Linking Numerical Scores with Sentiment Analysis of students' teaching and subject evaluation surveys: Pointers to Teaching Enhancements

Samuel Cunningham-Nelson, Mahsa Baktashmotlagh and Wageeh Boles
Queensland University of Technology
Corresponding Author Email: samuel.cunninghamnelson@qut.edu.au

CONTEXT
Student evaluations of teaching and subjects are systematically used at many educational institutions, to get students' feedback. They usually consist of a number of questions where students can anonymously provide their responses on 5 or 7-point Likert scales, with provision to provide free comments. There are several different ways of evaluating student satisfaction (Elliott, 2002) and the numerical scores are often taken as the main indicator of student satisfaction, although much insight can be gained from analysing the free text comments.

PURPOSE
The purpose of this study is to utilise students' free comments to gain better insights into problem areas, enabling targeted actions for curriculum design and teaching. Specifically, this study is seeking to answer two related research questions.

1. What is the correlation between numerical scores and students' comments?
2. Which teaching activities influence the selection of the numerical scores, and how can these be used to give pointers for teaching enhancement?

APPROACH
We analyse students' comments, using machine learning techniques to extract sentiment information. Students' comments from survey data obtained at QUT over 3 years, in almost 20 engineering subjects, with two surveys per subject per year. One of the surveys is conducted in the middle of semester (the Pulse survey), the other towards the end of semester (the Insight survey). A total of 2254 text responses have been analysed. Links between sentiment information and survey numerical scores are used to provide pointers to curriculum enhancement or the design and selection of appropriate teaching strategies.

RESULTS
From the analysis conducted so far, a clear link appears to be present between students' free comments and the numerical scores selected by students. There is a clear trend in the types of comments and phrases used when comparing positive and negative feedback. This trend provides confidence in the validity of the survey scores as a useful feedback mechanism. General areas for feedback have also been identified and extracted from the data. This can be applied on a subject by subject basis, or used at the whole-of-institution level.

CONCLUSIONS
A machine learning model has been designed which can predict student satisfaction, given students' free comments about a subject. This model gives a quantifiable score based on these free comments. Important recommendations will also be extracted from the data, by searching for key words (ie. tutorials, assignments, lectures, etc.), to address problem areas, or further enhance those that are judged to be positive and valuable by students. This initial model doesn't take into account negation words.

KEYWORDS
Machine Learning, Sentiment Analysis, Student evaluations of teaching, Surveys.

Introduction

Student evaluations of teaching are an important part of assessing student satisfaction in a particular class or course. The feedback from these student evaluations often consists of a score or ranking for several questions, and also the possibility of free text comments. Part of teaching is fulfilling the expectations of students, and feedback and ratings from evaluation allows that to be done (Cheong Cheng, 1997). From feedback, pointers and information can be determined and hopefully influence the teaching style and method of delivery. Using a single test item is one way of measuring student satisfaction (Elliott, 2002).

In its most basic form, student satisfaction for a unit can be rated as either good, neutral or bad. Sentiment analysis is the formal process of determining the opinion of students on a particular topic. Being able to predict opinion has many real applications from predicting the stock market (Mittal, 2012) to predicting the rating of upcoming blockbuster movies (Jong, 2011). These are two examples which the data being dealt with is not numbers, but textual statements. (Maas, 2011) shows a generic approach to using words as a vector and their application to sentiment analysis. All of these methods look at algorithms and search to find patterns and common connections.

Machine learning investigates algorithms that learn and improve in performance (Langley, 1996). Aside from the text analysis problem presented, machine learning can be used application such as: image recognition, handwriting identification, advertising, computer games and search engines such as Google.

Motivation

In this study, machine learning is used to provide a link between the satisfaction score given to a unit, and free text comment. This link will demonstrate that in fact there is a correlation between these pieces of information. This in turn means that, looking at a free comment given, the reader should be able to roughly determine how satisfied the student is with the unit. Linking satisfaction and the comments given by the students, will then lead into teaching recommendations, particularly for types of teaching activities. Having positive or negative words associated with a particular teaching activity will convey insight and possible modifications to teaching approaches to be made.

This paper consists of several sections. The first section looks at using machine learning algorithms to predict the scores. Since both pieces of information are available, the accuracy of the tested algorithm can be determined. Numerical results are detailed in this section. The next section considers students' comments in relation to teaching. From this, recommendations based on particular teaching activities can be inferred.

Predicting Scores Using Machine Learning

The Data Set

Using data from the Queensland University of Technology, data was obtained to be used for this study. QUT runs two student surveys titled 'Pulse' and 'Insight' as a part of the framework 'Reframe' for evaluating learning and teaching (QUT, 2015). The first survey, the Pulse survey, solicits students' feedback in the early weeks of the semester. The second survey, titled Insight, surveys students at the end of semester.

In each of these surveys, students are asked to rate their views on three statements. They respond to each of these questions on a 5 point Likert scale. The statements are,

1. This unit is providing me with good learning opportunities.
2. I am taking advantage of opportunities to learn in this unit.

3. I am satisfied with this unit so far.

Each statement was responded to on the Likert scale, 1 being disagree strongly, and 5 being agree strongly. After answering these questions, an open ended response was left optional, for students to respond to with feedback and suggestions asking them to “Please provide any further feedback you may have about this unit.”. This feedback provides recommendations and suggestions for teaching staff to read and take into consideration.

Data from the third statement and open ended responses was the focus of this study. The study and analysis presented in this paper used Pulse and Insight survey data from 19 units in the Science and Engineering Faculty, over a 3-year duration. The data consisted of 2254 responses which included the numerical and text feedback. These were used in the machine learning analysis.

Examples of student responses include,

- *Satisfaction rating:* 4. *Free response:* “Excellent work. Very helpful staff and lots of assistance given via video tutorials.”
- *Satisfaction rating:* 5. “Great structured unit. Very well organised and great learning environments”
- *Satisfaction rating:* 1. *Free response:* “The unit moved too quickly through concepts, and the content in the workshops has not helped with assessment items”
- *Satisfaction rating:* 2. *Free response:* “The assessment and the lectures didn't relate well - very confusing overall. Not enough practice problems.”

Machine Learning Techniques

Several machine learning techniques were explored to see if a correlation or relationship could be found between the students’ given satisfaction scores and their free text comments. Demonstrating this correlation will lead to the ability of exploring possible teaching technique suggestions or improvements. For this study, Linear Regression, Naïve Bayes and Support Vector Machines (SVMs) were explored. Naive Bayes is a probabilistic model which is based on the Bayes rule along with a simplifying conditional independence assumption. In this context, Naive Bayes classifier returns the score which has the maximum posterior probability given the comment, where in each comment, the words are assumed to be conditionally independent of each other. SVM is an instance based algorithm which uses instance data (support vector) to create a function that maximises the margin/distance between classes (scores in our context). Kernel Support Vector Machine (kSVM) is one type of SVM, which was explored in this study. Linear regression finds the best-fitting regression line through the points (features extracted from the comments) by minimising the sum of squared errors between the true and estimated predictions (scores).

Analysis of Survey Data

Data analysis of the survey data was an important part of the entire process. This section details that analysis process. Data was first gathered, information being formatted and pre-processed initially. The dictionaries used in the machine learning process were created. Data was rounded to a 3 point Likert scale, and the elements (or features) were extracted from the responses. Several machine learning methods were tested and implemented on the processed data. The results for these are summarised in the table at the end of this section. This rest of this section discusses the analysis process in more detail.

Initially the data was gathered and combined from each of the 19 units used. For each response, the important information was extracted. This consists solely of the students’ satisfaction rating, and the free text. Before working with the data, it is important that it was in the correct format for analysis. This procedure is called pre-processing, where the text data was then: converted to lowercase, removed included punctuation, and separated each response into individual words. Following this, data was then split into the division, 20% being used for testing and 80% for training. Common practice in machine learning applications, splitting the data in this proportion allows a large amount to be used for constructing the mode, and a smaller portion used to test the validity of the model.

Dictionaries, in computer fundamentals are an important part of performing an analysis. Just like a normal dictionary, but for a computer, the available dictionary details the possible list of words a computer program can recognise. For the analysis process, a dictionary of both positive and negative words (Hu, 2004) formed the starting point for the analysis. This dictionary contained two lists, one of positive words and one of negative words, totalling almost 7000 words. Since sentiment analysis looks at the 'positivity' or 'negativity' of a particular statement, these words were particularly important.

Two more dictionaries were considered; one dictionary considered of around 200 words that most commonly occurred in the responses, and were deemed important (leaving out non critical words, such as "and"). The other dictionary was one consisting of both the positive and negative words, as well as the 200 important words.

Examining the frequency which words occurred within the responses facilitated the second and third dictionaries to be created. The first dictionary contains only positive and negative words. The second dictionary containing the words that occurred most commonly, and were deemed to be important was constructed from the frequency in which the words occurred. The third combined dictionary was constructed from the distinctive entries which occurred in both dictionaries.

The final three dictionaries are referred to, and can be summarised as,

- External Dictionary – Consists of positive and negative words – Dictionary Size: 6789 words.
- Internal Dictionary – Consists of individually chosen words – Dictionary Size: 219 words. As discussed above, these words were chosen based on their merit.
- Combined Dictionary – Consists of a unique combination of External and Internal dictionary words – Dictionary Size: 7008 words.

After constructing the dictionaries, the data was rounded from the original 5-point Likert scale, to a 3-point one instead. This matches the object, and means the data is placed into either a positive, negative or neutral category.

The next part of the process completed is called feature extraction. The feature extraction process involves creating a list and checking when a word within one of the responses also occurs within one of the dictionaries. These occurrences were used for the machine learning process. Data from the feature extraction was also normalised before continuing. The data was normalised to ensure that all data sits on the same scale.

Using the different machine learning methods mentioned above, mathematical models were created for each. These models were trained using the training data, and then tested using the remaining 20% of responses. Results were once again rounded to sit within the 3 point Likert scale.

Preliminary results showed that the external dictionary performed better, obtaining a lower mean square error and increased classification accuracy. The external dictionary was used for the remainder of the results. Table 1 below shows a summary of the results for different machine learning methods. To produce this table, the methods were run 20 times/partitions each, and the results averaged. The average accuracy and standard deviation were calculated over 20 partitions, as this is a common practice in text categorization (Md Arafat Sultan, 2016).

Table 1: Testing and Training Data Accuracies

	Testing Data		*Training Data*	
Method	*Mean Squared Error*	*Percent Correctly Classified*	*Mean Squared Error*	*Percent Correctly Classified*
Linear Regression	0.586	39.67%	0.564	39.44%
kSVM	0.574	**57.70%**	0.517	62.29%
Naïve Bayes	0.821	53.54%	-	-

Table 1 lists two pieces of information: mean squared error (MSE) and percentage correctly classified. These pieces of information are listed for each machine learning method, as well as for both testing and training data. A lower mean squared error indicates better accuracy. A higher percentage correctly classified shows a better fit of the corresponding machine learning model. No results are included for training data for Naïve Bayes as this particular machine learning model does not use training data, it works using probabilities.

The testing data is the most important half of the table. The test data is data that was unseen by the trained model, and shows the ability of the model to correctly classify responses. The SVM method provided the highest percentage of correctly classified responses (highlighted above). Over 57% of free text comments entered were correctly matched to their corresponding satisfaction score stated by the students. This level of accuracy using the SVM machine learning method, and the amount of data obtained demonstrate a clear correlation between the types of words used by students in free text scores and their overall satisfaction rating. This level of correct classification (around 57%) is sufficiently comparable to what is achievable using textual data, as presented in published research (Md Arafat Sultan, 2016). Knowing that a link is present between the comments presented, and their weighting enables further analysis such as examining the keywords used.

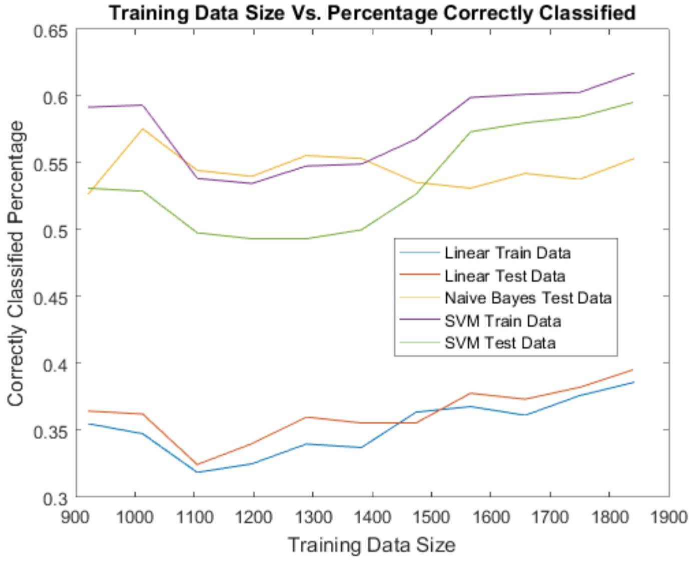

Figure 1: Training and test data size vs. correctly classified percentage

Figure 1 below also shows the percentage accuracy for each of the tested methods against the number of responses used as a part of the training data. The figure shows an increase in correctly classified samples as the training data size increases. With access to further text responses and scores, it is predicated that the classification accuracy will continue to increase.

Insights from the Analysis – Teaching Activities

After applying machine learning algorithms and establishing a relationship between the satisfaction scores and free text comments, teaching activities to be investigated were identified. Possible teaching activities or related words identified were: tutorial, tutor, assignment, lecture, lecturer, content, room, material, workshop, lab, assessment, prac, criteria and time. Identifying these words, the objective is to investigate which words impact students' satisfaction scores the most, and which ones provide insight into which teaching activities might need attention.

Table 2: Teaching Keywords Mentioned in Sentiment Analysis Categories

Keyword	*Low Score (1 or 2 out of 5)*	*Medium Score (3 out of 5)*	*High Score (4 or 5 out of 5)*
Tutorial(s) / Tute / Tut	31.86%	34.86%	27.87%
Tutor(s)	28.06%	17.86%	21.98%
Assignment(s)	11.60%	8.28%	8.84%
Lecture(s)	44.09%	41.18%	30.97%
Lecturer(s)	17.93%	14.81%	18.35%
Content	29.54%	28.32%	18.35%
Room	1.27%	1.09%	0.76%
Material	4.64%	5.45%	3.10%
Workshop(s)	6.33%	5.23%	2.42%
Lab(s)	5.06%	6.97%	5.82%
Assessment	0.42%	0.22%	0.15%
Prac(s) / Practical	8.44%	7.63%	7.70%
CRA / Criteria	2.95%	1.53%	0.68%
Time	24.89%	18.30%	13.14%

Table 2 displays each teaching keyword mentioned above, broken into three zones; low, medium and high scores. The three zones were determined from the given satisfaction scores for the unit, and the teaching keywords were extracted from the students' free textual comments. Each satisfaction score has an associated percentage that indicates the number of times a teaching keyword is mentioned in that particular zone. For example, 31.86% of students who have given the unit a low score (1 or 2 out of 5) mentioned the word tutorial (or similar). However, 27.87% of students who have given the unit a high score (4 or 5 out of 5) mentioned the word tutorial (or similar). This convention follows throughout the table.

From this table, conclusions can be drawn. For example, for the subset of data above;

- Lectures is a highly mentioned word, occurring in many student responses, both positive and negative. This could mean that lectures provide a pivotal role in students' satisfaction rating of a unit. This could also mean that lectures are the main mode of delivery.

- As a negative example, “content” is mentioned in much of the negative responses. If this table was focussing on a particular subject, this would indicate that students who gave the subject a low rating are likely to have had issues with the its content. Content for that unit might need to be closely examined, on its own, or in relation to other aspects such as lectures or tutorials, etc..
- The key word “Lecturers” has a high percentage mentioned in the high category, with many positive words being mentioned in close proximity. This could indicate that teaching strategies being applied in a certain unit are working well. This also means the potential for these lecturers to share their strategies with other for enhanced student experience.

The table shows the responses across all units, however applying the same analysis to individual units would provide valuable insights, and possible actions for teaching staff to take.

Conclusion

This research demonstrated that there was a correlation between the numerical satisfaction score given by students and their free text comments. Using an SVM machine learning classification method, 57% of free text responses were found to be correctly classified to their corresponding satisfaction score.

From this, using results either faculty wide or unit-by-unit, recommendations can be drawn for particular teaching activities. For example, from the chosen subset of student comments, if responses include positive impacts of the lecturers on student satisfaction scores, then, details of the teaching strategies those lecturers used can be shared with other lecturers to collectively enhance delivery or student-lecturer interactions. On the other hand, if content is associated with negative student comments, this signals the need to closely examine that content. Possible sources of problems in relation to unit content could be referring to how the content is organised, linked to pre-requisite knowledge, or difficulties arising from varied expectations between students and lecturers. However, in general, having established links between scores and comments can be a rich source of identifying specific actions that address problem areas, as well as further enhance those aspects that are working well, in the students’ views.

It must be noted here that student survey data are only one source of feedback on unit content and delivery, and should be used in conjunction with other teaching and learning evaluation strategies.

Future investigations will consider larger datasets, which is expected to improve the accuracy of the SVM machine learning model. It is also planned that negation words such as “not” will be taken into account to ensure that the correct positive or negative sentiment is assigned. Particular words might also influence the end result more than others. Using words with high correlation to the satisfaction score, would be another aspect to be explored to examine their effects on improving the classification accuracy.

References

Albertini, S., Zamberletti, A., & Gallo, I. (2014). Unsupervised feature learning for sentiment classification of short documents. *JLCL*, *29*(1), 1-15.

Póczos, B., Ghahramani, Z., & Schneider, J. (2012). Copula-based kernel dependency measures. *arXiv preprint arXiv:1206.4682*.

Cheong Cheng, Y., & Ming Tam, W. (1997). Multi-models of quality in education. *Quality assurance in Education*, *5*(1), 22-31.

Elliott, K. M., & Shin, D. (2002). Student satisfaction: An alternative approach to assessing this important concept. *Journal of Higher Education Policy and Management*, *24*(2), 197-209.

Erfani, S. M., Baktashmotlagh, M., Moshtahgi, M., Nguyen, V., Leckie, C., Bailey, J., & Ramamohanarao, K. (2016). Robust Domain Generalisation by Enforcing Distribution Invariance. In *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)*.

Hu, M., & Liu, B. (2004, August). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 168-177). ACM.

Jong, J. (2011). Predicting Rating with Sentiment Analysis.

Bach, F. R., & Jordan, M. I. (2002). Kernel independent component analysis. Journal of machine learning research, 3(Jul), 1-48.

Langley, P. (1996). *Elements of machine learning*. Morgan Kaufmann.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011, June). Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1* (pp. 142-150). Association for Computational Linguistics.

Sultan, M. A., Boyd-Graber, J., & Sumner, T. Bayesian Supervised Domain Adaptation for Short Text Similarity. In *North American Association for Computational Linguistics*.

Mittal, A., & Goel, A. (2012). Stock prediction using twitter sentiment analysis. *Standford University.*

Oymak, S., & Tropp, J. A. (2015). Universality laws for randomized dimension reduction, with applications. *arXiv preprint arXiv:1511.09433*.

QUT. (2015). Protocols: QUT's Evaluation Framework. Brisbane.

Acknowledgements

Data for this study was obtained from the QUT Pulse and Insight evaluation surveys. The authors would like to thank QUT students and staff for this information. Oliver Moulton, QUT student also contributed to the initial development and data analysis for this project.