SBE, Vol.24, No.1, 2021
ISSN 1818-1228
DOI: https://doi.org/10.29117/sbe.2021.0125

TEXT MINING DATA FROM STUDENTS TO REVEAL MEANINGFUL INFORMATION FOR EDUCATORS

Zainab M. AlQenaei
Assistant Professor
College of Business Administration
Kuwait University
zainab.alqenaei@ku.edu.kw

David E. Monarchi
Retired Full Professor
Information Systems
Leeds School of Business
University of Colorado at Boulder

ABSTRACT

Academic institutions adopt different advising tools for various objectives. Past research used both numeric and text data to predict students' performance. Moreover, numerous research projects have been conducted to find different learning strategies and profiles of students. Those strategies of learning together with academic profiles assisted in the advising process. This research proposes an approach to supplement these activities by text mining students' essays to better understand different students' profiles across different courses (subjects). Text analysis was performed on 99 essays written by undergraduate students in three different courses. The essays and terms were projected in a 20-dimensional vector space. The 20 dimensions were used as independent variables in a regression analysis to predict a student's final grade in a course. Further analyses were performed on the dimensions found statistically significant. This study is a preliminary analysis to demonstrate a novel approach of extracting meaningful information by text mining essays written by students to develop an advising tool that can be used by educators.

Keywords: educational data mining; post-secondary education; student learning; advising faculty; text mining; natural language processing.

1. Introduction

Academic advising is the term generally used to describe the process of guiding students throughout their academic years. Williamson *et al*. (2014) refer academic advising to the actions related to choosing courses, completing students' schedules, and tackling general issues in students' life.

There has been ongoing effort at universities and academic institutions to advise students at three different timings. First, advising related to the students' past when institutions might request students to complete an evaluation focusing on the instructor's teaching style, course content, delivery methods, and so on (Richardson 2005). If such evaluations were taken into careful consideration by educators, they will most likely serve as an advising tool, but primarily focusing on past courses. The students who completed the evaluation by the end of the semester will not benefit from the changes. Second, advising related to the students' present when academic feedback is given to students on their progress during a semester (see Matcha *et al.* 2019 and Kovanović *et al*. 2018). Finally, advising related to the students' future, which is primarily done during the student's entire academic program which guides them to what courses to take, what major/minor to select, or to simply provide an academic roadmap until graduation (see Santoso 2010 and Feghali *et al.* 2011).

With different timing of advising, research has been conducted proposing various advising tools for different objectives. For example, long ago Appleby (1989) discussed the use of an application for academic advising in which students go through two stages. The first stage is concerned with graduation requirements, and the second is concerned with future planning, that is the phase after a student graduates.

Kot (2014) studied the effect of centralized advising on students' GPAs and concluded that there exists a significant impact of advising on a higher GPA. In his ongoing research in higher education concerning student retention, Tinto discusses student success in his latest work (Tinto, 2012). He proposes a framework for effective institutional action to engage students in the learning process.

To test the effect of student advising on chances of academic success, Bahr (2008) addresses the sensitivity of advising timing. The author tests the effect of advising using two cases: successful remediation in math and transfer-seeking students. The author uses data from previous studies and applies logistic regression and discrete-time event history analyses. He concluded that advising is indeed beneficial to students' success. Similar to the objective of Bahr's research (2008), Freeman (2008) discusses the relationship between effective advising and students' success. The author states that academic advising must follow a hierarchy defined as (Freeman, 2008, p. 9): Exploration of life goals, values, abilities, interests, limitations; Exploration of vocational/career goals; Selection and design of academic major or program of study; Selection of courses; Scheduling of classes.

In order to advance the current research, in this study we use textual data from students in the form of essays projected in a dimensional vector space. As Massung and Zhai (2015, p. 571) state: "text data are produced by humans and thus contain rich information about people's preferences and opinions, making it a unique source of data from which we can mine knowledge about people; such knowledge is generally hard to obtain from other kinds of data."

To the best of our knowledge, this is the first work that targets the present time of advising, i.e., on current course(s) taken by a student, to extract meaningful information using singular value decomposition on students' essays and dimension profiling. Consequently, it offers the potential to affect

a student's success in a more immediate way than the others described previously. The approach could be used to develop an advising tool for faculty during a course to predict students' final grades among other objectives. Therefore, the research questions addressed in this study are:

Research Question 1: Can dimensions from a singular value decomposition of students' essays be used to predict a student's grade in a course?

Research Question 2: Can significant dimensions (from research question 1) be used to group students' essays to reveal different dominant topics?

To address these questions, we use a well-established mathematical technique to represent the students' essays in a dimensional vector space. The rest of the paper is organized as follows: Section 2 discusses related work; in Section 3 we explain the methods used in the analyses; in Section 4 we provide the results and implications of regression analysis, profiling the dimensions, and clustering algorithm; and we conclude in Section 5 by highlighting the limitations and future research.

2. Related Work

To bridge the research the gap, we present past research in the areas of predicting students GPA, automation of advising, learning analytics, and text mining students' data.

Predicting Students' GPA

Past research have used various variables to predict students' GPA. Long ago, standardized test scores were solely used for prediction purposes (Aleamoni and Oboler, 1975) before introducing other variables such as high school rank (Bai *et al*., 2014; Cohn *et al.*, 2004). Among those variables are enrollment data, demographic data, teachers' evaluations, and learning management systems (LMS) usage data. (Betts and Morell, 1999) found that there exists a significant relationship between a university GPA and the personal background of a student including, gender, ethnicity, and family income. Surprisingly, there exists a negative relationship between GPA and the proportion of teachers with advanced degrees. The authors also found a significant relationship between teacher experience, high school GPA, and SAT scores to university GPAs. In order to increase students' performance, (Al-Barrak and Al-Razgan, 2016) used educational data mining tools to predict the final GPA of students using their grades in previous course. The model suggested the most important courses during a study plan that could predict the final GPA. In comparison of different GPA prediction models, (Zhang and Wu, 2019) found that ID3 algorithm performance better than C4.5 and CART algorithms. Variables used to predict students' GPA were numerical including test scores, experimental scores, and number of solved questions. Students grades were also predicted using LMS logs such as Moodle by applying different data mining techniques (Figueira, 2017; Nasiri *et al.*, 2012; Walsh and Mahesh, 2017).

Automation of Advising

Santoso (2010) presents an approach to automate academic advising which mainly suggests courses for students to accelerate graduation. One of the advantages of automated academic advising suggested by the author is that it provides more consistent results since "the knowledge base is not susceptible to the human problems." (Santoso, 2010, p. 293).

Very similar to Santoso (2010), Al Ahmar (2011) introduces a prototype of an automated student advising expert system. The system helps students with a major in Information Systems select courses in each semester until graduation. The author found

more than 90% matching between the expert system and human advisors excluding special case students not taken into account using the expert system (i.e., students with low GPA or medical reports). Feghali *et al*. (2011) present a web-based decision support tool named "Online Advising". It uses technology to provide online academic advising students to supplement, but not replace, human advising. The software is mainly concerned with creating schedules and course planning through the years at college. These studies are examples of advising timing related to the future.

Learning Analytics

The field of learning analytics and educational data mining has emerged in the past few years. In a recent survey, Romero and Ventura (2020) review how educational data mining and learning analytics have been applied to educational data. Ferreira-Mello *et al.* (2019) also survey educational applications of text mining. The authors investigate the available text mining techniques, educational resources, and applications.

A large body of research has been also conducted in the areas of latent semantic analysis (LSA) and natural language processing (NLP). LSA is a subfield of NLP that deals with "using computers to determine and analyze the meaning of natural language." (Robson and Ray 2012, 2). For a current review of this work, see the recent study by Hirschberg and Manning (2015). Kovanović *et al.* (2015) introduce the term "content analytics" which is a subfield of learning analytics focusing on learning context. Such learning context can be of many different forms such as faculties' syllabi, students' discussion messages, and students' essays.

You (2016) also investigates the relationship between LMS data measures and course achievement. LMS data measures collected in the middle and at the end of course include (among others): duration of study time, total viewing time of instructional videos, and number of late submissions. All data used in You (2016) to predict course achievement are numerical data. According to You (2016, p. 24):

As considerable student data have become available in the education field, the attention to utilizing student data to improve academic success has increased. The use of analytic techniques in learning is called learning analytics, and learning analytics is defined as "the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs," according to the 1st Conference on Learning Analytics and Knowledge (Siemens 2010).

The key objectives of employing learning analytics involve identifying at-risk students by predicting student learning success, providing adequate interventions, and improving learning outcomes.

Studies using learning analytics for academic advising exist at different timings. There are studies of using learning analytics for advising related to the present. For example, to explore students' learning strategies, Matcha *et al*. (2019) explore the use of learning analytics to examine the characteristics on learning strategies. The authors use data from class preparation activities in a flipped classroom. Similarly, Jovanović *et al.* (2017) examine learning strategies in flipped learning settings and the association between learning strategies and course performance. Kovanović *et al.* (2018) uses automated text analytics for assessment of reflective writing by students. Similarly, there are studies of using learning analytics for advising related to the past (i.e. using data from courses already completed by students). For example, in Antonenko *et al.* (2012), the authors used cluster analyses of data from an online learning environment. The objective was to analyze different characteristics of

learning behaviors. In Dias and Diniz (2014) research, the authors' objective was to identify learning profiles using a sample of 36 undergraduate students. The main purpose of identifying students' profiles is to understand the key factors, weaknesses, and drawbacks of an online LMS. The methodology used included both qualitative (interview with students) and quantitative (content analysis software) methods.

Text Mining Students' Data

Essays written by students were analyzed in prior research for different objectives (summarizing, providing feedback, grading, etc.) than ours. Kovanović *et al.* (2015) use students' essays for summarizing purposes and providing feedback as part of the assessing process.

Carpenter *et al.* 2007 use students' essays for evaluating students' level of understanding. They propose an automated process to grade essay questions achieving results consistent with expert human graders in the field using knowledge representation methods (conceptual graphics) followed by cluster analyses.

Another study that uses students' essays is Hastings *et al.* (2012). The authors use text mining to score the essays automatically and, of all the techniques used, latent semantic analysis was superior. Crossley *et al.* (2016) also use data mining techniques focusing on text features to grade essays written by students. Pennebaker *et al.* (2014) analyze over 50,000-college admission essays to investigate the relationship between the essays and academic performance as measured by the student's GPA.

Foltz *et al.* (1999) used students' essays to develop a set of software tools to score the quality of essays using LSA. Another use of student essays was presented in Nguyen and Litman (2016) where the authors used 90 essays from an online corpus to extract new features to build a model for argument indicators and abstract over essay topics.

Other research mines chat messages written by students rather than essays written by them (Anjewierden *et al.* 2007). The authors use data mining tools on educational chat messages to guide and increase the awareness of learners. Their methodology was based on classifying chat messages into four pre-defined classes using 78 chat sessions and over 16,000 chat messages. Chen *et al.* (2016) used the Latent Dirichlet Allocation (LDA) topic modeling to automate the assessing and grading of students' reflection journals. Another piece of research which uses student written work is Klebanov *et al.* (2016). The authors use NLP to automate utility value evaluation of a piece of writing. Their proposed process measures to what extent a writing expresses utility value of the student.

3. Methods

The research data collection and methodology used in this paper is presented in Figure 1.

Figure 1. Research Methodology Steps

3.1 Data Description

The data included in the research (first step in Figure 1) consist of undergraduate students' essays at the College of Business Administration (Kuwait University) from

different courses and semesters. Courses were offered from the Quantitative Methods and Information Systems Department (QMIS). Since this study targets the present timing of advising, students were requested to write a one-page paper a few weeks before the end of the semester and submit it within a week, expressing their feelings toward the course in general, their performance in the course, their likes and dislikes, their impression of the final grade, and any personal views. Two bonus points were added to the overall grade of a student as an incentive to complete and submit the essay. Essays were written in English as a second language (spelling and grammar checked).

Students used word processor software to type in their essay and submit it online through a LMS, Blackboard. To ensure that the submitted material was not copied from an external source Safe Assign[1] was used (a software program to prevent plagiarism). Students had given their consent to use their essays and personal information for research purposes before data collection and analyses.

3.2 Sample Description

The essays used for this research were written by students enrolled in the following courses:
QMIS 130: Computer Based Business Applications (two sections – 70 students)
QMIS 240: Introduction to Management Information Systems (one section – 30 students)
QMIS 336: Data Communications and Networks (one section – 24 students)

Of the 124 requests, the response rate was 79.8%, yielding 99 essays that were used for further analyses.

All students enrolled at the College of Business Administration at Kuwait University must complete the courses QMIS 130 and QMIS 240 (QMIS 130 is a prerequisite of QMIS 240). Only students majoring in Information Systems could register in QMIS 336. Hence, all students completing QMIS 336 have also completed QMIS 130 and QMIS 240. In addition, exogenous variables were collected to test the effect of different factors influencing the final grade or style of writing. Those variables are student ID, expected grade, age, year in College, and major.

3.3 Document Representation

The second step as shown in Figure 2 is document representation. While there exist numerous ways to represent a textual document in a semantic space, the choice remains a challenge as it is dependent on the purpose of the study. For example, in Kowsari *et al.* (2019), the authors discuss the limitations and applications for different methods in text feature extraction and dimensionality reduction. A few feature extraction methods they mention are term frequency-inverse document frequency, bag of words, and Word2Vec (Waykole and Thakare, 2018). As stated by the authors "A problem with bag of words approach is that the words with higher frequency becomes dominant in the data. These words may not provide much information for the model. And due to this, problem domain specific words which do not have larger scores may be discarded or ignored. To resolve this problem, the frequency of the words is rescaled by considering how frequently the words occur in all the documents. Due to this, the scores for frequent words among all the documents are reduced." (Waykole and Thakare, 2018, p. 352). Waykole and Thakare (2018) found that Word2Vec is superior to the other two methods: it "takes a huge corpus of text as an input. It then creates a vector space, which is usually of hundreds of dimensions. Each distinctive word in the corpus is allotted with corresponding vector

[1] https://www.blackboard.com/teaching-learning/learning-management/safe-assign

in the space. The words with common contexts are placed in near proximity in vector space." (Waykole and Thakare, 2018, p. 352). Martinčić-Ipšić *et al*. (2019) also discusses different document representation models: bag of words, word2vec, doc2vec, and graph-of-words. The implementation of word2vec and doc2vec captures the words and documents in a semantic space which is similar to what is used in this paper. Thus, we combined the concept of the two techniques to represent the terms and essays written by students in a semantic space. The document representation methodology that was used in this research is explained next (AlQenaei, 2009):

Preprocessing

Preprocessing the documents involves removing stopwords, lemmatizing the remaining words, and deciding which parts of speech to use in the analysis (e.g. ignoring or retaining adverbs). stopwords include common words such as the, an, is, at, etc. They also include common domain-specific words. Lemmatization is the process of reducing multiple forms of a word to the same base. In this research, only nouns were retained as part of the lemmatization step; the other parts of speech were ignored. The preprocessing step was performed using SAS 9.4 Enterprise Miner, particularly the Text Parsing node.

Constructing the term frequency (TF) matrix

Next, we used SAS Enterprise Miner 9.4 Text Filter node to parse the essays to obtain the term by document frequency (TF) matrix.[2] This matrix has the terms as rows and the documents as the columns. The cells are the count (frequency) of a term in a document. The corpus originally contained 3,677 distinct words. After excluding all parts of speech except nouns, removing stopwords, and lemmatizing the remaining words, we were left with 662 terms. The number of essays (documents) is 99. We kept only nouns which appeared in two or more essays (approximately 2% of the total number of essays) and fewer than 90 essays (approximately 90% of the total number of essays). Our rationale for this is that terms which appear rarely provide little co-occurrence information, nor do terms which appear very frequently.[3]

Transforming the TF matrix

The 662 by 99 TF matrix was transformed using Mathematica© 10. The first step in the transformation process was weighting the counts using the log-entropy weighting method (log as the local weighting and entropy as the global weighting for each term). The local weighting (L) reflects the importance of a term within a document while the global weighing (G) does the same thing but across the whole corpus. The formulas used for the log-entropy weighting are (Hare and Lewis, 2005):

$L(i, j) = \log (tf_{ij} + 1)$

$$G(i) = 1 - \sum_{j=1}^{N} \left(\frac{tfij}{gfi} \log\left(\frac{tfij}{gfi}\right)\right) / \log N$$

where:

tf_{ij}: the frequency of term i in document j

gf_i: the frequency of the term i in the entire corpus

N: the number of documents (which in this case is 99).

The second step in the transformation process was normalization, which converts the document vectors to a standard length thereby compensating for the varying number of words in the documents. The

[2] The SAS Enterprise Miner Text Parsing node lemmatized the words and performed the POS analysis.

[3] We limited the results to nouns that appeared in two more essays to reduce the amount of noise in the TF matrix. The Text Parsing and Text Filter nodes in SAS identified a total of 2830 nouns and noun phrases, including lexemes and their variations (e.g., number and numbers). There was a total of 1214 nouns that occurred in only one essay. The remaining 1616 nouns consisted of 662 root nouns plus their variations. The 662 roots are the terms that we used in our analysis.

result of weighting and normalizing the TF matrix is the matrix A.

Decomposing A (the transformed TF matrix) using singular value decomposition, again using Mathematica 10:

Matrix A was decomposed into three matrices using singular value decomposition (SVD). Singular value decomposition (SVD) is a mathematical technique that decomposes a rectangular matrix into a linear combination of three matrices (Golub and Van Loan 1996). The decomposition of A results in matrices: U, S, and V.

$A_{m \times n} = U_{m \times r} S_{r \times r} V^T_{r \times n} \approx U_{m \times k} S_{k \times k} V^T_{k \times n} = \hat{A}_{m \times n}$

where:

U: left singular vectors corresponding to the terms

V: right singular vectors corresponding to the documents

S: singular values

^{T}: transpose

r: rank of A, which is ≤ min (m,n)

k: number of dimensions retained, $k \leq r$

Â: approximation of A using k dimensions

Each term is represented by a row in U. Similarly, each document is represented by a row in V. The critical point here is the issue of dimension reduction: that is, the reduction of the dimensionality of the vector space from r to k dimensions. The choice of k has been mainly a matter of judgment by researchers. Landauer *et al.* (2007) discuss the topic of choosing the optimal number of dimensions. They state that there is currently no method for automatically selecting k and give three reasons against having a general procedure to determine the optimal dimensionality: (1) it is task dependent, (2) it is content dependent, and (3) it is size dependent (p.82). As Deerwester *et al*. (1990, p. 398) also state "… we want a value of k that is large enough to fit all the real structure in the data, but small enough so that we do not also fit the sampling error or unimportant details. The proper way to make such choices is an open issue in the factor analytic literature." The SVD of text data has been used in many applications with k typically between 10 and 500 (Bingham *et al*., 2003; Bingham and Mannila, 2001; Dumais *et al.*, 1988; Gao and Zhang, 2005; Landauer and Dumais, 1997; Murray and Durrell, 2000; Zelikovitz and Hirsh, 2001). Fewer dimensions (less than 30) have been used in studies focused on clustering (Hotho *et al*., 2002) as opposed to information retrieval. The amount of variance accounted by the singular values were 78.5% in the first 40 singular values in our research, 62.9% in the first 20 singular values, 51.1% in the first 10 singular values, and 38.4% in the first 3 singular values. Given past research suggestions and amount of variance explained by the singular values, in this study we retained 20 dimensions and used them as inputs to the analyses discussed in subsequent sections.

Rotating the results of the decomposition

Finally, the U matrix multiplied by √S from the SVD was rotated using varimax rotation in SAS STAT 9.4 to assist in the interpretation of the dimensions retained for analyses. Then we applied this rotation to the V matrix multiplied by √S from the SVD (by using the rotation vectors output after rotating the U matrix) so that both matrices would still be represented in the same vector space.

Our belief is that it is much easier to label a dimension based on a set of terms rather than a set of documents[4]. Consequently, we rotated the weighted left singular vectors so that the coefficients of each term on the new dimensions are either very large (i.e., ± 1) or near zero.

[4] Trying to use a set of documents to label a dimension is confounded by the fact that an individual document may contain multiple themes, which are not readily separable. Admittedly, this issue also exists when trying to label a dimension based on the set of terms, which are closest to it, but we believe this is an approach which is easier to examine and critique. That is, we believe using a set of words is less opaque that using a set of documents. The documents introduce another layer of analysis because they require us to first identify the themes of the documents.

In a review of educational data mining and learning analytics in different studies, Aldowah *et al*. (2019) describe the different mining techniques used. The techniques identified are classification, clustering, visual data mining, statistics, association rule mining, regression, sequential pattern mining, text mining, correlation mining, outlier detection, casual mining, and density estimation (Aldowah *et al*. 2019, p. 21). After a review of the different techniques, we find the two that closely apply to the focus of our study and to answer our research questions are regression and clustering.

3.4 Statistical Treatment

Regression Analysis

The third step as shown in Figure 1 is regression analysis. Aldowah *et al*. (2019) explain regression as "a prediction technique used to determine the relationships between dependent variables (target field) and one or more independent variables, as well as determining how such relationships can contribute to individuals' learning outcomes." They find that regression analysis can help predict students' academic performance in university courses (Aldowah *et al*. 2019, p. 25). To answer the first research question and determine whether dimensions contributed significantly to predicting the final grade of a student in a given course, we run a stepwise linear regression using IBM SPSS Statistics. Moreover, we split the dataset into a training set to build the model and then applying it to the testing set for evaluation purposes. The values of the first 20 dimensions of the essays are the independent variables, and the final grade of the students in the course is the dependent variable. The final grade was computed as a continuous variable transforming the letter grades to the point scale used at Kuwait University linearly ranging from A (4.00) to F (0.00).

Dimension Profiling

The fourth step as shown in Figure 1 is dimension profiling. The dimensions retained after the document processing step were used as independent variables for the regression analysis, then those dimensions found statistically significant to predict the final grade of a student were used to cluster the essays. We first need to name the dimensions to discover meaningful insights of the results. Various naming methodologies exists when analyzing textual documents such as topic modeling, content analysis, grounded theorizing, and NLP approaches (Hannigan *et al*., 2019). We decided to follow a simple approach that aligns with our document representation methodology. To assign names to the dimensions retained, we profiled each dimension by constructing a scree plot of the highest loading terms on each dimension after the loadings were normalized. This technique served as a visualizing tool for the dimension-naming process. The process has been used to name dimensions following factor analysis or principal component analysis (AlQenaei and Monarchi, 2016).

Clustering

The fifth step as shown in Figure 1 is clustering. Aldowah *et al*. (2019, p. 24) state that "clustering in higher education might still be considered as an effective technique to group students based on their learning characteristics, individual learning style preferences, academic performance, and behavioral interaction. It can also be used to explore collaborative learning patterns and to boost the retention rate that would allow institutions to identify at risk students at an early stage."

Of the numerous clustering algorithms available, we decided to use Ward's method, an agglomerative hierarchical clustering algorithm, to cluster the essays in order to find structure and patterns within a group of essays. Our choice was based on the fact that Ward's algorithm – a minimum-variance

method based on the distance between two clusters - is a well-established clustering choice in similar educational contexts to extract patterns and define the optimal number of clusters (see Matcha *et al.* (2019) and Jovanović *et al.* (2017)).

To answer the second research question, we used the results of the multiple regression analysis to decide which dimensions to retain. Based on the beta weights, we retained only the most significant dimensions. The corresponding beta weights of those dimensions were multiplied by the rotated essay coefficients (refer to step e. above). Finally, the weighted coefficients were used to cluster the essays. We explored cluster profiles using the exogenous variables collected from students. This was done by first finding the distribution of the number of essays in each cluster. Moreover, the average GPA of students within each cluster was calculated. The loading of significant dimensions for each cluster was found. Finally, the distribution of the number of essays written by students in the three different classes across the clusters was illustrated. The analyses of this information are discussed in the next section.

4. Results and Discussion

This section demonstrates the results the regression analysis, dimension profiling and clustering, and followed by practical implications to discuss the usefulness of the results for both faculty and students

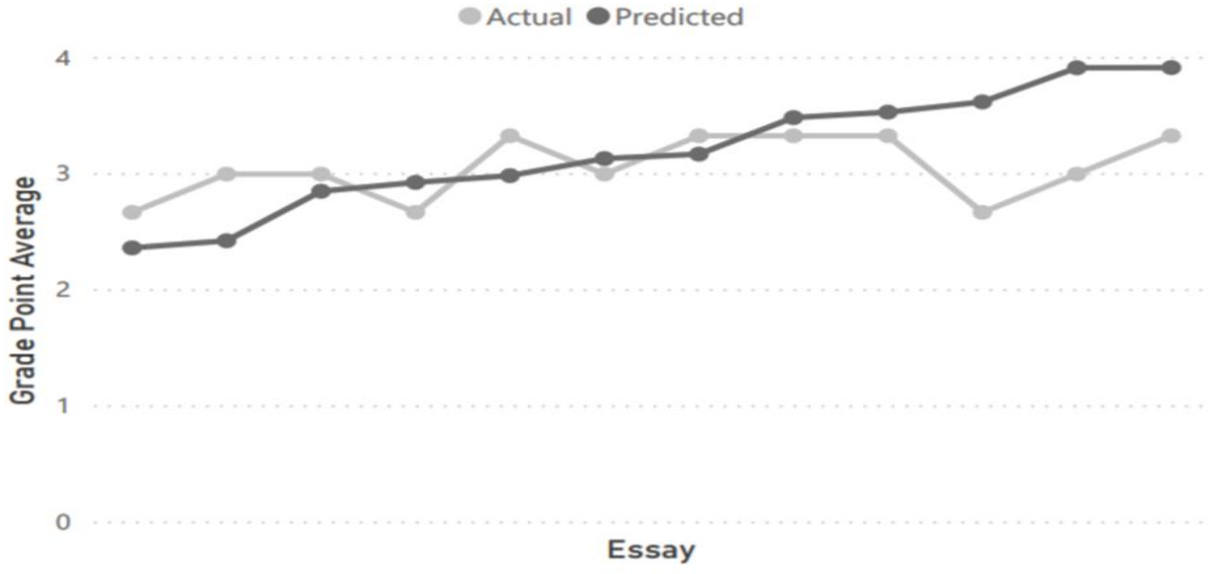
Regression Analysis

The results of the regression analysis indicate that the dimensions from a singular value decomposition of essays written by students can be used to predict a student's grade in a course. This answers our first research question. A significant regression equation was found ($F(1,93) = 4.587$, $p < 0.035$), with an R^2 of 0.258, an adjusted R^2 of 0.218, a PRESS statistic of 64.65. Students' predicted final grades is equal to 3.093 – 2.399 (Dimension 5) - 1.857 (Dimension 2) + 1.667 (Dimension 11) + 1.690 (Dimension 15) – 1.555 (Dimension 20). Refer to Appendix A for the model's summary.

After splitting the dataset into a training set to build the model and a testing set for evaluation purposes, the results of the t-test indicate that there is no statistically significant difference between the predicted grades and the actual grades of students; $t(29)=2.045$, $p = 0.913$.

Figure 2, illustrates a sample of those differences for students with actual grades between 2.50 and 3.50.

Figure 2. Sample of Actual vs. Predicted GPA



This is an interesting finding, as the results could be used as an advising tool during the present timing of a course. Students could be informed of their predicted grades before the end of the semester to be able to increase their performance in the remainder of the course. Educators could use such a model during a course as a predictive tool for the final grade and advise students who were predicted to do poorly. Similarly, in a study by Chen and Ward (2019), the authors use classification and regressions models to predict students' numerical grades and found that the liner regression model performed better in prediction. Our proposed methodology is different in the type of data used as we start with textual data in the form of essays written by students, while Chen and Ward (2019) use numerical data generated from an auto-grading system.

Dimension Profiling

Those dimensions (2, 5, 11, 15, and 20) found statistically significant in predicting student's final grades were profiled using scree plots as described above. That is, names were assigned to each dimension to interpret clusters. Refer to Appendix B for

the profiles of dimensions 2, 5, 11, 15, and 20. Both the absolute values and the signs the terms were used in creating a name for the dimension. As an example, Figure 3 displays the profile of dimension 5. Based on the terms highly loaded on the dimension - internet, graphics, image, web, and networking - the dimension was described as the topics covered in class related to the Internet and web, types of graphics and images, and networking.

The terms highly loaded on dimension 2 are interest subject, job, subject, carriers' package, and future. Accordingly, dimension 2 was described as the importance of the subject with respect to students finding it interesting and practical for their future job. The terms highly loaded on dimension 11 are memory, device, chip, and data. This dimension was described as the different topics related to PC hardware and software. The terms highly loaded on dimension 15 are chapter, email, lecture and announcement. Therefore, dimension 15 was described as the different tools and resources used to assist in the teaching process. Those tools are chapters of the textbook, using email, lecturing, and announcements posted in Blackboard and announced at the beginning of each class. Other tools were used during the course but based on the text mining process performed on students' essays, the tools associated with dimension 15 were found to be the most useful.

Finally, only two terms loaded highly on dimension 20: article and terms, and thus this dimension was described as web-based scholarly articles assigned to students related to the material being covered in class. The descriptions assigned to dimensions 2, 5, 11, 15, and 20 are listed in Table 1.

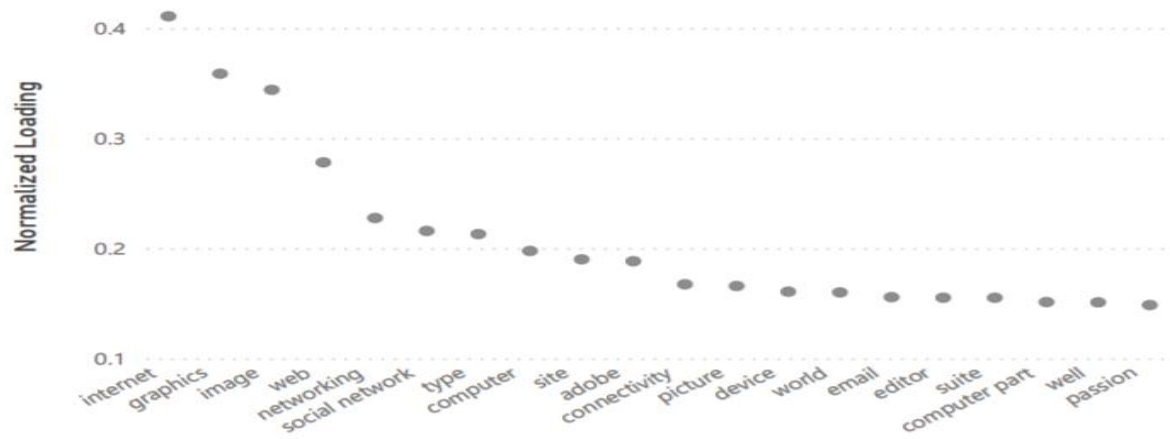


Figure 3. Term Loading on Dimension 5

Table 1. Naming of Dimensions

Dimension	Name
2	The importance of the subject, students finding it interesting and practical for their future job.
5	Topics covered in class related to the Internet and web, types of graphics and images, and networking.
11	Topics related to PC hardware and software.
15	Different tools and resources used to assist in the teaching process (chapters of the textbook, using email, lecturing, and announcements posted in Blackboard and announced at the beginning of each class).
20	Web-based scholarly articles assigned to students related to the material being covered in class.

In a study by Harrak *et al.* (2018), the authors collect textual data from students before a class begins and use a clustering algorithm to find different students' profiles. The data collected are in the form of questions posted by students and were used to find whether those questions are related to students' performance and learning behavior. The terms used to name the clusters were extracted using automatic annotation of clusters. It is different from our approach in that the authors first identified keywords that represent the values in each dimension. In our research, the dimension-profiling step was fully automated using the term loadings on each dimension.

Clustering

Based on the results of the regression analysis, dimensions 2, 5, 11, 15, and 20 were used to cluster students' essays. The result of the Ward's clustering algorithm is illustrated in the dendrogram (Figure 4). (The numbers at each junction/cluster are from SAS Proc Cluster.) As an agglomerative algorithm, Ward's algorithm always generates n-1 clusters from n leaves. The leaves are considered clusters of size one. We have 99 leaves, so consequently we have 197 clusters, with the final cluster consisting of all the leaves.

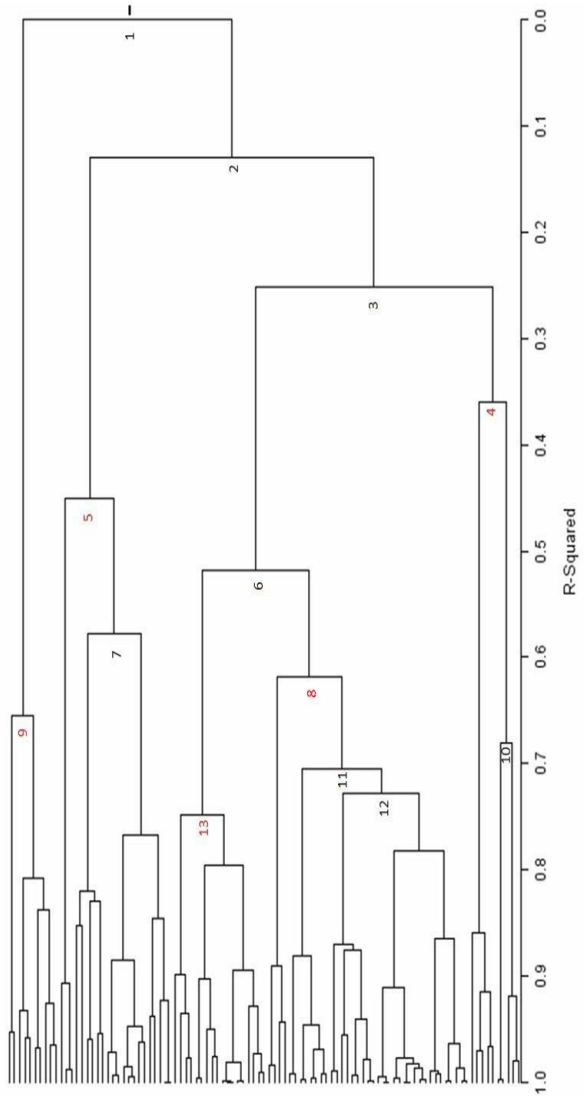


Figure 4. Dendrogram from Ward's Algorithm

In brief, Ward's method is to minimize the total variance within clusters. Initially, each leaf is a cluster, so the variance within a cluster is zero, and the total variance is zero. The variance is computed as the sum of the squared distances from the centroid of the cluster.

One of the perennial problems in cluster analysis is deciding what constitutes a "good" cluster. At one level, the answer is predetermined by the choice of algorithm used to perform the clustering, and the similarity (or dissimilarity) measure used in the algorithm. In Ward's algorithm, our criterion for choosing clusters for further examination from the dendrogram was to start at the root, cluster 1, and proceed down each branch recursively until we found a cluster, which had a "well-defined" pattern across the five retained dimensions. Our implementation of "well-defined" was a cluster in which the absolute value of at least one of the loadings on the dimensions exceeded 0.2 while the R^2 was less than 0.8. The purpose of the first criterion is to ensure that the cluster has sufficient loading on at least one dimension to aid in interpreting the cluster. The purpose of the second criterion is to avoid the trivial extreme of a cluster consisting of only a few essays. (R^2 is one for each individual essay considered as a singleton cluster; R^2 is zero for the single cluster consisting of all essays.)[5] Thus, five clusters were used to answer the second research question. Their numbers are shown in red in Figure 4.

Further analyses confirm that the dimensions found significant (from research question 1) could be used to cluster students' essays to reveal different dominant topics. The distribution of the number of essays in each cluster will assist in finding patterns across clusters. Moreover, the average GPA of students within each cluster were found. Figure 5 shows the distributions and GPA: 10 essays in Cluster 4 with an average GPA of 2.07, 18 essays in Cluster 13 with an average GPA of 2.70, 22 essays in Cluster 5 with an average GPA of 2.82, and 10 essays in Cluster 9 with an average GPA of 3.37. The remaining 39 essays are all found in one cluster, which is Cluster 8 with an average GPA of 3.29. Cluster 8 consists of all the essays, which were not incorporated into the other four clusters during the agglomeration process.

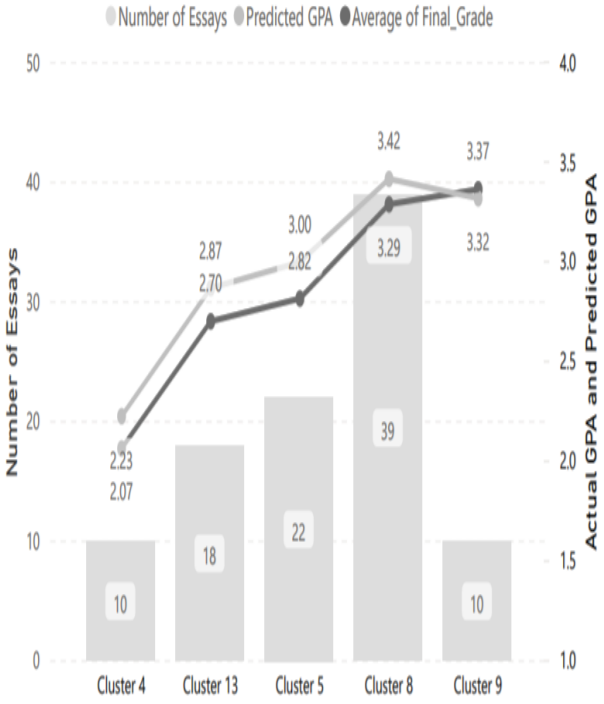


Figure 5. Distribution of Essays and GPA across Clusters

The loading of dimensions 2, 5, 11, 15, and 20 for each cluster is depicted in Figure 6. The distribution of the number of essays written by students in the three different courses across the clusters is depicted in Figure 7.

[5] Since our work is exploratory, there is no "correct" answer here. "There is no definitive answer since cluster analysis is essentially an exploratory approach; the interpretation of the resulting hierarchical structure is context-dependent and often several solutions are equally good from a theoretical point of view." (http://stats.stackexchange.com/questions/3685/where-to-cut-a-dendrogram).

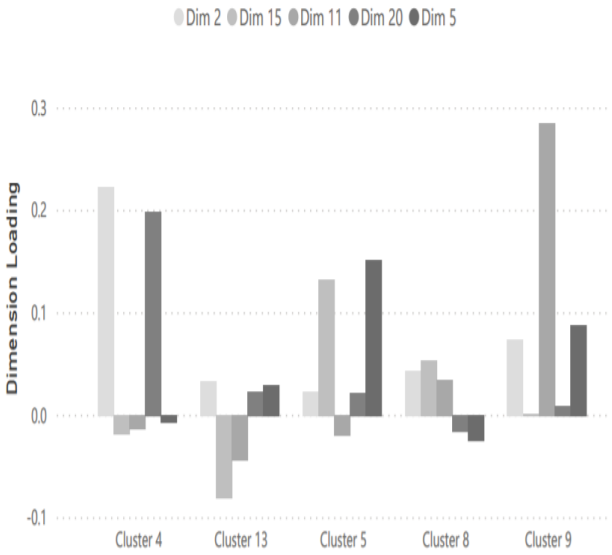


Figure 6. Dimension Loading across Clusters

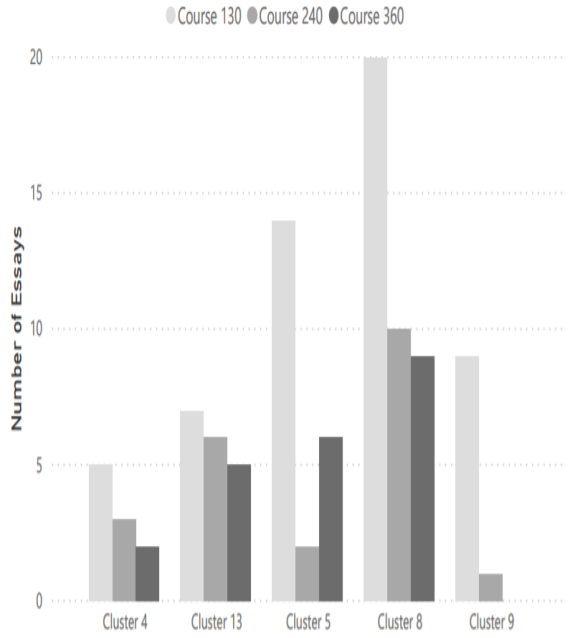


Figure 6. Number of Essays per Cluster

The previous analysis of each cluster will brief educators with insightful information about their students and topics using different terminologies Cluster 4 consists of essays written by students with low GPA on average (Figure 5). Hence, students assigned to this cluster should be academically advised early in the course since there is higher probability of those students getting a low score based on the regression model. Cluster 4 also has a relatively high loading on dimensions 2 and 20 (Figure 6 and Table 1). Dimension 2 loading indicates that essays in this cluster were about the importance of the subject taken and that they find it interesting and practical. The loading on dimension 20 also indicates that the students mentioned the web-based articles assigned to them, which are related to the material being covered in class (Table 1).

Cluster 5 consists of essays written by students with an average GPA of 2.82 (Figure 5). This cluster has a relatively high loading on dimensions 5 and 15 (Figure 6) which implies that essays in this cluster discussed topics covered in class related to the Internet and Web, types of graphics and images, and networking. The essays in cluster 5 also discuss different tools and resources used to assist in teaching (Table 1). The topics are core concepts in QMIS 130, and, from Figure 10, the highest percentage of essays within Cluster 5 is those that are written by QMIS 130 students.

Cluster 8 contains no crisp profile (Figure 6). Similarly, cluster 13 has relatively low loadings across all dimensions (Figure 6). This pattern of near zero loadings on all dimensions suggests that neither group of students has a defined profile in the context of these dimensions. The findings of Cluster 9 are rather interesting. The profile of the cluster has a high loading on dimension 11 (Figure 6). It consists of essays discussing topics related to PC hardware and software (Table 1). Also cluster 9 has the highest average GPA of all of the clusters (Figure 5) and consists of students only taking QMIS 130 and QMIS 240 (Figure 7). None of the essays written by students majoring in Information Systems were found in cluster 9. The nature of both courses QMIS 130 and QMIS 240 heavily depends on memorizing concepts versus applying concepts and analyzing cases. It could be inferred that the

profile of students in cluster 9 is more technical and practical than theoretical.

Implications

The approach used in this study represents the first steps to develop an advising tool for educators. First, it allows an instructor to identify students with no clear interest profiles (e.g. a student that falls into cluster 8 or 13). A follow-up discussion with those students could help identify the reason(s) behind those profiles. Based on the feedback, a faculty member could adjust the course settings, discuss importance of the course delivered, or emphasize the content of the course.

Second, the approach reveals topics by groups of students with different GPAs. An instructor could easily identify the topics discussed by students with high, low, or average GPA. For example, students in cluster 4 have low GPAs and the instructor could discuss the importance of the subject taken and the web-based articles assigned to them during the semester. Students in cluster 5 have an average GPA and discuss topics related to the content of the course in addition to the different tools and resources used in teaching. Students in cluster 9 have the highest GPA and may relate better to the technical topics discussed in the class. Thus, depending on cluster to which a student essays belongs and the GPA of that student, a faculty member could facilitate a more fruitful discussion to increase the knowledge base of students.

Third, it reveals top topics relevant to students in different classes. For example, none of the students that fall in cluster 9 has taken a QMIS 360 course yet and is more interested in non-theoretical topics. In addition, cluster 5 consists of students mainly registered in QMIS 130 and focus on topics such as the Internet and the web. This could identify the topics in which students are highly interested in. Given that information, a faculty member could accommodate the topics discussed in class according to the course delivered to meet the learning objectives.

Finally, the approach could reveal different tools and techniques used by the instructors that students find interesting in the teaching process (e.g. the terms highly loaded on dimension 15 are email, lecture, announcement, homework, and textbook). Those tools and techniques should be critically evaluated by faulty members on a regular basis to check whether they are used appropriately to enhance the learning of students.

5. Conclusion

This research uses a well-established mathematical technique to represent students' essays in a dimensional vector space. The results would serve as an advising tool for faculty members to predicts students' grades, profile topics discusses, and explore patterns within a group of essays. This would lead to a more customizable course delivery that would engage students during a semester. The aim of the research was to answer whether dimensions from a SVD of students' essays can be used to predict students' grades. Moreover, it investigates the possibility of revealing topics from clustered essays using significant dimensions extracted from the SVD.

The methodology of this study started with text mining essays written by students in three different undergraduate courses at the College of Business Administration at Kuwait University. Students were asked to write about their feelings toward the course in general, their performance in the course, their likes and dislikes, their expectation of their final grade, and any personal views. The essays were collected before the end of the semester.

The essays and terms were projected in a 20-dimensional space. The 20 dimensions were used as independent variables in a regression analysis. The five dimensions found

significant in predicting the final grades of students were used to cluster the essays. Ward's clustering algorithm was used, and five clusters were chosen for further analyses. The profiles of the clusters are illustrated in Figures 5, 6, and 7. Some notable differences exist among clusters and were discussed in the previous section.

The uniqueness of this study lies in the data used, automation of methodology, and the overall approach followed to text mine students' data. In the literature, there exists numerous related work to our study. For example, in a study by Bouchet *et al*. (2013), the authors collected numerical data (durations, proportions, and frequencies related to scoring, reading, and exam attempts) from students and used clustering techniques to find learning profiles of students. The data collected to cluster students were then used to find profiles and name the clusters based on the means and standard deviations of the numerical data used. The general approach is quite similar to our study; our data was however different in that we started with textual data in the form of essays written by students. As mentioned earlier, Harrak et al. (2018) used manual identification of keywords for each dimension whereas in our study dimension profiling was automated.

One of the limitations of this study lies in the fact that the results were based on a relatively small sample size of students enrolled in selected courses at Kuwait University. As a future work, the research method would be replicated after collecting a large sample size. The findings of the study may not be generalizable due to the choice of courses, university, instructor, etc. It would be necessary to replicate this study across different semesters in order to investigate whether the research model could be used as an advising tool for subsequent years.

Another limitation is related to the methodology used in this study. It needs further development to ensure replicability. The analyses heavily depended on the numbers of dimensions retained, weighting method used in the transformation process, clustering algorithm chosen, etc. Further work must be performed to test the effect of using different choices, such as including verbs with the nouns or using a different clustering algorithm. Furthermore, new student essays may or may not represent new patterns, dimensions, or clusters. Therefore, further work would include an automated process to replicate the analyses using new data. Finally, future work will include analyses on LMS usage data to complement the results of the current analyses.

This study is a continuation of a research stream in using essays written by students to develop advising tools for educators. The findings of the study will contribute to a better understanding of different student profiles across different kinds of courses.

References

Al-Barrak, M. A., and Al-Razgan, M. (2016). Predicting Students' Final GPA Using Decision Trees: A Case Study. *International Journal of Information and Education Technology*, 6(7), 528.

Aldowah, H., Al-Samarraie, H., and Fauzy, W. M. (2019). Educational Data Mining and Learning Analytics for 21st Century Higher Education: A Review and Synthesis. *Telematics and Informatics*, 37, 13-49.

Al Ahmar, M. A. (2011). A Prototype Student Advising Expert System Supported with an Object-Oriented Database. *International Journal of Advanced Computer Science and Applications (IJACSA)*, Special Issue on Artificial Intelligence, 100-105.

AlQenaei, Z. M. (2009). An Investigation of The Relationship between Consumer Mental Health Recovery Indicators and Clinicians' Reports Using Multivariate Analyses of the Singular Value Decomposition of a Textual Corpus, (Doctoral dissertation). University of Colorado at Boulder, Colorado, USA. Retrieved from http://gradworks.umi.com/33/66/3366570.html

AlQenaei, Z. M., and Monarchi, D. E. (2016). Semantic Dimension Naming (SDN): A Process for Naming Dimensions in a Semantic Space. *Advances in Computer Science and Engineering*, 16(3/4), 61.

Anjewierden, A., Kolloffel, B., and Hulshof, C. (2007). Towards educational data mining: Using data mining methods for automated chat analysis to understand and support inquiry-learning processes. *In International Workshop on Applying Data Mining in e-Learning* (ADML 2007).

Antonenko, P. D., Toy, S., and Niederhauser, D. S. (2012). Using cluster analysis for data mining in educational technology research. *Educational Technology Research and Development*, 60(3), 383-398.

Appleby, D. C. (1989). The microcomputer as an academic advising tool. *Teaching of Psychology*, 16(3), 156-159.

Bahr, P. R. (2008). Cooling out in the community college: What is the effect of academic advising on students' chances of success? *Research in Higher Education*, 49(8), 704-732.

Bai, C. E., Chi, W., and Qian, X. (2014). Do College Entrance Examination Scores Predict Undergraduate GPAs? A tale of two universities. *China Economic Review*, 30, 632-647.

Betts, J. R., and Morell, D. (1999). The determinants of undergraduate grade point average: The relative importance of family background, high school resources, and peer group effects. *Journal of human Resources*, 268-293.

Bingham, E., Kabán, A., and Girolami, M. (2003). *Topic Identification in Dynamical Text by Complexity Pursuit*. Neural Processing Letters, 17(1), 69-83.

Bingham, E., and Mannila, H. (2001). Random projection in dimensionality reduction: applications to image and text data. *In Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* (pp. 245-250), ACM.

Bouchet, F., Harley, J. M., Trevors, G. J., and Azevedo, R. (2013). Clustering and profiling students according to their interactions with an intelligent tutoring system fostering self-regulated learning. *Journal of Educational Data Mining* (JEDM), 5(1), 104-146.

Carpenter, S. L., Delugach, H. S., Etzkorn, L. H., Farrington, P. A., Fortune, J. L., Utley, D. R., and Virani, S. S. (2007). A knowledge modeling approach to evaluating student essays in engineering courses. *Journal of Engineering Education*, 96(3), 227-239.

Chen, Y., Yu, B., Zhang, X., and Yu, Y. (2016). Topic modeling for evaluating students' reflective writing: a case study of pre-service teachers' journals. *In Proceedings of the Sixth International Conference on Learning Analytics and Knowledge*, ACM, 1-5.

Chen, H., and Ward, P. A. (2019, November). Predicting student performance using data from an auto-grading system. *In Proceedings of the 29th Annual International Conference on Computer Science and Software Engineering* (pp. 234-243).

Cohn, E., Cohn, S., Balch, D. C., and Bradley Jr, J. (2004). Determinants of undergraduate GPAs: SAT scores, high school GPA and high-school rank. *Economics of Education Review*, 23(6), 577-586.

Crossley, S., Allen, L. K., Snow, E. L., and McNamara, D. S. (2016). Incorporating learning characteristics into automatic essay scoring models: What individual differences and linguistic features tell us about writing quality. *JEDM-Journal of Educational Data Mining*, 8(2), 1-19.

Deerwester, S., Dumais, S. T., Furnas, G., and Landauer, T. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41(6), 391-407.

Dias, S. B., and Diniz, J. A. (2014). Towards an Enhanced Learning Management System for Blended Learning in Higher Education Incorporating Distinct Learners' Profiles. *Educational Technology and Society*, 17(1), 307-319.

Dumais, S. T., Furnas, G. W., Landauer, T. K., Deerwester, S., and Harshman, R. (1988). Using latent semantic analysis to improve access to textual information. *In Proceedings of the SIGCHI conference on Human factors in computing systems, ACM*, 281-285.

Feghali, T., Zbib, I., and Hallal, S. (2011). A Web-based Decision Support Tool for Academic Advising. *Educational Technology and Society*, 14 (1), 82–94.

Ferreira-Mello, R., André, M., Pinheiro, A., Costa, E., and Romero, C. (2019). Text mining in Education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(6), e1332.

Figueira, Á. (2017, October). Mining Moodle logs for grade prediction: a methodology walk-through. *In Proceedings of the 5th International Conference on Technological Ecosystems for Enhancing Multiculturality* (pp. 1-8).

Foltz, P. W., Laham, D., and Landauer, T. K. (1999). The intelligent essay assessor: Applications to educational technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning*, 1(2), 939-944.

Freeman, L. C. (2008). Establishing Effective Advising Practices to Influence Student Learning and Success. *Peer Review*, 10(1), 12.

Gao, J., and Zhang, J. (2005). Clustered SVD Strategies in Latent Semantic Indexing. *Information Processing and Management*, 41(5), 1051-1063.

Golub, G. H., Van Loan, C. F. (1996). Matrix Computations. *The Johns Hopkins University Press.*

Hannigan, T. R., Haans, R. F., Vakili, K., Tchalian, H., Glaser, V. L., Wang, M. S. and Jennings, P. D. (2019). Topic modeling in management research: Rendering new theory from textual data. *Academy of Management Annals*, 13(2), 586-632.

Hare, J. S., and Lewis, P. H. (2005, July). On image retrieval using salient regions with vector spaces and latent semantics. *In International Conference on Image and Video Retrieval* (pp. 540-549). Springer, Berlin, Heidelberg.

Harrak, F., Bouchet, F., Luengo, V., and Gillois, P. (2018, March). Profiling students from their questions in a blended learning environment. *In Proceedings of the 8th International Conference on Learning Analytics and Knowledge* (pp. 102-110).

Hastings, P., Hughes, S., Magliano, J. P., Goldman, S. R., and Lawless, K. (2012). Assessing the use of multiple sources in student essays. *Behavior Research Methods*, 44(3), 622-633.

Hirschberg, J., and Manning, C. D. (2015). Advances in natural language processing. *Science*, 349(6245), 261-266.

Hotho, A., Maedche, A. and Staab, S., 2002. Ontology-Based Text Document Clustering. *KI*, 16(4), pp.48-54.

Jovanović, J., Gašević, D., Dawson, S., Pardo, A., and Mirriahi, N. (2017). Learning Analytics to Unveil Learning Strategies in a Flipped Classroom. T*he Internet and Higher Education*, 33(4), 74-85.

Klebanov, B. B., Burstein, J., Harackiewicz, J. M., Priniski, S. J., and Mulholland, M. (2016). Enhancing STEM Motivation through Personal and Communal Values: NLP for Assessment of Utility Value in Student Writing. *In Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, 199-205.

Kot, F. C. (2014). The impact of centralized advising on first-year academic performance and second-year enrollment behavior. *Research in Higher Education,* 55(6), 527-563.

Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., and Brown, D. (2019). Text classification algorithms: A survey. *Information*, 10(4), 150.

Kovanović, V., Joksimović, S., Mirriahi, N., Blaine, E., Gašević, D., Siemens, G., and Dawson, S. (2018, March). Understand students' self-reflections through learning analytics. *In Proceedings of the 8th international conference on learning analytics and knowledge* (pp. 389-398).

Landauer, T. K., and Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2), 211.

Landauer, T., McNamara, D. S., Dennis, S., and Kintsch, W. (2007). *Handbook of Latent Semantic Analysis*. Mahwah, NJ, Lawrence Erlbaum Associates.

Martinčić-Ipšić, S., Miličić, T., and Todorovski, L. (2019). The Influence of Feature Representation of Text on the Performance of Document Classification. *Applied Sciences*, 9(4), 743.

Massung, S. and Zhai, C., (2015). SyntacticDiff: Operator-based transformation for comparative text mining", In Big Data (Big Data). *The 2015 IEEE International Conference on Systems, Man, and Cybernetics* (pp. 571-580).

Matcha, W., Gašević, D., Uzir, N. A. A., Jovanović, J., and Pardo, A. (2019, March). Analytics of learning strategies: associations with academic performance and feedback. *In Proceedings of the 9th International Conference on Learning Analytics and Knowledge* (pp. 461-470).

Murray, D., and Durrell, K. (2000), "Inferring demographic attributes of anonymous Internet users", In Web Usage Analysis and User Profiling (pp. 7-20). Springer Berlin Heidelberg.

Nasiri, M., Minaei, B., and Vafaei, F. (2012, February). Predicting GPA and academic dismissal in LMS using educational data mining: A case mining. In 6th National and 3rd International conference of e-Learning and e-Teaching (pp. 53-58). IEEE.

Nguyen, H., and Litman, D. J. (2016). Improving Argument Mining in Student Essays by Learning and Exploiting Argument Indicators versus Essay Topics. *In FLAIRS Conference*, 485-490.

Pennebaker, J. W., Chung, C. K., Frazee, J., Lavergne, G. M., and Beaver, D. I. (2014). When small words foretell academic success: The case of college admissions essays. *PloS one*, 9(12), e115844.

Richardson, J. T. (2005). Instruments for obtaining student feedback: A review of the literature. *Assessment and Evaluation in Higher Education*, 30(4), 387-415.

Robson, R., and Ray, F. (2012). Applying Semantic Analysis to Training, Education, and Immersive Learning. *In the Inter-service/Industry Training, Simulation and Education Conference (I/ITSEC)* (No. 1).

Romero, C., and Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(3), e1355.

Santoso, P. B. (2010). The Development of a Case-Based Reasoning System in Relational Model using Group technology for Academic Advising. *International Journal of Academic Research*, 2(6), 287-294.

Siemens, G. (2010). About: 1st international conference on learning analytics and knowledge. (Retrieved from https://tekri.athabascau.ca/analytics/about). Accessed 24 June 2016

Tinto, V. (2012). Completing college: Rethinking institutional action. University of Chicago Press.

Walsh, K. R., and Mahesh, S. (2017). Exploratory Study Using Machine Learning to Make Early Predictions of Student Outcomes. *Twenty-third Americas Conference on Information Systems*, Boston, MA, Volume: 23.

Waykole, R. N., and Thakare, A. (2018). A Review of Feature Extraction Methods for Text Classification. *International Journal of Advance Engineering and Research Development (IJAERD)*, 5(04).

Williamson, L. V., Goosen, R. A., and Gonzalez Jr, G. F. (2014). Faculty Advising to Support Student Learning. *Journal of Developmental Education*, 38(1), 20-24.

You, J. W. (2016). Identifying significant indicators using LMS data to predict course achievement in online learning. T*he Internet and Higher Education* 29, 23-30.

Zhang, Y., and Wu, B. (2019, May). Research and application of grade prediction model based on decision tree algorithm. *In Proceedings of the ACM Turing Celebration Conference-China* (pp. 1-6).

Zelikovitz, S., and Hirsh, H. (2001). Using LSI for text classification in the presence of background text. *In Proceedings of the tenth international conference on Information and knowledge management*, ACM, 113-118.

Appendix A

Stepwise Regression Output

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics R Square Change	F Change	df1	df2	Sig. F Change
1	.277	.077	.067	.85970	.077	8.064	1	97	.006
2	.363	.132	.114	.83791	.055	6.110	1	96	.015
3	.418	.175	.149	.82127	.043	4.930	1	95	.029
4	.470	.221	.188	.80221	.046	5.568	1	94	.020
5	.508	.258	.218	.78733	.037	4.587	1	93	.035

1. Predictors: (Constant), Dim_5
2. Predictors: (Constant), Dim_5, Dim_2
3. Predictors: (Constant), Dim_5, Dim_2, Dim_11
4. Predictors: (Constant), Dim_5, Dim_2, Dim_11, Dim_15
5. Predictors: (Constant), Dim_5, Dim_2, Dim_11, Dim_15, Dim_20

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	5.960	1	5.960	8.064	.006[b]
	Residual	71.691	97	.739		
	Total	77.651	98			
2	Regression	10.250	2	5.125	7.300	.001[c]
	Residual	67.401	96	.702		
	Total	77.651	98			
3	Regression	13.575	3	4.525	6.709	.000[d]
	Residual	64.076	95	.674		
	Total	77.651	98			
4	Regression	17.158	4	4.290	6.666	.000[e]
	Residual	60.493	94	.644		
	Total	77.651	98			
5	Regression	20.002	5	4.000	6.453	.000[f]
	Residual	57.649	93	.620		
	Total	77.651	98			

a. Dependent Variable: Final_Grade
1. Predictors: (Constant), Dim_5
2. Predictors: (Constant), Dim_5, Dim_2
3. Predictors: (Constant), Dim_5, Dim_2, Dim_11
4. Predictors: (Constant), Dim_5, Dim_2, Dim_11, Dim_15
5. Predictors: (Constant), Dim_5, Dim_2, Dim_11, Dim_15, Dim_20

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients			Correlations			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	3.046	.091		33.385	.000					
	Dim_5	-2.218	.781	-.277	-2.840	.006	-.277	-.277	-.277	1.000	1.000
2	(Constant)	3.165	.101		31.295	.000					
	Dim_5	-2.370	.764	-.296	-3.104	.003	-.277	-.302	-.295	.993	1.007
	Dim_2	-1.937	.784	-.236	-2.472	.015	-.212	-.245	-.235	.993	1.007
3	(Constant)	3.117	.101		30.717	.000					
	Dim_5	-2.415	.749	-.302	-3.225	.002	-.277	-.314	-.301	.993	1.007
	Dim_2	-1.891	.768	-.230	-2.461	.016	-.212	-.245	-.229	.993	1.007
	Dim_11	1.632	.735	.207	2.220	.029	.205	.222	.207	.998	1.002
4	(Constant)	3.043	.104		29.280	.000					
	Dim_5	-2.319	.732	-.290	-3.166	.002	-.277	-.310	-.288	.990	1.010
	Dim_2	-1.748	.753	-.213	-2.322	.022	-.212	-.233	-.211	.986	1.014
	Dim_11	1.712	.719	.217	2.382	.019	.205	.239	.217	.996	1.004
	Dim_15	1.748	.741	.216	2.360	.020	.236	.236	.215	.989	1.011
5	(Constant)	3.093	.105		29.561	.000					
	Dim_5	-2.399	.720	-.300	-3.332	.001	-.277	-.327	-.298	.987	1.013
	Dim_2	-1.857	.741	-.226	-2.507	.014	-.212	-.252	-.224	.982	1.019
	Dim_11	1.667	.706	.212	2.362	.020	.205	.238	.211	.995	1.005
	Dim_15	1.690	.727	.209	2.323	.022	.236	.234	.208	.988	1.012
	Dim_20	-1.555	.726	-.192	-2.142	.035	-.176	-.217	-.191	.992	1.008

a. Dependent Variable: Final Grade

Appendix B

Term Loadings on Dimension 2, 5, 11, 15, and 20

The figures below illustrate the positive and negative loadings which refer to the coefficient of the value in the factor (dimension) pattern for a term. For example, the coefficient of Dimension 15 for the term "career" is -0.23. Those figures were used to name the dimensions retained in this study further analyses.

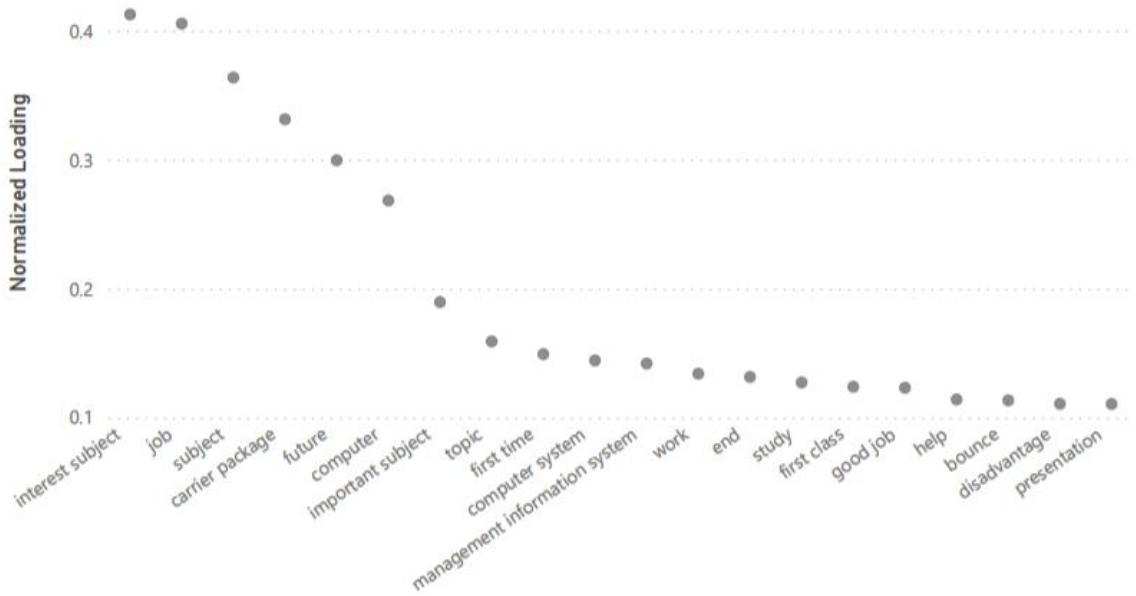


Figure 7. Term Loading on Dimension 2

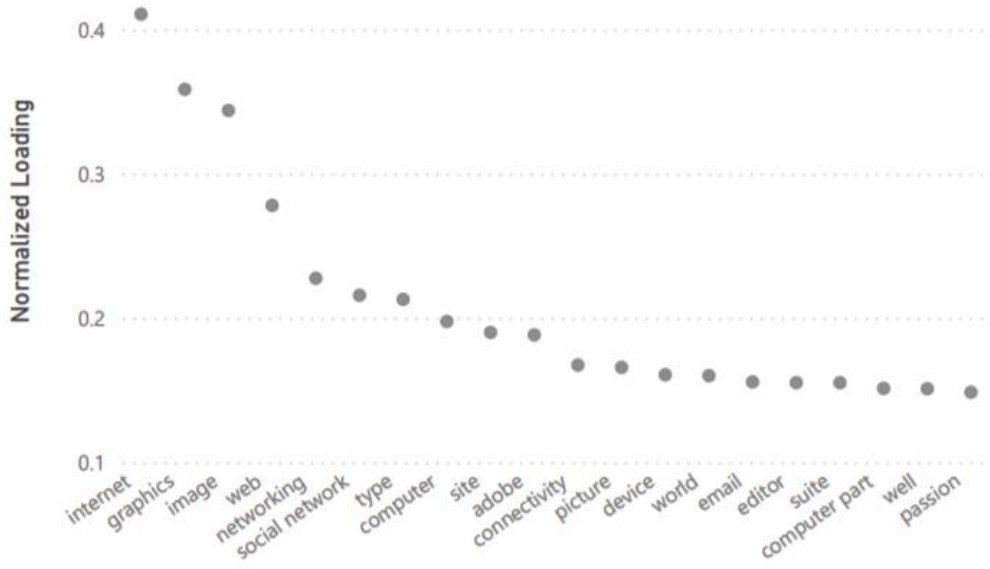


Figure 8. Term Loading on Dimension 5

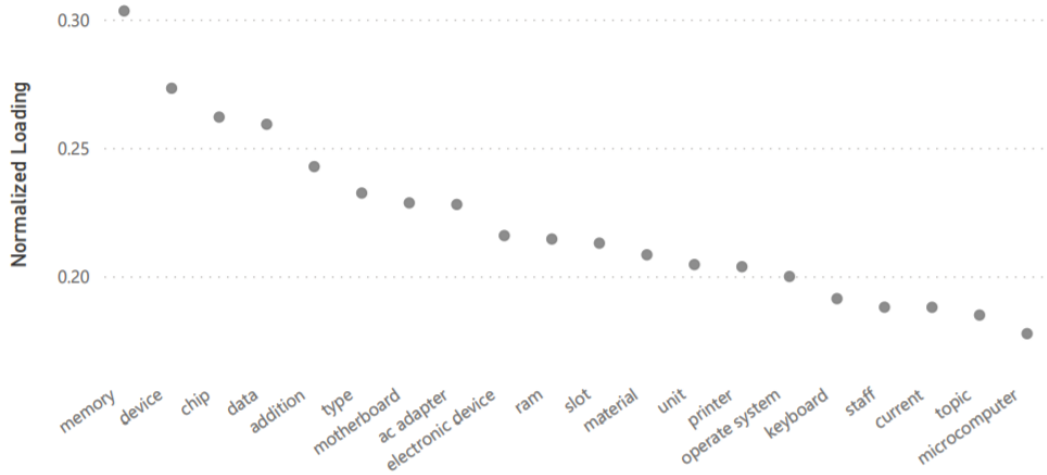


Figure 9. Term Loading on Dimension 11

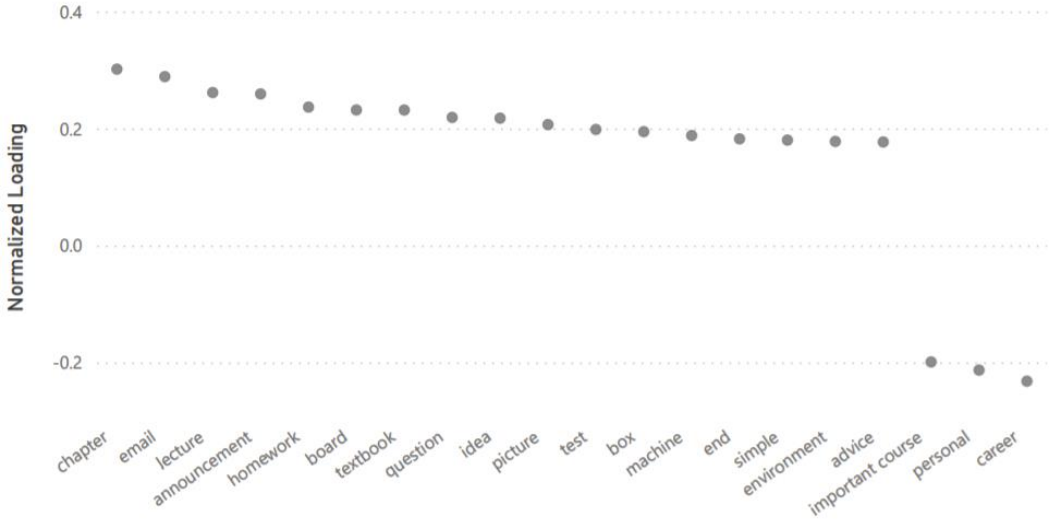


Figure 10. Term Loading on Dimension 15

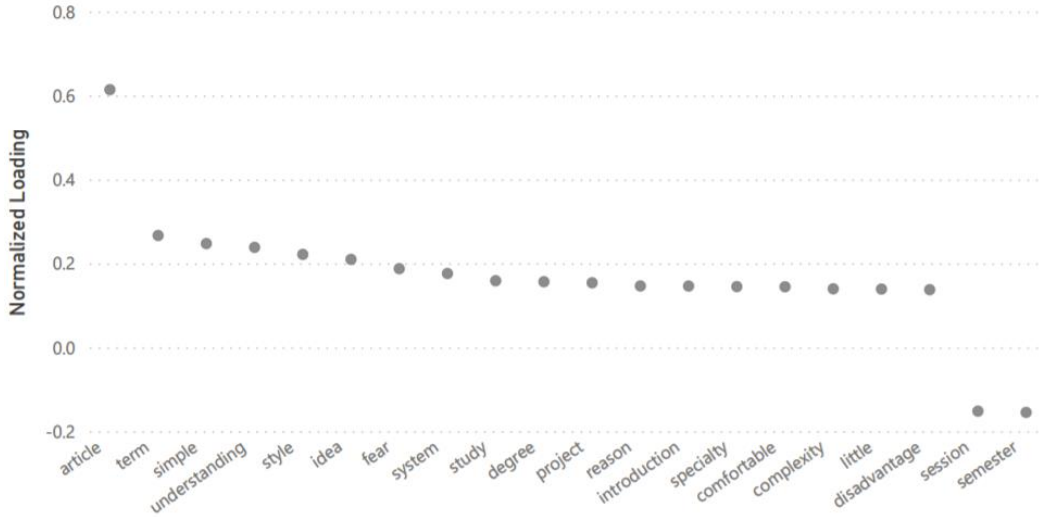


Figure 11. Term Loading on Dimension 20

Zainab M. AlQenaei is an Assistant Professor at the College of Business Administration, Kuwait University. Her current research interests include text mining and business analytics. She received a Ph.D. in Business Administration from the University of Colorado, a Master of Business Administration from the University of Pittsburgh, and Bachelor in Computer Engineering from Kuwait University. She is a member of the Chief Science Officer steering committee at the Kuwait Foundation for the Advancement of Sciences, a member of the Informatics Academy team at His Highness Sheikh Salem Al-Ali Al-Sabah Informatics Award, and a co-founder of a local IT awareness campaign Be Smart and Safe.

David E. Monarchi is a retired Full Professor of Information Systems at the Leeds School of Business, University of Colorado at Boulder. He has published in a variety of peer-reviewed journals including CACM, JASA, MISQ, JMIS, Decision Sciences, The American Statistician, and DSS. His current research interests are in the areas of text mining and network analysis. He is now an independent consultant working broadly in the area of Business Intelligence.