Leveraging Large Language Models for Analysis of Student Course Feedback

Zixuan Wang
zwan843@aucklanduni.ac.nz
University of Auckland
Auckland, New Zealand

Paul Denny
University of Auckland
Auckland, New Zealand
paul@cs.auckland.ac.nz

Juho Leinonen
University of Auckland
Auckland, New Zealand
juho.leinonen@auckland.ac.nz

Andrew Luxton-Reilly
University of Auckland
Auckland, New Zealand
andrew@cs.auckland.ac.nz

ABSTRACT

This study investigates the use of large language models, specifically ChatGPT, to analyse the feedback from a Summative Evaluation Tool (SET) used to collect student feedback on the quality of teaching. We find that these models enhance comprehension of SET scores and the impact of context on student evaluations. This work aims to reveal hidden patterns in student evaluation data, demonstrating a positive first step towards automated, detailed analysis of student feedback.

CCS CONCEPTS

• **Social and professional topics → Computing education**.

KEYWORDS

Student Evaluation of Teaching, Large Language Model, Student Feedback, Natural Language Processing

ACM Reference Format:

Zixuan Wang, Paul Denny, Juho Leinonen, and Andrew Luxton-Reilly. 2023. Leveraging Large Language Models for Analysis of Student Course Feedback. In *16th Annual ACM India Compute Conference (COMPUTE '23), December 09–11, 2023, Hyderabad, India.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3627217.3627221

1 INTRODUCTION

Student feedback on teaching (through a tool such as the Summative Evaluation Tool — SET — analysed in this study) is a commonly used method of evaluating the quality of course delivery. Although the use of student evaluations for teachers and courses is widely accepted [4], a significant body of research indicates that these scores may not accurately measure teacher professional competence [4, 9, 10]. Despite these concerns, collating student perceptions of teaching can provide useful feedback for teachers that may be used for continuous professional development.

Analyzing broad, qualitative data and scores to draw insights is challenging. Student comments are important as they highlight specific strengths and areas of improvement not captured by scores. However, manual analysis of these comments is laborious and time-consuming, limiting their usefulness in large educational settings.

Large language models like ChatGPT recently demonstrated excellent text analysis capabilities. These models process and generate natural language, crucial for qualitative feedback analysis. We explored deploying these models to classify and categorise students' course comments for a more efficient, comprehensive analysis providing valuable insights into teaching practices.

Our study assessed the feasibility of using large language models to classify students' course comments. Combining qualitative student comments with quantitative SET scores can offer a more holistic view of the teaching and learning process. We explored these models' usefulness in understanding how course features impact SET scores across disciplines, promising insights into patterns that traditional statistical analysis may overlook. Hence, our research questions are:

RQ1 Can large language models be effectively deployed to classify and categorize students' course comments, thereby providing a deeper understanding of SET scores?

RQ2 How can large language models contribute to understanding the differential impact of course characteristics on SET scores between Computer Science and other disciplines?

2 RELATED WORK

Despite their common use for teaching improvement, research indicates scores from student evaluation of teaching are not valid indicators of teaching competence [4, 9, 10]. Qualitative feedback from students is considered a viable alternative to quantitative evaluation [3], and may result in context-specific perspectives on student experience that are more relevant to improving teaching and learning outcomes [8], but due to the difficulty of analysing qualitative data manually, there is significantly less research on the value of qualitative teaching evaluation methods.

A large language model (LLM) is an advanced computer model capable of processing and generating natural language using deep learning algorithms [11]. These models can learn various tasks, including recognition, search, translation, prediction, speech, generative text, and bots, among others [2]. Kant et al. used unsupervised pre-training and fine-tuning to achieve good results on difficult text

COMPUTE '23, December 09–11, 2023, Hyderabad, India

ACM ISBN 979-8-4007-0840-4/23/12…$15.00
https://doi.org/10.1145/3627217.3627221

classification tasks using a large language model [5]. This raises the potential for text classification of student feedback.

Using the RoBERTa (Robust Optimised BERT Pre-training Method) model, Cunningham et al. examined and discussed a method for identifying and removing unacceptable comments from student evaluations of teaching. Their results showed that the method successfully identified and removed unacceptable comments, reduced the need for manual review and allowed students to revise comments [1].

Rybinski et al.'s study [7] examined using advanced Natural Language Processing (NLP) models, specifically the BERT algorithm, to analyze over 1.6 million student comments from the US and UK and evaluate teaching quality. They sought to establish NLP models as an alternative to traditional Likert scale-based SETs. While NLP models accurately predicted university ratings and teaching quality, predicting main topics in student comments was more challenging, and they often amplified existing biases in the data, like simpler course bias and tutor gender or rank.

2.1 Differences in Computer Science Course Student Evaluation

Morgan et al. [6] explored student engagement differences between computing and non-computing courses through literature review and academic discussions. Their study, examining past research and using various tools to assess participation, revealed lesser engagement among computing students and a deficiency in the understanding of student engagement among computing instructors. Educational methods in computing education appeared insufficient for enhancing student engagement. Preference for individual learning and independent reasoning over collaborative work and communication was common among these students, possibly due to large class sizes, limited interaction opportunities, typical classroom resource constraints, and challenges in collaboration, communication, and forming learning communities. The authors, however, remained uncertain about the reasons for the disparity in student engagement between computing and other subjects.

3 QUALITATIVE SET COMMENTS ANALYSIS

We analyzed 8832 text comments from 2944 students across 272 courses in the Science Faculty at one University. The comments related to aspects of the course that were helpful, aspects that were most challenging, and areas that could be improved. To maintain confidentiality, we locally deployed a large language model, Llama, for in-depth analysis without compromising student confidentiality.

3.1 Methodology

We applied the Llama 13b model for Text Classification on our comment dataset using the Pandas library and GPT4ALL Python Binding. We used the nine topics from the Likert-scale student evaluation data — Accessibility, Collaboration, Communication, Clarity, Relevance, Feedback, Community, Engagement, and Quality — as labels for text classification. We didn't provide Llama with specific topic definitions, meaning there's no assured correlation between these topics and the nine quantitative section questions.

Due to computational and time limits, we focused on classifying 2075 comments from 31 Computer Science courses and randomly

Table 1: Example of some student comments and classification results by Llama

Question	Comments	Llama Classification
3.Challenges (OLE)	Information regarding the assignments is not easily accessible. Even for A6, basic instructions on how to run the program were missing. Only when someone asked on Ed did Mano tell us. He wrote 4 steps in the instructions to run the program, why not just include that in the assignment page?	Accessibility
2.Areas to Improve	The lectorial style tuts were a bit off-putting and didn't really invite open discussion as much as a normal tutorial would, maybe holding that class in a smaller room that is more personal. also at the beginning of the course have a lab-like session for setting up and or installing your own VM, and of course, having the uni-provided VM not have connection problems would be very cool but that is completely understandable.	Collaboration
1.Helpful Aspects	I would really like to emphasize how valuable the work put in by the tutor, has been for my learning. He is really great at explaining concepts in a way that is easy to understand, and always encourages students to ask questions and try to engage with the content without worrying about not understanding or being wrong.	Communication
2.Areas to Improve	create a student community which have student from the nationality to know each other	Community

selected 1766 more comments for a control group. Ensuring the pre-trained model's accuracy, we excluded comments less than 150 words long. Ultimately, this included 491 comments from Computer Science courses and 632 from the control group. LLAMA assigned multiple labels to most comments without a defined label limit.

The prompt we provided to LLaMa is *"For each of the next student reviews, categorize and label them. Classify them as [Accessibility, Collaboration, Communication, Clarity, Relevance, Feedback, Community, Engagement, Quality] Each student review will be preceded by the code for the course. Please only show me the label of the review."*

3.2 Accuracy

In our study, we used a lower temperature for precision and implemented the pre-trained Llama model, specifically gpt4all-l13b-snoozy, as acquiring a substantial training set for specific student comment classification was not feasible. We believe the model's generic linguistic patterns and features garnered during pre-training enable it to manage untrained tasks. The gpt4all-l13b-snoozy model demonstrated reliable accuracy in classifying our student assessment data upon visual observation of the results (see Table 1).

In this case, however, we evaluated the accuracy of Llama 13b for classifying student reviews by manually annotating 70 student reviews for a course. It is worth noting that the accuracy of the test set is not guaranteed as the annotators are not trained. We chose to have multiple annotators manually annotate at the same time, and eventually compiled a list of the most accurate annotations. Each comment was manually tagged with the two most relevant tags and compared to the LLAMA tags. Out of 140 manual tags for 70 comments, 109 annotations were identical to the annotations given by LLAMA. Despite the small size of the test set, there was evidence of LLAMA's accuracy in this student comment classification task (Accuracy = 77.86%).

3.3 Comparison of Computer Science courses with other courses

Upon completing the categorization of the student SET (Student Evaluation of Teaching) comments, we analyzed the results of the classifications for Computer Science and Science courses. The findings reveal that, regarding helpful aspects, students most frequently mentioned "Quality," "Clarity," and "Relevance" in comments about Computer Science courses. Conversely, in science course feedback, the most frequently mentioned elements were "Quality," "Engagement," and "Collaboration" (Table 2). In terms of areas for improvement, students most frequently cited "Clarity," "Quality," and "Relevance" in their comments about Computer Science courses, while for Science courses, the most frequently mentioned aspects were "Relevance," "Clarity," and "Quality". As for challenges, both in the Computer Science and all Science courses, students most frequently referred to "Quality," "Clarity," and "Feedback".

Compared to other Science disciplines, Computer Science courses had about half the proportion of comments regarding helpful aspects marked as "collaborative", "communicative" and "engagement". Also similar to Morgan et al.'s findings [6], a much lower proportion of student comments on registering helpful aspects appeared to be related to Engagement. The most frequently cited helpful aspects by computing students for their learning were "quality", "clarity"and "relevance". Students do not seem to find sufficient engagement and collaboration opportunities in CS courses, perhaps because these courses tend to focus more on individual learning and independent thinking.

Students in Science courses stress the significance of "Engagement" and "Collaboration", reflecting the practical and team-based nature of these subjects. However, in Computer Science, students value comprehensive material, clear instructions, and relevant content, which aligns well with it being an application-oriented discipline. These contrasting priorities suggest different academic disciplines may require specific pedagogical approaches to meet student expectations and enhance learning outcomes.

3.4 Student Comments Summary

Large Language Models (LLMs) can efficiently process and analyze large volumes of student feedback data, providing teachers with summarised feedback. They can identify key themes, understand student perspectives on the course, its strengths and challenges, and student needs and expectations. Such insights can enhance teaching strategies, course design, and personalised support. We submitted all course comments exceeding 50 words to our localized model to mimic a scenario where the model functioned as a teacher, efficiently extracting information from SET comments.

The LLM effectively differentiated the three comment categories (Helpful Aspects, Areas to Improve, Challenges) and summarised key points from student feedback. However, it has limitations; it gathers summaries based on training data patterns and consequently might not understand specific domain terminology and background knowledge and might fail to account fully for the contextual and semantic relationships in comments. Given the "Black Box" issue [7], and LLM's anthropomorphic summarization nature, accuracy confirmation for automated summarization is challenging. The potential of LLMs to summarize student feedback offers promising support for educational research and teaching enhancement if combined with human expert involvement to ensure result accuracy and sound interpretation.

4 DISCUSSION

The analysis of the qualitative SET comments using a large language model (LLAMA) provided deeper insights into the students' feedback and further understanding of SET scores.

Text Classification with LLAMA: We used the LLAMA 13b model to classify student SET comments. It reliably categorized comments into topics (77.86% accuracy) such as Accessibility, Collaboration, Communication, Clarity, Relevance, Feedback, Community, Engagement, and Quality. Utilizing large language models like LLAMA enables efficient processing of large volumes of student feedback, yielding insights into strengths, improvement areas, and challenges.

Comparison of Computer Science and Science Courses: Classification results highlighted differing aspects between Computer Science and Science courses. For Computer Science, students frequently cited "Quality," "Clarity," and "Relevance" as helpful, needing improvement, and as challenges. Science students emphasized "Quality," "Engagement," and "Collaboration" as helpful, "Relevance," "Clarity," and "Quality" as needing improvement, while challenges centred around "Quality," "Clarity," and "Feedback."

Student Comments Summary: The large language model efficiently identified key feedback themes when summarising student comments, providing valuable insights for teaching strategies, course design, and personalised support. However, it may struggle to understand domain-specific terminology and capture the full context and semantic relationships in comments. Thus, human involvement and expertise are essential to guarantee the accuracy and interpretation of results.

Large language model analysis helped understand how course characteristics impact student evaluation scores differently in Computer Science versus other disciplines. Quantitative findings were echoed qualitatively, with Computer Science courses scoring lower in collaboration, communication, and engagement. Student comments placed high importance on aspects like quality, clarity, and relevance, aligning with the discipline's focus on individual learning and independent thinking. Contrarily, Science students valued engagement and collaboration, reflecting the practical, hands-on nature of these subjects. These findings illustrate how large language models can provide insights into discipline-specific impact of course characteristics on SET scores, assisting educators in understanding and addressing unique challenges and expectations.

Despite SET scores being questioned as valid teaching competence indicators [4, 9, 10], and student comments within them largely neglected due to analysis difficulties, large language models can efficiently process significant volumes of SET data. They rapidly identify and categorize key information, reducing manual processing burdens and generating comment summaries for teachers. While the 'Black Box' problem and traditional accuracy assessment challenges may limit interpretability, ethical and student privacy matters also need consideration. Still, large language models offer a new analysis method for classroom student comments, supporting educational research, teaching enhancement, and potential fairness in teacher evaluations.

Table 2: Text Classification Result of CS and all Science Courses

Label	Helpful Aspects				Areas to Improve				Challenges (OLE)			
	CS (num = 126)		Science (num = 197)		CS (num = 152)		Science (num = 178)		CS (num = 213)		Science (num = 257)	
	Number	Percentage	Number	Percentage	Number	Percentage	Number	Percentage	Number	Percentage	Number	Percentage
Accessibility	26	20.63%	34	17.26%	26	17.11%	44	24.72%	53	24.88%	80	31.13%
Collaboration	19	15.08%	87	**44.16%**	11	7.24%	27	15.17%	7	3.29%	26	10.12%
Communication	20	15.87%	62	31.47%	15	9.87%	24	13.48%	32	15.02%	38	14.79%
Clarity	54	**42.86%**	70	35.53%	87	**57.24%**	73	**41.01%**	117	**54.93%**	107	**41.63%**
Relevance	52	**41.27%**	83	42.13%	63	**41.45%**	76	**42.70%**	61	28.64%	57	22.18%
Feedback	34	26.98%	66	33.50%	46	30.26%	41	23.03%	68	**31.92%**	61	**23.74%**
Community	3	2.38%	21	10.66%	5	3.29%	7	3.93%	1	0.47%	2	0.78%
Engagement	33	26.19%	99	**50.25%**	22	14.47%	28	15.73%	21	9.86%	35	13.62%
Quality	79	**62.70%**	125	**63.45%**	75	**49.34%**	58	**32.58%**	125	**58.69%**	114	**44.36%**

Note that we utilised a single semester's student SETs data from one university, so generalising results to other scenarios requires caution. Future research should broaden the dataset, incorporating SET feedback from different institutions, disciplines, and timeframes, and explore the benefits of fine-tuning language models specifically for student feedback analysis tasks.

5 CONCLUSIONS

Our study found that Computer Science students prioritised "Quality," "Clarity," and "Relevance," whereas Science students highlighted "Quality," "Engagement," and "Collaboration." We suggest that fostering Engagement and Collaboration could enhance teaching and learning efficacy and student satisfaction in Computer Science courses. Effective strategies could include creating an interactive learning environment, promoting student collaboration, providing prompt feedback and guidance, encouraging interaction, and investing in teacher training and professional development.

Our analysis underscored the impact of course attributes on student perceptions, with Stage 2 courses receiving consistently lower scores due to their complexity. Theoretical courses, especially online ones, saw higher satisfaction compared to programming courses and smaller classes had higher scores in collaboration, communication, relevance, feedback, community, engagement, and quality.

Overall, we find that large language models like LLAMA show potential for efficient text classification and summarisation of student comments, offering valuable insights for teachers and researchers.

REFERENCES

[1] Samuel Cunningham-Nelson, Melinda Laundon, and Abby Cathcart. 2021. Beyond satisfaction scores: visualising student comments for whole-of-course evaluation. *Assessment & Evaluation in Higher Education* 46, 5 (July 2021), 685–700. https://doi.org/10.1080/02602938.2020.1805409 Publisher: Routledge _eprint: https://doi.org/10.1080/02602938.2020.1805409.

[2] Robert Godwin-Jones. 2023. Emerging spaces for language learning: AI bots, ambient intelligence, and the metaverse. (Feb. 2023). https://hdl.handle.net/10125/73501 Publisher: University of Hawaii National Foreign Language Resource Center.

[3] Leonid Grebennikov and Mahsood Shah. 2013. Student voice: using qualitative feedback from students to enhance their university experience. *Teaching in Higher Education* 18, 6 (Aug. 2013), 606–618. https://doi.org/10.1080/13562517.2013.774353 Publisher: Routledge _eprint: https://doi.org/10.1080/13562517.2013.774353.

[4] Henry A. Hornstein. 2017. Student evaluations of teaching are an inadequate assessment tool for evaluating faculty performance. *Cogent Education* 4, 1 (Jan. 2017), 1304016. https://doi.org/10.1080/2331186X.2017.1304016 Publisher: Cogent OA _eprint: https://www.tandfonline.com/doi/pdf/10.1080/2331186X.2017.1304016.

[5] Neel Kant, Raul Puri, Nikolai Yakovenko, and Bryan Catanzaro. 2018. Practical Text Classification With Large Pre-Trained Language Models. arXiv:1812.01207 [cs.CL]

[6] Michael Morgan, Jane Sinclair, Matthew Butler, Neena Thota, Janet Fraser, Gerry Cross, and Jana Jackova. 2018. Understanding International Benchmarks on Student Engagement: Awareness and Research Alignment from a Computer Science Perspective. In *Proceedings of the 2017 ITiCSE Conference on Working Group Reports (ITiCSE-WGR '17)*. Association for Computing Machinery, New York, NY, USA, 1–24. https://doi.org/10.1145/3174781.3174782

[7] Krzysztof Rybinski and Elzbieta Kopciuszewska. 2021. Will artificial intelligence revolutionise the student evaluation of teaching? A big data study of 1.6 million student reviews. *Assessment & Evaluation in Higher Education* 46, 7 (Oct. 2021), 1127–1139. https://doi.org/10.1080/02602938.2020.1844866 Publisher: Routledge _eprint: https://doi.org/10.1080/02602938.2020.1844866.

[8] Carly Steyn, Clint Davies, and Adeel Sambo. 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 (Jan. 2019), 11–24. https://doi.org/10.1080/02602938.2018.1466266 Publisher: Routledge _eprint: https://doi.org/10.1080/02602938.2018.1466266.

[9] Wolfgang Stroebe. 2020. Student Evaluations of Teaching Encourages Poor Teaching and Contributes to Grade Inflation: A Theoretical and Empirical Analysis. *Basic and Applied Social Psychology* 42, 4 (July 2020), 276–294. https://doi.org/10.1080/01973533.2020.1756817 Publisher: Routledge _eprint: https://doi.org/10.1080/01973533.2020.1756817.

[10] Bob Uttl, Carmela A. White, and Daniela Wong Gonzalez. 2017. Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation* 54 (Sept. 2017), 22–42. https://doi.org/10.1016/j.stueduc.2016.08.007

[11] Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: Story Writing With Large Language Models. In *27th International Conference on Intelligent User Interfaces (IUI '22)*. Association for Computing Machinery, New York, NY, USA, 841–852. https://doi.org/10.1145/3490099.3511105