Extraction of Students’ Feedback in University Services using Logistic Regression and Chi-Square Automatic Interaction Detection (CHAID) Algorithm

Maryli F. Rosas 1, Shaneth C. Ambat 2, Ace C. Lagman 3

1 Computer Studies Department, De La Salle University – Dasmariñas, City of Dasmariñas, Philippines.
2 Graduate School, FEU Institute of Technology, Manila, Philippines.
3 Information Technology, FEU Institute of Technology, Manila, Philippines.
Email: mfrosas@dlsud.edu.ph, shaneth_ambat@yahoo.com, aclagman@feutech.edu.ph

Abstract - The ultimate goal of every Higher Education Institution is to provide quality education to their students. Quality of Education can be measured through the student’s satisfaction level based on their overall university experience. Students are the primary consumers in every Higher Education Institution services. Hence, their feedback or insights on their overall university experiences are significant for the university’s continuous quality improvement. The use of machine learning algorithm concepts can determine and extract useful information which can be used as reference in formulating necessary policies suited to improve university’s academic services. Our study focuses on the application of logistic regression through equation development and CHAID Algorithm through rules development. The study also uses text analytics concepts in which sentiment scores are computed to classify the comments according to satisfaction level.

Key words - Data Mining, CHAID, regression, sentiment, text analytics, University Services

A shorter version of this paper with title ‘Data Mining of Students’ Response in the University Services using Chi-Square Automatic Interaction Detection (CHAID) was presented at 1st IEEE International Conference on Knowledge Innovation and Invention (ICKII), July 23 to July 27, 2018, Jeju Island Korea[2].

I. INTRODUCTION

Higher Education targets to develop complex theoretical, abstract and analytical reasoning capabilities in the alumni [3]. Quality education can be determined thru the quality of services that were given to the students. It is every organization’s effort to continuously improve their quality of services. In order to measure students’ reactions, evaluation forms can be used by educational institutions [4]. In the university setting, it is very vital for them to know their performance through gaining insights from the students in a form of an exit interview.

Data mining in education enables data-driven decision-making for improving current educational practice and learning materials [5]. Another paper justified the capabilities of data mining techniques in context of higher education by offering a data mining model for higher education system [6].

The general objective of this study is to develop and implement an improvement plan based on the significant patterns produced by the CHAID algorithm to improve the quality of services of the university.

This study could be a significant endeavor in monitoring and escalating the effectiveness and efficiency of the units contributing to the over-all quality of student services.

The study investigated on how students’ experiences affect their satisfaction level by using data mining techniques particularly the implementation of regression and decision tree analysis with Chi-square Automatic Interaction Detector (CHAID) algorithm. Moreover, this study implemented Sentiment Classification Technique using supervised Machine Learning approach [7] that will correlate to CHAID algorithm.

This paper would like to address the following research questions:
1. How logistic regression determines significant attributes in students satisfaction towards academic services?
2. How to extract predictive data models using Decision Tree Algorithm under Chi-square Automatic Interaction Detector (CHAID method?)
3. How acceptable the software developed using software evaluation instrument as perceived by IT Expert in terms of the following criteria
   a. Functionality
   b. Usability
   c. Reliability
   d. Performance
   e. Security
   It is worth noting that on the study of student satisfaction is not yet fully utilized. And the implication of the result of such would help the Higher Education Institution to improve on their quality of education. Hence, the researcher is motivated to explore and discover more of this area.

The researchers used Knowledge Discovery in Database in which machine learning algorithm was applied to generate model. In this study, the researcher used logistic regression analysis to determine the significant attributes. Through executing Logistic Regression analysis model, the Sentiment,
Comment, College and Area are the significant attributes that contribute to the satisfaction level of the student.

The researchers further use of text analytics to identify and extract useful information and patterns. The comments analyzed according to its sentiments. All comments that have less than 1 are classified as negative. And all of these negative comments were subject for the creation of a knowledge base of recommendations. In addition, applying Sentiment analysis, the comments were extracted to generate key phrases. These key phrases were also used in the creation of the knowledge base of the recommendations.

The developed Online Exit Interview application applies the CHAID algorithm and text analytics. The Data mining report of the online Exit Interview Application shows that most of the students were satisfied with the faculty area which validates the result of the mined 2013 to 2016 data shown in the histogram. Most of the students commented “excellent professors”. In addition, in the heat-map result, the student services should likely be improved.

Further, Software Development Model was also used in the study carefully following the four phases of Prototyping which will be later discussed in the research methodology.

II. LITERATURE REVIEW

According to Dejaeger, K. et.al, in order to measure students' reactions, evaluation forms can be used by educational institutions. The results of the evaluation forms can be used for internal and external communication [4]. He also furthered that it is important to identify which aspects for the most part need to be managed to maximize the students' overall satisfaction with the course.

As the main objective of data mining in education is largely to improve learning, measurements are more difficult to obtain, and must be estimated through improved performance. In general, data mining in education enables data-driven decision-making for improving current educational practice and learning materials. However, there are many more specific objectives in EDM depending on the viewpoint of the final user and the problem to resolve as cited by Romero and Ventura.

Data mining is a heart of knowledge discovery process. Using data mining, according to him, people can find useful patterns from large volumes of data and interpret them for useful knowledge and information [8]

The purpose of one scholarly work is to analyze three separate constructs (demographics, study habits, and technology familiarity) that can be used to identify university students’ characteristics and the relationship between each of these constructs with student achievement. The research survey was used to collect data, and the data were analyzed using the chi-squared automatic interaction detection (CHAID) algorithm. That is why in their study, they preferred to use the CHAID algorithm to obtain a decision tree that indicated the most significant breaks in the prediction of academic achievement of university students, and accordingly, the aimed to reveal the characteristics of successful and unsuccessful students who are pursuing their higher education. Further, they used logistic regression to determine whether the results of the CHAID algorithm were compatible with the results of the logistic regression.[9]

Another paper used CHAID Decision Trees to Formulate Pathways for the Early Detection of Metabolic Syndrome in Young Adults.[10]

Diaz-Pérez and Bethencourt, on the other hand examined the tourists’ expenditure pattern in Melaka City as their study. To model interaction of domestic and international tourists in five tourism sectors: accommodation, food and beverages, transportation, entertainment and shopping Chi-square Automatic Interaction Detection (CHAID) was used. Hence, the conducted CHAID analysis has clearly illustrated the expenditure pattern of domestic and international tourists in Melaka [11].

An ensemble algorithm of data mining decision tree (DT)-based Chi-square Automatic Interaction Detection (CHAID) is widely used for prediction analysis in variety of applications.[12]

This review of related literature shows several scholarly works on the use of data mining in various areas like marketing, medicine, bioinformatics, geosciences and education. Most of the literature used decision trees, logistic regression and CHAID. As mentioned by one author, to yield a better data mining result, a study should use a combination of different algorithms. With such motivation, the researcher had used the three algorithms mentioned above.

III. RESEARCH METHODOLOGY

In this section, the researcher used two main procedures (i) Knowledge Discovery in Database concept (ii) Software development model.

The researchers used Knowledge Discovery in Database (KDD), which consist of nine (9) procedures to extract data model using decision tree algorithm. One major procedure in the KDD process is the Data Mining process. (In this stage, it entails the application of machine learning algorithm to generate a data model. A data model can be used for predictive and profiling data sets. To determine the predictive model, decision tree algorithm with Chi-square Automatic Interaction Detector (CHAID) method. CHAID is a tool used to discover the relationship between variable.

A. Chi-square Automatic Interaction Detector (CHAID) Algorithm

Following the steps in Chi-square Automatic Interaction Detector (CHAID) Algorithm discussed by Koehler,K.J., yield to this resulting tables and tree.

Data comes from the sentiments of the seven(7) colleges: College of Science and Computer Studies (CSCS), College of Liberal Arts and Communication (CLAC), College of Business Administration and Accountancy (CBAA), College
of Tourism and Hospitality Management (CTHM), College of Criminal Justice Education (CCJE), College of Education (CoED), and College of Engineering Architecture and Technology (CEAT).

Table 1 shows the result of following Steps 1, cross tabulate the response variable with each of the explanatory variables.

<table>
<thead>
<tr>
<th>TABLE I. COMPUTED PEARSON X² TEST</th>
</tr>
</thead>
<tbody>
<tr>
<td>Column</td>
</tr>
<tr>
<td>CBAA</td>
</tr>
<tr>
<td>CCJE</td>
</tr>
<tr>
<td>CEAT</td>
</tr>
<tr>
<td>CLAC</td>
</tr>
<tr>
<td>COED</td>
</tr>
<tr>
<td>CSCS</td>
</tr>
<tr>
<td>CTHM</td>
</tr>
</tbody>
</table>

After the response variable has been cross-tabulate with each of the explanatory variables, compute Pearson x² tests for independence for each allowable sub-table to achieve Step 2.

As a rule, if x² value is not significant, combine the column categories (and repeat step 2 if the new table has more than two columns).

Followed by Step 3, if x² is significant, implement split and return to step 2. And Step 4 simply finds the most significant of these optimally merged explanatory variables and ends with step 5 using the most significant variable in step 4 to split the node with respect to the merged categories for that variable.[13] After satisfying the steps, Columns 6 and 7 will be merged as shown in Table II.

<table>
<thead>
<tr>
<th>TABLE II. MERGED INSIGNIFICANT COLUMNS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Column</td>
</tr>
<tr>
<td>CBAA</td>
</tr>
<tr>
<td>CCJE</td>
</tr>
<tr>
<td>CEAT</td>
</tr>
<tr>
<td>CLAC</td>
</tr>
<tr>
<td>COED</td>
</tr>
<tr>
<td>CSCS</td>
</tr>
<tr>
<td>CTHM</td>
</tr>
</tbody>
</table>

Both CBAA and CLAC has same values, therefore, we can merge the two rows shown in Table III.

<table>
<thead>
<tr>
<th>TABLE III. MERGED ROWS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Column</td>
</tr>
<tr>
<td>CBAA, CLAC</td>
</tr>
<tr>
<td>CCJE</td>
</tr>
<tr>
<td>CEAT</td>
</tr>
<tr>
<td>COED</td>
</tr>
<tr>
<td>CSCS</td>
</tr>
<tr>
<td>CTHM</td>
</tr>
</tbody>
</table>

Table IV shows the final merged columns and rows. Values that are below 0.05 are significant while values over 0.05 are not significant. Thus, comments of CEAT students has significant value in terms of Administrators. While all Colleges has significant comments over Faculty area. Lastly, students from CBAA, CLAC, CEAT and COED has significant sentiments in terms of Student Services.

<table>
<thead>
<tr>
<th>TABLE IV. MERGED COLUMNS by AREA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Column</td>
</tr>
<tr>
<td>CBAA, CLAC</td>
</tr>
<tr>
<td>CCJE</td>
</tr>
<tr>
<td>CEAT</td>
</tr>
<tr>
<td>COED</td>
</tr>
<tr>
<td>CSCS</td>
</tr>
</tbody>
</table>

where: C1 = Administrator, C2 = Facilities, C3 = Faculty, C4 = Quality of Education; and C5 = Student Services.

B. Software Development Model

The Prototyping Model is a systems development method (SDM) in which a prototype (an early approximation of a final system or product) is built, tested, and then reworked as necessary until an acceptable prototype is finally achieved.
from which the complete system or product can now be developed.

The researcher chose this kind of software model due to its benefits such as improved system usability, improved design quality, improved maintainability, reduced development effort and a closer match to users’ real needs as cited in the book of Sommerville [14].

The prototyping model involves four phases as shown in Fig. 2 and discussed further below:

1. Prototyping Plan: The researcher reviewed the existing manual procedure and other core processes involved in the different units that have significant role in the improvement of the quality of education.

   The researcher gathered reports and other forms needed in the system. Interviews with the associate deans, deans, and other academic and student service units were conducted. The gathered data were reviewed to be able to create a plan for the development of the online system for the use of the essential units. An entity-relationship diagram was constructed to aid the researcher in analyzing the database.

2. Outline Definition: In this phase, the data gathered were analyzed based on the result of evaluation of the existing process. The researcher designed an interface that is user friendly. A centralized database was designed so as to lessen data duplication and inconsistencies.

3. Executable Prototype: The actual program coding was done in this phase. Using Visual C# as programming language, Microsoft .net for framework as the front end and MYSQL for the backend.

4. Evaluation Report: The proposed system was presented to the end users for evaluation. Clients such as the associate deans, deans and other academic and student service offices’ feedback were taken in to consideration. Their feedback is important in the development of an effective system.

   Based on the customer’s feedback, the prototype was polished. Suggestions and comments served as a basis for refining the prototype.

C. Conceptual Framework

Fig. 3 shows the Conceptual Framework of this study. Graduating students across the seven (7) colleges such as from the College of Science and Computer Studies (CSCS), College of Liberal Arts and Communication (CLAC), College of Business Administration and Accountancy (CBAA), College of Tourism and Hospitality Management (CTHM), College of Criminal Justice Education (CCJE), College of Education (CoED), and College of Engineering Architecture and Technology (CEAT) have to undergo an exit interview.

The exit interview contains students’ insights, comments, suggestions and their satisfaction level on different areas and facilities of the university. The acquired data coming from the colleges’ exit interview were analyzed using appropriate data mining techniques. In this study, classification technique using decision tree was used. Applying classification technique undergoes four different steps such as data cleansing, data transformation, pattern discovery and interpretation of results. Hence, the gathered data first underwent data cleansing so that missing values, and inconsistencies were handled. After which data transformation was implemented to derive a standard and usable format. Once a standard format of data is derived, it can now be used to find out patterns. In discovering patterns, different data mining algorithms may be applied.

The researcher used the most common data mining algorithms such as the decision tree and regression algorithm. Decision tree proved its superiority over other techniques as they have noticed from most of the literature reviews [15]. After applying all the algorithms, the results were interpreted and be used as inputs for the desired improvement plan of the five key areas such as Administrator, Facilities, Faculty, Quality of Education and Student Services. As we all know, educational processes entail continuous quality improvement in a lot of interrelated areas. Quality of Education would concentrate on the course on the knowledge and skills gained of the students. Facilities, on the other hand, would concentrate on the satisfaction of students in terms of the
utilization of the physical infrastructures such as the library, laboratories (computer, science, architecture, and all the specialized laboratories), gym, canteens, parking space that would contribute to the learning of the students during their stay. Moreover, the services experienced by the students from different servicing units like registrar, accounting, student wellness, security, admission, and other related offices that in a way add up to the success of the students.

TABLE V. DETAILED ACCURACY BY CLASS

<table>
<thead>
<tr>
<th>Class</th>
<th>0</th>
<th>1</th>
<th>Weighted Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP Rate</td>
<td>0.995</td>
<td>0.985</td>
<td>0.992</td>
</tr>
<tr>
<td>FP Rate</td>
<td>0.011</td>
<td>0.005</td>
<td>0.008</td>
</tr>
<tr>
<td>Precision</td>
<td>0.989</td>
<td>0.995</td>
<td>0.992</td>
</tr>
<tr>
<td>Recall</td>
<td>0.995</td>
<td>0.989</td>
<td>0.982</td>
</tr>
<tr>
<td>F-Measure</td>
<td>0.992</td>
<td>0.992</td>
<td>0.992</td>
</tr>
<tr>
<td>MCC</td>
<td>0.984</td>
<td>0.984</td>
<td>0.984</td>
</tr>
<tr>
<td>ROC Area</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>PRC Area</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Teachers, on the other hand, may use the result to focus on what is really needed to improve their teaching methodology. Administrators, on a broader perspective, may use the 'mined' data to redesign or enhance their policies on areas that need more attention.

With the use of the data mining, it helped the university identify the present and future requirements of students and their preferences to enhance their satisfaction levels.

IV. RESULTS AND DISCUSSION

The researcher used logistic regression analysis to determine the significant attributes. These data from seven (7) attributes (Area, College, Department, Program, Comment, Sentiment, and Section) underwent data cleansing and the cleansed data were used as the training set in the Weka data mining tool. Further, the researcher identified the dependent variable Sentiment and the independent variables are the Area, College, and Comment. Since the dependent variable is dichotomous in nature, logistic regression is appropriate regression analysis to conduct.

A. Logistic Regression

The researcher then used WEKA Data Mining tool to implement the Logistic Regression – R1.0E-8-M-1-num-decimal-places 4. The R refers to set the ridge in the log-likelihood where default of 1.0E-8 was used. Further, in running the model, the researcher used training set for the test options. Below shows the result of logistic regression.

Correctly Classified Instances 7246 99.1788%
Incorrectly Classified Instances 60 0.8212%
Kappa statistic 0.9836
Mean absolute error 0.0111
Root mean squared error 0.0744

Relative absolute error 2.2264%
Root relative squared error 14.8767%
Total Number of Instances 7306

Kappa statistics measures the agreement of prediction with the true class. The result of Kappa statistics above is 0.9836 which according to Landis and Koch (1977) scheme, a kappa value of 0.81 to 1 is as almost perfect agreement.

The mean absolute error measures how close forecasts or predictions are to the eventual outcomes. An average of the absolute error is only 0.0111 which is forecast with almost minimal error.

Lower values of Root Mean Squared Error (RMSE) indicate better fit. Since the model generates a minimal value of 0.0744, it indicates that the model fits.

The result of Relative Absolute Error of 2.2264% and Root Relative Squared Error (RRSE) is only 14.8767% indicates that the model is good.

Table V shows the result of TP or True Positive rate and Recall of both classes (0.995 for negative sentiment and 0.989 for positive sentiment) is a good indication that the class was captured correctly. Precision results of class 0 (negative sentiment) is 0.989 while 0.995 for class 1 (positive sentiment) simply indicates that almost all of the predicted positive are actually positive. While ROC or Receiver Operating Characteristic measures discrimination, that is the ability of the test to correctly classify those negative and positive sentiments. Since the result of ROC area is 1 and that represents a perfect test.

In the confusion matrix, 3,673 instances are correctly classified as negative sentiments, and only 19 were misclassified as negative sentiments. Moreover, there are 3,573 instances that are correctly classified as positive sentiments and 41 are incorrectly classified.

B. CHAID Algorithm

The researcher used one of the major components of data mining which is Classification. By using classification, the researcher analyzed a set of data and generate a set of grouping rules which can be used to classify future data. The target attribute Sentiments are classified according to its predictors Area, Comment and Section.

The researcher used one of the oldest tree classification methods, the CHAID or Chi-square Automatic Interaction Detector.

Using IBM SPSS, Fig. 4 shows the CHAID rule set. This shows that if comments are less than 39, then the predicted Area is Administrator and Facility (Area 1 and 2). If the predicted Area is 1 or Administrator, then it would be negative. And If the predicted Area is 2 or Facility, then the if predicted colleges would be CEAT or CLAC or CSCS or CTHM (1, 4, 6 or 7), then the predicted sentiment is negative. Else if the predicted college is CEAT or COED (3 or 5) then the predicted sentiment is also negative.
If the comment is greater than 39 and comment is less than or equal to 94 (mostly comments for facilities) and then if the College is CBAA or CLAC or CTHM (1,4,7) then it is predicted as negative sentiment. Otherwise, if the college is CCJE or CEAT or COED or CSCS (2, 3, 5 or 6) then the predicted sentiment is positive.

If the comment is greater than 94 and comment is less than or equal to 144 (still majority comments are on facilities and 3 comments on faculty) and then if the College is CBAA or CCJE or CEAT or CLAC or COED (1, 2, 3, 4, 5) then the predicted sentiment is negative. Moreover, if the College is CSCS or CTHM then the predicted sentiment is also negative.

If the comment is between 145 and 159 (most comments on effectiveness and excellent professors) then it is predicted as positive comments. This means across all colleges, students have positive sentiments as regards to the faculty.

If comment is between 160 and 178 (most comments on professors’ teaching methodology and personality attributes) then it is predicted as negative sentiment. This means the students across all colleges have negative insights as regard to some professors’ teaching methodology and personality attributes.

If the comment is greater 178 but less than or equal to 208 (mostly on professors’ personality attributes and curriculum) then if Area is equal to Faculty (3) then if College is CBAA or CLAC or COED or CSCS or CTHM (1, 4, 5, 6, 7) then the predicted sentiment is negative. And, if the college is CCJE or CEAT still the predicted sentiment is negative. Otherwise, If the Area is Quality of Education (4) then if the College is CBAA or CEAT or CLAC or CTHM (1, 3, 4, 7) then the predicted sentiment is positive. More so, if the College is CCJE or COED or CSCS (2, 5, 6) then the predicted sentiment is still positive. This means that all colleges have predicted positive sentiments on the Area of Quality of Education.

If the comment is from 209 up to 265 (mostly comments about knowledge gained and applicability of learning) and If the College is CBAA or CEAT or CLAC or COED or CTHM...
If the comment is greater than 265 but less than or equal to 310 (mostly comments on the curriculum) and if the College is CBAA or CEAT or CLAC (1,3,4) or If the College is CCJE, CSCS and CTHM (2,6,7) or if the College is COED (5) then the predicted sentiment is negative. This shows that all students from different colleges have negative sentiments on the Quality of Education.

If the comment is from 311 up to 354 (quality of education and some student services) and if the College is CBAA or CEAT or CLAC (1,3,4) then the predicted sentiment is positive. Otherwise, if the College is CCJE or COED or CSCS or CTHM (2,5,6,7) then it is predicted with positive sentiment. Still, this means that across all colleges, students are satisfied with the Quality of Education.

If the comment is from 354 onwards (mostly about student services) and if the College is CBAA or CLAC (1 or 4) then it is predicted negative sentiments. Else if the College is CEAT or COED (3 or 5) then the predicted sentiment is negative. And if the College is CSCS or CTHM (6 or 7) then the predicted sentiment is still negative. This concludes that six (6) out of seven (7) colleges have negative sentiments on the student services of the university.

V. CONCLUSIONS

After following the steps in Chi-square Automatic Interaction Detector (CHAID) Algorithm discussed by Koehler,K.J, the results of it corroborates the results yield in the Fig. 4 generated by IBM SPSS. Thus, students’ feedback on Faculty and Student Services gained large significance to students overall learning experience. This means that students are more concerned on the learning experience gained from their teachers. Also, the services given to them by the university deeply matters.

Students from all colleges have keen focus on Faculty and Student Services area. Based on the results of this study, the university gained positive sentiments for the faculty area, which means students are highly satisfied. While, most students coming from four (4) out of 7 Colleges are not satisfied with regards to Student Services. This would encourage the university to focus more of its improvement in this area.

The study used the ISO25010:2011 Quality model to evaluate the software. The Evaluation Questionnaire used by the proponent was adapted from Gatpandan and Ambat [16]. The result of the evaluation was interpreted using weighted mean and standard deviation. A 5-point Likert scale was used with the following rate: 5 – Strongly Agree, 4 – Agree, 3 – Slightly Agree, 2 – Disagree, and 1- Strongly Disagree. Table VI shows the results of the Software Evaluation.

To conclude, the average computed standard deviation results to 0.49 and the weighted mean for all the criteria is 4.72 which can be interpreted that the system gains an overall very high quality rating.

ACKNOWLEDGMENT

The authors of this research wish to express their utmost gratitude and appreciation to our parents, parents-in-law, husband and kids for their unending support in every journey of this research. Thank you Dr. Johnny A. Ching, Dr. Marco Saez, Dr. Olivia Legaspi, Dr. Paulino and Dr. Mengvi Gatpandan, for their utmost support. Special thanks to the University Research Office, DLSU-Dasmariñas. And most especially, above all, to the Heavenly Father for His guidance and provisions.

REFERENCES


MARYLI F. ROSAS et al: EXTRACTION OF STUDENTS’ FEEDBACK IN UNIVERSITY SERVICES USING . .


[9] B. Baran, and E. Kılıç, “Applying the CHAID Algorithm to Analyze How Achievement is Influenced by University Students’ Demographics”, Study Habits, and Technology Familiarity. (Submitted November 20, 2013, Revised September 1, 2014, Accepted October 8, 2014)


