Beyond the Arc: Leveraging Machine Learning for Predictive Sports Analytics in College Basketball

Coy Evans
Georgia Southern University
Savannah, Georgia, USA
coyevans13@gmail.com

ABSTRACT

This paper presents a novel machine learning framework designed to predict college basketball game outcomes. By leveraging a diverse set of statistical metrics—including player performance, team dynamics, and historical game data—the proposed model demonstrates improved accuracy over traditional forecasting methods. The methodology integrates data preprocessing, feature engineering, and advanced algorithms to capture complex patterns inherent in basketball games. The findings not only highlight the model's potential in college basketball but also pave the way for its adaptation across various basketball sports, providing valuable insights for coaches, analysts, and sports enthusiasts alike.

CCS CONCEPTS

• **Social and professional topics** → **Computing education**; **CS1**; *Student assessment*; • **Applied computing** → **Education**; **Collaborative learning**; *Interactive learning environments*; *Learning management systems.*

KEYWORDS

https://github.com/coyevans13/DS$_M$L$project_c$olab−$integration.git$

ACM Reference Format:
Coy Evans. 2021. Beyond the Arc: Leveraging Machine Learning for Predictive Sports Analytics in College Basketball. In *2021 ACM Southeast Conference (ACMSE 2021), April 15–17, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3409334.3452042

1 INTRODUCTION

Cyberlearning, as defined by the National Science Foundation (NSF), is the use of networked computing, communications technologies, and innovative digital tools not only to learn (e-learning) but also do something that supports learning. Besides enabling new educational experiences using advanced learning technology, cyberlearning has changed what and how people learn. Cyberlearning research has its roots in multiple disciplines, including learning sciences, computer and information science and engineering, and cognitive and social sciences. Cyberlearning has enabled instructors to leverage emerging technologies to design transformative student engagement activities, emphasize continuous improvement, and measure learning outcomes. This research study considers unraveling the growing trends in innovative learning technology and the advancements required in computation and technology to broaden access and increase students' depth of learning.

Education technology has revolutionized learning by improving the fundamental infrastructure in terms of hardware (e.g., desktop, laptop, tablet, and smartphone) and Learning Management Systems (LMSs) (e.g., Canvas, Moodle, D2L, and Blackboard). The digital online content can now be hosted using cloud-based educational tools. The trend in cyberlearning research has drifted from engineering the benefits of face-to-face teaching methodologies and improving learning engagement strategies to applying cognitive, behavioral and computational sciences to improve learning outcomes across diverse people, institutions and settings. This technological advancement has resulted in innovative design, development, and analysis leading to better learning experiences.

The US Department of Education recognizes approximately 6, 900 universities, and out of those, approximately 3, 500 universities have adopted LMSs for teaching purposes. Table 1 shows the usage of LMSs in Spring 2019, Spring 2020, and Fall 2020 in US universities. 36.7% of institutions actively use the Canvas LMS. Nevertheless, Hassan et al. show that there are a variety of different modalities for usage, and a variety of ideas about "active" use [**?**].

Zhou reported the following useful key findings after an extensive meta-analysis of research studies on online discussions in higher education settings: (i) there is a lack of research studies related to discussion component in some of the academic disciplines in higher education settings, (ii) the application of more qualitative methods than quantitative methods in the online discussion studies indicate the complexity of the topic, (iii) applying SNA measures generate findings from a new perspective and a new direction for future research, (iv) there are only fewer studies on investigating students' perceptions and performances, indicating a need for more empirical studies, and (v) there is a potential relationship between the quality of discussions and the level of critical thinking/cognitive engagement demonstrated in the course discussions [**?**].

Figure 1 shows the LMS usage data across the world (data collected from [**?**]). It is evident that e-Learning (learning online) has transformed the way that we look at knowledge and skill acquisition. There is an increasing demand for LMSs to systematically implement and manage e-Learning.

The purpose of this paper is to present a prototype, Canvas Online Discussion Analyzer (CODA), which is capable of grading and analyzing students' performance in the graded discussion boards in Canvas LMS. CODA helps visualize the student interactions and

ACMSE 2021, April 15–17, 2021, Virtual Event, USA

ACM ISBN 978-1-4503-8068-3/21/04
https://doi.org/10.1145/3409334.3452042

Table 1: LMS Usage - *Spring 2019, **Spring 2020 and ***Fall 2020 (Data provided by Edutechnica [?])

	Blackboard	D2L	Canvas	Moodle	Sakai	Others
*Number of Institutions	1062	381	1050	607	85	257
Percentage	31%	11%	30%	18%	3%	8%
**Number of Institutions	973	369	1147	573	72	244
Percentage	28.4%	10.8%	33.4%	16.7%	2.1%	7.1%
***Number of Institutions	994	406	1290	570	63	209
Percentage	26.8%	11.5%	36.7%	16.2%	1.8%	5.9%

computes the grades based on either only participation of students in discussion the course content or using complex network measures to provide more insights about the behavioral aspects of the students (e.g., team-building, leadership, facilitating discussions, relevancy).

The two-fold objectives of this study are (i) to analyze the characteristics of student participation on online discussion forums using Canvas LMS, and (ii) verify the positive impact that these discussions create on students' learning outcomes of the CS1 programming course. This study makes the following contributions to support this objective:

- Application of text analysis algorithms on discussion forum content on Canvas LMS to highlight frequently discussed terms and identify student sentiments on course topics.
- Visualization of student collaboration in Canvas LMS discussion forums as the course progressed to understand student learning behavior.
- Identification of central students and topics in the discussion to assess participation on Canvas LMS discussion forums.
- Preparation of metrics for assessment based on SNA and text analysis algorithms for grading discussions on Canvas LMS.

The remainder of the article is structured as follows: Section 2 provides a brief description of the background work in collaborative learning, network modeling, and mining interaction networks using network analysis and text mining techniques. Section 3 describes the system architecture by presenting the overview, grading metrics used in the prototype, and the CODA tool development. Section 4 provides the results and a detailed discussion of the prototype. Finally, Section 5 concludes the article with the threats to validity and the future work directions of research.

2 BACKGROUND

Educational data analysis has been widely studied by many academic researchers using various learning management platforms such as Canvas, Moodle, Blackboard, etc. Online discussion forums in these LMSs encourage students to learn collaboratively and facilitate data collection and analysis for cognitive and pedagogical benefits. Besides the cognitive benefits of positive impact on student achievement, effort, persistence, motivation, and peer interaction, collaborative learning also promotes social skills needed for future professional work in the STEM fields [**?**].

Various research on quantifying students' performance in collaborative learning environments have used learning analytics, Educational Data Mining (EDM) and SNA techniques to model and analyze the student interactions/collaboration in the course discussions. The fundamental challenge in evaluating the effectiveness of discussions largely depends on developing efficient computational methods that are capable of quantitatively measuring the interaction patterns in the network and computing weighs on the edges [**?**].

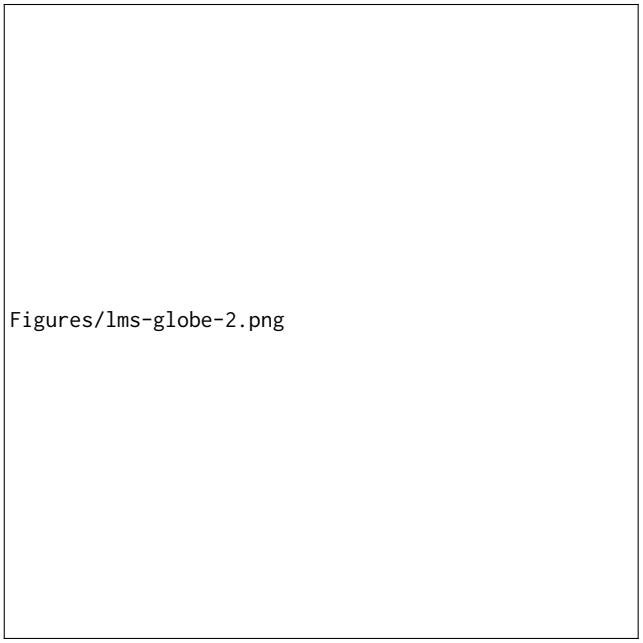

Figure 1: LMS Usage Across the Globe (Based on 2019 Data Provided by ClientStat [?])

The major performance evaluation measures like individual grades, group grades, and the final grades were studied by many researchers. EDM techniques were used to predict the performance of students based on the participation data collected from online discussion forums [**?**]. Classification techniques were used to forecast student performance [**?**] [**?**]. The relative importance of each student can be computed using a modified PageRank (PR) algorithm [**?**] [**?**]. These computation and prediction results helped the instructors to intervene and use alternative techniques to support students who were potentially at the risk of failing or dropping out. One such decision support tool to predict students' grades was developed by Kotsiantis [**?**].

2.1 Network Analysis

Graph, a mathematical structure is used to model the pairwise relations (called edges) between the objects (called vertices or nodes)

in a network in many applications. A graph $G = (V, E)$ consists of a finite set of V vertices and E edges, where an edge $e = (u, v)$ in the edge set E connects the vertices u and v from the vertex set V. The activity in a network due to the central nodes can be determined using the centrality measures, which are numeric values computed for that node.

The studies on centrality measures address the challenges of communication in small groups and hypothesized a relationship between structural centrality and influence in group processes. Various measures of centrality have been proposed to quantify the importance of an actor in a social network, and these concepts have been helpful for analyzing the characteristics of the network structure and functionalities [**?**].

2.2 Text Mining

It is possible to derive patterns that provide valuable information by using traditional data mining algorithms and information retrieval algorithms such as text categorization, text clustering, concept extraction, text comparison, sentiment analysis, text summarization, topic tracking, visualization, predictive analytics, etc. [**?**].

The following text mining techniques are used in this research.

- **Text Classification** also called text categorization is used to categorize text data based on their content. It uses supervised learning algorithms to learn from known examples (labelled training data set) and then performs classification of unknown examples (unlabeled test data set).
- **Keyword Extraction** involves tokenization, filtering, lemmatization, and stemming of the text data to collect important keywords used. [**?**] [**?**].
- **Sentiment Analysis** uses subjective information in online conversations to understand the social sentiment. Sentiment analysis is used in various domains like reviewing a service or product, monitoring opinion about a policy change by a political candidate, monitoring customer support performance [**?**] [**?**] [**?**] [**?**].

3 SYSTEM ARCHITECTURE

The prototype developed - Canvas Online Discussion Analyzer (CODA) is used to analyze the messages posted on the discussion forums by the students and provides useful insights to the instructor. The development, application, and results obtained from the prototype are described in the following subsections.

3.1 Data Used for the Study

The CODA prototype was tested on the data collected from two online discussion board posts of four sections consisting of a total of 102 students enrolled in the introductory programming course (*CS*1) in Fall 2018 at a medium-sized public university. The same instructor taught the four sections. The grades are computed for the two online discussion boards created on the Canvas *LMS*. The posts made by the students were about discussing classroom experiences, sharing their learning, asking/answering questions, sharing challenges, giving opinions, and posting comments on messages posted by other students in the discussion thread.

3.2 Overview

Student academic data and graded discussion data is gathered from Instructure's Canvas [**?**] and stored on a cloud_hosted Neo4j database called GrapheneDB [**?**] as a graph. Neo4j [**?**] is an ACID-compliant transactional database used to store and process graphs. Students and Discussion Titles are represented as the nodes in the graph. Each message is classified as "QUESTIONS", "COMMENTS" or "DISCUSSES" using naive Bayes classifier from Machine Learning Library in PHP [**?**] to denote the type of the message posted for creating edges between the nodes. The training dataset was created using the TP-GraphMiner framework [**?**] and manually labeling discussion posts along with the keywords used. Once the graph database is created, SNA metrics are applied to quantify students' participation and highlight the central actors in each discussion. These SNA metrics are also used to compute grades and award them to participating students. The interaction networks are visualized as vector images on the browser. Keywords are extracted from each message posted on the threaded discussions using the PHP implementation of Rapid Automatic Keyword Extraction algorithm (RAKE) [**?**] [**?**] and recurring keywords are visualized as a wordcloud using Anychart library [**?**]. The keywords are compared with the topics in the course curriculum and messages containing those topics are classified based on the student sentiment as "Positive", "Negative" or "Neutral" using ParallelDots API [**?**]. CODA is developed based on the architecture shown in Figure 2.

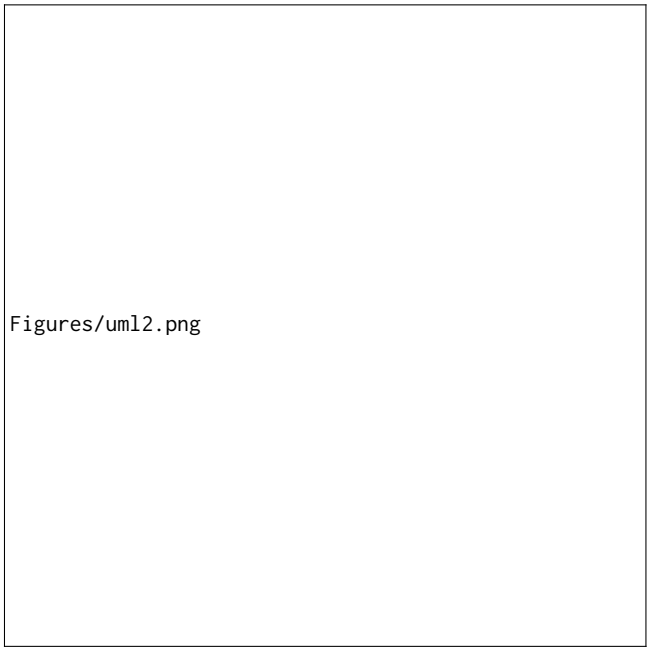

Figure 2: UML Use Case Diagram of Canvas Online Discussion Analyzer: Dotted Arrow (- ->) Represents *«include»*; Right Arrow (→) Represents *«extend»* Relationships

The following subsections discuss the APIs, packages, and libraries used by CODA for performing its functions.

3.2.1 Data Collection - Canvas REST API. Student academic data and graded discussion data is gathered from Canvas API [**?**] [**?**] using the following three endpoints:

- **List users in a course**

  GET /api/v1/courses/:course_id/students

  Here, the path variable *course_id* is the id of the course and the endpoint returns the paginated list of discussion topics for this course.
- **List Discussion Topics**

  GET /api/v1/courses/:course_id/discussion_topics

  Here, the path variable *course_id* is the id of the course and the endpoint returns the paginated list of students enrolled in this course.
- **Get the detailed Discussion Topic**

  GET /api/v1/courses/:course_id/discussion_topics/:topic_id/view

  Here, the path variable *course_id* is the id of the course and the path variable *topic_id* is the id of the desired discussion topic. This endpoint returns a cached structure of the discussion topic, containing all entries, their authors, and their message bodies.

3.2.2 Data Storage - GrapheneDB. GrapheneDB [**?**] is a Neo4j hosting provider on which users can create a database and connect their application to the database. The connection can be made using either Bolt Protocol or Hypertext Transfer Protocol (HTTP). After establishing a connection, ready-to-use code snippets are provided for popular Neo4j drivers. GraphAware is an enterprise grade PHP client for Neo4j [**?**].

3.2.3 Data Manipulation and Visualization. The following five libraries are used for processing the data and visualizing the obtained results:

- **Neo4j Labs Graph Algorithms Library** is developed and maintained by Neo4j Labs for Neo4j version 3.x. It contains a number of algorithms which can be accessed via Cypher projections.
- **Rake-PHP-Plus Package** [**?**] is developed for keyword extraction in PHP.A set of stop-words are defined in the package, but the user has the liberty to include additional stop-words as needed.
- **PHP-ML Library** provides Machine learning algorithms such as association rule learning, classification, regression, etc. in PHP. CODA uses naive bayes algorithm on PHP-ML Library to classify messages posted as “QUESTIONS", “COMMENTS" or “DISCUSSES".
- **ParallelDots API** provides a wide range of text analysis algorithms such as sentiment analysis, keyword extraction, semantic analysis, emotion analysis, etc.
- **Anychart** is a JavaScript library that provides frameworks for data visualization [**?**] and can be used to generate word-cloud, charts for visualizing sentiment analysis, and charts for visualizing correlation of grades with network measures.

3.2.4 Front-end. CODA is a PHP-based application developed using the following front end technologies:

- **JQuery** is a JavaScript library that makes HTML document traversal and manipulation, event handling, animation, and Ajax easier to use.
- **Bootstrap** is a CSS framework designed to make responsive web-pages.

3.3 Evaluation Metrics

The network measures used for this study are defined in the previous articles published in SIGCSE and FIE conference by the authors [**?**] [**?**]. These articles include teh experimental results of the proposed prototype and a detailed description how these metrics help to quantify student participation.

CODA uses Neo4j Centrality (to analyze the interaction network) and keyword extraction (to analyze the messages) algorithms and provides the following metrics to the instructor for awarding grades to the students:

(1) **Participation in the discussion:** Degree centrality (C_D) is a measure of the activeness of a node. If the student has posted a message i.e., participated in the discussion forum, C_D will be ≥ 1. So, the grade awarded to a student s for participating in a discussion is calculated as:

$$grade_1(s) = \begin{cases} 1, & \text{if } C_D(s) \geq 1 \\ 0, & \text{otherwise} \end{cases}$$

where $C_D(s)$ is the C_D of student s.

(2) **Relative participation:** A participating student thoroughly reads and understands the message they reply to or the responses they receive. C_D is dependent on the size of the network over which it is calculated. In order to award grades that measure how a student participates relatively to the other students participating in a discussion, normalization is performed. This grade awarded can be calculated as:

$$grade_2(s) = \frac{C_D(s) - C_{Dmin}}{C_{Dmax} - C_{Dmin}}$$

where $C_D(s)$ is the C_D of student s, C_{Dmax} is the maximum C_D in the network, and C_{Dmin} is the minimum C_D in the network.

(3) **Facilitating communication between peers:** The betweenness centrality (C_B) of a node scales with the number of pairs of nodes as implied by the summation indices. Normalization can be performed to scale C_B between 0 and 1. The grade awarded to a student s for facilitating communication between peers is calculated as:

$$grade_3(s) = \frac{C_B(s) - C_{Bmin}}{C_{Bmax} - C_{Bmin}}$$

where $C_B(s)$ is the C_B of student s, C_{Bmax} is the maximum C_B in the network, and C_{Bmin} is the minimum C_B in the network.

(4) **Leadership and team-based qualities:** Closeness centrality (C_C) shows us how the students worked together as a team to solve problems. The grade awarded to a student s on how central they are in communicating with their peers is calculated as:

$$grade_4(s) = C_C(s)$$

where $C_C(s)$ is the C_C of student s.

(5) **Validity of the post (number of keywords):** The amount of information in a post made by a student is an important criterion to determine how valid the post is. Tracking the keywords posted by a student s can help evaluate the validity of that student's post and thereby award points to students whose contribution is relevant to the discussion. The number of keywords in a valid post may vary, so they must be normalized for precision. The grade awarded to a student s on the validity of their post is calculated as:

$$grade_5(s) = \frac{K(s) - K_{min}}{K_{max} - K_{min}}$$

where $K(s)$ are the keywords posted by a student s, K_{min} is the minimum number of keywords in the discussion thread and K_{max} are the maximum number of keywords.

(6) **Connectedness:** The connectedness of a node in a network can be measured by computing the clustering coefficient (CC). The grade awarded to a student s on their connectedness is calculated as:

$$grade_6(s) = CC(s)$$

where $CC(s)$ is the CC of a student s.

One or many grading metrics can be selected from ($grade_{1-6}(s)$) as needed by the instructor and can be aggregated by taking an average. These computed scores are then shown in a table and exported to a CSV file, as shown in Figure 3.

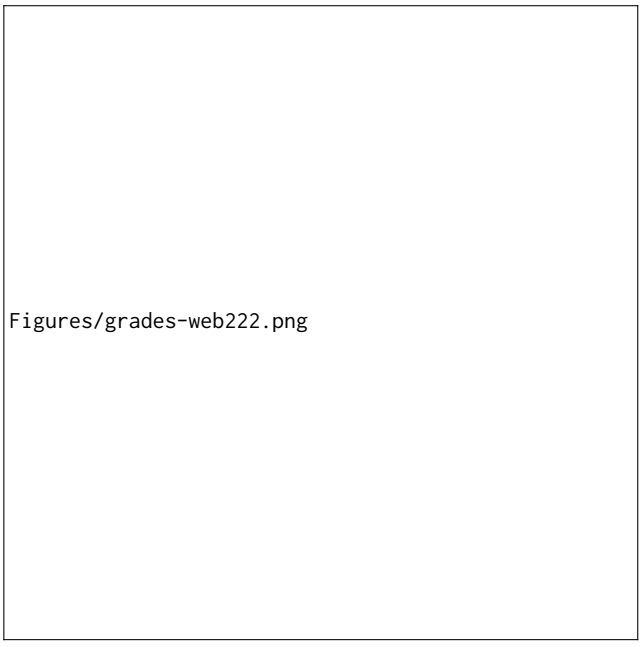

Figure 3: Selecting Grading Metrics

3.4 CODA Workflow and Experimental Results

CODA can be embedded in Canvas by using the Redirect Tool [**?**] developed by Instructure. The Redirect Tool is available on the Edu App Center. CODA can be included in the course navigation section (on the left side of the page) as shown in Figure 4.

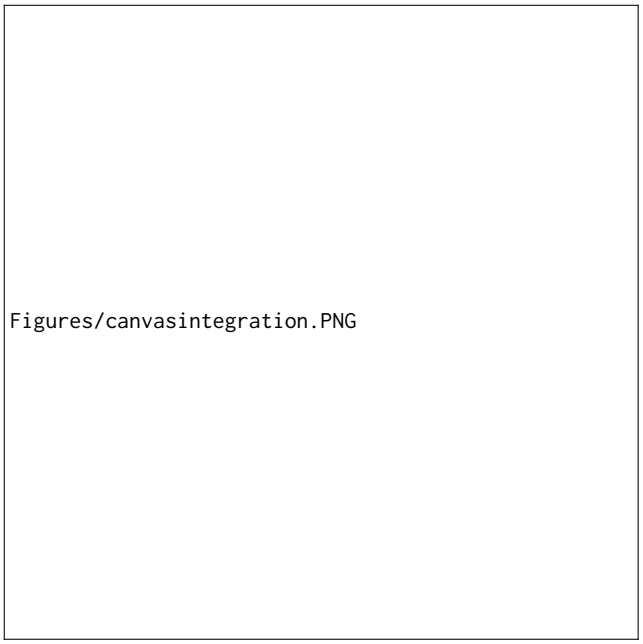

Figure 4: CODA App on Canvas

On entering the course ID and the authorization token and successfully logging in, CODA creates sections for the graded discussions that are being analyzed. The responsive web-design of CODA using Bootstrap enables users to access the tool from their mobile devices also. For each section, it provides buttons to visualize student interaction network, view highlights of the discussion, and compute and award grades to each participating student based on a set of parameters (Figure 5). The following subsections describe the functionalities and the results obtained.

3.4.1 Visualization of student interaction network. The student interaction network in the graded discussions of the course can be visualized as Scalable Vector Graphics (Figure 6). Here, *blue* color represents the student nodes and *green* color represents the discussion nodes. There are 4 types of edges in the network: TAKES, QUESTIONS, DISCUSSES, and COMMENTS. Each student that takes part in the discussion forum has a TAKES edge. The QUESTIONS, DISCUSSION, and COMMENTS edges were classified using Naive Bayes classifier in the PHP-ML Library as described in Section 3.2.3. A similar visualization of the student's interaction data computed using the descriptive statistics on the duration of participation in the discussion(s), course LMS usage, and the performance(grade) of the student in the course provides information about student progress. Instructor intervention is recommended for the students that have the least scores in all these parameters.

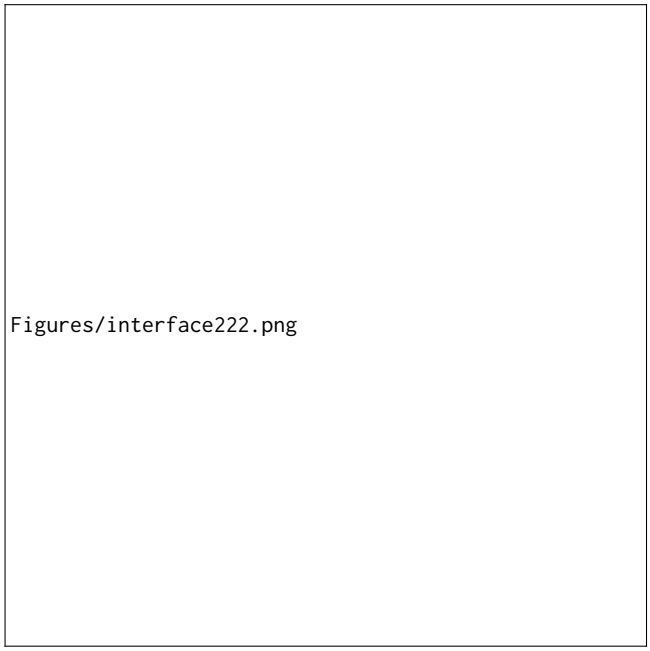

Figure 5: CODA Interface After Clicking a Discussion Section (Title: Loops, Methods, Arrays and Sample Exam 3 Discussion)

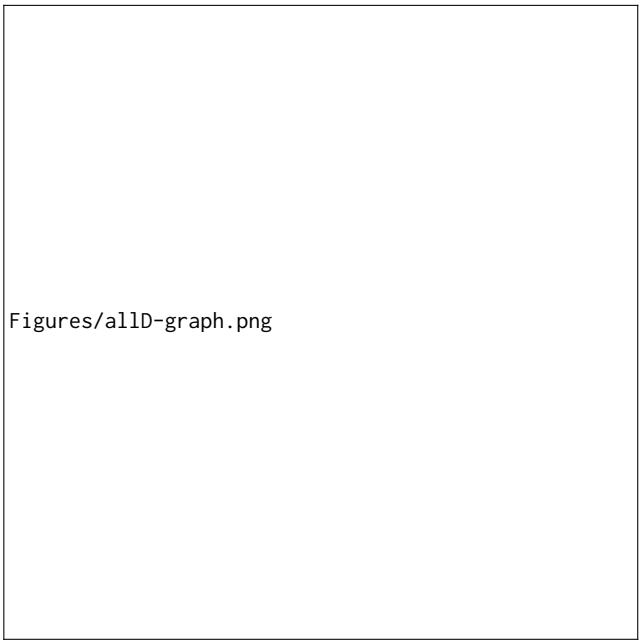

Figure 6: Visualization of Student Interaction Network Obtained from Discussion Board Data

3.4.2 Wordcloud of Trending Topics. CODA provides a button to visualize the commonly discussed topics as a wordcloud. On clicking the button 'Show Wordcloud,' the user is prompted to enter a number between the minimum and maximum frequency of words in the selected discussion. After entering a number, a wordcloud of commonly discussed words with a frequency greater than the input number is displayed using AnyChart Library (Figure 7) .

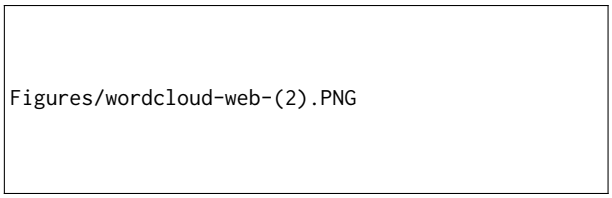

Figure 7: Wordcloud of Words Occurring More Than 10 Times in Discussion #1 (Title: Loops, Methods, Arrays and Sample Exam 3 Discussion)

3.4.3 Student Sentiment on Syllabus Topics. The terms from the CS1 curriculum were stored as an array of strings and sentiment analysis was performed on the messages posted on the discussion forums containing keywords from the syllabus. CODA provides a button to view sentiment analysis in the 'All Discussions' section. On clicking the button 'View Sentiment Analysis', a stacked bar chart shows the student sentiment on topics in the syllabus (Figure 8). *Light blue* shows positive sentiment, *dark blue* shows negative sentiment, and *orange* shows neutral sentiment towards a particular topic. On mouse hover, the bar chart created using AnyChart shows the percentage of students that display a particular sentiment towards that topic.

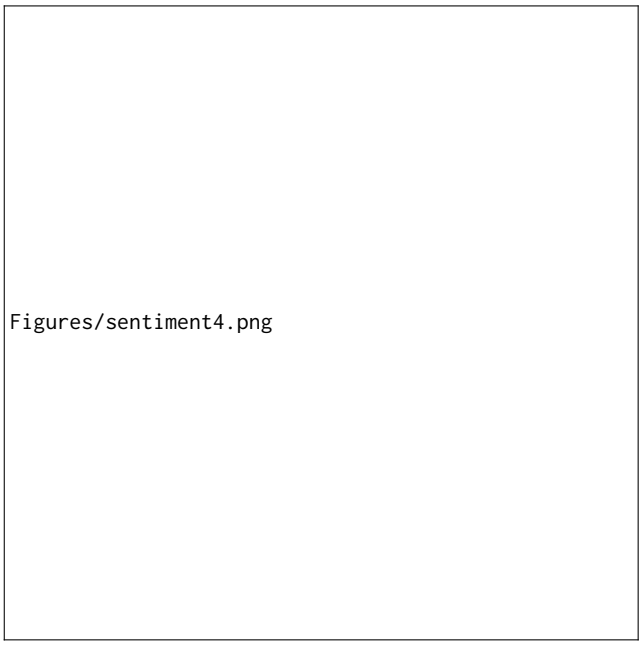

Figure 8: Student Sentiment on Syllabus Topics

3.4.4 Highlights of the discussion. On clicking the button 'View Highlights,' centrality measures are computed using Neo4j Graph

Algorithms Library. CODA tells the number of students that participated in the discussion and presents a *hall of fame* highlighting the active students based on their participation (highest degree centrality, participation log - duration of participation), friendliness (highest betweenness centrality), leadership qualities (highest closeness centrality), verbosity (maximum number of keywords used), and connectedness (highest clustering coefficient) (Figure 9).

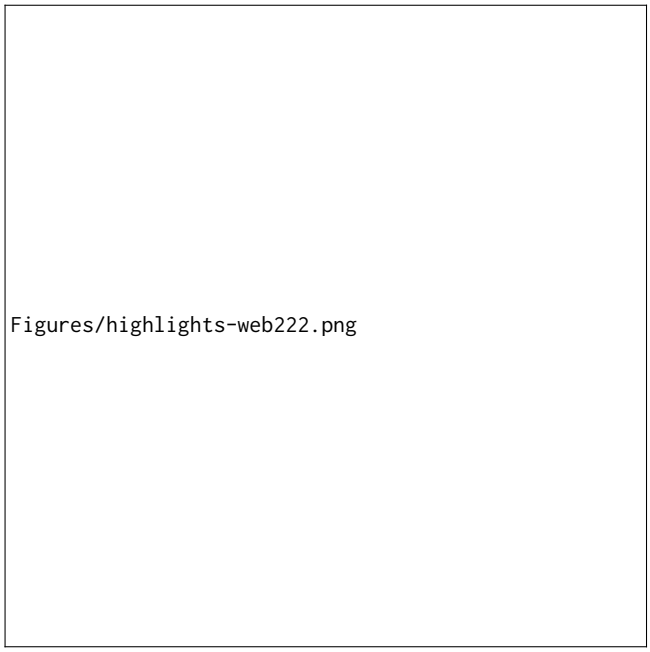

Figure 9: Highlights of Discussion #1 (Title: Loops, Methods, Arrays and Sample Exam 3 Discussion)

3.4.5 Computation of Grades. CODA enables the instructor to compute grades based on the six metrics described in Section 3.3. It provides the instructor with a list of checkboxes containing the grade computation metrics. After selecting the grading criteria and clicking the button 'Compute Grades', the user gets a prompt to enter the total awarded points. The default value of the total awarded points is set as 2 unless the user enters another value. CODA then generates a table containing the students' names and awarded points (Figure 10). This table can then be exported as a CSV file by clicking on the 'Export' button.

3.4.6 Additional Features. To assist non-technical users, CODA has tooltips that briefly describe every SNA measure used. It provides the user with highlights of the discussion and bar charts showing the correlation between the graph algorithms and students' final grades. Summary statistics provides the user with information on the categories of messages posted in the discussion. The tool also reports student access computer logs, their behaviors defined using descriptive data on the frequency of participation, duration of participation, and the correlation between participation in discussions and completion of assignments, which helps the instructor to gather more insights about students' perceptions on learning and pay attention to the non-successful students.

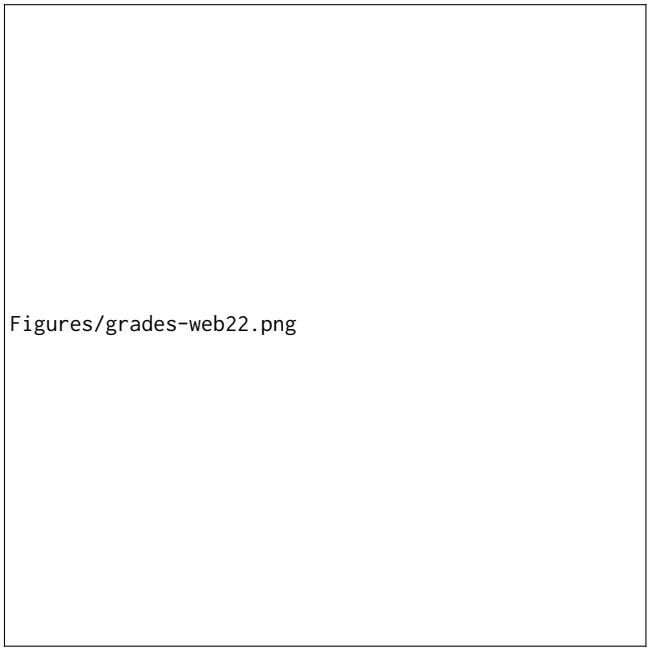

Figure 10: Grades Awarded for Participation in Discussion #1 (Title: Loops, Methods, Arrays and Sample Exam 3 Discussion)

4 DISCUSSION

This study aims to analyze the participation metrics of students on online discussion forums on Canvas LMS and verify its positive impact on students' learning outcomes and other pedagogical and cognitive benefits in the CS1 programming course. We summarize the valuable insights that CODA offers to the students and the instructors, the computational technique used, and the benefits in Table 2.

CODA provides information on the structural evolution due to interactions among students in Canvas discussion forums using social network analysis and text analysis techniques. It provides snapshots of information and visualization of students' participation in the discussion forums and their interactions. CODA provides various metrics to assess student participation thereby helps instructors reduce the time-consuming task of manually reading every discussion post and monitoring student participation to measure individual participation and even unrealistic in courses with high enrollment. The grading metrics are normalized to avoid loss of precision and maintain equality. Besides providing various grading metrics to assess student participation based on centrality measures and the validity of a message posted, CODA offers many useful insights.

Using keyword extraction and sentiment analysis, CODA creates a summary of the discussed topics, which gives the instructor a quick view of what is under discussion. Moreover, CODA recognizes the leading students in the discussions based on centrality measures and keyword usage. Centrality metrics correspond to the influence, leadership abilities, connectedness, and friendliness in the student network.

Table 2: Useful Insights, Computation Techniques, and Benefits of CODA Tool

Insights Offered by CODA	Techniques Used	Benefits
Discussion Grades	Centrality metrics Clustering Coefficient	Participation(relative), facilitating and leadership/team-based qualities, post validation(use of keywords), connectedness
Summary of topics discussed	Keyword extraction and sentiment analysis	Course content mapping, student sentiments, preparedness to topics
Trending topics	Keyword extraction	Visualization - word cloud
Challenging topics	Sentiment analysis	Challenging topics
Influence (connectedness), leadership abilities(verbosity), and friendliness	Centrality metrics	Link dynamics & team-based characteristics

CODA provides consolidated discussion board summaries that help instructors reduce the time-consuming process of reading all the student post threads and summarizing discussions in courses with high enrollment. Additionally, CODA captures the sentiment displayed by students in their posts by performing sentiment analysis so that the instructor can address topics that the majority of the students find challenging. This systematic analysis of every discussion in the collection allows for a better assessment of the teaching-learning process.

The experimental results from the previous studies conducted by the authors showed a positive correlation between the number of discussion posts made by students and their academic performance in terms of the final grade [**?**]. These results demonstrated the need for efficient EDM tools to model interaction data and evaluate student performance in course discussions constructed from active student collaborations. The manual data collection, processing, and analysis in the authors' previous studies also motivated to develop a prototype to grade the students' contribution in posting useful and relevant information without having to read the long threads of discussion posts and thereby gather helpful insights about characterizing student learning behaviors for pedagogical benefits.

5 CONCLUSION AND FUTURE WORK

This paper highlights graph modeling of linear discussion threads on Canvas LMS as a collaboration network for academic performance evaluation. The web-based prototype (CODA) was developed to mine discussion board data on Canvas LMS using graph mining and text analysis algorithms. The outcomes of the research are demonstrated using the following functionalities: (i) wordclouds showing what the trending topics in the discussions are and a bar chart showing which topics students find difficult, (ii) grading metrics for awarding grades, and (iii) visualizing the student interaction network. This prototype can improve instructional outcomes by providing the instructor with a birds-eye view of the educational phenomena from a student's perspective and exploring the potential relationship between the instructional requirements on the quality of discussion and critical thinking or cognitive engagement. The linear and temporal nature of the discussion threads leads to students being biased and posting replies only to the latest posts. This poses a threat to the validity of constructing efficient network models. Another threat to validity is the relevance of students' posts when measuring and analyzing student participation. Although the evaluation metrics based on the validity of the post described in Section 3.3 address this issue, during the computation of C_D we assumed every post is relevant to the discussion.

The future work of this research will include validation of the tool by manually analyzing individual samples of data to verify that CODA tool accurately represents student participation. The tool can be customized to be used in other computer science courses and other LMSs. The authors will consider performing a user study to evaluate the proposed prototype that would potentially improve instructional outcomes. Further analysis will include extracting additional features from the students' demographic data, such as their majors, gender, cultural background, etc., and exploring their learning patterns in other courses in which they are enrolled. Inferences drawn from such a study would enable clustering students based on their learning behavior and finding the evolutionary properties of these temporal systems. In the future, causation analysis and Natural Language Processing algorithms will be used on the student interaction network to study the sentiments and semantics of the content posted. Such an exploratory study of discussion forums combined with psychometric instruments is needed to gather more insights into students' mental health and behavioral conditions in unprecedented COVID-like unprecedented situations.

REFERENCES