Kashihara, A. et al. (Eds.) (2024). Proceedings of the 32nd International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Developing a Feedback Analytic Tool to Support Instructor Reflection

Feng LIN[a*], Chenchen LI[a], Rebekah Wei Ying LIM[b] & Yew Haur LEE[c]
[a]*Teaching & Learning Centre, Singapore University of Social Sciences, Singapore*
[b]*College of Interdisciplinary & Experiential Learning, Singapore University of Social Sciences, Singapore*
[c]*Business Intelligence & Analytics, Singapore University of Social Sciences, Singapore*
*linfeng@suss.edu.sg

Abstract: Gathering student feedback on teaching is an important way for instructors to learn about students' experiences and engage in reflective practices to improve their teaching. Such feedback usually includes numerical ratings and qualitative comments. While numerical patterns can be quickly discerned, analyzing and interpreting qualitative feedback can be time-consuming and challenging, especially in situations involving large classes or multiple courses. To help instructors effectively reflect on the qualitative feedback, we have developed an analytical tool to visualize this feedback. This paper outlines the methodologies and approaches we developed for creating the feedback analytics tool.

Keywords: Text mining, student feedback, instructor, feedback analytic tool

1. Introduction

Student feedback on teaching is a cornerstone in higher education (Spooren et al., 2013), offering instructors invaluable insights into students' learning experiences and their teaching efficacy. Student feedback typically comprises numerical ratings and qualitative comments. The former relies on structured data, commonly sourced from responses to Likert-scale questions, while the latter is drawn from unstructured textual responses elicited by open-ended questions. While numerical patterns can be quickly discerned, interpreting qualitative feedback poses a time-consuming challenge for instructors, especially in situations involving large classes or multiple courses. To help instructors learn the patterns of students' feedback more efficiently and to prompt their reflection for improving teaching practices, we have developed new approaches to mine students' qualitative feedback. Subsequently, we developed a new feedback analytic tool to visualize these mined data. This paper aims to outline the approaches we developed to mine and visualize students' qualitative feedback.

Given the importance of qualitative feedback, there has been growing interest in developing analytic approaches to extract and visualize student qualitative feedback. Overall, previous studies have primarily concentrated on two facets of student qualitative feedback: sentiments and topics. For example, Wook et al. (2020) developed OMFeedback system to collect student feedback and then used lexicon-based approach to detect sentiment. Misuraca and colleagues (2021) used opinion mining techniques to analyze the sentiments in student qualitative feedback. Hujala and colleagues (2020) developed a model of topic-modelling approach using Latent Dirichlet Allocation (LDA) method to analyze the topics of students' open-ended feedback in higher education institutions.

A small number of studies also developed visualization tools based on text mining. For example, Cunningham-Nelson et al. (2019) used Latent Dirichlet Allocation (LDA) model to extract topics and aspects from course feedback, and they also conducted sentiment analysis at the idea level (break-down from course comment) using Afinn lexicon. In the end, the topics and sentiments were visualized via bar charts. Similarly, Gottipati and team (Gottipati et al., 2018; Pyasi et al., 2018) mined topics in students' qualitative feedback using LDA model and

conducted sentiment analysis using Textblob-Improved. Then they developed application tools such as independent desktop tool and webpage to visualize the mined results.

Generally, prior research has predominantly employed unsupervised techniques, such as LDA, to extract topics from student feedback. While unsupervised methods facilitate automated identification of latent topics within the data, the resulting topics are often challenging to interpret (Madzík & Falát, 2022), as they are typically derived from frequently appearing words (Chauhan & Shan, 2021). To provide meaningful and informative categories to instructors for reflection, we have developed a new approach that integrated content analysis and natural language processing (NLP) to mine the topics in student qualitative feedback. Meanwhile, instead of using lexicon-based approach or traditional machine learning models, we also applied large language model (LLM) to mine sentiments. Eventually, we developed an innovative and comprehensive feedback dashboard using Microsoft Power BI to visualize the mined topics and sentiments. In the following, we will describe the approach we developed to mine and visualize student qualitative feedback.

2. Methods

2.1 Data source

The data used in this research comprises responses from undergraduate students to end-of-course evaluation surveys collected between 2020 and 2022 at a university in Singapore. The survey consisted of several mandatory Likert-scaled questions and one optional open-ended question, inviting students to provide suggestions to the instructors. Our research specifically focuses on mining and visualizing students' responses to this open-ended question.

2.2 Data preparation

A total of 324,461 pieces of feedback were retrieved from the university database between 2020 and 2022. The feedback was processed at two levels: feedback level and feedback sentence level. At the feedback level, blanks and uninformative feedback (e.g. NA, Nil, and some words or symbols that are not comprehensible) were identified and removed. Feedback written in Chinese was removed as we only focused on English feedback. After feedback-level processing, only 33,791 pieces of feedback deemed useful were retained. Some feedback involves multiple sentences, and they were tokenized into sentences.

Following feedback tokenization, the uninformative feedback sentence and uninformative information (e.g. email address and website) within the sentences were further removed. For emoticons and emojis, to retain intactly sentimental information, they were replaced using plain text with corresponding meaning during the processing. Figure 1 shows the data preparation process.

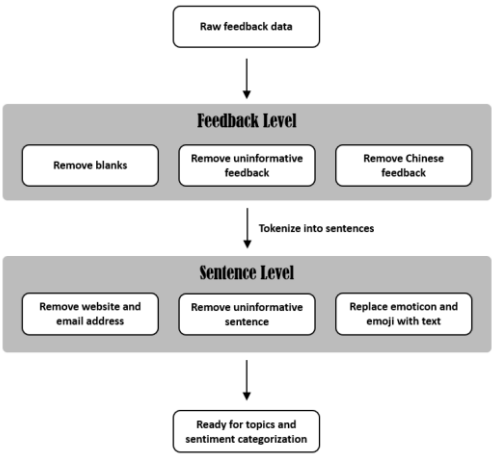

Figure 1. Workflow of Feedback Data Preparation.

2.3 Topic categorization

Informed by literature (e.g., Steyn et al., 2019) and our actual data, we developed an initial coding book consisting of 5 main categories and 81 tentative subcategories (See details in Table 1). Responses that did not fit into any of the five main categories were grouped under "Others" (Figure 5). These subcategories were not developed all at once but were constantly refined through an iterative process. After an initial coding book was developed, a rule-based approach using SPSS Modeler was employed to categorize the 3-semester (i.e., Jan 2020, Jul 2020 and Jan 2021) feedback sentences into the pre-defined subcategories. As illustrated in Table 2, the rules consist of words, short phrases, and logical symbols. These rules enable us to automatically identify feedback sentences relevant to each subcategory.

Table 1. *Definition of Main Category*

Main category	Definition	Number of subcategories
Instructor	Comment on the characteristics of instructor	27
Teaching	Comment on the teaching styles/approaches	35
Content and materials	Comment on the topics covered and course materials	6
Time	Comment on the pace and timing of the classes	4
Assessment	Comment on the assessment approach and materials	9

Table 2. *Example of Categorization Rule in SPSS Modeler*

Main category	Subcategory	Rule
Instructor	Approachable	* approachable * & !(no approachable | be more *)
Time	Pace issue	(class pace | teaching pace | pace | pace of the lesson) & (fast | faster | better | quicken)

Note: Meaning of the symbols in coding rules: "&" means "and"; "|" means "or"; "*" means there are additional words; "!" means "exclude".

The accuracy of categorization for all feedback sentences was manually checked. Accuracy (in the form of precision) was determined by calculating the percentage of sentences correctly captured by the rules. Figure 2(a) shows the precision range of all subcategories. It shows that the precision of all subcategories was above 0.7, and 32 subcategories were above 0.9, which indicated good model accuracy. To further evaluate the performance of the rule-based model, we apply the model to a new semester's data (Jul 2021). The precision of categorization was again manually checked. As shown in Figure 2(b), the precision of most subcategories was above 0.7, with only 3 subcategories between 0.6-0.7. This indicated good reliability of the model. Based on the results, refinements were made to some categories and their rules to improve the model accuracy.

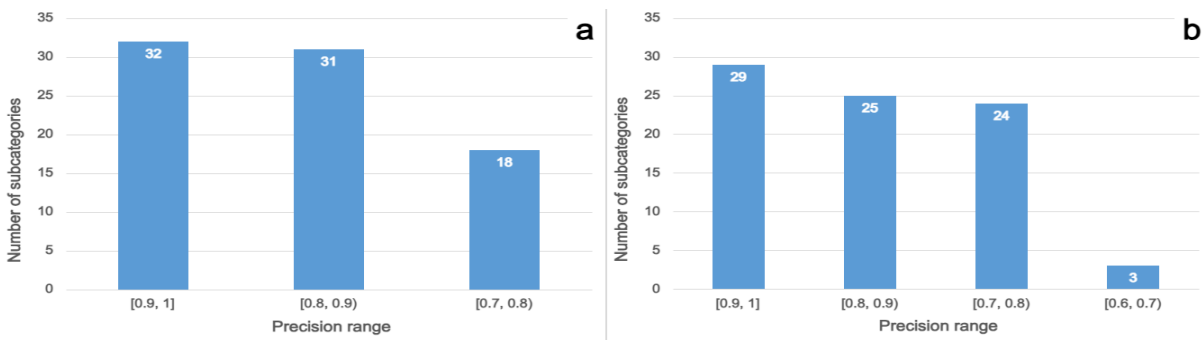

Figure 2. Number of subcategories in different precision ranges (a) data from Jan 2020, Jul 2020, and Jan 2021); (b) data from semester July 2021.

2.4 Sentiment analysis

We used a supervised machine learning approach to mine sentiments in student feedback. To ensure the training dataset represents the entire population, we used stratified sampling with schools as the basis. After that, all the sampled sentences (7,504) were manually coded with sentiment categories (i.e., positive, neutral or negative). To ensure the reliability of the human rater, two raters coded the same 503 feedback sentences. Cohen's kappa (κ) was calculated to assess the inter-rater reliability. Result shows κ is 0.829, indicating almost perfect agreement between the two raters (Landis & Koch, 1977). These coded data were split into 80% for training and 20% for testing.

Pre-trained large language model was used to perform this sentiment categorization task. The whole sentiment categorization model architecture included a Bidirectional Encoder Representations from Transformers (BERT) layer and a neural network (NN) layer. Pre-trained BERT model (bert-base-uncased version from Hugging Face) was used for comprehending the context through word embeddings (Devlin et al., 2019). Subsequently, the NN framework was performed as classifier for final sentiment categorization. Eighty percent of coded data was used to finetune pre-trained BERT model during the sentiment categorization model training process.

Model performance was evaluated based on the testing data (20% of the coded data). Figure 3 shows the normalized confusion matrix resulting from the evaluation. The result indicates that sentiment categorization model could predict correctly 95.53% positive feedback sentences and 88.93% negative feedback sentences, but only 43.33% for neutral. The model did not perform well on categorizing neutral feedback because it was difficult to distinguish slightly negative sentiment from neutral. The other reason might relate to the data imbalance issue (Abonizio et al., 2021), as there was much lesser data from the neutral category during the model training. In terms of recall, precision and f1-score for each sentiment category, all these metric values of both positive and negative were above 0.85, indicating that the sentiment model is reliable and robust to be applied to new data for positive and negative categories. However, the neutral was not as accurate as the positive and negative categories. This outcome is of lesser concern within our context since instructors usually focus on positive and negative categories for insights.

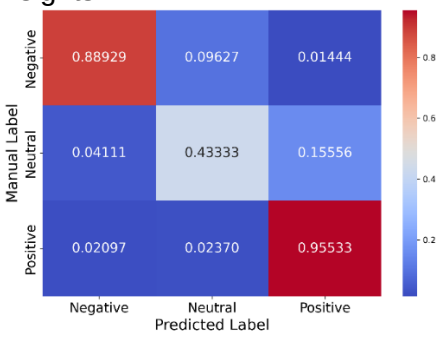

Figure 3. Normalized confusion matrix.

2.5 Development of feedback dashboard

After mining the topics and sentiments, a dashboard using Microsoft Power BI was developed. It consists of several interactive features to facilitate teachers' reflection on their teaching. Figure 4 shows an example of one instructor's dashboard interface. The bubbles on the left side show the mined topics of students' qualitative feedback. Big bubbles represent the main categories, and small bubbles within each big bubble represent subcategories. The size of the bubbles represents the frequency and counts of categories, with bigger indicating more frequently mentioned topics. The feedback percentage metric card on the right side shows the percentage of students giving useful feedback. The ratio of full-time and part-time students is shown as a donut chart. Moving downwards, the sentiment bar chart illustrates sentiment distribution across the main categories. To mitigate negative sentiments while still highlighting areas for improvement, we renamed the "negative" category as "areas for

improvement". Through these visualizations, instructors could easily identify the main topics and sentiments in students' qualitative feedback. The filters on the top area can be manipulated to filter data based on semester, course code, student type, and group code, etc. The dashboard also has an interactivity feature. For instance, when the instructor selects a category, the sentiment graph adjusts accordingly. Additionally, the lower-right section displays detailed feedback related to the selected category.

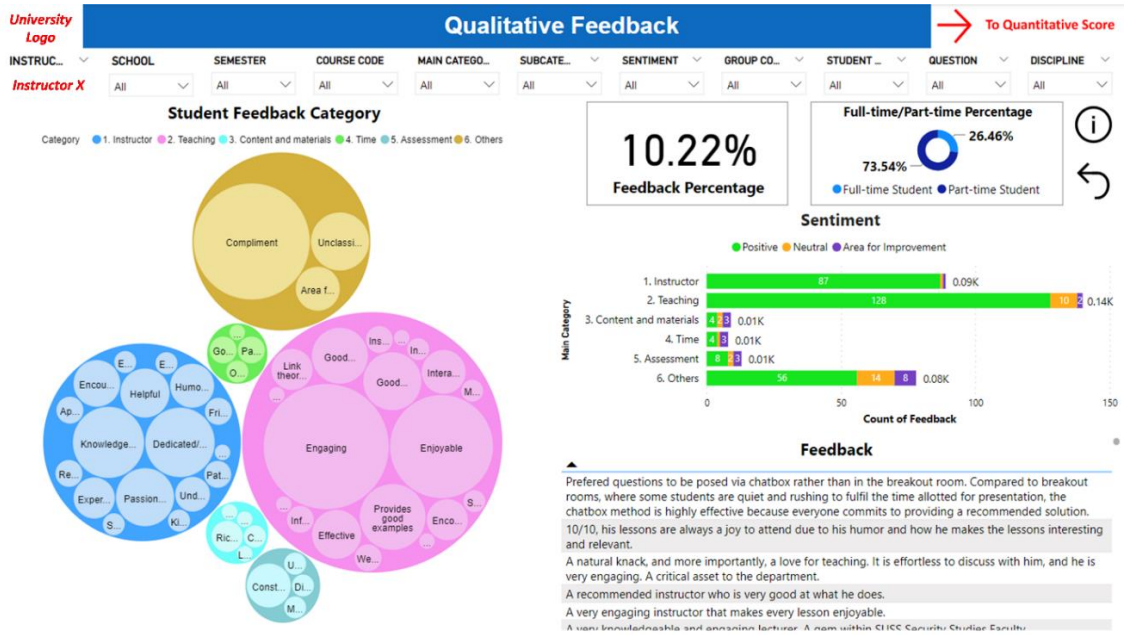

Figure 4. Dashboard Demo of Instructor X.

3. Discussion and conclusion

In this paper, we described the new approaches we developed to mine and visualize the topic and sentiment in student qualitative feedback. These approaches were developed through the cross-disciplinary application of text mining and learning sciences. This research will contribute to the literature in four ways: 1) Previous research has been mainly using unsupervised methods (e.g., topic modeling) to mine the topics in student qualitative feedback (e.g., Hujala et al., 2020; Pyasi et al., 2018). While this approach is straightforward to implement, the topics generated are usually challenging to interpret. Our study shows the potential for combining content analysis with a rule-based mining approach to extract topics from students' feedback. This approach allows for the predefined topics to be derived from both learning theories and actual data, enabling the generation of meaningful insights essential for instructors' reflection. 2) We developed a coding book to capture the patterns of student qualitative feedback. This coding book can be further adapted in other studies to understand student feedback. It can also serve as the basis for categorizing training data for future machine-learning applications. 3) Most prior research relied on lexicon-based methods or traditional machine-learning techniques for sentiment analysis (e.g., Wook et al., 2020). However, these approaches involve extensive text preprocessing and are also not effective at capturing contextual meaning. In this study, we employed the state-of-the-art language model BERT to mine the sentiments in students' qualitative feedback. This new approach can be further used in future research for mining sentiments. 4) At last, we developed a new comprehensive and interactive feedback dashboard using Microsoft Power BI to visualize student qualitative feedback. The developed dashboard could enable instructors to effectively explore the topics and sentiments in student qualitative feedback. Through its intuitive interface and visualizations, the tool enables instructors to quickly identify their strengths and areas for improvement, encouraging ongoing enhancement of their teaching practices.

This is an ongoing project. Moving forward, we will continue to finetune the mining approach and feedback dashboard. Once the tool is fully developed, we will research how

instructors utilize the feedback dashboard to reflect on their teaching practices. Collecting student feedback for teaching is a common practice in higher education institutions, and our goal is to provide insights into the techniques for extracting and visualizing student feedback. This can serve as a useful reference for other researchers and institutions in developing their own feedback analytics tools, ultimately supporting instructors in improving their reflective practices within their institutions.

Acknowledgements

This work was supported by the University Strategic Plan Fund. We thank Campus IT Services and Business Intelligence & Analytics departments for providing the data and assistance. We also thank the management and colleagues for providing helpful comments on the project. Any opinions, findings, and errors are those of the authors, and do not reflect the views of the university. This project has paved the way for securing funding for a new research project through the MOE Tertiary Education Research grant (MOE2023-TRF-031).

References

Abonizio, H. Q., Paraiso, E. C., & Barbon, S. (2021). Toward text data augmentation for sentiment analysis. *IEEE Transactions on Artificial Intelligence, 3*(5), 657-668.

Chauhan, U., & Shah, A. (2021). Topic modeling using latent Dirichlet allocation: A survey. *ACM Computing Surveys, 54*(7), 1-35.

Cunningham-Nelson, S., Baktashmotlagh, M., & Boles, W. (2019). Visualizing student opinion through text analysis. *IEEE Transactions on Education, 62*(4), 305-311.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Gottipati, S., Shankararaman, V., & Lin, J. R. (2018). Text analytics approach to extract course improvement suggestions from students' feedback. *Research and Practice in Technology Enhanced Learning, 13*, 1-19.

Hujala, M., Knutas, A., Hynninen, T., & Arminen, H. (2020). Improving the quality of teaching by utilising written student feedback: A streamlined process. *Computers & education, 157*, 103965.

Landis J. R., Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174.

Madzík, P., & Falát, L. (2022). State-of-the-art on analytic hierarchy process in the last 40 years: Literature review based on Latent Dirichlet Allocation topic modelling. *Plos One, 17*(5), Article e0268777.

Misuraca, M., Scepi, G., & Spano, M. (2021). Using Opinion Mining as an educational analytic: An integrated strategy for the analysis of students' feedback. *Studies in Educational Evaluation, 68*, 100979.

Pyasi, S., Gottipati, S., & Shankararaman, V. SUFAT: An analytics tool for gaining insights from student feedback comments. (2018). *Proceedings of the 2018 Frontiers in Education Conference 48th FIE: San Jose, CA, October 3-6*, 1-9.

Spooren, P., Brockx, B., & Mortelmans, D. (2013). On the validity of student evaluation of teaching: The state of the art. *Review of Educational Research, 83*(4), 598-642.

Steyn, C., Davies, C., & Sambo, A. (2019). Eliciting student feedback for course development: the application of a qualitative course evaluation tool among business research students. *Assessment & Evaluation in Higher Education, 44*(1), 11-24.

Wook, M., Razali, N. A. M., Ramli, S., Wahab, N. A., Hasbullah, N. A., Zainudin, N. M., & Talib, M. L. (2020). Opinion mining technique for developing student feedback analysis system using lexicon-based approach (OMFeedback). *Education and Information Technologies, 25*, 2549-2560.