Journal of Effective Teaching in Higher Education

Using Text Mining and Data Mining Techniques for Applied Learning Assessment

Jessica Cook, University of North Carolina Wilmington
Cuixian Chen, University of North Carolina Wilmington, chenc@uncw.edu
Angelia Reid-Griffin, University of North Carolina Wilmington

Abstract. In a society where first hand work experience is greatly valued many universities or institutions of higher education have designed their Quality enhancement plan (QEP) to address student applied learning. This paper is the results of a university's QEP plan, called Experiencing Transformative Education Through Applied Learning or ETEAL. This paper will highlight the research that was conducted using text mining and data mining techniques to analyze a dataset of 672 student evaluations collected from 40 different applied learning courses from fall 2013 to spring 2015, in order to evaluate the impact on instructional practice and student learning. Text mining techniques are applied through the NVivo text mining software to find the 100 most frequent terms to create a document-term matrix in Excel. Then, the document-term matrix is merged with the manual interpretation scores received to create the applied learning assessment data. Lastly, data mining techniques are applied to evaluate the performance, including Random Forest, K-nearest neighbors, Support Vector Machines (with linear and radial kernel), and 5-fold cross-validation. Our results show that the proposed text mining and data mining approach can provide prediction rates of around 67% to 85%, while the decision fusion approach can provide an improvement of 69% to 86%. Our study demonstrates that automatic quantitative analysis of student evaluations can be an effective approach to applied learning assessment.

Keywords: Text mining, data mining, applied learning assessment, short answer questions, student evaluation

Text mining, sometimes referred to as text data mining, is the action of obtaining patterns or interesting knowledge from text-based documents. Text mining can become very complicated and time-consuming when original text documents lack structure (Tan, 1999). The process of text mining consists of two main phases: refining the original text documents to some chosen form and extracting knowledge from the text documents through patterns (Delgado, 2002). Mining a text-based document after it has been refined to the chosen form finds critical patterns and relationships seen across all documents (Tan, 1999).

Student evaluations of teaching (SET) are seen from two different perspectives: informal and formal (Scriven, 1967; Stake, n.d.). Formal evaluation is done by conducting standardized testing of students. This study will focus heavily on the informal perspective of student evaluations. An informal student evaluation is perceived as informal based on its casual observation and subjective bias/judgment. The reason for focus on informal perspective is provide a personalize approach to evaluating course and learning objective that educator

emphasizes in the course. One study revealed that most educators feared that scorers would not pay adequate attention to the characteristics that the educator deems most important. The best teachers continually utilize what is learned from student evaluations to improve their teaching practices (Ramsden, 2003).

Educational Evaluations: SETs

Student evaluation of Teaching (SET) is a common tool used in numerous institutions of higher education to provide evidence of teaching effectiveness and reflection of students' learning (Wagner, Rieger, & Voorvelt, 2016). In terms of evaluating effectiveness of teaching, students are positioned to be intuitively knowledgeable of information on actual effectiveness. Oftentimes students lack information on how to assess teaching effectiveness which is problematic when SET scores are used for promotions and contracts renewals (Boring, 2017). Because SETs have a history of being biased in areas of race/ethnicity and gender (Boring, 2017; Wagner, Rieger, & Voorvelt, 2016) is the reason why this study focusing on the informal perspective of SETs and how it measures the instructional practices and student learning in 40 different applied learning courses from fall 2013 to spring 2015.

The SETs are typically designed as a rating form for students to rank the instructor and/or course based on numerous specific characteristics of effectiveness (Uttl, White, Gonzalez, 2017). They are administered at the end of the semester and are often optional for student to complete. However, some higher education institutions have implemented required completion of SETs to improve response rates of the instrument (Boring, 2017).

Students have been reported of not showing any objection to filling out evaluations and are often honest (Douglas & Carroll, 1987; Gal & Gal, 2014). According to Gal and Gal's (2014) study on knowledge bias of student evaluations in an Economics course, students believed their role in evaluating courses is special, as it positions them to provide feedback that is reflective of the teaching quality. Other claims of student evaluations being reliable than other teacher effectiveness measures, such as peer ratings and observations is supported by other researchers (Heller & Clay, 1993; Fike, Fike, & Zhang (2015). The research by Galbraith, Merrill, & Kline's (2012) on the student evaluation of teaching effectiveness (SETE) validity in measuring student outcomes in business classes, found "student rating of learning outcome problem from different statistical perspectives, resulted in a high degree of consistency with respect to validity (p. 368). Recent studies on SET indicate that students demonstrate some bias in terms of teacher background and behaviors rather than quality of course instruction (Wagner, Rieger, & Voorvelt, 2016). Often, students believe that evaluations are effective and that teachers value the input from student evaluations and do not rank based on personal biases or grade. Students also believe that evaluations are a critical way to improve/adjust faculty teaching methods and improve quality of course (Scriven, 1967; Wagner, Rieger, & Voorvelt, 2016). A study found that students prefer mid-semester evaluations over those that take place at the end of the semester, because they are able to see the change being applied from the evaluations (Abott et al., 1990).

Student evaluation is a strong measure of how effective a faculty's teaching practices are and can reflect student learning (Beleche, Fairris, & Marks, 2012). It is important that students are motivated to actively participate and provide honest input that contributes to the success of evaluation systems. Research conducted by Chen et al. (2003) found that students consider improvement in the implemented teaching practices to be the most attractive outcome of the evaluation system. The second most attractive outcome is seeing change to improve the course content. Chen et al. (2003) finds that students are more motivated to participate in evaluations when they believe their feedback is seen as meaningful. The quality of student evaluations is essential in obtaining meaningful student feedback to provide areas of opportunity to improve teaching methods and effectiveness.

Teaching and learning in higher education are inextricably and elaborately linked. Good teachers continually use what they learn from their students to improve their own practice. The assumption that the primary goal of teaching is to improve student learning and teaching, leads to the argument that a reflective approach would be effective. Thus, student evaluation is an essential aspect to improve faculty teaching methods and course content leading to increased student learning (Ramsden, 2003). The role SETs have in providing feedback in higher education aids in student satisfaction of course and retention and completion at the institution. When student course evaluations are matched with student specific objectives for courses there can be positive, statistically significant associations between students' learning and the course evaluation (Beleche, Fairris, & Marks, 2012).

As there are numerous studies that have been conducted on student evaluation of faculty instruction using quantitative, meta-analyses practices (Evans, 2013; Uttl, White, & Gonzalez, 2017; Zhao & Gallant, 2012), this study provides a timely and unique approach to using text mining and data mining techniques in examining the validity and reliability of student evaluations in accessing teacher effectiveness and student learning. Taking into account previous literature on student evaluation we are able to use this practice in providing a thorough critique of assessment and gain insight on the extent to which the classroom environment or other related factors affect student evaluation of faculty instruction in the applied learning courses (Zhao & Gallant, 2012).

Abd-Elrahman et al. (2010) considers the automatic text mining techniques as a good method to investigate student course evaluation in a qualitative, open-ended manner. These techniques aim to identify unrevealed aspects affecting student learning process and develop a quantitative tool for these aspects. After preprocessing, each evaluation is categorized with the negative and positive comments made regarding the course. Then text mining is utilized to create two major groups: one for positive words and one for negative words. This study shows that the written responses from the student's courses can be analyzed through text mining to understand the effectiveness of teaching.

Applied Learning Assessment

Like many universities, the higher education institution in this study aims to engage students in the research process or in creative scholarly activity in meaningful ways. Following such commitment, among the Quality Enhancement Plan, the Experiencing Transformative Education through Applied Learning (ETEAL) program has been initiated to have a positive impact on student learning with an applied learning experience in three areas: critical thinking, thoughtful expression, and inquiry. The ETEAL supported pedagogy initiatives offer many great opportunities, resource and funds for faculty to explore innovative pedagogies in applied learning, and/or implement high-impact pedagogies in new disciplines, promote the involvement of undergraduate students in faculties' scholarly and creativity work, and enrich the interdisciplinary collaboration across campus. Since fall 2013, over a hundred ETEAL-supported initiatives have been implemented campus wide. Enormous efforts have been made to promote applied learning among departments of traditional sciences, social sciences, humanities, arts, etc.

After three years since the ETEAL initiatives started, it is pressing to review the assessment data to evaluate its impact on instructional practice and student learning. Such data includes faculty survey, student survey, and scores of student artifacts from ETEAL-supported initiatives, as well as from non-ETEAL supported *Exploration Beyond the Classroom* (EBC) activities in classes, projects, internships, study-abroad and etc. Therefore, it is critical to formulate and evaluate the influence of applied learning experiences to determine analytically whether the ETEAL-supported applied learning techniques are effective in comparison to non-ETEAL Exploration Beyond the Classroom experiences. The statistical analysis outcomes will provide scientific evidence of student learning and program effectiveness, with assessment foci on both student learning outcome and program outcome. By comparing the assessment data from ETEAL and non-ETEAL Exploration Beyond the Classroom (EBC), we aim to determine whether there is any statistically significant difference among ETEAL and EBC in terms of student learning and program effectiveness, and discover the related factors if such a difference exists. Specially, applied learning courses at the university are assessed by student evaluations completed throughout the length of the course. At the start of the semester, students complete an intention reflection articulating their expectations, the purpose, and/or goals of the experience in terms of personal educational development (EBC 1). Upon completion of the course, students submit a final reflection synthesizing: (i) knowledge drawn from their coursework to address challenges involved in the experience (EBC 2), (ii) the impact of the experience on personal educational development (EBC 3A), and (iii) the impact of the experience in the profession or in the field of study (EBC 3B). A sample of guidance for both the initial reflection and the final reflection for ETEAL supported pedagogy initiatives is illustrated in Appendix A.

In order to evaluate the impact on instructional practice and student learning, all student evaluations are manually interpreted and scored on a scale 0 to 4 based on a provided scoring rubric by scorers who must first go through a mandatory training process. A sample of the scoring rubric is illustrated in Appendix B. For the training,

each scorer is required to participate in two parts of an event. The first part consists of a five-hour session during which the rubric is reviewed and each person begins scoring with a partner. The second part consists of completion of the scoring of student work on one's own, this can last up to approximately 5 hours. At the end of the event, each scorer is asked to provide feedback regarding the process and rubric for continual improvement in the scoring process. It is mandatory that scorers attend at least one event, but are invited to attend as many as they like. Scorers are allowed to pick from events covering topics including student critical thinking skills, student-written communication skills, and student evaluation skills. It is noted that the human manual scoring process is very complicated and time consuming.

It is believed that the more in-depth evaluation leads to a better understanding of instructional practice and student learning outcomes. Therefore, even though intensive human manual scoring to analyze student evaluations is important, automatic quantitative analysis of student evaluation can be an alternative efficient approach to analyze students' text response. In this paper, both text mining and data mining techniques are investigated on students' text-based course evaluation to identify unrevealed aspects of instructional practice and student learning and develop a quantification tool to formulate and evaluate the influence of applied learning experiences.

Data Gathering and Cleaning

All original PDF files are provided by the institution's General Education Assessment Office. These PDF files cover student evaluations of applied learning experiences from both ETEAL and EBC courses, consisting of scanned handwritten documents and scanned typed documents. As a pre-processing step, the answers from the original scanned PDF files are transcribed into .txt files by three students and a faculty member, which proved to be a very time-consuming process. Many issues come with the case of scanned handwritten files, including sloppy handwriting and faded handwriting. For some files, human judgment is used to best make out the writing that is illegible or has become extremely faded after being scanned in as a PDF file. In the case of scanned typed files, a PDF file converter is used to convert the PDF files into a document that could easily be copied and pasted into a .txt file. The PDF file converter can only convert one file at a time, so it is a time-consuming process. A drawback of using the PDF file converter is spelling and grammatical errors that are caused by the converter program being used. To fix these errors, each file is manually checked for spelling and grammar mistakes. A few of the original PDF files are not used because they are written in a different language (e.g. in French).

Our final dataset consists of 672 student evaluation .txt files. All student evaluations are collected from the cycle of two academic years (fall 2013-spring 2015). Among them, part of the student evaluations are collected from 21 different courses for the academic year of fall 2013- spring 2014, while the rest are from 19 different courses during the academic year of summer 2014 -spring 2015. These

courses include traditional sciences, social sciences, humanities, and arts. All, but four, of the applied learning courses covered are ETEAL-supported courses.

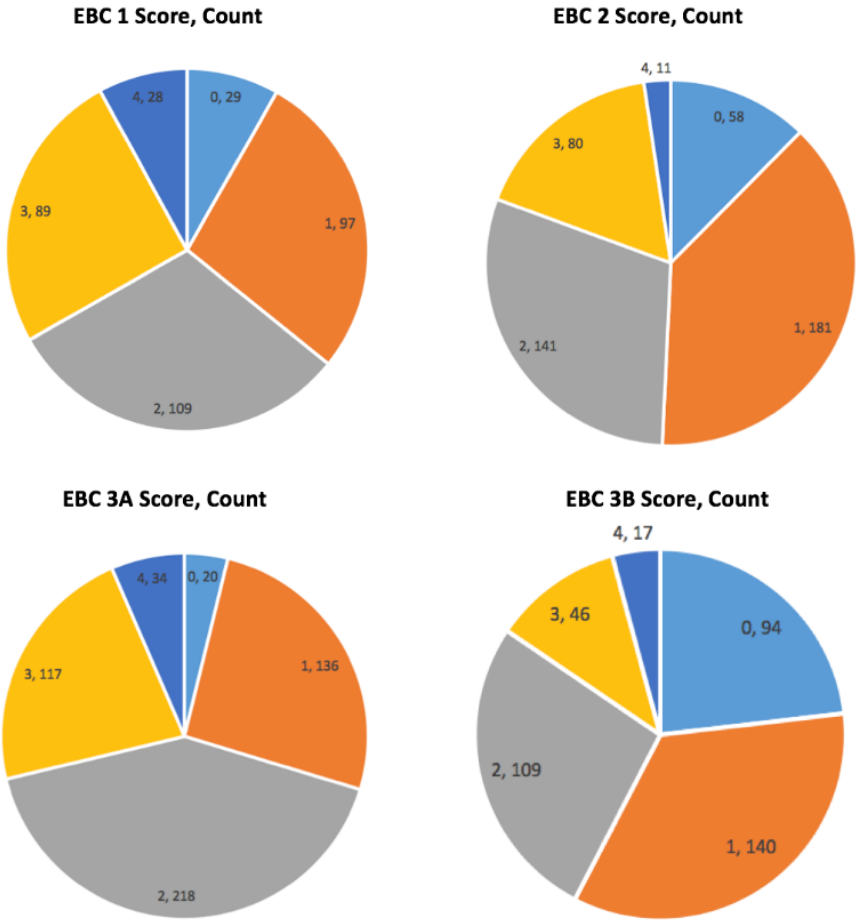

Figure 1: Each pie chart shows the distribution of the scores all student evaluations received for each category of EBC 1, EBC 2, EBC 3A, and EBC 3B. The notation used above shows the score received, and a count of the student evaluations that received that score. For example, (1, 97) represents 97 student evaluations receive a score of 1.

As mentioned previously, student evaluations are scored on four separate criteria. In this study, pie charts for EBC 1, EBC 2, EBC 3A, EBC 3B are created respectively to better visualize the manual perceived scores, which are shown in Figure 1. It is clear that most student evaluations are scored with a 1 or 2. It is noted that the student evaluations are scored based on human manual scoring of the provided scoring rubric. Also, when a student evaluation is scored as 0, this can either imply the student evaluation was written poorly or that no student evaluation is ever received.

Methodology of Text Mining Techniques

Text mining techniques are performed on the cleaned student evaluation data, by using both the statistical programming language of R and NVivo. The characteristics, including strengths and weakness of both software will be compared in detail below.

Challenges on Text Mining with R

In our text mining investigation, we begin analyzing student evaluations in the statistical computing software R. In order to perform text mining analysis, 21 required packages must first be installed in R. A directory is set up where all original .txt files are loaded into R to begin analysis. Next, the files are loaded from the directory as the source of the files making up the corpus. The function Corpus in R uploads all the files. To begin with, these files are named original documents so they can later be used for comparison. To prepare for text analysis, more pre-processing of the documents needs to be done. First, all numbers and punctuation are removed from the original documents. When numbers, punctuations, and stop words are removed, they are replaced by a white space where the word, number, or symbol have originally been in the corpus. In order to remove this white space, we use a command in R that strips any extra remaining white space. All text characters in the documents are converted to lowercase characters. Next, all English stop words are removed. English stop words are common words found in the English language. There exist 174 common stop words in the English language. Before moving forward to stemming and stem completion, it is important to check all student evaluations for spelling errors. This may seem trivial, yet it is essential in order to yield an accurate result. Correcting a spelling error in R requires a new line of code for each correction. To avoid this, all evaluations are manually checked for spelling errors and updated.

Table 1: Term Frequency Table

Least frequent terms									
1	2	3	4	5	6	7	8	9	10
3256	1127	682	444	334	244	195	172	138	114
Most frequent terms									
1690	1737	1743	1755	1956	2006	2216	2467	2625	3190
1	1	1	1	1	1	1	1	1	1

Note: This table provides a brief summary into the frequency distribution of terms appearing in the student evaluations for the least frequent and most frequent terms by R. For example, this table is interpreted as there ae 3,256 terms that only appear once in the evaluations. On the other extreme end, there is one term that appears 3,190 times.

Lastly, in the pre-processing phase, stemming and stem completion is done on all documents. Stemming is the process of reducing words to their base form. Sometimes a word is stemmed to a phrase that is not a base form itself and stem completion completes the phrase back to a base form. Stem completion uses a dictionary created by the original documents. For example, "argue", "argued", "argues", and "arguing" reduce to the stem "argu". Then, R refers to the dictionary

to stem complete "argu" back to a base form. At this stage, R tends to have difficulties with stemming and stem completion. To list a couple examples, "many" is stemmed into "maniac" and "really" is stemmed into "reallife." Outputting both the results after stemming and the stem completion into an Excel file allows us to compare with the original documents and find where mistakes are made.

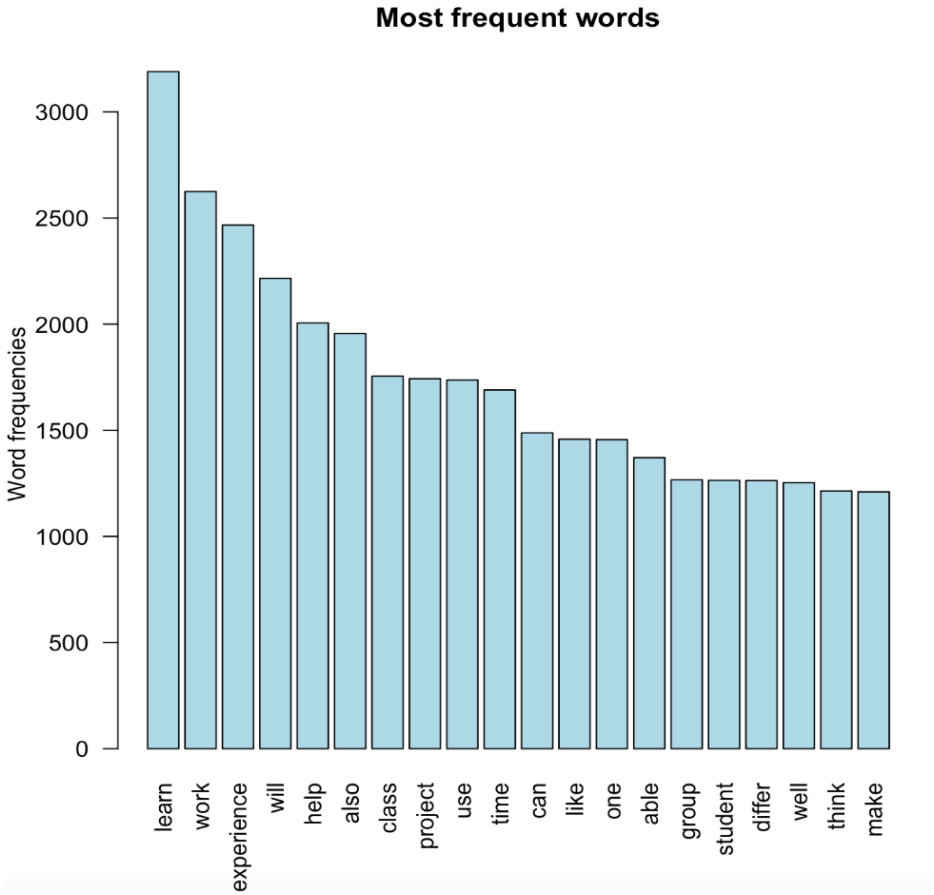

Figure 2: A bar graph of the 20 most frequent terms by R. This graph allows for a better visualization of the terms that are appearing most frequently throughout the evaluations.

A *document-term matrix (dtm)* is obtained, as a matrix with the 672 student evaluations as the rows and the terms found in the student evaluations as the columns. Each cell in the matrix is a frequency/count. Inspecting the dtm shows the distribution of the terms and the percentage of sparsity found in the matrix. To obtain the distribution of term frequencies, the dtm must be converted into a regular matrix and then the sum of columns is taken. Ordering the term frequencies allows a list to easily be created showing the least and most frequent terms, with a sample shown in Table 1, for easier interpretation. At first inspection of the dtm, it is revealed that the dtm contains 98% sparsity. Sparsity refers to infrequent terms occurring in the student evaluations. For example, in Table 1, there are 3,256 terms that only appear once in the student evaluations. R has a function to remove

a selection of sparse terms. After sparse terms are removed, the dtm now contains 37% sparsity, which is a huge improvement.

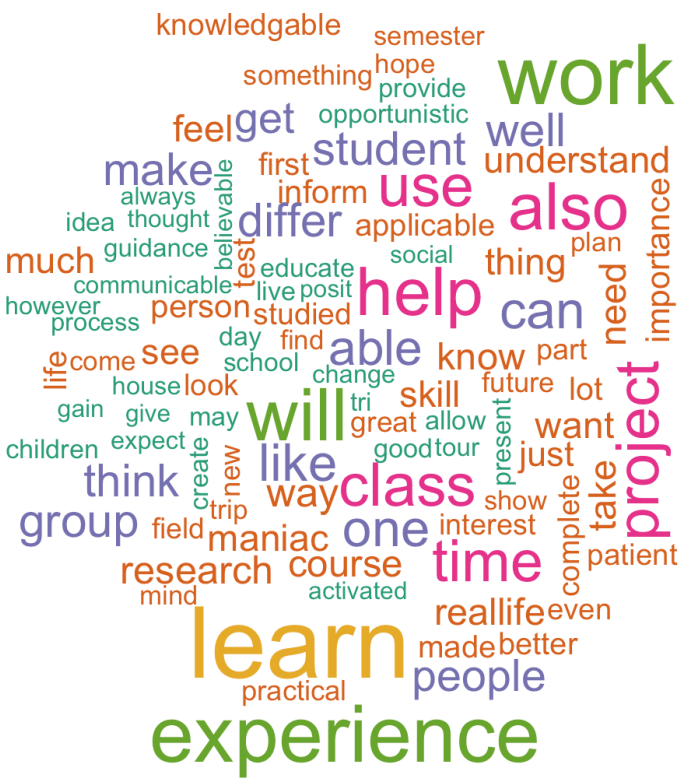

Figure 3: A visual representation of the word cloud produced by R.

After all the pre-processing is done and the dtm is created, we can analyze the data to get better visual representations. Figure 2 represents the 20 most frequent words found in all student evaluations. It appears that some of these words may be deemed insignificant for what we are interested in (e.g., *will*, *also,* and et al.). To better understand the significance of these terms, it is important to look at the context in which the terms are used. A better visualization of the most frequent terms is shown in the word cloud produced by R in Figure 3.

Alternative Text Mining with NVivo

As previously mentioned, R lacks an approach to efficiently looking at the context of a term and performed poorly in the pre-processing stage of stemming and stem completion. These fallbacks in R steer us away from the software, and introduce us to the qualitative analysis software called NVivo. NVivo has the ability to create a flowchart of a term over all the student evaluations. This allows a deeper look at the context of the term in question over that achieved by human manual interpretation, when categorizing a term as significant or insignificant. NVivo also has the ability to produce word clouds with a chosen number of significant terms faster and more efficiently than R. NVivo has the option to group together like terms (stemming and stem completion) by just simply clicking a button. This fixes

the mistakes caused by R, after running stemming and stem completion on all student evaluations.

Hereafter, NVivo is used as the primary software for all textual based analysis. The NVivo option of word count query is used to produce the word cloud shown in Figure 4, which includes the 100 most frequent words that appear in all student evaluations. In the word cloud, different words are depicted with different color and font size. The font size is directly related to the frequency of the 100 most frequent words found. NVivo also has a word count query that allows us to search for each of the most frequent words across all student reflections and provides a count of how many times these words appears in that reflection. This function was used to generate the document-term matrix for all data mining classification techniques conducted below.

Figure 4: A word cloud produced quickly and efficiently by NVivo that shows the 100 most frequent words found in all student evaluations.

As mentioned above, we use the "stemmed words" option to group together the like terms so no one term appears more than once in the word cloud. We see very similar results when comparing the word clouds produced by R and NVivo. Table 2 illustrates a count of the term and the terms that are grouped together under a given term to present a better idea on how NVivo performs in stemming and stem completion. It is important to note that the "experience" is included in the word cloud by R, whereas "experiments" is shown in the NVivo word cloud. Note that in table 2, "experience" is grouped together with the term "experiments". NVivo is able to quickly produce a flow chart of the context of the term used in all evaluations. However, it is a large flowchart that requires time to shift through. Hereafter, it is assumed that all most frequent terms are used in a positive and significant context.

Table 2: Word frequency provided by NVivo

Word	Count	Similar Words
Learn	3181	Learn, learned, learning, learns
Works	2641	Work, worked, working, workings, works
Experiments	2288	Experience, experiences, experiment, experimented, experimenting, experiments
Helps	2000	Help, helped, helpful, helping, helps

Note: NVivo software has a "stemmed words" option that groups together like terms when calculating word frequency.

R is able to produce a document-term matrix (dtm) which is a matrix including a count of the number of times a term appears in each of the student evaluations. NVivo has a similar function under its word search query option. This option allows the user to input the term and NVivo produces a list of all evaluations the term is located and a count of the term is located in that individual evaluation. A drawback of this option in NVivo is that NVivo does not include the evaluations where the term is not found in. As a result of this, a difficulty is created when generating a larger matrix that includes all evaluations as rows and the most frequent terms as columns. A count of the most frequent term is included in each cell. Another drawback of this NVivo option is that it only allows the user to search for one word at a time. Due to these drawbacks, the matrix had to be entered manually, which proves to be a time-consuming process. Once this document-term matrix is created for 100 most frequent terms and a matrix is created in Excel to lay out how many times these 100 most frequent words occur in each individual student evaluation, the data mining techniques are used to further analyze the student evaluations quantitatively.

Methodology of Data Mining Techniques

In this paper, after the document-term matrix (dtm) is obtained from the text mining techniques, we first consider four different classifiers to access the classification, including Random Forest, K-nearest neighbors (KNN), and Support Vector Machines (SVM) with Linear Kernel and Radial Kernel.
Suppose there are n observations: $(x_1, y_1),\ (x_2, y_2), \ldots, (x_n, y_n)$, where $x_i \in R^d$, and $y_i \in \{0, 1\}$ representing a score of Low or High. Random forest is a statistical classifier developed by Breiman (2001). Random forest builds a number of decorrelated decision trees, and then uses the mode of the predictions from the decision trees as the model output. Breiman (2001) suggests that as the number of the trees in the forest increases, the generalization error of random forest converges almost surely to a limit. Thus, the weak but unbiased decision trees produce relative efficient

predictions. In order to decorrelate the trees, a random sample of predictors is chosen from the full set of predictors at each split in a tree.

Let n be the number of data observations and let d be the number of predictors to be selected. Suppose the number of decision trees to be built is N_t, with minimum node size n_{node}. The algorithm for random forest for classification is as following:

(1) Draw a bootstrap sample of size n from the training observations.

(2) With the bootstrapped data, grow a tree by repeating the following steps:

i. Select m variables at random from the d predictors.

ii. Find the best variable among the m selected variables, as well as the best split point for classification.

iii. Split the node into two descendent nodes with each node resulting from the classification.

iv. Stop growing the tree when the minimum node size n_{node} is reached for all terminal nodes.

(3) Repeat steps (1) and (2) N_t times to obtain the a collection of trees $\{T_i\}_{i=1}^{N_t}$.

(4) For any input vector x, let $G_i(x)$ be the class prediction from the ith Random Forest tree. The prediction from the random forest is $G(x)$= mode of $\{G_i\}_{i=1}^{N_t}$.

The K-nearest neighbors classifier is memory-based. Given a query point, say x_0, assume we find K training points closest in distance to the given point x_0 among n observations, say (x_1^*, y_1^*), (x_2^*, y_2^*), (x_K^*, y_K^*), which satisfies that

$$\left|| x_1^* - x_0 \right|| \leq \left|| x_2^* - x_0 \right|| \leq \cdots \leq \left|| x_n^* - x_0 \right||,$$

where $||\cdot||$ represents the Euclidean distance. Let $H(x_0)$ be the class prediction for the query point x_0. Then $H(x_0)$ = mode of $\{ y_1^*, y_2^*, \dots, y_K^* \}$, by the majority vote of its K nearest training points. K can take any integer within the sample size. To determine the best K for our experiments, 5-fold Cross-Validation (CV) is applied to choose a K value in order to minimize the Cross-Validation prediction error: $min_K CVerror(K)$.

The technique of Support Vector Machines is considered as a method of classifying the data into the newly created High/Low variable. In the binary setting, suppose there are n observations: (x_1, y_1), $(x_2, y_2), \dots, (x_n, y_n)$, where $x_i \in R^d$, and $y_i \in \{-1, 1\}$. SVM aims to find a separable hyperplane that best separates the two classes and produces a lower error of classification. The optimal hyperplane is the hyperplane that passes the farthest from all training observations with a maximum margin separating hyperplane $w \cdot x + b = 0$ in the feature space through a quadratic programming:

$$min_{w,b} \frac{1}{2} ||w||^2 + C \sum_{i=1}^{n} \quad \xi_i, \text{ subject to } y_i(w \cdot x + b) \geq 1 - \xi_i \text{ and } \xi_i \geq 0, \ \forall\, i,$$

where $||\cdot||$ represents the l_2 vector norm, w is the normal vector to the hyperplane and the parameter $\frac{b}{\|w\|}$ determines the offset of the hyperplane from the origin. The constant $C > 0$ is a "cost" parameter which must be carefully tune for the "counts" of feature points $\sum_{i=1}^{n} \xi_i$ which lie within the margin or on the wrong side of the hyperplane. In the case that the data is linearly separable, we select two parallel hyperplanes that separate the two classes of data. When selecting these two parallel hyperplanes, we want the distance between them maximized.

Geometrically, the distance between the two parallel hyperplanes defined above is represented as $\frac{2}{\|w\|}$. We want to maximize the distance between the two parallel hyperplanes which is achieved by minimizing $\|w\|$.

To extend the method of SVM to cases in which the data is not linearly separable, we can consider a kernel function $\kappa(x, x')$:

$$\kappa(x, x') = \Phi(x) \cdot \Phi(x') = exp\,(-\gamma||x - x'||^2),\ \forall\, x, x' \in R^d,$$

where γ is a positive constant, Φ is a function to map the training examples into some feature space $\mathcal{F}$ such that $\Phi: R^d \mapsto \mathcal{F}$.

Furthermore, 5-fold Cross-Validation is considered to evaluate the performance of these four different classifiers. In 5-fold Cross-Validation, the dataset is randomly divided into five folds with approximately equal size. Then one fold is held out and treated as a validation set, while the remaining four folds are treated as a training set to build a classification system. This procedure is repeated five times, with a different fold of observations treated as a validation set until all folds have been used as a test dataset.

Ensemble Learning to Improve Classification Performance

To further improve the overall performance, ensemble learning by fusing multiple predictive decisions to make a final decision could be a potential way to get a more robust decision (Polikar, 2006; Moreno-Seco et al, 2006). For example, the classifier ensembles with different combination techniques have been widely explored in recent years. These methods have been shown to potentially reduce the error rate in the classification tasks compared to an individual classifier in a broad range of applications. In the decision fusion with ensemble-based systems, it is important to consider the diversity of decisions to be fused, with respect to diverse classifiers. In our analysis, we consider fusing independent classifiers among Random Forest, K-nearest neighbors, and Support Vector Machines with radial kernel.

For the ith observation x_i, let $G(x_i)$, $H(x_i)$, $J(x_i)$ be the class predictions from Random Forest, K-nearest neighbors, and Support Vector Machines respectively. Then the final class predictions for Ensemble Learning is given by $F(x_i)$ = mode of $\{G(x_i),\ H(x_i),\ J(x_i)\}$.

Results of Data Mining Techniques

Using the document-term matrix (dtm), data mining techniques can now be applied to classify these student evaluations into two categories of High or Low. All data mining techniques are performed in R. In order to achieve that, first, a new response variable of High and Low is created for EBC1, based on both the distribution of scores shown in the pie charts in Figure 1 and the criteria of the applied learning scoring rubric shown in Appendix B. Then repeat this procedure for the rest of EBC 2, EBC 3A, and EBC 3B, creating four new response variables. With these factors in mind, all student evaluations that received a score of 2 or below

are classified in the Low class and student evaluations receiving a score of a 3 or 4 are classified as High. The aforementioned document-term matrix (dtm) is then merged with the student evaluation corresponding EBC 1, EBC 2, EBC 3A, and EBC 3B scores with four response variables of High or Low.

Random Forest, K-nearest neighbors, and Support Vector Machines (with either linear or radial kernel) are all considered as the classification techniques using the 5-fold Cross-Validation to analyze the free-style text of student evaluations. Each classification method is run on all EBC category of High or Low that a student evaluation receives (EBC 1, EBC 2, EBC 3A, and EBC 3B) respectively. The overall accuracies are shown in Table 3. Comparing to the EBC2, EBC3A and EBC3B, the overall prediction accuracy for student reflection EBC 1 scores indicates the lowest accuracy of around 65-68%. On the other hand, the EBC 2 student evaluation scores have a stronger accuracy of 78-82%. EBC 3A scores hold around a 70-73% overall prediction accuracy, and EBC 3B scores have the highest overall prediction accuracy of 83-85%. A graphical visualization of the overall accuracies produced from each method of classification is shown in Figure 5. From Figure 5, it is interesting that the classification results from EBC1 illustrate outliers consistently for all four classification methods applied. The possible reason why the overall accuracies for EBC 1 is lower than EBC 2, EBC 3A and EBC 3b is that students' expectations, the purpose, and/or goals of the experience in terms of personal educational development can be at a larger range of terms used and/or less associated to the terms from the document-term matrix.

Table 3: Prediction rates table.

	EBC 1	EBC 2	EBC 3A	EBC 3B
Random Forest	0.671	**0.813**	0.723	0.840
K-nearest neighbors	**0.674**	0.809	**0.728**	**0.845**
SVM with Linear Kernel	0.657	0.784	0.701	0.838
SVM with Radial Kernel	0.668	0.807	0.707	0.845
Ensemble Learning	**0.685**	**0.817**	0.707	**0.856**

Note: 5-fold Cross-Validation is run on the four models of Random Forest, K-nearest neighbors, and Support Vector Machines with either Linear or Radial Kernel. Each classification method is run four different times using each of the EBC scores among EBC1, EBC2, EBC2A, and EBC3B) as the response variable. The overall accuracies from the five folds is shown in the table. Ensemble Learning accuracies after the method of decision fusion is used to combine the classifier methods of KNN, Random Forest, and SVM with a radial kernel. Note: For KNN, a different K (the number of neighbors) is chosen each time after running 5-fold Cross-Validation to determine the best K.

Decision fusion is further considered as a method aiming to improve the classification performance. Random Forest, K-nearest neighbors, and Support Vector Machines with radial kernel are used in the decision fusion approach. It is important in decision fusion that all methods are independent of one another and an odd number of methods are used so that there are no ties created. It is revealed

that the majority of the misclassified observations in the data are the observations with a ground truth of High that are misclassified as Low. Decision fusion is once again run on all EBC scores. The accuracies are illustrated in Table 4. Overall prediction accuracies are improved slightly. It is shown that the decision fusion approach results in higher accuracy than any of individual classifiers for EBC1, EBC2, and EBC3B. For EBC3A, even though the decision fusion approach does not lead to the highest accuracy, the accuracy is still competitive comparing to the individual classifiers. These results indicate that decision fusion with ensemble is effective in this text mining task.

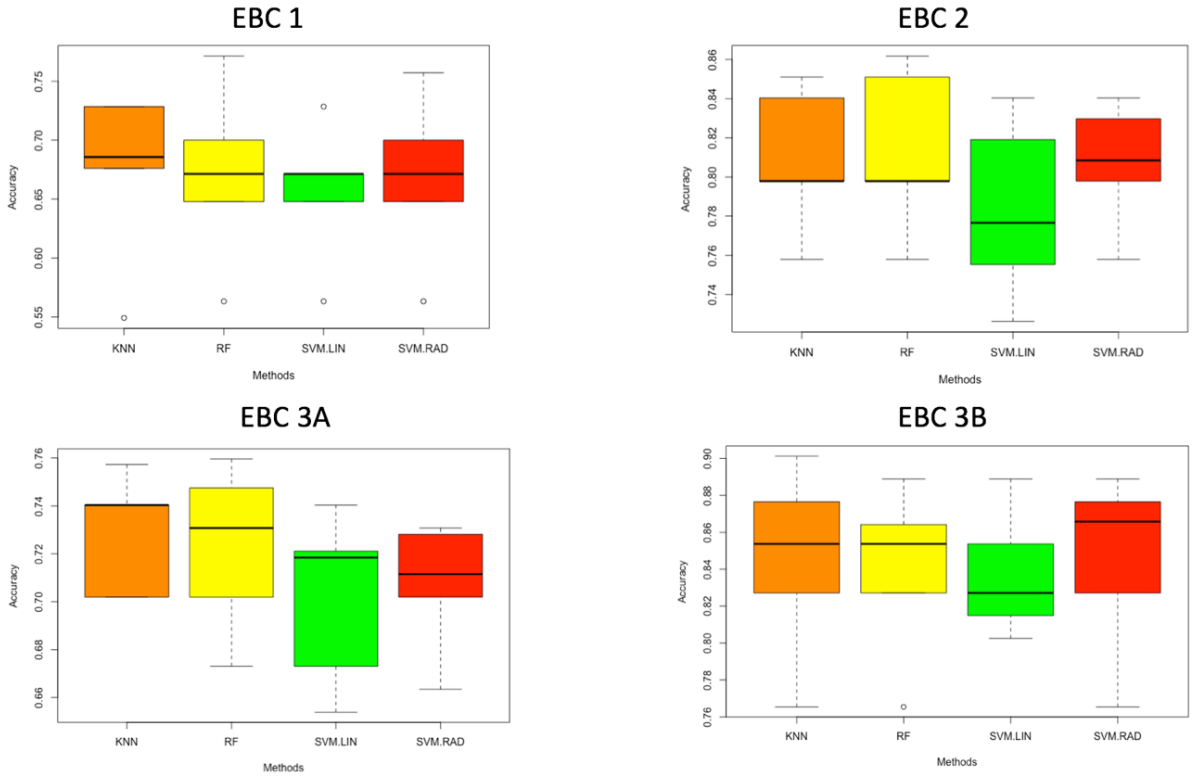

Figure 5: Boxplots to the overall accuracies for the four classification methods of Random Forest, K-nearest neighbors (KNN), and Support Vector Machines with Linear and Radial Kernel using EBC1, EBC2, EBC3A, and EBC3B as response variables.

Conclusion and Recommendations

Our dataset from two academic years over fall 2013-spring 2015 is studied systematically to provide the preliminary analysis results. The results of our experiments show that text mining is a promising technique to analyze the open-ended free-style text based student reflections quantitatively, and automatically. Text mining can be an effective way to analyze text responses and how a student evaluation will score quantitatively which reveals how well a course and/or instructor is performing. Analyzing these text based student evaluations into

quantitative information allows one to gain additional insights to evaluate student performance, instructor performance, and course performance. One can also gain a deeper understanding of individual schools at the university, departments, and majors as well, and eventually evaluate the impact on the implemented instructional practice and the student learning outcomes.

Data mining classification methods show promising overall prediction accuracies for all EBC scores of student evaluations. Decision fusion is a method implemented to further improve classification accuracies, and while it does so, no strong change is made in the overall prediction accuracy of student reflections that are classified or misclassified by the three classification methods used. While accuracies held steady after decision fusion, the method does allow for a deeper understanding of the data being analyzed. Decision fusion reveals the individual student evaluations that are misclassified which can reveal what departments or majors have the most incorrect classifications and the performance or motivation of the students in those departments or majors on evaluations. Providing faculty and administrators with this information to be able to interpret results more critically and be able to make rational and fair decisions in terms of teaching effectiveness (Hou,Lee, & Gunzenhauser, 2017).

Analyzing student evaluations by terms is a significant way to analyze the applied learning program as a whole, as well as the effectiveness of the applied learning program on overall student learning. Analyzing text based student evaluations provide additional insights. For example, a higher EBC score is associated with greater student performance and/or understanding of the course. Then we associate that these motivated students will provide a more in-depth and meaningful evaluation. On the other side, this well-trained system of text mining and data mining can be applied to the future applied learning student evaluations. In this case, the new student evaluations will be pre-processed in the same way of text mining as described previously and fed into the data mining system. Consequently, the scores of EBC1, EBC2, EBC3A, and EBC3B will be produced automatically. This framework can be an efficient way to provide quick preliminary analysis on the program evaluation of instructional practice and student learning.

Abd-Elrahman et al. (2010) support the claim of automatic text mining techniques as a good method to investigate open-ended student course evaluation. This study extends the original two categories of negative and positive comments made regarding the course for each evaluation into four categories of benchmark, milestone-I, milestone-II, and capstone. Data Mining techniques are incorporated in our study for quantitative analysis. The promising overall prediction accuracies demonstrate that such automatic quantitative analysis of student evaluations can be an effective approach to applied learning assessment.

Hou, Lee and Gunzenhauswer (2017) support the claim of these evaluations as instruments that can support transformative decisions in improving quality of teaching. By valuing the contributions of students and faculty in this process could help in preventing erroneous decisions based on some biased student feedback.

References

Abbott, R.D., Wulff, D.H., Nyquist, J.D., Ropp, V.A. & Hess, C.W. (1990). Satisfaction with processes of collecting student opinions about instruction: The student perspective. *Journal of Educational Psychology*, *82,* 201-206.

Abd-Elrahman, A., Andreu, M., & Abbott, T. (2010). Using text data mining techniques for understanding free-style question answers in course evaluation forms. *Research in Higher Education Journal*, *9*, 11–21.

Beleche, T., Fairris, D., & Marks, M. (2012). Do course evaluations truly reflect student learning? Evidence from an objectively graded post-test. *Economics of Education Review*. *31*, 709-19.

Boring, A. (2017). Gender biases in student evaluations of teaching. *Journal of Public Economics.* 145. 27-41.

Chen, Y., & Hoshower, L. B. (2003). Student evaluation of teaching effectiveness: An assessment of student perception and motivation. *Assessment & Evaluation in Higher Education*, *28*(1), 71-88.

Cohen, K. B., & Hunter, L. (2008). Getting started in text mining. *PLoS Computational Biology*, *4*(1), e20. doi:10. 1371/journal.pcbi.0040020

Cronbach, L. J. (1963). Course improvement through evaluation. *Teachers College Record*, *64*, 672-683.

Fike, D. S., Fike, R., & Zhang, S. (2015). Teacher qualities valued by students: A pilot validation of the teacher qualities (T-Q) instrument. *Academy of Educational Leadership Journal, 19*(3), 115-125.

Gal, Y., & Gal, A. (2014). Knowledge bias: Is there a link between students' feedback and the grades they expect to get from the lecturers they have evaluated? A case study of Israeli colleges. *Journal of the Knowledge Economy, 5*(3), 597-615. doi:10.1007/s13132-014-0188-5

Galbraith, C. S., Merrill, G. B., & Kline, D. M. (2012). Are student evaluations of teaching effectiveness valid for measuring student learning outcomes in business related classes? A neural network and bayesian analyses. *Research in Higher Education, 53*(3), 353-374. doi:http://dx.doi.org.liblink.uncw.edu/10.1007/s11162-011-9229-0

Delgado, M., Matrín-Bautista, M.J., Sánchez, D., & Vila, M.A. (2002, September). Mining text data: special features and patterns. *Proceedings of EPS Exploratory Workshop on Pattern Detection and Discovery in Data Mining*, London.

Douglas, P.D. & Carroll, S.R. (1987). Faculty Evaluations: Are college students influences by differential purposes? *College Student Journal*, *21*(4).

Evans, C. (2013). Making sense of assessment feedback in. higher education. *Review of Educational Research. 83*(1), 70-120. doi: 10.3102/0034654312474350

Hastie, T., Tibshirani, R. & Friedman, J. (2016). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* New York: Springer.

Heller, H.W., & Clay, R.J. (1993). Predictors of teaching effectiveness: The efficacy of various standards to predict the success of graduates from a teacher education program. *ERS Spectrum, 11*, 7-11.

Hou, Y., Lee, C., & Gunzenhauser, M.G. (2017). Student evaluation of teaching as a

disciplinary mechanism: A Foucauldian analysis. *The Review of Higher Education, 40*(3), 325-352.

Marsh, H.W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and utility. *Journal of Educational Psychology*, *76*(5), 707-754.

Marsh, H.W. (1987). Students' evaluation of university teaching: Research findings, methodological issues and directions for future research. *International Journal of Educational Research*, *11*(2), 253-388.

Moreno-Seco, F., Inesta, J. M., de León, P. J. P., and Micó, L. (2006). Comparison of classifier fusion methods for classification in pattern recognition tasks. In *Structural, Syntactic, and Statistical Pattern Recognition* (pp. 705-713). Berlin, Germany: Springer-Verlag.

Polikar, R. (2006). Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine, 6*(3), 21-45.

Ramsden, P. (2003). *Learning to teach in higher education* (2nd ed.) London: Routledge Falmer.

Scriven, M. (1967). The methodology of evaluation. In R.W. Tyler, R. M. Gagné, & M. Scriven (Eds.), *Perspectives of curriculum evaluation*. Chicago, IL: Rand McNally.

Stake, R. E. (n.d.). The Coutenance of Educational Evaluation. Center for Instructional Research and Curriculum Evaluation, University of Illinois.

Tan, A. (1999). Text mining: The state of the art and the challenges. In *Proceedings, PAKDD '99 Workshop on Knowledge Discovery from Advanced Databases* (KDAD '99).

Uttl, B., White, C., & Gonzalez, D. (2017). Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation*, *54*, 22-42.

Wagner, N., Rieger, M., & Voorvelt, K. (2016). Gender, ethnicity and teaching evaluations: Evidence from mixed teaching teams. *Economics of Education Review, 54*, 79-94.

Zhao, J. & Gallant, D. (2012). Student evaluation of instruction in higher education: Exploring issues of validity and reliability. *Assessment & Evaluation in Higher Education*, *37*(2), 227-235.

Appendix A

Here is a sample of guidance for both the initial reflection and the final reflection for ETEAL supported pedagogy initiatives: (Note: SLO represents Student Learning Outcome)

Intention reflection prompts *(at the start of the semester)*:
Explains in depth the purpose for engaging in the experience and directly links it to personal educational development through expected educational outcomes. Your intention reflection will be typed in 1 page, by answering the following questions.
(SLO1) a. Articulate your expectation from, and the reason for participation in this project.
(SLO1) b. Examine and explain what you hope to gain from this experience in terms of personal, educational, and/or career goals.
(SLO1) c. Explain what statistical methods, presentation and communication skills, and use of technology you hope to learn from this project.
(SLO1) d. Explain the impact (on others or on the field) that you hope to make through this project.

Final reflection prompts *(upon completion of the course)*:
(SLO2) Summarize the relevant theories, ideas and skills you were able to apply in this project.
(SLO2) Demonstrate how you apply what you learnt from other courses to complete this project.
(SLO3) Summarize your team work and/or leadership experience through this project.
(SLO3) Over the several presentation occasions, explain how you address questions from people of different fields, and lessons you have learn to improve your oral presentation and communication skills.
(SLO3) Summarize the significance of your work in the field from this project.
(SLO3) Summarize a personal challenge and how you overcome it during this project.

Appendix B

Here is a sample of the scoring rubric for human manual scoring (Revised October 2014):

Student Work Product Number _____ Scorer_____

Applied Learning Critical Reflection Scoring Rubric

Evaluators are encouraged to assign a zero to any work sample or collection of work that does not meet benchmark (cell one) level performance.

	Benchmark 1	Milestone 2	Milestone 3	Capstone 4	Score
Intention [EBC 1]	Identifies a purpose for engaging in the experience without discussing personal educational development.	Identifies the purpose for engaging in the experience and mentions personal educational development, but does not link these.	Explains the purpose for engaging in the experience, and discusses the link(s) to personal educational development.	Explains in depth the purpose for engaging in the experience and directly links it to personal educational development through expected educational outcomes.	
Application of Knowledge[1] [EBC 2]	Makes vague references to knowledge drawn from previous or current coursework, but does not demonstrate how it was used in the applied learning experience.	Refers to knowledge drawn from previous or current coursework and provides some insight into how it was useful in the applied learning experience.	Connects previous or current coursework and provides concrete evidence of how it affected performance the applied learning experience.	Connects and extends previous or current coursework and synthesizes it in an innovative way within the applied learning experience.	
Reflection [EBC 3a]	Describes own performance in general or abstract terms, without indicating impact or significance on personal educational development.	Identifies at least one strength and/or challenge highlighted by the experience, and indicates a somewhat broader perspective about personal educational development.	Evaluates strengths and challenges encountered in the experience, and reveals broader perspectives about personal educational development.	Envisions a future direction for growth and/or application of strengths, and reveals significantly broader perspectives about personal educational development.	
Evaluation of Impact [EBC 3b]	Provides a vague or oversimplified statement of the impact of the experience on others or on the profession/field.	Provides a specific example of the impact of the experience on others or on the profession/field.	Discusses the results of the experience, providing concrete examination of its impact on others or on the profession/field.	Considers the results of the experience with a thoughtful evaluation of its impact on others or on the profession/field.	

[1]Modified from Transfer dimension of Foundations for Lifelong Learning VALUE Rubric

EBC 1. The student will articulate their expectations, the purpose, and/or the goals of the experience in terms of their personal educational development. [Thoughtful Expression]
EBC 2. The student will synthesize knowledge drawn from their coursework to address the issues/challenges/questions involved in the experience. [Critical Thinking, Foundational Knowledge, Inquiry]
EBC 3. The student will communicate the impact or significance on their personal educational development *and* on others in the profession or in the field at the *conclusion* of the experience. [Critical Thinking]
NOTES: