal Component Analysis (PCA) to identify the measure that explains the highest variance.

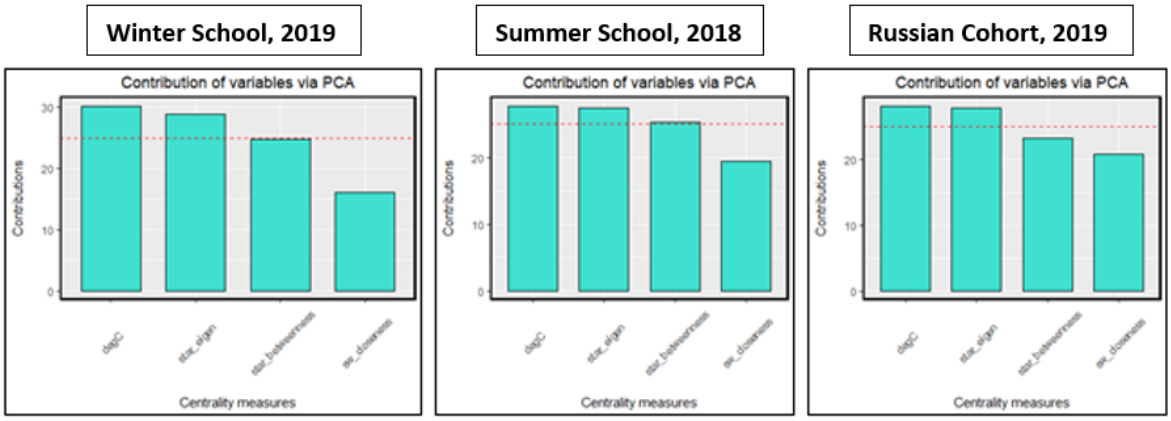


Figure 24. Contribution of different centrality measure based on PCA for each cohort

The result of PCA illustrates in ***Figure 24*** that Degree centrality and Eigenvector centrality may have sufficient explanatory power to evaluate the students' performance.

Pearson correlation of different centrality measures

The Pearson correlation is calculated among various measures for each cohort to test the hypothesis:

	Degree Centrality	Closeness	Betweenness	Eigenvector	Words_count
Degree Centrality	1	0.6368003	0.9198131	0.9625202	0.9111444
Closeness	0.6368003	1	0.4022015	0.646138	0.611013
Betweenness	0.9198131	0.4022015	1	0.8332505	0.7623324
Eigenvector	0.9625202	0.646138	0.8332505	1	0.9680857
Words_count	0.9111444	0.611013	0.7623324	0.9680857	1

Figure 25. Pearson correlation for Cohort 1

	Degree Centrality	Closeness	Betweenness	Eigenvector	Words_count
Degree Centrality	1	0.75622274	0.96033884	0.99165117	0.00625931
Closeness	0.756222736	1	0.63510861	0.76458811	-0.07235411
Betweenness	0.960338837	0.63510861	1	0.93169173	-0.07577916
Eigenvector	0.991651173	0.76458811	0.93169173	1	0.03957237
Words_count	0.006259314	-0.07235411	-0.07577916	0.03957237	1

Figure 26. Pearson correlation for Cohort 2

	Degree Centrality	Closeness	Betweenness	Eigenvector	Words_count
Degree Centrality	1	0.636685	0.9621701	0.8395251	0.7358558
Closeness	0.636685	1	0.4522624	0.8934562	0.5595366
Betweenness	0.9621701	0.4522624	1	0.700869	0.6507767
Eigenvector	0.8395251	0.8934562	0.700869	1	0.7264535
Words_count	0.7358558	0.5595366	0.6507767	0.7264535	1

Figure 27. Pearson correlation for Cohort 3

For cohort 1 and cohort 3, the results depict strong positive correlation between Degree centrality measures and Eigenvector centrality measure, which stands for determining influential nodes/students of the blog. Both these measures also have high positive correlation with number of meaningful words. Additionally, Degree of Centrality have high correlation with closeness and Betweenness centrality for Cohort 1. For Cohort 2, Eigenvector centrality illustrates higher positive correlation with meaning words than Degree of Centrality. There is also high positive correlation between Degree Centrality and Eigenvector Centrality. Based on the correlation results in ***Figure 25,26,27*** and previous PCA analysis in ***Figure 24***, it may be reasonable to consider Degree of Centrality to assess the students' engagement in the discussion forum and assign scores based on it. In order to test our hypothesis, correlation test (significance level) has been conducted between **Degree of Centrality** and **number of words** for each cohort. The significant P-value in 5% confidence interval for each cohort supports our null hypothesis that High frequency of meaningful words leads to higher degree of interaction.

Limitation of the results

In the sentiment analysis, few students/users have zero score, as their posted words did not hit any similarity measures with any words of the used lexicon. Moreover, students' behavior/communication pattern may change if they know the approach of the assessment. Many students may try to post irrelevant texts frequently to obtain higher score. This limitation also supports the idea of developing subject specific lexicon to assess the quality and relevance of the students' posts in the blog.

In summary, the main advantage of the work is its simplistic approach, easy to interpret and most importantly can be replicated for any blog discussion.

5.3 LMS (Moodle) Data Analysis

Moodle data from "Computer Science" course at Tomsk State University of Control Systems and Radioelectronics, Russia is used for this analysis. The main objective of this analysis is

to predict a learning path and classification of students based on the LMS data. The hypothesis is:

"Frequency of activity in LMS predicts learners' final score."

Data Preprocessing

The data was extracted from the LMS (Moodle) as a form of log. The log contains observation of 93,893 instances with lots of missing values, which makes it a classic problem of big data analysis. As illustrated in the architecture diagram (**Figure *11*.** Architecture of LMS data Analysis), data preprocessing was the primary step to start the analysis. The data preprocessing was done using R. The main challenge of data preprocessing was handling missing data and identifying the parameters for data modeling. As a first step, KNN imputation was used to fill out the missing value. It is popular and effective method for missing data handling as it works with discrete, continuous, categorical and ordinal all kind of data. Since the overall grade was not available, Students having quiz grade more than 10% considered to be completed the course. The dependent variable was constructed as a categorical variables based on this logic having value of 0 (not complete/fail) and 1 (complete/pass). The next steps was to identify the independent/predictor variables, which could be aggregated against grade for further modeling. In order to do so, total 405 number of unique user id was identified. Each user id represents a student, which means total 405 students participated the course. Then, the variables depicted in Table 16 is aggregated for each students to prepare the final dataset.

Table 16. Variables used in the analysis

Variables	Description
Lesson View	Total number of times each student performed different lesson
Course Page View	Adding up the number of times each student view the course page. Assuming this reflects the number of times they used the LMS for learning in this course
Downloads	counting the number of course materials download by each student

Table 17 continues. Variables used in the analysis

Variables	Description
Link View	Total number of link (mainly video lectures) view by each student
Quiz Performed	Counting number of quiz performed by students

The final dataset contains a record of 405 observations with 6 rows. The data analysis is also conducted in R.

Data Analysis and Results

As a first step of analysis, multivariate regression model is developed aiming to see causal relationship between students' course completion (pass) and activity in the LMS (variables illustrated in Table 16). The regression model equation:

$$Completed = \beta_o + \beta_1 lesson\ view + \beta_2 courseapge\ view + \beta_3 downloads + \beta_4 link\ view + \beta_5 quiz\ performed + \epsilon$$

In below **Figure *28*** illustrates the results of regression model,

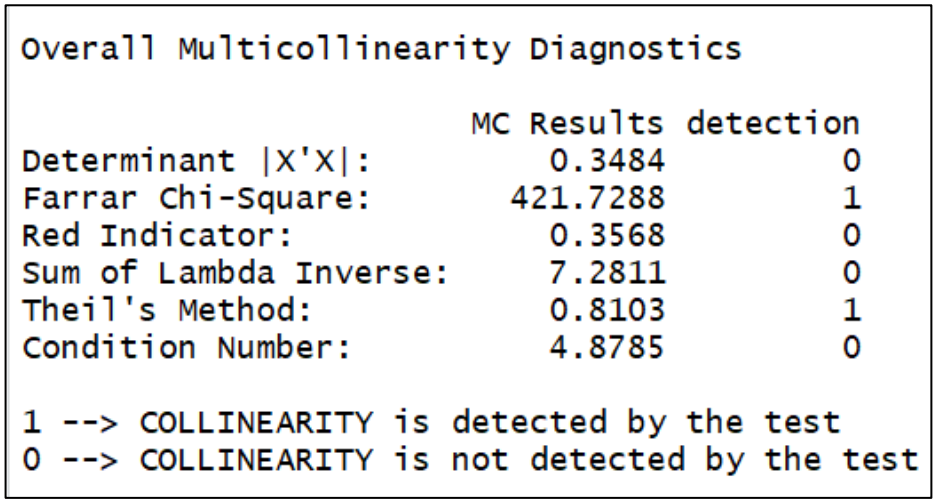
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.01220    2.14106   3.275 0.001149 **
lesson      -0.00453    0.01800  -0.252 0.801408
CourseView   0.06875    0.04743   1.449 0.148003
download     1.30573    0.33473   3.901 0.000113 ***
linkView     0.07463    0.07365   1.013 0.311577
Quiz         0.90150    0.19968   4.515 8.36e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 24.64 on 398 degrees of freedom
Multiple R-squared:  0.1739,    Adjusted R-squared:  0.1635
F-statistic: 16.76 on 5 and 398 DF,  p-value: 4.957e-15
```

Figure 28. Results of Regression Model

From the results, it seems that only "Downloads" and "Quiz performed" is significant for 5% confidence interval. The significant F-value implies that all the predictor value together can explain the variability of the dependent variable (students' course completion). However, the low R square value (16.35%) suggests that the model may not be suitable to use for predicting students' course completion.

The insignificant value of the variables and low R square value led to checking multicollinearity of variables. It is a phenomenon when the independent variables are highly correlated (linear functions) with each other. It means that they explain same thing. Multicollinearity violates the basic assumption of regression. We conducted Farrar-Glauber test to check overall and individual diagnostics of multicollinearity. **Figure *29*** illustrates the result of multicollinearity test.



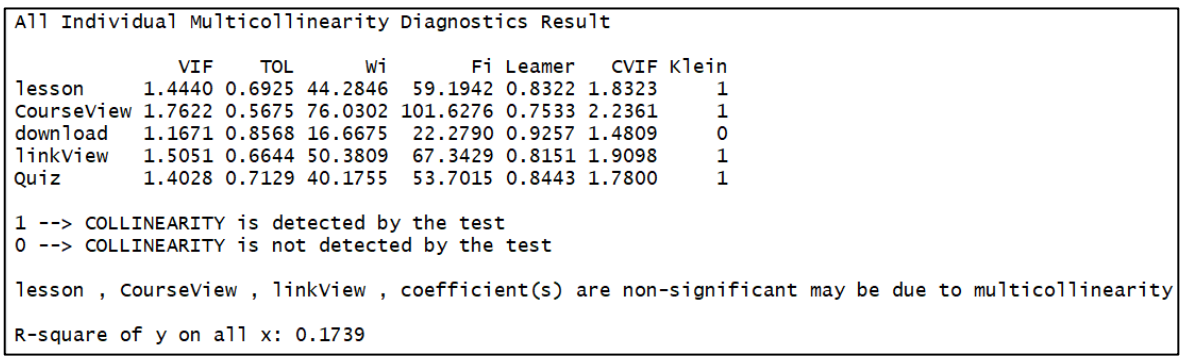
```
Overall Multicollinearity Diagnostics

                          MC Results detection
Determinant |X'X|:            0.3484         0
Farrar Chi-Square:          421.7288         1
Red Indicator:                0.3568         0
Sum of Lambda Inverse:        7.2811         0
Theil's Method:               0.8103         1
Condition Number:             4.8785         0

1 --> COLLINEARITY is detected by the test
0 --> COLLINEARITY is not detected by the test
```

Figure 29. Overall multicollinearity test

The high chi-square statistics value implies the presence of multicollinearity. In order to identify the variables dealing with multicollinearity, we conducted individual diagnosis of multicollinearity. The result of the test (**Figure *30***) suggests that “Lesson View”, “Course Page View” and “Link View” may be non-significant due to multicollinearity.



```
All Individual Multicollinearity Diagnostics Result

              VIF    TOL      Wi       Fi Leamer   CVIF Klein
lesson     1.4440 0.6925 44.2846  59.1942 0.8322 1.8323     1
CourseView 1.7622 0.5675 76.0302 101.6276 0.7533 2.2361     1
download   1.1671 0.8568 16.6675  22.2790 0.9257 1.4809     0
linkView   1.5051 0.6644 50.3809  67.3429 0.8151 1.9098     1
Quiz       1.4028 0.7129 40.1755  53.7015 0.8443 1.7800     1

1 --> COLLINEARITY is detected by the test
0 --> COLLINEARITY is not detected by the test

lesson , CourseView , linkView , coefficient(s) are non-significant may be due to multicollinearity

R-square of y on all x: 0.1739
```

Figure 30. Individual multicollinearity test

Therefore, it will not be statistically correct to build a linear regression model with all 5 variables highlighted in Table 1 as it violates the basic assumption of linear regression. It is possible to build a model excluding the variables with multicollinearity but it will limit the

scope of the study. Since the objective of the study is to predict students' course outcome (pass or fail) based on LMS activity, it is necessary to build a model with the inclusion of all identified variable. In order to do so, we employed classification tree model because of it is simplistic in nature and not affected by multicollinearity. The behind principle of classification tree is discussed briefly in chapter 3. The dataset is divided into training set (70%) and test set (30%). The result of the analysis (illustrated below in **Figure *31***) exhibits a learning path of students based on their activity on LMS to predict course outcome (pass or fail).

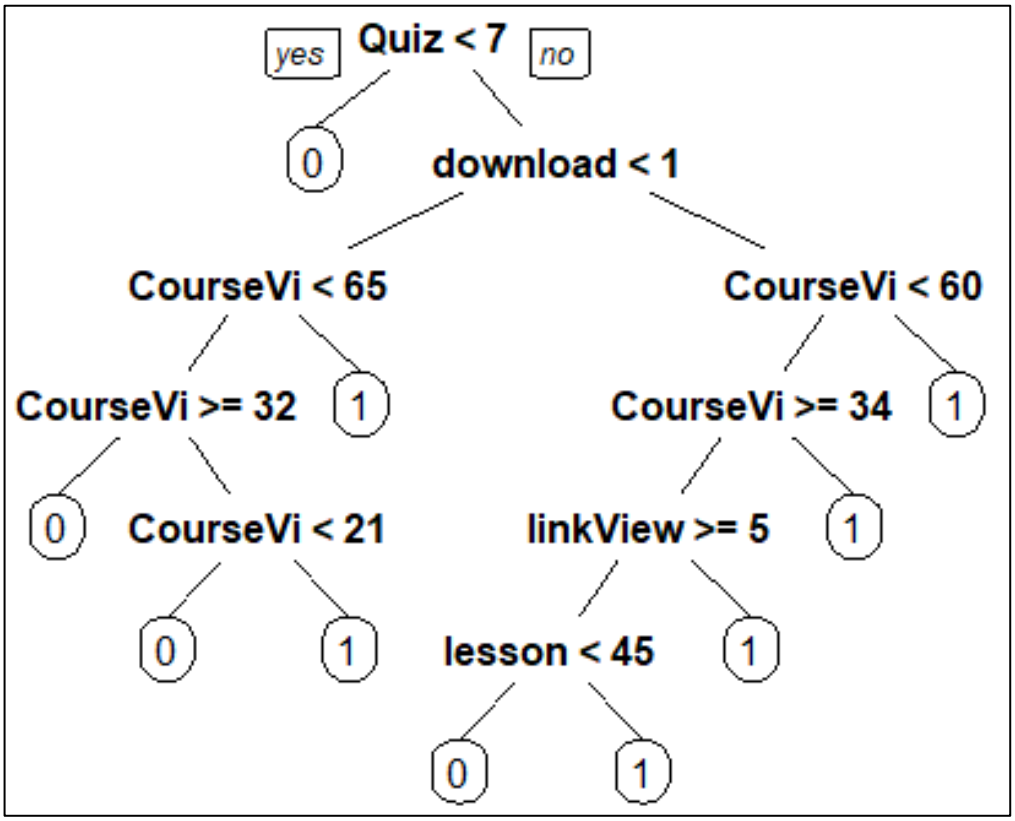


Figure 31. Learning path of students to predict course outcome (Pass or Fail)

It is evident from the classification tree that "Number of Quiz performed" depicts the main criteria (root node) to predict students' outcome. The value of this variable is also significant in the regression analysis (**Figure *28*.** Results of Regression Model) as well. The performance of the classification is evaluated on the test set by using confusion matrix. In the below Table 18 depicts the result of the classification tree. The high accuracy rate of 70% justifies the use of such an approach to predict the students' course outcome. The result also suggests in favor of our hypothesis that the frequency of LMS activity predicts learners' final score.

Table 18. Performance measure of classification tree

	No	Yes
No	52	19
Yes	17	33
Accuracy	0.702493	
Sensitivity (True Positive Rate)	0.6346154	
Specificity (True Negative Rate)	0.7536232	

In the next step, K-means clustering algorithm is applied to cluster the successful students based on their LMS activity pattern. These clustering provide insights about the major trend in students' activity pattern that influences the chances of success. The number of optimal cluster (K value) is determined as 4 based on the elbow method illustrated in **Figure *32***.

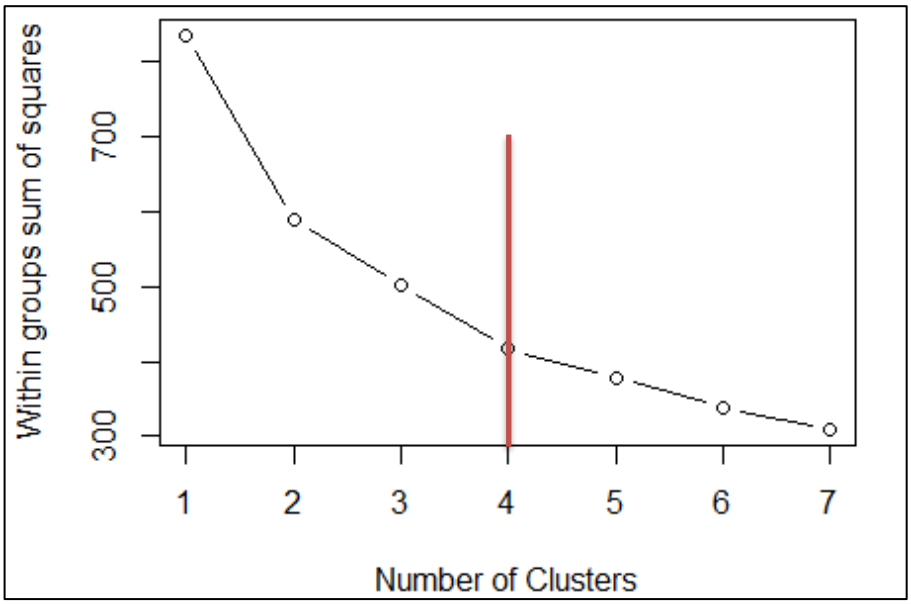


Figure 32. Elbow method to evaluate number of clusters

The results of K-means clustering for the successful students is depicted in Table 19. It seems that the majority of successful students obtained better grades with less quiz attempt. Moreover, the number of lesson attempts also depicts significant influence on the success.

Table 19. Cluster of students based on K-means clustering

Group	Lesson View	Course Page View	Downloads	Link View	Quiz Performed	Number of Students
1	30.35135	35.24324	2.603604	16.90991	9.342342	111
2	51.78261	89.04348	3.304348	20.69565	24.434783	23
3	109.63636	73.45455	17.181818	34.09091	11.909091	11
4	250.90909	86.54545	3.136364	52.72727	12.863636	22

Limitation of the results

One of the major limitations of the results is the usage of Moodle log data. The Moodle log consists records of learners' activity. It does not contain the explanation or logic for the significance of certain variables over others. Further research consisting survey, focus group discussion of the students should be conducted to determine the reasons for the significance. Another limitation is the pedagogic design or learning format of the course. It was more like a blended course rather than entirely online. Therefore, there were several in-class activities, which may also have influence on learners' success but was not possible to capture by the LMS data. In order to obtain better results, this method should be applied on fully online course.

5.4 Result Summary and Data Informed Decision

The above section describes the objective, limitation and implication of results for each method. The conformity of hypothesis and explanation on the research questions illustrated in Table 8, also discussed in details. This section highlights how the results of each method can be mapped with the "Learning analytics and assessment framework" illustrated in **Figure *1*** and help the instructors/teachers to take data driven decision.

The problem statement and the objective of the work was discussed in chapter 1. Then, we try to evaluate the usability of the proposed framework based on three research questions and hypotheses illustrated in Table 1. Research questions. In chapter 2, we discussed the background, different dimensions and essence of adopting learning analytics in digital education. We reviewed state of art with respect to learning analytics and highlighted our systematic approach (Table 4. Selection Criteria of Papers) of selecting relevant important

studies. Based on the literature review, we presented the relevant works using tabular format in Table 5 and Table 6 into two types: methods and research goals. So, it becomes easily understandable to the readers. In Chapter 3, we discussed briefly the theory and behind principles of the used methodologies so that the readers can easily understand the interpretation of results. In chapter 5, we evaluated the framework (**Figure *1*.** E-learning/flipped learning analytics and assessment framework) based on the research questions and hypotheses associated with each method.

Method 1 (Video Analytics) focuses on improvement of course contents mainly video lectures based on learners' engagement after each cohort. The results exhibit that the learners' have interest in concept related videos for this subject matter. If we aim for 50% audience retention rate, we can improve the video lectures by limiting the concept related video lectures under 6 minutes. However, we need to keep in mind the design recommendation from "nose, body and tail" analysis section. It is evident from method 1 analysis and interpretation of the results, the teachers can take data driven decision for course content improvement (video lectures) after each cohort by simply iterating the method 1 analysis.

Method 2 (Discussion Forum Analysis) focuses on assessment of students' active engagement and interactivity of course discussion forum. Teachers' can immediately get insights on the interactivity of the course discussion forum for each cohort by looking into the **Figure *20*.** Social Network Graph of each cohort. This picture also allows the teacher to improve the course discussion forum for the next cohort since they already have an understanding how and to what extent the course discussion forum performed in previous cohorts. The scores from the centrality measures (illustrated in **Figure *21***, **Figure *22*** and **Figure *23***) allows the teacher to evaluate the performance and active engagement of the students in quantitative measure. For an example, if teacher wants to provide 10% marks/grade to the students based on their interactivity in the discussion forum, they simply can take the score from Degree Centrality measure and convert it into 10 %.

Method 3 (LMS Data Analysis) focuses on predicting students' course outcome based on their activity in the LMS. The predicted learning path based on the classification tree illustrated in **Figure *31*** exhibits that Quiz activity plays most influential role on learners'

success. Based on this result, the instructors can introduce more quiz activity in the pedagogic design of the course in order to increase learners' success rate. However, the learners' LMS activity pattern analysis in Table 19 depicts that majority of the successful students (cluster group 1: total 111 students) have not participated in the LMS activity a lot but successfully completed course. This result put the applicability and usability of the LMS in question. It also indicates that classroom activity may have greater influence in students' learning rather than the online activity. It also strongly suggests that the LMS (Moodle) was not properly utilized for this course. The instructors should reconsider the overall LMS activities and their usage in the course pedagogic design if they want to offer this course digitally for next cohort.

6 CONCLUSION

This research to practice work speculate about the applicability and usability of "Learning Analytics Framework" in digital education. Adoption of such framework can be beneficial to the teachers, instructors and instructional designers for continuous improvement and delivery of digital course contents based on a data driven approach. The main strength of the study is that the author not only propose the framework but also put it into practice. The author applied the method on two courses: "Systematic Creativity and TRIZ basics" at Lappeenranta University of Technology, Finland and "Computer Science" at Tomsk State University of Control Systems and Radioelectronics, Russia. The framework (**Figure *1*.** E-learning/flipped learning analytics and assessment framework) is evaluated based on the research questions and hypotheses associated with each method. The results from each method is mapped as input to the corresponding block of the framework and data driven decision is taken to improve the course contents and pedagogic design as an output. The result from Method 1 (Video Analytics) depicts that 1 unit change in video length, audience retention changes by -1.6777 unit for this VBL. It also suggests that the instructors should focus on improving the concept related video lectures by shortening the length as these videos have higher audience interest. Method 2 (Discussion Forum Analysis) focuses on assessment of students' active engagement in online discussion. From the social network graph for different cohort illustrated in **Figure *20***, the instructors can immediately speculate about the degree of interaction of the course discussion forum and take necessary action for the next cohort. The analysis also provides a set of measured score for each students, which allow the teachers to assign grade for active communication in an online course. The results from Method 3 (LMS Data Analysis) allows the teachers to measure the usability of LMS for the course and predict a learning path for the students using their activity data from the LMS.

The major advantage of this work is that the methods and analysis results are easy to use and interpret. The methods also perform well on smaller datasets. Moreover, the work can be easily replicated so that the instructors and instructional designer can adopt it without prior depth understanding on the methods. This work provides a solid background to conduct

further research on this subject matter and adoption of standard learning analytics in digital education.

The future work may include applying other machine-learning algorithm and statistical techniques to evaluate the proposed method. For video analytics, data can be collected in a controlled environment so that the analysis can be done only for the internal students. Subject specific lexicon can be developed to analyze the relevance of learners' post and meaningful words. It is also possible to develop a web app where teachers can upload their blog data and get the score of the students on their active engagement based on centrality measures. This web app can be used an assessment tool for course discussion forum. Support Vector Machine (SVM) can also be applied to analyze the social network graphs. In order to predict students' learning path based on the LMS data, it is possible to evaluate the application of random forest and convolution neural network (CNN). All of these mentioned algorithms require large number of observation. Unavailability of data/Large sample size is one of the major challenges in educational data mining and adopting learning analytics. The first step of the future work can be formulating a standard for data harvesting in digital education. The work was partially supported by CEPHEI project of ERASMUS+ EU framework which is an ongoing project focusing on digitalization of industrial innovation related contents. The authors hope to conduct further research and extend this work in future with possible support from this project.

LIST OF REFERENCES

Abdous, M., He, W. and Yen, C. J. (2012) 'Using data mining for predicting relationships between online question theme and final grade', *Educational Technology and Society*.

Adamopoulos, P. (2013) 'What makes a great MOOC? An interdisciplinary analysis of student retention in online courses', in *International Conference on Information Systems, ICIS 2013*.

Albert, M. and Beatty, B. J. (2014) 'Flipping the Classroom Applications to Curriculum Redesign for an Introduction to Management Course: Impact on Grades', *Journal of Education for Business*. doi: 10.1080/08832323.2014.929559.

Ali, L. *et al.* (2012) 'A qualitative evaluation of evolution of a learning analytics tool', *Computers and Education*. doi: 10.1016/j.compedu.2011.08.030.

Antonova, N., Shnai, I. and Kozlova, M. (2017) 'Flipped classroom in the higher education system: A pilot study in Finland and Russia', *New Educational Review*, 48(2), pp. 17–27. doi: 10.15804/tner.2017.48.2.01.

Ash, K. (2012) 'Educators View "Flipped" Model With a More Critical Eye: Benefits and drawbacks seen in replacing lectures with on-demand video', *Education Week*. doi: 10.3115/991365.991409.

Barla, M. *et al.* (2010) 'On the impact of adaptive test question selection for learning efficiency', *Computers and Education*. doi: 10.1016/j.compedu.2010.03.016.

Baron, J., Willis, J. and Lee, R. A. (2010) 'Creating higher education academic and information technology resources in an international context', *Computers in the Schools*. doi: 10.1080/07380569.2010.523885.

Barthélemy, M. (2004) 'Betweenness centrality in large complex networks', *European Physical Journal B*, 38(2), pp. 163–168. doi: 10.1140/epjb/e2004-00111-4.

Beichner, R. J. *et al.* (2007) 'The Student-Centered Activities for Large Enrollment Undergraduate Programs (SCALE-UP) Project', *Physics*, 1(1), pp. 1–42. doi: 10.1093/schbul/sbp059.

Bergmann, J. and Sams, A. (2009) 'Remixing Chemistry Class: Two Colorado Teachers Make Vodcasts of Their Lectures to Free Up Class Time for Hands-On Activities', *Learning & Leading with Technology*.

Bienkowski, M., Feng, M. and Means, B. (2012) 'EDM-00: Enhancing teaching and learning

through educational data mining and learning analytics', in *USA 2018-01-11 77p*.

Blei, D. M. D. *et al.* (2014) 'Online learning for latent dirichlet allocation', *Naacl*. doi: 10.1.1.100.1089.

Bonacich, P. (2002) 'Power and Centrality: A Family of Measures', *American Journal of Sociology*. doi: 10.1086/228631.

Bonacich, P. (2007) 'Some unique properties of eigenvector centrality', *Social Networks*, 29(4), pp. 555–564. doi: 10.1016/j.socnet.2007.04.002.

Bryson, C. and Hand, L. (2007) 'The role of engagement in inspiring teaching and learning', *Innovations in Education and Teaching International*, 44(4), pp. 349–362. doi: 10.1080/14703290701602748.

Campbell, J. P., DeBlois, P. B. and Oblinger, D. G. (2007) 'Academic Analytics: A New Tool for a New Era', *EDUCAUSE Review*.

Chaplot, D. S., Rhim, E. and Kim, J. (2015) 'Predicting student attrition in MOOCs using sentiment analysis and neural networks', *CEUR Workshop Proceedings*, 1432(June), pp. 7–12.

Chen, C. M. and Chen, M. C. (2009) 'Mobile formative assessment tool based on data mining techniques for supporting web-based learning', *Computers and Education*. doi: 10.1016/j.compedu.2008.08.005.

Clow, D. and Makriyannis, E. (2011) 'iSpot analysed: Participatory learning and reputation', in *Proceedings of the 1st International Conference on Learning Analytics and Knowledge*. doi: 10.1145/2090116.2090121.

Cole, J. *et al.* (2014) *Using Moodle, Igarss 2014*. doi: 10.1007/s13398-014-0173-7.2.

Cole, J. and Foster, H. (2007) *Using Moodle: Teaching with the Popular Open Source Course Management System, Journal of Chemical Information and Modeling*. doi: 10.1017/CBO9781107415324.004.

Cormack, A. (2016) 'A data protection framework for learning analytics', *Journal of Learning Analytics*. doi: 10.18608/jla.2016.31.6.

Cormier, D. and Siemens, G. (2010) 'Through the Open Door: Open Courses as Research, Learning, and Engagement', *Educause*.

Creavity LAB, LUT (2015) *TRIZ Youtube Channel*. Available at: https://www.youtube.com/channel/UCqr4R5hyHjs1ve-4znD0asQ?view_as=subscriber (Accessed: 25 June 2019).

Critz, C. M. and Knight, D. (2013) 'Using the flipped classroom in graduate nursing

education', *Nurse Educator*. doi: 10.1097/NNE.0b013e3182a0e56a.

Currier, A. and Fishman, E. (no date) *Understanding Audience Retention*. Available at: https://wistia.com/learn/marketing/understanding-audience-retention (Accessed: 23 February 2019).

David, R. (2018) *When Learning Analytics Violate Student Privacy*. Available at: https://campustechnology.com/Articles/2018/05/02/When-Learning-Analytics-Violate-Student-Privacy.aspx?Page=1 (Accessed: 6 July 2019).

Dejaeger, K. *et al.* (2012) 'Gaining insight into student satisfaction using comprehensible data mining techniques', *European Journal of Operational Research*. doi: 10.1016/j.ejor.2011.11.022.

Dekker, A. (2002) 'Applying Social Network Analysis Concepts to Military C4ISR Architectures', *Connections*.

Dekker, G. W., Pechenizkiy, M. and Vleeshouwers, J. M. (2009) 'Predicting Students Drop Out: A Case Study', in *the 2nd International Conference on Educational Data Mining*.

Dougiamas, M. and Taylor C, P. (2003) 'Moodle: Using learning communities to create an open source course management system', in *In EdMedia: World Conference on Educational Media and Technology. Association for the Advancement of Computing in Education (AACE)*.

Early, S. L. (2011) 'Book Review Student Engagement Techniques: A Handbook for College Faculty', *Journal of the Scholarship of Teaching and Learning*, 11(1), pp. 155–157.

Eriksson, T., Adawi, T. and Stöhr, C. (2017) '"Time is the bottleneck": a qualitative study exploring why learners drop out of MOOCs', *Journal of Computing in Higher Education*. doi: 10.1007/s12528-016-9127-8.

Faust, K. and Wasserman, S. (1995) *Social Network Analysis Methods and Applications Structural Analysis in the Social Sciences: Methods and Applications, Structural Analysis in the Social Sciences*.

Ferguson, R. (2013) 'Learning analytics: drivers, developments and challenges', *International Journal of Technology Enhanced Learning*, 4(5/6), p. 304. doi: 10.1504/ijtel.2012.051816.

Ferguson, R. *et al.* (2016) 'Setting Learning Analytics in Context: Overcoming the Barriers to Large-Scale Adoption', *Journal of Learning Analytics*. doi: 10.18608/jla.2014.13.7.

Ferrara, E. (2018) 'Measurement and Analysis of Online Social Networks Systems', in *Encyclopedia of Social Network Analysis and Mining*. doi: 10.1007/978-1-4939-7131-

2_242.

Ferreri, S. P. and O'Connor, S. K. (2013) 'Redesign of a large lecture course into a small-group learning course', *American Journal of Pharmaceutical Education*. doi: 10.5688/ajpe77113.

Fournier, H., Kop, R. and Sitlia, H. (2011) 'The value of learning analytics to networked learning on a personal learning environment', in *Proceedings of the 1st International Conference on Learning Analytics and Knowledge - LAK '11*. doi: 10.1145/2090116.2090131.

Freeman, L. C. (1978) 'Centrality in social networks conceptual clarification', *Social Networks*. doi: 10.1016/0378-8733(78)90021-7.

Freeman, L. C. (2006) 'A Set of Measures of Centrality Based on Betweenness', *Sociometry*. doi: 10.2307/3033543.

Galagan, P. A. (2001) 'Mission E-possible: The cisco E-learning story', *Training & Development*.

Galway, L. P. *et al.* (2014) 'A novel integration of online and flipped classroom instructional models in public health higher education.', *BMC medical education*. doi: 10.1186/1472-6920-14-181.

Giannakos, M. N. *et al.* (2013) 'Analytics on video-based learning', in. doi: 10.1145/2460296.2460358.

Giannakos, M. N., Chorianopoulos, K. and Chrisochoides, N. (2015) 'Making sense of video analytics: Lessons learned from clickstream interactions, attitudes, and learning outcome in a video-assisted course', *International Review of Research in Open and Distance Learning*.

Giesbers, B. *et al.* (2013) 'Investigating the relations between motivation, tool use, participation, and performance in an e-learning course using web-videoconferencing', *Computers in Human Behavior*. doi: 10.1016/j.chb.2012.09.005.

Greller, W. and Drachsler, H. (2012) 'Translating Learning into Numbers: A Generic Framework for Learning Analytics Multimodal Learning Analytics for Collaborative Learning Understanding and Support View project', 15, pp. 42–57. Available at: http://groups.google.com/group/learninganalytics.

Guo, P. J., Kim, J. and Rubin, R. (2014) 'How video production affects student engagement', in *Proceedings of the first ACM conference on Learning @ scale conference - L@S '14*. doi: 10.1145/2556325.2566239.

Hanneman, R. A. and Riddle, M. (1998) 'Introduction to Social Network Methods',

Network. doi: 10.1109/78.700969.

Heilesen, S. B. (2010) 'What is the academic efficacy of podcasting?', *Computers and Education*. doi: 10.1016/j.compedu.2010.05.002.

Hollander, E. L., Saltmarsh, J. and Zlotkowski, E. (2011) 'Indicators of Engagement', in. doi: 10.1007/978-1-4615-0885-4_3.

Hone, K. S. and El Said, G. R. (2016) 'Exploring the factors affecting MOOC retention: A survey study', *Computers and Education*. doi: 10.1016/j.compedu.2016.03.016.

Huang, S. and Fang, N. (2013) 'Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models', *Computers and Education*. doi: 10.1016/j.compedu.2012.08.015.

Jensen, J. L., Kummer, T. A. and Godoy, P. D. D. M. (2015) 'Improvements from a flipped classroom may simply be the fruits of active learning', *CBE Life Sciences Education*. doi: 10.1187/cbe.14-08-0129.

Jordan, K. (2014) 'Initial trends in enrolment and completion of massive open online courses', *International Review of Research in Open and Distance Learning*.

Jordan, K. (2015) 'Massive open online course completion rates revisited: Assessment, length and attrition', *International Review of Research in Open and Distance Learning*.

Kapil, S. and Chawla, M. (2017) 'Performance evaluation of K-means clustering algorithm with various distance metrics', in *1st IEEE International Conference on Power Electronics, Intelligent Control and Energy Systems, ICPEICES 2016*. doi: 10.1109/ICPEICES.2016.7853264.

Khribi, M. K. (2013) 'A web mining based approach for automatic student model discovery', in *2013 4th International Conference on Information and Communication Technology and Accessibility, ICTA 2013*. doi: 10.1109/ICTA.2013.6815287.

Khribi, M. K., Jemni, M. and Nasraoui, O. (2009) 'Automatic recommendations for e-learning personalization based on web usage mining techniques and information retrieval', *Educational Technology and Society*.

Kim, J. *et al.* (2014) 'Understanding in-video dropouts and interaction peaks inonline lecture videos', in. doi: 10.1145/2556325.2566237.

Kizilcec, R. F., Piech, C. and Schneider, E. (2013) 'Deconstructing disengagement: analyzing learner subpopulations in massive open online courses', in *LAK '13: Proceedings of the Third International Conference on Learning Analytics and Knowledge*.

Kolowich, S. (2013) 'Coursera Takes a Nuanced View of MOOC Dropout Rates.', *The*

chronicle of higher education.

Lage, M. J. and Platt, G. (2000) 'The internet and the inverted classroom', *Journal of Economic Education*. doi: 10.1080/00220480009596756.

Leong, C. K., Lee, Y. H. and Mak, W. K. (2012) 'Mining sentiments in SMS texts for teaching evaluation', *Expert Systems with Applications*. doi: 10.1016/j.eswa.2011.08.113.

Li, N. *et al.* (2012) 'A Machine Learning Approach for Automatic Student Model Discovery', *EDM 2011 4th International Conference on Educational Data Mining*.

Lin, C. F. *et al.* (2013) 'Data mining for providing a personalized learning path in creativity: An application of decision trees', *Computers and Education*. doi: 10.1016/j.compedu.2013.05.009.

Lin, F. R., Hsieh, L. S. and Chuang, F. T. (2009) 'Discovering genres of online discussion threads via text mining', *Computers and Education*. doi: 10.1016/j.compedu.2008.10.005.

Macfadyen, L. P. and Dawson, S. (2010) 'Mining LMS data to develop an "early warning system" for educators: A proof of concept', *Computers and Education*. doi: 10.1016/j.compedu.2009.09.008.

Mak, S. F. J., Williams, R. and Mackness, J. (2010) 'Blogs and forums as communication and learning tools in a MOOC', in *Proceedings of the 7th International Conference on Networked Learning*.

Martin, S. K., Farnan, J. M. and Arora, V. M. (2013) 'Future: New strategies for hospitalists to overcome challenges in teaching on today's wards', *Journal of Hospital Medicine*. doi: 10.1002/jhm.2057.

Mason, G. S., Shuman, T. R. and Cook, K. E. (2013) 'Comparing the effectiveness of an inverted classroom to a traditional classroom in an upper-division engineering course', *IEEE Transactions on Education*, 56(4), pp. 430–435. doi: 10.1109/TE.2013.2249066.

Mason, R. (2014) *Elearning: The Key Concepts*, *Elearning: The Key Concepts*. doi: 10.4324/9780203099483.

Matlab (no date) *K-means Clustering*. Available at: https://se.mathworks.com/help/stats/kmeans.html (Accessed: 28 June 2019).

McDonald, K. and Smith, C. M. (2013) 'The Flipped Classroom for Professional Development: Part I. Benefits and Strategies', *The Journal of Continuing Education in Nursing*. doi: 10.3928/00220124-20130925-19.

McLaughlin, J. E. *et al.* (2013) 'Pharmacy student engagement, performance, and perception in a flipped satellite classroom', *American Journal of Pharmaceutical Education*. doi:

10.5688/ajpe779196.

McLaughlin, J. E. *et al.* (2014) 'The flipped classroom: A course redesign to foster learning and engagement in a health professions school', *Academic Medicine*. doi: 10.1097/ACM.0000000000000086.

Missildine, K. *et al.* (2013) 'Flipping the Classroom to Improve Student Performance and Satisfaction', *Journal of Nursing Education*. doi: 10.3928/01484834-20130919-03.

Mohamed Chatti., A. *et al.* (2012) 'A reference model for learning analytics', *International Journal of Technology Enhanced Learning*, 4(5–6), pp. 1–22.

Moodle Statistics Page (no date). Available at: https://moodle.net/stats/ (Accessed: 26 June 2019).

Mostow, J. *et al.* (2005) 'An Educational Data Mining Tool to Browse Tutor-Student Interactions : Time Will Tell !', *Proceedings of the Workshop on Educational Data Mining (2005)*.

Onah, D. F. O., Sinclair, J. and Boyatt, R. (2014) 'Dropout rates of massive open online courses: Behavioural patterns', in *EDULEARN14 Proceedings*.

Pantò, E. and Comas-Quinn, A. (2013) 'The challenge of open education', *Journal of E-Learning and Knowledge Society*.

Pardo, A. and Siemens, G. (2014) 'Ethical and privacy principles for learning analytics', *British Journal of Educational Technology*. doi: 10.1111/bjet.12152.

Pardos, Z. A. *et al.* (2016) 'Affective States and State Tests: Investigating How Affect and Engagement during the School Year Predict End-of-Year Learning Outcomes', *Journal of Learning Analytics*. doi: 10.18608/jla.2014.11.6.

Pierce, R. and Fox, J. (2012) 'Vodcasts and active-learning exercises in a "flipped classroom" model of a renal pharmacotherapy module', in *American Journal of Pharmaceutical Education*. doi: 10.5688/ajpe7610196.

Plasencia, A. and Navas, N. (2014) 'MOOCs, the flipped classroom, and khan academy practices: The implications of augmented learning', in *Innovation and Teaching Technologies: New Directions in Research, Practice and Policy*. doi: 10.1007/978-3-319-04825-3_1.

Prinsloo, P. and Slade, S. (2013) 'An evaluation of policy frameworks for addressing ethical considerations in learning analytics', in. doi: 10.1145/2460296.2460344.

Prober, C. G. and Khan, S. (2013) 'Medical education reimagined: A call to action', *Academic Medicine*, 88(10), pp. 1407–1410. doi: 10.1097/ACM.0b013e3182a368bd.

Rabbany, R. *et al.* (2014) 'Collaborative learning of students in online discussion forums: A social network analysis perspective', *Studies in Computational Intelligence*. doi: 10.1007/978-3-319-02738-8_16.

Rajaraman, A. and Ullman, J. D. (2011) *Mining of massive datasets, Mining of Massive Datasets*. doi: 10.1017/CBO9781139058452.

Romero-Zaldivar, V. A. *et al.* (2012) 'Monitoring student progress using virtual appliances: A case study', *Computers and Education*. doi: 10.1016/j.compedu.2011.12.003.

Romero, C. *et al.* (2008) 'Data mining algorithms to classify students', in *1st Int. Conf. on Educational Data Mining (EDM'08)*.

Romero, C. *et al.* (2009) 'Applying Web usage mining for personalizing hyperlinks in Web-based adaptive educational systems', *Computers and Education*. doi: 10.1016/j.compedu.2009.05.003.

Romero, C. and Ventura, S. (2013) 'Data mining in education', *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. doi: 10.1002/widm.1075.

Romero, C., Ventura, S. and García, E. (2008) 'Data mining in course management systems: Moodle case study and tutorial', *Computers and Education*. doi: 10.1016/j.compedu.2007.05.016.

Rosvall, M. and Bergstrom, C. T. (2008) 'Maps of Information Flow Reveal Community Structure In Complex Networks', *Proceedings of the National Academy of Sciences of the United States of America*. doi: 10.1073/pnas.0706851105.

Sams, A. and Bergmann, J. (2012) *Flip your classroom*. doi: 10.1111/teth.12165.

Santos, O. C. and Boticario, J. G. (2012) 'Affective issues in semantic educational recommender systems', in *CEUR Workshop Proceedings*.

Schlairet, M. C., Green, R. and Benton, M. J. (2014) 'The flipped classroom strategies for an undergraduate nursing course', *Nurse Educator*. doi: 10.1097/NNE.0000000000000096.

Sclater, N. (2016) 'Developing a code of practice for learning analytics', *Journal of Learning Analytics*. doi: 10.18608/jla.2016.31.3.

Scott, J. *et al.* (2015) 'Social Network Analysis: An Introduction', in *The SAGE Handbook of Social Network Analysis*. doi: 10.4135/9781446294413.n2.

Siemens, G. (2013) 'Learning Analytics: The Emergence of a Discipline', *American Behavioral Scientist*. doi: 10.1177/0002764213498851.

Siemens, G. and Baker, R. (2012) 'Learning analytics and educational data mining: Towards communication and collaboration', in *Proceedings of the second international conference*

on learning analytics and knowledge. doi: 10.1145/2330601.2330661.

Stewart, S. A. and Abidi, S. S. R. (2012) 'Applying social network analysis to understand the knowledge sharing behaviour of practitioners in a clinical online discussion forum', *Journal of Medical Internet Research*, 14(6). doi: 10.2196/jmir.1982.

Sun, P. C. *et al.* (2008) 'What drives a successful e-Learning? An empirical investigation of the critical factors influencing learner satisfaction', *Computers and Education*. doi: 10.1016/j.compedu.2006.11.007.

Suraj, P. and Roshni, V. S. K. (2016) 'Social network analysis in student online discussion forums', *2015 IEEE Recent Advances in Intelligent Computational Systems, RAICS 2015*, (December), pp. 134–138. doi: 10.1109/RAICS.2015.7488402.

Taboada, M. *et al.* (2011) 'Lexicon-based methods for sentiment analysis', *Computational linguistics*, 37(2), pp. 267–307. doi: 10.1162/COLI_a_00049.

Thai-Nghe, N., Horváth, T. and Schmidt-Thieme, L. (2011) 'Factorization models for forecasting student performance', in *Proceedings of the 4th International Conference on Educational Data Mining*.

Thakare, Y. S. and Bagal, S. B. (2015) *Performance Evaluation of K-means Clustering Algorithm with Various Distance Metrics - ProQuest*, *International Journal of Computer Applications*.

Traphagan, T., Kucsera, J. V. and Kishi, K. (2010) 'Impact of class lecture webcasting on attendance and learning', *Educational Technology Research and Development*. doi: 10.1007/s11423-009-9128-7.

Tsai, Y.-S. *et al.* (2018) 'The SHEILA Framework: Informing Institutional Strategies and Policy Processes of Learning Analytics', *Journal of Learning Analytics*. doi: 10.18608/jla.2018.53.2.

Tucker, B. (2012) 'The Flipped Classroom: Online instruction at home frees class time for learning', *Education Next*. doi: 10.1017/CBO9781107415324.004.

Verbert, K. *et al.* (2012) 'Dataset-driven research for improving recommender systems for learning', in. doi: 10.1145/2090116.2090122.

Vihavainen, A., Luukkainen, M. and Kurhila, J. (2012) 'Multi-faceted support for MOOC in programming', in *Proceedings of the 13th annual conference on Information technology education - SIGITE '12*. doi: 10.1145/2380552.2380603.

Wen, M., Yang, D. and Rosé, C. (2014) 'Sentiment Analysis in MOOC Discussion Forums: What does it tell us?', in *Proceedings of 7th International Conference on Educational Data*

Mining (EDM2014), 4 - 7 July 2014, London, UK.

Widmer, E. D. ;Lafarg. L.-A. (1999) 'Boudedness and Connectivity of Contemporary Families: a case study', *Connections*.

Wise, A., Zhao, Y. and Hausknecht, S. (2013) 'W11-Learning analytics for online discussions: a pedagogical model for intervention with embedded and extracted analytics', *\ldots Conference on Learning Analytics \ldots*. doi: 10.1145/2460296.2460308.

Xu, D. and Jaggars, S. S. (2014) 'Performance Gaps between Online and Face-to-Face Courses: Differences across Types of Students and Academic Subject Areas', *The Journal of Higher Education*. doi: 10.1080/00221546.2014.11777343.

Young, T. *et al.* (2014) 'The Flipped Classroom: A Modality for Mixed Asynchronous and Synchronous Learning in a Residency Program', *Western Journal of Emergency Medicine*. doi: 10.5811/westjem.2014.10.23515.

Zappe, S. E. *et al.* (2009) 'Flipping the classroom to explore active learning in a large undergraduate course', *ASEE Annual Conference and Exposition, Conference Proceedings*. Available at: https://pennstate.pure.elsevier.com/en/publications/flipping-the-classroom-to-explore-active-learning-in-a-large-unde-2.

Zhuoxuan, J., Yan, Z. and Xiaoming, L. (2015) 'Learning Behavior Analysis and Prediction Based on MOOC Data', *Journal of Chemical Information and Modeling*. doi: 10.1017/CBO9781107415324.004.