Making Sense of Student Success and Risk through Unsupervised Machine Learning and Interactive Storytelling

Ahmad Al-Doulat*, Nasheen Nur*, Alireza Karduni, Aileen Benedict, Erfan Al-Hossami, Mary Lou Maher, Wenwen Dou, Mohsen Dorodchi, and Xi Niu

University of North Carolina at Charlotte , NC, USA
{adoulat,nnur,akarduni,abenedi3,ealhossa,
m.maher,wdou1,mdorodch,xniu2,}@uncc.edu

Abstract. This paper presents an interactive AI system to enable academic advisors and program leadership to understand the patterns of behavior related to student success and risk using data collected from institutional databases. We have worked closely with advisors in our development of an innovative temporal model of student data, unsupervised k-means algorithm on the data, and interactive user experiences with the data. We report on the design and evaluation of FIRST, Finding Interesting stoRies about STudents, that provides an interactive experience in which the advisor can: select relevant student features to be included in a temporal model, interact with a visualization of unsupervised learning that present patterns of student behavior and their correlation with performance, and to view automatically generated stories about individual students based on student data in the temporal model. We have developed a high fidelity prototype of FIRST using 10 years of student data in our College. As part of our iterative design process, we performed a focus group study with six advisors following a demonstration of the prototype. Our focus group evaluation highlights the sensemaking value in the temporal model, the unsupervised clusters of the behavior of all students in a major, and the stories about individual students.

Keywords: Sensemaking in Learning Analytics, Data Storytelling, Unsupervised Machine Learning, Data Visualization, Interactive User Experience, Human-centered Design

1 Introduction

As artificial intelligence in education becomes increasingly prominent, there is a growing need to consider augmented intelligence. This is the idea that artificial intelligence can and should be used to enhance human intelligence and abilities rather than attempt to replace it. The 2016 National Artificial Intelligence Research and Development Strategic Plan stated that "the walls between humans

* These authors contributed equally

and AI systems are slowly beginning to erode, with AI systems augmenting and enhancing human capabilities. Fundamental research is needed to develop effective methods for human-AI interaction and collaboration" [1]. Popenici and Kerr further emphasize the importance of recognizing education as a "human-centred endeavor" and the idea that "solely rely[ing] on technology is a dangerous path, and... that humans should identify problems, critique, identify risks, and ask important questions..." [2]. Therefore, we should take on a human-centered approach in the era of AI. Human-centered AI is a viewpoint that AI systems and algorithms "must be designed with an awareness that they are part of a larger system involving humans" [3]. AI research should not just be technological, but humanistic and ethical as well [4]. One aspect of human-centered AI is to create systems that help humans understand the system itself [3]. Therefore, the goal is not simply to provide results through a black-box model. The focus is to help users understand those results and how those results are derived.

We explore sensemaking in Learning Analytics (LA) as an example of human-centered AI and present how we address this challenge for advisors that are presented with large amounts of data and analytics about their students. LA is an interdisciplinary field that emerged to make sense of unprecedented amounts of data collected by the extensive use of technology in education. LA brings together researchers and practitioners from two main fields: data mining and education [5]. Effective presentation of analytical results for decision making has been a major issue when dealing with large volumes of data in LA [6]. Many systems for early alerts on student performance provide results without providing necessary explanations as to how the system derived those results. If an early warning system gives a result that is inconsistent with the expectations of a teacher or an advisor, and there is no information to explain how the system arrived at the prediction, it can easily cause educators to discount or mistrust the prediction [7]. Human sensemaking relies on developing representations of knowledge to help serve a task, such as decision-making, and on the design of AI approaches to better aid these tasks. We discuss the design, implementation, and evaluation of an interactive system designed to help advisors better understand student success and risk. In contrast to many LA systems designed to support student awareness of their performance or to support teachers in understanding the students' performance in their courses, our interactive system is designed to support advisors and higher education leadership in making sense of students' success and risk in their degree programs. Our approach to interactive sensemaking has three main parts: (1) a temporal student data model, (2) data analytics based on unsupervised learning, and (3) storytelling about the student experience.

2 Related Work

In this section, we review related research in two interdisciplinary threads: (1) sensemaking in LA, and (2) data storytelling techniques.

2.1 Sensemaking in Learning Analytics

Sensemaking is process of understanding connections to anticipate their trajectories and to act effectively [8]. Van et al. [9] stated that sensemaking is a core component of LA dashboard interventions, as the purpose of these tools is to provide users with the ability to become aware of, reflect upon, and make data-based decisions. Echeverria et al. [6] proposed a learning design-driven data storytelling approach where they support user sensemaking by directing the user's attention to the critical features of the students' data using visualizations with data storytelling components. Their user study suggests that adding storytelling elements to the LA dashboards has the potential to help users make sense of the critical features of students' data with less effort. CALMSystem [10] is another example of a LA system that supports sensemaking, awareness, and reflection. It was developed on top of an intelligent tutoring system to give a learner insight into the learner model. Klein et al. [11] proposed a model of student sensemaking of LA dashboards to show how data and visualization inform user sensemaking and action. Verbert et al. [11] introduced a LA system for learners and teachers visualizing learning traces with four distinguished stages for the process model - (i) *awareness* is only concerned with the students' data presented using various visualizations, (ii) *reflection* focuses on usefulness and relevance of the queries by the users, (iii) *sensemaking* is concerned with users' responses in the reflection process and the creation of new insights, and (iv) *impact* is concerned with the induction of new meaning or changing behavior by the users. Additionally, researchers made contributions to better prediction and sensemaking of student progress trajectories. Learning Management Systems (LMSs) storing students' temporal data have been leveraged in various works to analyze students' progression throughout their whole program [12, 13, 14, 15, 16] and within a course level [12, 17, 18, 19].

2.2 Sensemaking with Data Storytelling

Stories are capable of conveying essential information to users more naturally and familiarly for them [20]. Data storytelling aims to make data more understandable and memorable by human users by presenting data in the form of stories. Several research studies created natural language presentations of tabular or numeric data ranging from summarizing statistical results [21, 22], stock market trends [23], and environmental data [24]. Many applications of Natural Language Generation (NLG) have been used to generate stories from data to promote the user sensemaking. Notable examples of tools that generate textual forecast from structured data include the Forecast Generator (FoG) [25], MULTIMETEO [26], and the SumTime system [27]. Such systems increase interpretability and reduce routine writing tasks performed by human forecasters. NLG is also used in medicine. TOPAZ [28], creates reports of blood cell and drug dosages for lymphoma patients. It uses a schema-based generation system that generates a textual report read by clinicians. Other systems that generate medical reports include Suregen [29], Narrative Engine [30], and STOP [31]. These

systems tend to facilitate the users' sensemaking of homogeneous data through brief textual summaries. FIRST is capable of generating stories to support advisors' sensemaking of complex, temporal, and heterogeneous student data.

3 FIRST: Design and Implementation

The goal of FIRST is to better communicate analytics results by guiding the user through sensemaking tasks and interactive LA. Sensemaking tasks consist of information gathering, developing insights, and performing knowledge discovery [32]. In the sensemaking process, domain experts such as the educational leaders, teachers, and academic advisors decide on the existing challenges and expected outcomes for their institution. Most of the learning management tools involve data scientists in the knowledge discovery process to design the student data model, analytics approach, visualizations, and a reporting system to understand students' patterns of success or failure. Next, domain experts design intervention methods based on the analytics. The analytical process, essential to knowledge discovery, needs substantial data science skills. Domain experts do not engage in the discovery process since the analytical model is a black box to them. In FIRST, domain experts can select features from the temporal data model, see the stories about students, and explore which factors are major contributors to a student's performance and behaviors.

3.1 Interface Design

Our system is designed to allow advisors to engage in sensemaking by interacting with temporal data, reviewing aggregate analytics, and reading stories. Figure 1A shows the interface for the user to select the student features in the temporal model. The selected features are used when generating stories for each student. The user can change their preferred features at any point, which will consequently change the content of the stories. It is also possible for the system to automatically generate stories based on what it selects as the most appropriate features. However, allowing the user to select the features is important to sensemaking. Figure 1B shows the user experience with the results of unsupervised learning, and Figure 1C shows the user experience for interacting with the automatic story generator. FIRST differs from existing LA tools in the following ways:

- The user can leverage their insights about student behavior and participate in model construction, giving them the flexibility to change the features to be used in the analytic models and automatically generated stories.
- The user is presented with automatically generated stories to complement the results from analytic models.

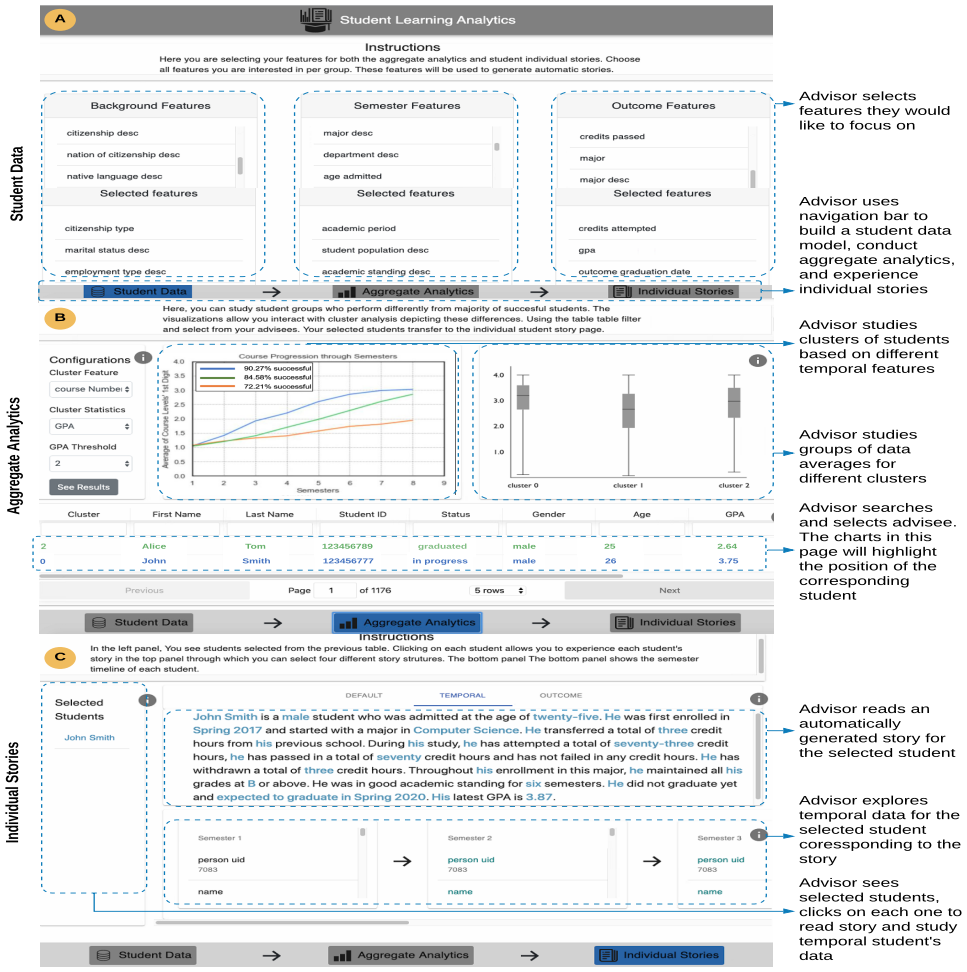

Fig. 1. Interface design for FIRST

3.2 Temporal Data Model

FIRST uses a temporal data model that uses time segments to group heterogeneous sources of data and form sequences of information for each student [16]. This allows the analytic models to consider the temporal dependencies of students throughout their enrollment. The temporal model gives flexibility in defining the duration of the temporal node, contextualizing information within a node, and interpreting sequences of nodes as stories. The data model contains one sequence per student that starts with their enrollment and ends with when the student graduates or leaves the university. Each node in a sequence represents a period (e.g., a single semester) and contains a vector of features (variables, such as courses taken in that semester). There are three types of temporal nodes for

each student: the background node with demographic information, the semester node with semester-wise activities and information, and the outcome node with the value of the performance variable. The student data model is shown in Figure 2A.

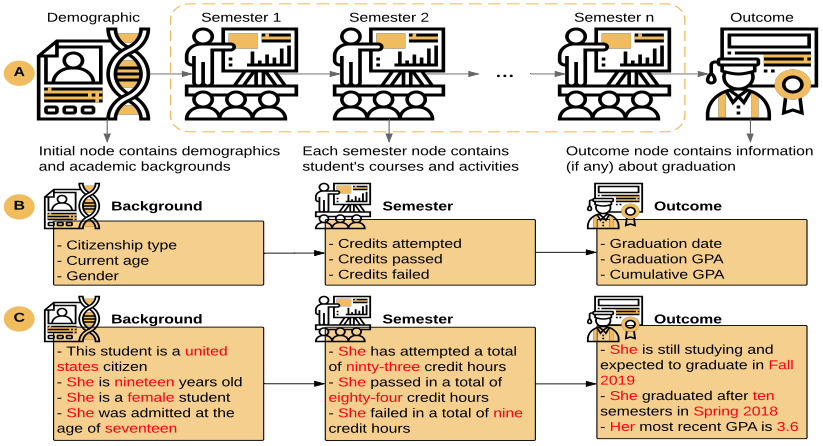

Fig. 2. Components for generating students' stories: (A) temporal data model, (B) selected student features, and (C) examples of sentences in the story

3.3 Unsupervised Learning

FIRST uses unsupervised learning to identify patterns of student behavior and then maps that behavior onto performance. The user can select from options for the student performance variable, such as GPA, and can select filters to include a subset of the total population, such as male or female students or a period of time. Figure 1B shows the results of clustering all students according to their course progression with the performance variable of GPA, where 2.0 is a minimum value to be successful. Course progression is an example engineered temporal feature, which is the average value of the first digit of a course level for each semester. For example, if a student took three courses with levels 1200, 1212, and 3000 in his/her first semester, this feature will take a value of 1.7 (average of 1, 1, and 3) for the first semester. We then formed a 2D (two-dimensional) feature vector for each student in which each row has the values for one of the engineered features for each semester. We used the K-means clustering algorithm [33] on several engineered features and found that course progression, for example, was able to cluster students with high "purity" in terms of the defined outcome variable. We used the elbow method [34] to determine the optimal number of clusters. We analyzed each cluster to see if they were "coherent" in terms of student performance. For example, after we applied the K-means approach to the "course progression" feature, the result could separate the successful and risky student reasonably clearly. Our primary hypothesis for this feature is that it should be either increasing or steady along the semesters for those successful

students. If it is decreasing or steady for a long time, the student did not progress to higher-level courses or the student was repeating lower-level courses.

Figure 1B presents the clustering results with 3 clusters for the engineered feature "Course Progression Through Semesters". In the blue cluster with 483 students, successful students are the most dominant with a percentage of 90.27%. As we see the intercept and the slope of this blue line in Figure 1B, it has a higher average course level in each semester compared to the other two clusters. In addition, the average course level is consistently increasing. This suggests that this cluster of students consistently takes courses at a higher level and starts to progress early on. The green cluster also has a higher percentage of successful students than the orange cluster. If we compare their intercepts and slopes, the green line stays above the orange one and makes more "linear" progression than the orange counterpart. In this analysis, we define student success as obtaining the final GPA last semester higher than 2.0. If we changed the GPA threshold, the clustering results would be different. The user can select each cluster and further review the data for each student who belongs to that cluster. The bar chart shows the average GPA for each cluster. The user can select an individual student or groups of students in the analytic interface and review their temporal data. The selected students in exploring the analytic results are saved and available on the storytelling page.

We use clustering since more students are successful than unsuccessful:: a supervised learning approach could overfit and impose an accuracy paradox due to a higher number of majority class examples caused by the imbalance. Equalizing class membership by adjusting the within-class imbalance and using random sampling can introduce unrealistic patterns in the data [35]. We use clustering to separate and classify samples. The clustering results provide insight into the engineered features that discriminate on percentages of successful students compared to students at risk. This classification describes characteristics of cohorts of students and how they behave in the clusters. In the future, we will consider a guided re-sampling and classification method to overcome over-fitting. For this reason we adopted an unsupervised clustering approach to find patterns of student behavior that map onto success criteria. In the future, we plan to incorporate the cluster results into a predictive model to apply our knowledge about patterns of behavior in cohorts of students to develop early alerts or predictions for individual students.

3.4 Student Stories

FIRST automatically generates stories for each student using the features selected in the temporal data model. These stories present a summary of the student's experience in a narrative. Figure 1C shows the user experience for interacting with the student stories. When the user selects a student from the left panel, the timeline and story sections are updated. The storytelling algorithm uses user-selected and standard features. The stories are generated from the data in the temporal model shown in Figure 2. Figure 2A shows the nodes in the temporal data model, Figure 2B shows the features selected from each node, and

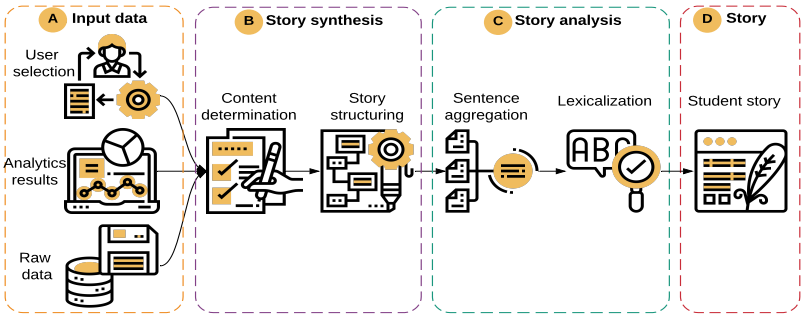

Fig. 3. Process of story generation

Figure 2C shows the sentences that are constructed from each feature. The text in black is from a predefined template while the text in red is generated from the features. After generating the sentences for each of the selected features, these sentences are used to generate the story as discussed below. An example of a generated story can be shown in Figure 1C. Figure 3 illustrates the 3 stages in the process of generating stories: raw data source and user selection inputs, story synthesis, and story analysis. We describe each stage of story generation.

Data Source As shown in Figure 3A, the input data for story generation comprises: (i) features in the temporal data model, (ii) the results of the analytics, and (iii) the user selected features and outcome. The features in the data model are used in a set of predefined template rules, the analytics results are used to compare the current student with other similar students, and the user-selected variables are used to make the story more customized for the user.

Story Synthesis The goal of this stage is to determine and sort the content presented in the student's story. As shown in Figure 3B, synthesis has two tasks: content determination and story structuring.

- Content Determination: this is the task of choosing what is included in the story. The selection is based on these factors:
  - user-selected features: we include the features selected by the user as illustrated in Figure 1A.
  - performance rules: we identified a set of rules that either inspect any sudden changes of the students' performance over time (e.g., A student's GPA suddenly dropped or increased), or abnormal information compared to most students (e.g., the number of attempted, passed, or failed courses for a semester is higher, or the number of D-scored courses is higher).
  - comparison with other similar students: we used clusters to look for students that are similar and successful to inspect if the student per se is an outlier in terms of some variables.
- Story Structuring: this is the task of deciding the order of information in which it is presented to the reader. We order the information based on the

student temporal data model, in which the story starts with the background information about the student, then with the semester information, and ends with the outcome information.

Story Analysis This stage improves the language of the stories so they are more human-readable and coherent. As shown in Figure 3C, this includes 2 tasks: sentence aggregation and lexicalization.

- Sentence Aggregation: Clusters multiple pieces of the same kind of information together into a single sentence instead of several ones. For instance, if we have a set of candidate sentences as "student achieved an A in the course X", and "student achieved B in course Y", these sentences should be aggregated into one sentence "student maintained all his grades at B or above".
- Lexicalization and Linguistic Realization: Lexicalization is choosing the proper words and phrases to transform the data into natural language text. Linguistic realization is inserting punctuation, functional words and other elements required for the text to be fluid and coherent.

4 USER STUDY - FOCUS GROUP

A focus group study was conducted with the goal of learning what users find important in a tool to support advising. In the focus group session, we demonstrated FIRST and then asked questions about the value of the student data model, analytics, and storytelling. We recruited six professional and faculty advisors whom are already familiar with multiple tools that provide data, analytics, and risk scores for the students that they advise. A focus group study was selected for its effectiveness in collecting user opinions and attitudes through group discussion and dynamic conversations. Some preliminary questions were asked to collect information related to the current technology used during advising and the useful features of those tools. The participants revealed that they often ignored the risk score provided by the analytics in their advising tool because the process behind the calculation is not clear to them. They mentioned that although the student reports generated by the existing tool were useful, they would like more flexibility to customize the information for different cohorts of students. The group discussed that one goal for such tools is to be prepared for advising before the student arrives for the advising appointment. FIRST was demonstrated to the group with scenarios for specific students. The participants asked questions about the system and the facilitator demonstrated additional interactive features. Then the participants were asked to answer questions to assess the sensemaking they performed through the demonstration: (i) What insights were you able to gain about students through viewing this tool? (ii) What are the differences between what you learned about the students from the analytics versus the stories? (iii) What is the value of the analytics results and the stories? (v) How can the student stories help you with advising? And (vi) Can you think of other good predictors(features) of student success? Two researchers reviewed the transcript and identified emerging themes independently

and through discussion they agreed on three higher-level themes. These three high-level themes were then used to revisit and code the transcript according to the themes.

- **Selecting Features for Student Models:** Participants appreciated that they could select the features they thought should be part of a predictive model of risk or part of the student story. They also like a number of features that were included, such as students' financial need status, family life, housing options, and mailing addresses. Many expressed surprise that the University actually had a lot of data that would be useful for advising that was not available in the other tools.
- **Value of Aggregate Analytics and Temporal Data:** Participants agreed that aggregate analytics is essential for understanding students, especially a targeted group of students. They found the presentation of the student data as a temporal progression is useful since it presents the overall students' progression through semesters.
- **Value of Student Stories:** The participants agreed that student stories were useful and effective to provide a high-level overview or snapshot of the student. They mentioned that the stories would be helpful for understanding a specific student quickly. They agreed that stories provide a good understanding of students in terms of their demographic information as well as their academic performance. One participant said: "I like the stories the best - knowing that the story was created using analytics is reassuring". One comment to extend FIRST is the suggestion to tell stories about groups of students that lie in a single cluster.

5 Conclusions and Future Work

In this paper, we present FIRST, an interactive LA system designed to support advisors using a temporal data model, unsupervised models, and storytelling. FIRST enables the advisor to select specific features, review the aggregate analytics based on unsupervised learning algorithms, and interact with stories about specific students. The student stories are automatically generated using user-selected features, the features that indicate significant changes, and additional data about the student using rules that present a more complete story. The process for generating stories has 3 stages: sourcing the data, selecting and structuring story components, and text-processing the sentences. A focus group study was conducted to evaluate FIRST and gather feedback. The participants highlighted the sensemaking value of storytelling and the increased access to student data compared to other tools. The aggregate analysis was reported to be enhanced by the storytelling since the user can switch between the story and the visual analytics. The results of the focus group confirm our hypothesis that storytelling complements dashboard-style analytics. In the future, we plan to do a longitudinal study of the use of FIRST to learn more about the changes in the advisors' understanding of their students with and without FIRST.

Bibliography

[1] STRATEGIC PLAN. The national artificial intelligence research and development strategic plan. 2016.

[2] Stefan AD Popenici and Sharon Kerr. Exploring the impact of artificial intelligence on teaching and learning in higher education. *Research and Practice in Technology Enhanced Learning*, 12(1):22, 2017.

[3] Mark O Riedl. Human-centered artificial intelligence and machine learning. *Human Behavior and Emerging Technologies*, 1(1):33–36, 2019.

[4] Wei Xu. Toward human-centered ai: a perspective from human-computer interaction. *interactions*, 26(4):42–46, 2019.

[5] Dragan Gašević, Vitomir Kovanović, and Srećko Joksimović. Piecing the learning analytics puzzle: A consolidated model of a field of research and practice. *Learning: Research and Practice*, 3(1):63–78, 2017.

[6] Vanessa Echeverria, Roberto Martinez-Maldonado, Roger Granda, Katherine Chiluiza, Cristina Conati, and Simon Buckingham Shum. Driving data storytelling from learning design. In *Proceedings of the 8th international conference on learning analytics and knowledge*, pages 131–140, 2018.

[7] Robert F Murphy. Artificial intelligence applications to support k–1 2 teachers and teaching. *RAND Corporation, DOI: https://doi.org/10.7249/PE315*, 2019.

[8] Gary Klein, Brian Moon, and Robert R Hoffman. Making sense of sensemaking 2: A macrocognitive model. *IEEE Intelligent systems*, 21(5):88–92, 2006.

[9] Mark Van Harmelen and David Workman. Analytics for learning and teaching. *CETIS Analytics Series*, 1(3):1–40, 2012.

[10] Alice Kerly, Richard Ellis, and Susan Bull. Calmsystem: a conversational agent for learner modelling. In *International Conference on Innovative Techniques and Applications of Artificial Intelligence*, pages 89–102. Springer, 2007.

[11] Katrien Verbert, Erik Duval, Joris Klerkx, Sten Govaerts, and José Luis Santos. Learning analytics dashboard applications. *American Behavioral Scientist*, 57(10):1500–1509, 2013.

[12] Kimberly E Arnold and Matthew D Pistilli. Course signals at purdue: Using learning analytics to increase student success. In *Proceedings of the 2nd international conference on learning analytics and knowledge*, pages 267–270, 2012.

[13] Alfred Essa and Hanan Ayad. Student success system: risk analytics and data visualization using ensembles of predictive models. In *Proceedings of the 2nd international conference on learning analytics and knowledge*, pages 158–161, 2012.

[14] Kwok Tai Chui, Dennis Chun Lok Fung, Miltiadis D Lytras, and Tin Miu Lam. Predicting at-risk university students in a virtual learning environ-

ment via a machine learning algorithm. *Computers in Human Behavior*, page 105584, 2018.

[15] Nasheen Nur, Noseong Park, Mohsen Dorodchi, Wenwen Dou, Mohammad Javad Mahzoon, Xi Niu, and Mary Lou Maher. Student network analysis: A novel way to predict delayed graduation in higher education. In *International Conference on Artificial Intelligence in Education*, pages 370–382. Springer, 2019.

[16] Mohammad Javad Mahzoon, Mary Lou Maher, Omar Eltayeby, Wenwen Dou, and Kazjon Grace. A sequence data model for analyzing temporal patterns of student data. *Journal of Learning Analytics*, 5(1):55–74, 2018.

[17] Annika Wolff, Zdenek Zdrahal, Drahomira Herrmannova, Jakub Kuzilek, and Martin Hlosta. Developing predictive models for early detection of at-risk students on distance learning modules. 2014.

[18] Samuel PM Choi, Sze Sing Lam, Kam Cheong Li, and Billy TM Wong. Learning analytics at low cost: At-risk student prediction with clicker data and systematic proactive interventions. *Journal of Educational Technology & Society*, 21(2):273–290, 2018.

[19] Cristóbal Romero, Sebastián Ventura, and Enrique García. Data mining in course management systems: Moodle case study and tutorial. *Computers & Education*, 51(1):368–384, 2008.

[20] Arturo Nakasone and Mitsuru Ishizuka. Storytelling ontology model using rst. In *Proceedings of the IEEE/WIC/ACM international conference on Intelligent Agent Technology*, pages 163–169. IEEE Computer Society, 2006.

[21] Leo Ferres, Avi Parush, Shelley Roberts, and Gitte Lindgaard. Helping people with visual impairments gain access to graphical information through natural language: The igraph system. In *International Conference on Computers for Handicapped Persons*, pages 1122–1130. Springer, 2006.

[22] Lidija Iordanskaja, Myunghee Kim, Richard Kittredge, Benoit Lavoie, and Alain Polguere. Generation of extended bilingual statistical reports. In *COLING 1992 Volume 3: The 15th International Conference on Computational Linguistics*, 1992.

[23] Karen Kukich. Design of a knowledge-based report generator. In *Proceedings of the 21st annual meeting on Association for Computational Linguistics*, pages 145–150. Association for Computational Linguistics, 1983.

[24] Bernd Bohnet, François Lareau, Leo Wanner, et al. Automatic production of multilingual environmental information. In *EnviroInfo (2)*, pages 59–66, 2007.

[25] Eli Goldberg, Norbert Driedger, and Richard I Kittredge. Using natural-language processing to produce weather forecasts. *IEEE Expert*, 9(2):45–53, 1994.

[26] José Coch. Interactive generation and knowledge administration in multimeteo. In *Proc. 9th International Workshop on Natural Language Generation (INLG-98), Aug.*, 1998.

[27] Somayajulu Sripada, Ehud Reiter, and Ian Davy. Sumtime-mousam: Configurable marine weather forecast generator. *Expert Update*, 6(3):4–10, 2003.

[28] MG Kahn, LM Fagan, and LB Sheiner. Combining physiologic models and symbolic methods to interpret time-varying patient data. *Methods of information in medicine*, 30(03):167–178, 1991.

[29] Dirk Hüske-Kraus. Suregen-2: A shell system for the generation of clinical documents. In *Demonstrations*, 2003.

[30] Mary Dee Harris. Building a large-scale commercial nlg system for an emr. In *Proceedings of the Fifth International Natural Language Generation Conference*, pages 157–160, 2008.

[31] Ehud Reiter, Roma Robertson, and Liesl M Osman. Lessons from a failure: Generating tailored smoking cessation letters. *Artificial Intelligence*, 144 (1-2):41–58, 2003.

[32] Daniel M Russell, Mark J Stefik, Peter Pirolli, and Stuart K Card. The cost structure of sensemaking. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*, pages 269–276, 1993.

[33] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

[34] Andrew Ng. Clustering with the k-means algorithm. *Machine Learning*, 2012.

[35] Adam Nickerson, Nathalie Japkowicz, and Evangelos E Milios. Using unsupervised learning to guide resampling in imbalanced data sets. In *AISTATS*, 2001.