Pedagogically Driven Topic Extraction on Module-Based Student Survey Data

C.W Chan*, L.C Ho and C.M Acebedo

Temasek Polytechnic / School of Informatics & IT, Singapore

*CHAN_chong_wei@tp.edu.sg

Abstract

Institutions of higher learning collect surveys on course modules so that students can provide open-ended qualitative feedback. However, it is challenging to fully comprehend major concerns of students from a pedagogical point of view when reading hundreds of seemingly diverse student responses. Generally, it would be more useful to have qualitative feedback framed according to a pedagogical-driven taxonomy which is well understood by educators. The taxonomy includes sub-topics for assessments, projects, assignments, content, teaching plan, pace, difficulty, and student preferences. For example, a sub-topic for student preference can be about dissatisfaction with specific methodology such as e-learning or flipped classrooms. This paper explores the use of Large Language Models (LLMs) to automatically tag students' qualitative feedback. LLMs attain good model outcome with task-agnostic fine-tuned performance learning. This ensures that fewer samples of survey responses for each topic in the taxonomy, are required for the LLM to learn, relative to non-LLM approaches. Using the proposed methodologies, the qualitative responses can now be automatically tagged and organised in pedagogically meaningful topics and further merged with other relevant student information to be rendered visually as dashboards for easy understanding. Dashboards are foundational in helping stakeholders improve their course design, student engagement, and pedagogical approach, across the different semesters. The stakeholders can come from a diverse group such as lecturers, pedagogy designers and program administrators.

Keywords: s*tudent survey, topic extraction, pedagogical taxonomies, large language models, natural language processing, dashboarding, learning analytics*

1. Introduction

Student surveys are de-rigueur for assessing the teaching effectiveness and student learning in Institutes of Higher Learning. Survey results are analysed without much difficulty when responses are structured into meaningful categories such as 'agree/disagree' or according to a Likert scale. What is more difficult to analyse are the student responses to open-ended questions like 'What are the possible areas for improvement?'. Verbatim feedbacks are rich in information because this data represents the voices of the students.

Analysis of open-ended qualitative data is complicated due to the messiness caused by 3 main factors. First, the responses are typically large in volume ranging from a few hundred to a few thousand entries, depending on the context of the analysis. Second, a response can be overwhelming to comprehend particularly when a student has a long commentary consisting of a litany of topics. Third, student responses are very diverse in nature as different students have varying preferences, aversions and needs.

Technically, it is not impossible to analyse such data. A data analyst can add structure to this messiness by methodically tagging each student's comment to a relevant topic. Such data entry methods are infeasible for an organization. This lack of structure in the analysis meant that course surveys are best left to course lecturers who will form their own overall impressions of student learning needs, usually by looking out for responses that stood out when browsing through the comments.

We propose letting AI (Artificial Intelligence) tag pertinent student comments to relevant topics as well as tag less useful comments as non-comments. The tagged data can then be organized in a visual dashboard, to allow interested stakeholders to identify the major areas of student concerns readily. The dashboard user can focus on the areas of interest by drilling-down on the more salient topics, to read the verbatim comments. Educators with different background in terms of responsibilities and pedagogical needs can use the dashboards for better insights on the areas of student concerns about the subject. The insights gained help the educators to take the appropriate actions on the pedagogy and instructional design of the subject in the next course offering.

This paper is structured as follows. We first present as Section 2, Background of the survey. Section 3 is the Literature Review, highlighting work that were done in this area previously. Section 4 details the methodology

and the data. A description of the survey data from Temasek Polytechnic is made without compromising the institution's data governance policy. This section also describes the issues typical to such survey data. The taxonomy as well as how this taxonomy was developed is elaborated. This section then describes the data preparation efforts primarily in terms of annotating the text, and the algorithms used for developing the AI model. Section 5 presents the results and discussion. These include the identification of the best algorithm and a sample dashboard. Section 6 presents the conclusion and Section 7 ends with areas for future work.

2. Background

This study involved student survey response from the School of Informatics and IT (IIT), Temasek Polytechnic, The Teaching Evaluation & Subject Survey is administered at the end of each semester and all students are strongly encouraged to respond to the survey for each of their course module. The Subject Survey consists of five questions with Likert-scale responses and two open-ended questions. More specifically, the open-ended question of interest was: "Identify area(s) in which the subject could further improve, to better support student learning".

3. Literature Review

The use of Natural Language Processing (NLP) techniques to classify responses to open-ended student surveys for dashboarding purposes using traditional Machine Learning (ML) was proposed by Gottipati et al. (2018).

Large Language Models (LLM) are now considered state-of-the-art over traditional machine learning approaches proposed above. Fine-tuning a pretrained LLM such as BERT (Devlin et al., 2019) to specific tasks has produced state-of-the-art results in text classification, due to its ability to differentiate semantic nuances in human expressions.

This approach allows the conception of more granular topics, in the taxonomy development phase.

4. Methods

The work done on the student survey could be broken down into five stages:

4.1. Data Exploration

The dataset obtained from the survey was considered small for NLP purposes. The limited size of the dataset was further constrained by the fact that majority of students did not provide any valid comments to the open-ended question as shown in Figure 1.

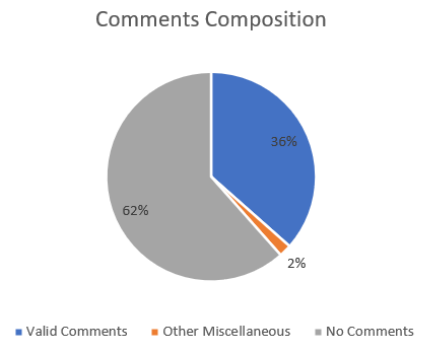

Figure 1: Composition of Relevant and Irrelevant Comments

Only 36% of the comments were classified as 'Valid Comments'. The biggest group of non-valid comments included 'No Comments', associated with responses such as: blank, "NIL", "N.A.", "idk (I don't know)", "Can't think of anything" or something of the nature that is positive like "This course is good. I have nothing further to add" or "The lecturer is great".

The remaining group of non-valid comments belonged to the "Other Miscellaneous" category. These comments were limited in value due to the following reasons:

a) Hard to interpret or are non-actionable. Examples are: 'lesson', 'teaching environment', 'hh', where such comments were deemed as too cryptic to be useful.
b) Non-systemic in nature. This refers to student's concerns that is very rare and non-repeatable across other different student cohorts and communities. Some examples of such comments are: "more stern and louder", "XYZ topic being taught does not appear in the examination", "prefer to focus more on XYZ topic and drop ABC topic"

4.2. Taxonomy Development

The main component enabling the survey data to actionable use, is the establishment of a good taxonomy that leads to a pedagogical outcome in the areas of content, assessments, assignments, pace, difficulty, instructional levels, and student preferences. The following rules were followed when developing the taxonomy:

a) Close consultation with the operational users of the taxonomy, such as course chairs and lecturers, to ensure that student concerns could be mapped to a pedagogical need.
b) Sufficient granularity of topics to facilitate analysis but not too high such that it is unable to stand by itself as a class. For example, in certain comments, it is not possible to know if a student was referring to an assignment or a project. So "Assignments, Projects Not Enough Time" would be a better topic. To have a distinction between assignments and assessments, a separate topic "More Time For Assessments/Test" was also introduced.

Initially, unsupervised Machine Learning (ML) was used to hasten the development of the taxonomy. Unsupervised ML essentially requires no human intervention in the topic discovery process. Techniques such as Latent Dirichlet Allocation (LDA) or BERTopic (a hierarchical clustering algorithm that makes use of embeddings from BERT) were used. However, unsupervised ML did not work well due to two reasons. First, the dataset is highly unbalanced in nature. Some topics could be attributed to a few hundred students while others could only be attributed to a handful of students. Topics with low attribution had low chances of discovery. Second, the diversity of what was said in the comments was huge and it is important to distinguish the useful comments from the less useful ones (i.e. noise). Unsupervised ML is highly susceptible to noise. In the end, it was not practical to use unsupervised learning. Human knowledge and intervention were needed to develop the topics for the taxonomy and to prepare the data.

An examination of valid comments led to the development of the following taxonomy of 59 topics:

- Assignments Close Deadlines.
- Assignments, Projects Not Enough Time
- Assignments/Assessments Time Consuming/Workload
- Assignments/Assessments Are Too Difficult
- Better Content Organization
- Change Component Weightage
- Class Scheduling Issues
- Clearer Or Better Labsheet
- Coding Related Challenges
- Content Less Wordy/More Concise
- Focus On The Basics, Knowledge Gap Exists
- Group Mixing Issues
- Hard to Follow, Understand-Concepts
- Issues With Group Work
- Issues With Practicals and Assessments
- Issues With Presentation
- Issues With Self-Learning
- Issues With E-Learning
- Issues With Flipped Classroom
- Issues With Subject Teaching Plan
- Lecturer Go Through Materials Together
- Lecturers More Feedback/Consultation
- Lecturers Revise More
- Lesson Times Are Too Long
- Lesson Times Are Too Short
- Marking Rubric Is Unclear
- More choices In tools, techniques, projects
- More Engaging/Interactive Lecturers
- More Face To Face (F2F) Contact Time
- More Gamification Activities
- More Group Work /Discussions
- More (Home Based Learning)
- More Help, Clarity On Projects, Assignments, Tests
- More or Better Guides/Notes/Slides
- More Or Better Videos
- More Practice/Hands-On/Lab Activity
- More Quizzes/Mock Test
- More Relevant Content
- More Templates for Projects/Assignments
- More Theory To Build Foundational Understanding
- More Time For Assessments/Test
- More Time For Labs/Practice
- More Time/Attention On Certain Topic
- Need Explanations
- Need More Examples
- Prefer To Be Non-Graded Subject/Component
- Provide Answers To Aid Learning
- Provide Summary/Cheatsheet
- Release Materials and Info Earlier
- Subject Content Not-Up-To-Date, Buggy
- Subject Is Too Hard
- Subject Is Too Simple
- Teaching Pace Can Be Faster
- Teaching Pace Can Be Slower/Reduce Workload
- Technical Issues With Tools,Techniques
- Too Basic, More Depth On Certain Topic
- Too Boring
- Too Much Content
- Uninterested In the Subject

4.3. Data Preparation and Annotation

Manual effort to label each student's comment with an appropriate topic, was required to prepare sufficient high-quality data for the AI model to learn from. The following rules were followed for data labelling:

a) Each student comment was assigned a label according to the topic defined in the taxonomy.
b) Extraneous phrases or sentences in the comments not relevant to any topic in the taxonomy were removed.
c) Any single sentence in a comment resulting in multiple topics, were treated as follows:

   - Disambiguation by separation. If a student was talking about two different topics within a single sentence, it would be helpful to disambiguate the sentence by separating that sentence into two entries each with its own relevant topic. For example, sentence like "I find the lessons boring and that there isn't enough time to finish the practical test", could be split into two entries - "I find the lessons boring" and "I find that there isn't enough time to finish the practical test".
   - Multi-label scenario. If a student were talking about two different topics within a single sentence and topics are related to one other, that sentence should be preserved whole and tagged with multiple labels. To illustrate, "The slides could show more examples of students work from the previous batches", is a sentence which should be preserved whole. It would be provided two labels namely, "Need More Examples" and "More or Better Guides/Notes/Slides"

d) Student comments labelled as "Other Miscellaneous" were omitted for model training.

4.4. Model Development

Candidate machine learning models were adapted from Tunstall,et al. in Dealing With Few To No Labels (2022). The listing of the algorithms is sorted by the levels of computational cost, from the least to the most expensive. The aim was to understand the trade-offs between the model performance and the computational cost in model development and use. The use of LLMs would require the use of specialized hardware such as Graphic Processing Units (GPUs).

a) Naïve Bayes Classifier is a relatively simple probabilistic classifier that computes the probability of texts belonging to a class. This classifier's drawback is the simple assumption that relationships between words in a sentence does not matter, thereby losing the ability to differentiate text with contextual and semantic nuances.

b) Zero-Shot Machine Learning uses a pre-trained LLM model to figure out the relationship between the topic label and the student comment based on semantics. This method does not require any manual effort for text annotation as the model need not train on the text to learn about that relationship.

c) Nearest Neighbour Embedding. This approach uses a LLM to translate a piece of text into representative embeddings. The nearest neighbour algorithm then uses these embeddings to find a boundary space which encapsulates all student comments within a specific topic, as provided by the topic label.

d) Fine-Tune BERT. BERT is a LLM which can capture the contextual semantics in sentences well. BERT is fine-tuned when it is trained by learning the relationship between the annotated labels and text.

All the models were developed by adopting 70% of the data for training, 15% for validation and 15% for testing. The micro and macro averaged F1 scores were then calculated for each model on the test data, to estimate the model performance for different models.

4.5. Dashboard Development

A prototype dashboard design was created to enable practical use of the model. The survey also included another question: "Overall I am satisfied with this subject". The dashboard links the data on student satisfaction to the topics. The stakeholders wanted to understand the concerns of students with low satisfaction rating.

5. Results and Discussion

The model performance improved when developing the models on a successively bigger datasets with more training examples as shown in Figure 3 and Figure 4. By referring to the F1 scores which denotes how good the model is, the Fine-Tuned BERT model is the best performing and the Naïve Bayes Classifier is the least performing model.

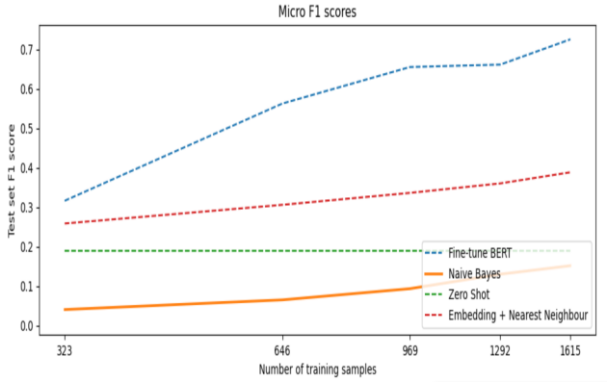

Figure 2: Model Performance based on Micro Avg F1 Scores

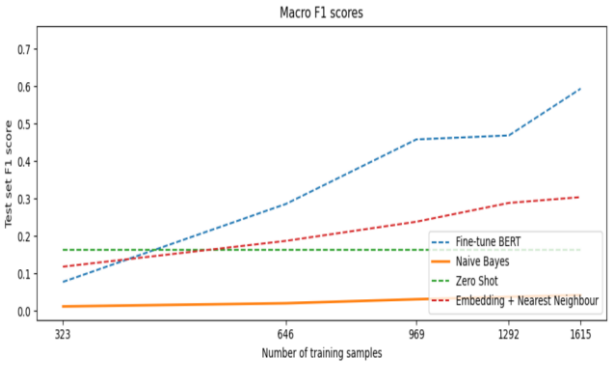

Figure 3: Model Performance based on Macro Avg F1 Scores

Table 1 shows that topics with more labelled instances for the models to train on (i.e., > 20) have significantly higher F1 score denoting better model performance.

Table 1: Macro Average F1 Scores of Topics with Few Labels and Many Labels

Avg F1 score of topics with many labelled instances	0.707
Avg F1 score of topics with few labelled Instances	0.407

The fine-tuned BERT model was given a threshold for the prediction probability at 0.5. Each comment was parsed into sentences for sentence level prediction and the following student comment (as illustration) was used to test the model:

"I could not run my code as the versions given in the lessons are outdated. I wish that the deadline for the final project could be extended by a week as I don't have enough time. Some my group members did not contribute much to the project, and I have to do most of the work"

The model was able provide the following tags:

- Assignments, Projects Not Enough Time
- Issues With Group Work
- Coding Related Challenges
- Subject Content Not-Up-To-Date, Buggy

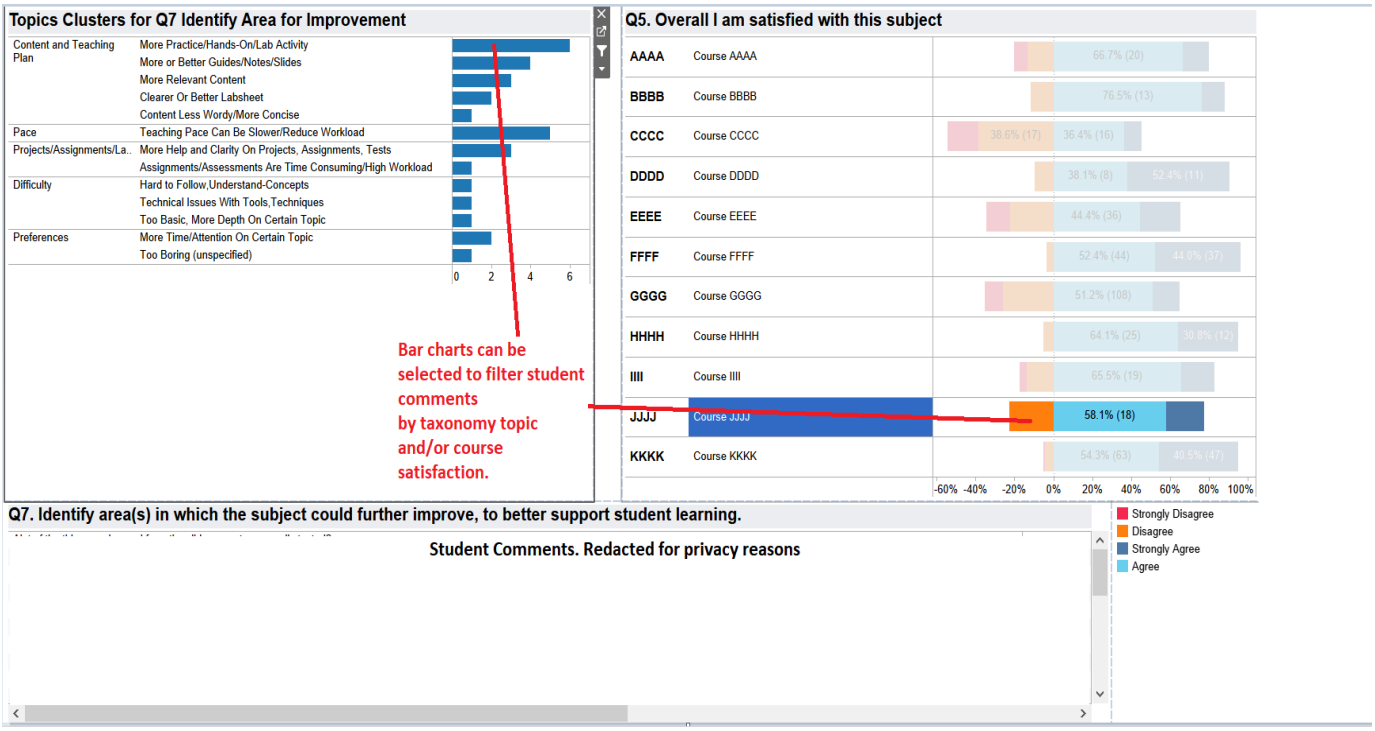

Figure 4: Dashboard linking topics for Q7 and Likert scores of Q5

Figure 4 provides a concept design of how a dashboard prototype can provide insights to improve student satisfaction by linking satisfaction levels with student concerns. The insights are for illustration only. A dashboard user could select Course JJJJ, followed by the orange band beside the course listing. This action zooms in on students with low course satisfaction in Course JJJJ. The foremost topic after zooming in, is shown as "More Practice/Hands-On/Lab Activity". The user could then select this foremost topic to zoom in on all the verbatim comments associated with that topic.

6. Conclusion

Data preparation proved to be the first hurdle requiring human understanding to determine initially which comments were useful or were irrelevant. The next hurdle was to label the relevant comments with the appropriate topic.

There are different NLP techniques in use for topic extraction. This paper identified the fine-tuning of BERT as the most suitable approach to map student comments to a pre-defined taxonomy.

Even though this technique required higher computational expense to train and to use, it offered exceptional model performance.

The above technique showed the possibility of using AI to label student survey data for dashboarding, to be available to stakeholders interested in monitoring student concerns according to a pedagogical definition, given the taxonomy.

7. Future Work

We wish to explore the use of various active learning strategies to keep the AI model relevant and up to date in line with the evolving needs of future cohorts of students.

To improve the model performance for topics which have fewer labels due to fewer student comments, the use of few-shot learning as suggested by Tunstall et al. in Efficient Few-Shot Learning (2022) and data augmentation with CHATGPT as suggested by Dai et al. (2023).

References

Dai, H., Liu, Z., Liao, W., Huang, X., Cao, Y., Wu, Z., Zhao, L., Xu, S., Liu, W., Liu, N., Li, S., Zhu, D., Cai, H., Sun, L., Li, Q., Shen, D., Liu, T., & Li, X. (2023). AugGPT: Leveraging ChatGPT for Text Data Augmentation. ArXiv. /abs/2302.13007

Devlin, J., Chang, M., Lee K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational
Linguistics.

Gottipati, S., Shankararaman, V., and Lin, J. (2018). Text analytics approach to extract course improvement suggestions from students' feedback. In Research and Practice in Technology Enhanced Learning (2018) 13:6

Tunstall, L., Werra, L. von, and Wolf, T. (2022). Dealing With Few To No Labels. In Natural Language Processing with Transformers, Revised Edition (chapter 9). O'Reilly Media, Inc. https://learning.oreilly.com/library/view/natural-language-processing

Tunstall, L., Reimers, N., Jo, U. E., Bates, L., Korat, D., Wasserblat, M., & Pereg, O. (2022). Efficient Few-Shot Learning Without Prompts. *ArXiv*. /abs/2209.11055