11th International Conference on Higher Education Advances (HEAd'25)
Universitat Politècnica de València, València, 2025
DOI: https://doi.org/10.4995/HEAd25.2025.20043

Applying Generative AI to Analyze Statements from Student Surveys

Axel Böttcher, Sebastian Dünnebeil, Veronika Thurner
HM Munich University of Applied Sciences, Department of Computer Science and Mathematics, Lothstraße 64, D-80335 München, Germany.

How to cite: Böttcher, A.; Dünnebeil, S.; Thurner, V. (2025). Applying Generative AI to Analyze Statements from Student Surveys. In: 11th International Conference on Higher Education Advances (HEAd'25). Valencia, 17-20 June 2025. https://doi.org/10.4995/HEAd25.2025.20043

Abstract

Applications based on Generative Artificial Intelligence (GenAI) already permeate many areas of higher education. In this paper, we discuss using GenAI to assist staff in extracting information from qualitative data gathered by student surveys.

Manually analysing surveys of large groups of students usually is a tedious task. In particular, categorising and evaluating open-ended questions can be very time consuming. As a means for effectively processing and analysing large amounts of student feedback, we experimented with using GenAI for this task. In this work, we describe a simple, inexpensive and easily adaptable setup for analysing student statements, and report on our experience with this approach. The goals are (1) to let GenAI assign each student statement to an adequate content category and to accurately determine its sentiment, and (2) to use GenAI to produce summaries of statements.

We implemented our approach twice, using two different GenAI products, namely the commercial OpenAI API and a self-hosted Llama model. Comparing the results, we find that both are valuable in the process of evaluating students' statements with respect to their sentiments and towards categorizing and summarizing them.

Keywords: *Student surveys, sentiment analysis, generative AI, evaluation, educational data mining.*

1. Introduction

Every academic year during the summer term, our university's quality management department performs an anonymous survey regarding various aspects of student satisfaction, situation, and well-being. The survey comprises several Likert-scaled questions as well as open questions, where students can freely formulate statements with their opinions and concerns. In order to be able to establish a context for the students' statements, the survey requests students to specify the department and study program they are enrolled in. Surveys are performed using an online evaluation system (evasys in our case). The completed survey forms are collected centrally for the whole university. Then, each department is provided with access to the survey results of their respective students.

This approach easily leads to hundreds of completed forms thus. Evaluating Likert-scaled questions is highly efficient when the results are available electronically. However, drawing valuable feedback from a large amount of more or less well-formulated answers to open questions can be a highly laborious task (Gray, 2023). Therefore, we tried to use readily available GenAI-tools to support the evaluation of freely formulated statements. It is quite easy to create results in a straightforward manner by feeding data to a tool like ChatGPT. This, however, requires carefully validating whether the reported results do in fact reflect the content of students' verbal feedback in a sufficiently precise way, as GenAI models have been observed to occasionally hallucinate (Farrelly & Baker, 2023).

Therefore, we started an analysis where we compared machine generated ratings for a manageable sample of statements from students' surveys against human expert rating. The questions we wanted to answer in this context are:

1. Can the use of GenAI provide an effortless yet meaningful overview of survey results?
2. Does a low-cost, data privacy compliant and self-hosted GenAI model perform worse than the commercial OpenAI variant?

2. Related Work

Much work has been done on various AI-based methods and techniques for sentiment analysis, reviewed for example by Jin et. al. (2023). A more recent review of this rapidly evolving field, focusing on the use of the transformer models that we consider for our work, can be found in Bashiri and Naderi (2024).

Shaik et. al. (2023) give a good overview of the available techniques for sentiment analysis and opinion mining, and they address many different areas of application in educational contexts. They discuss several algorithms and tools as of 2022. For more recent reviews of analyses of student feedback see, for example, the technically oriented review given by Zyout and Zyout (2024) or a more content-oriented review by Sunar & Khalid (2024). Dake & Gyimah (2023) present a use-case that is very similar to what we focus on in this paper. Their analysis of qualitative feedback from students with respect to sentiment is undertaken with the objective of facilitating the analysis of qualitative feedback. Our work, however, focusses on using existing and already trained GenAI-tools for the analysis of students' statements.

Finally, Schulhoff et al. (2024) discuss prompting techniques in detail, as another important aspect of using GenAI.

3. Approach

The student survey, which was carried out across our university in the summer term of 2024, resulted in a total of 148 completed questionnaires for our faculty in four computer science-related Bachelor's and three Master's degree programs. This corresponds to a participation of 8.6% of our students. The questionnaire consists of approximately 50 Likert-scaled questions and 15 open questions. In this paper, we focus on the evaluation of two open questions[1] that we consider as prototypical, namely:

1. *"What is the main reason why you would or would not recommend your degree program?"* This question asks for personal opinions, leading to answers comprising statements with both positive and negative sentiments, mixed in any order. Therefore, on each statement, we perform a GenAI-based sentiment analysis and categorization, thus attempting to categorize the items within the statement with respect to content, and to identify if the feedback is positive, negative or neutral.
2. *"In relation to which aspect of your studies would you need more flexibility?"* In contrast to the first question, this second question inherently unveils predominantly negative aspects. Therefore, we used GenAI to cluster which changes need to be addressed most urgently.

For our analysis, we use two GenAI-tools, namely:

1. The commercial API of OpenAI 3.5 (see: https://platform.openai.com/docs/overview). Tests showed that using API version 4.0 did in fact not produce better results, even though this version is newer and much more expensive; so we were happy to use version 3.5.
2. Meta's open-source AI model Llama 3.3, self-hosted via Docker container. We used the model out of the box and did not do any additional pre-training whatsoever.

3.1. Categorization and Sentiment Analysis of Statements

Analyzing a student's answer (or statement) to survey question 1 comprises two distinct operations: Firstly, the statement is decomposed into its constituent semantic entities. A semantic entity is a sequence of words that forms a single unit of meaning. Secondly, each semantic entity is analyzed in terms of assigning a category and a sentiment. We conducted some experiments to refine the prompt based on the findings reported by Schulhoff et al. (2024). In the end, we came up with the following, which performs both operations of our requirement in one prompt. The words in bold must be replaced with the item and statement to be analyzed, i.e., question and student answer. Triple quotation marks are an essential part of the prompt:

[1] Please note that for this paper, all questions, student answers, and prompts were translated from German using the support of deepl.com

> *The following statement is a student's answer to this evaluation item: """***item***""": """***statement***""". If this statement contains several semantic entities, then divide it into its semantic entities. Ignore dashes and numbering.*
> *Assign each entity a sentiment 'positive', 'negative', 'neutral' and use 'unclear' if the entity is not clearly positive, negative or neutral.*
> *Furthermore, assign exactly one of the categories 'content', 'teaching quality', 'study conditions', 'requirements', 'personal opinion' to each semantic entity.*
> *For each semantic entity, return only a single line in csv format of the form "semantic entity;category;sentiment". I don't need any detailed explanations.*

Although it is inherent to GenAI models that a repetition of the process does not necessarily lead to identical results, we observed none to only very little differences; this might be due to the lengthy prompt and comparatively short statements. However, the models make mistakes, as we will discuss in the results section.

To automate the analysis, we built a small program that reads students' statements from a file, generates the prompts, feeds them to the respective model and writes the results back to a file in csv format (comma separated values). This format can be post-processed further to produce textual or graphical summaries.

3.2. Automated Summaries of Student Answers

A second use of GenAI in our evaluation process was to use the tool to summarize items in the survey. We used it for items that did not leave too much room for emotion, such as the item *"In what aspect of your studies would you need more flexibility? (Please provide keywords)"*. Here, we only expect aspects that need improvement from the students' point of view. Assigning sentiments gives misleading results. The subsequent prompt, with the responses of the students incorporated, yielded constructive outcomes:

> *I work as a dean at a university. Summarize the following list of requests from our students for more flexibility and give me a list of the changes that need to be addressed most urgently.*

4. Results

Corresponding to section 3, we discuss the results for both sentiment analysis and for automatically generated summaries.

4.1. Results of Analyzing Statements with respect to Sentiment and Category

Figure 1 gives an overview of the analysis process for student replies using GenAI. The GenAI-tools we used made mistakes in both respects: decomposing questions into semantic units and assigning categories and sentiments. As an example for an erroneous decomposition, the statement *"Practical relevance and good professors + interesting subjects"* was handled as one statement by OpenAI, rather than being segmented into its three atomic semantic units as was done by Llama. On the other hand, the statement *"Lectures with a small number of students, making it easier to form study groups"* was split by both tools, leading to the senseless statement *"making it easier to form study groups"*.

Furthermore, OpenAI did not ignore dashes or numberings in students' statements, even though this was explicitely requested by the prompt. Even worse, in some cases OpenAI reported numbers or dashes as separate semantic entities. Therefore, we decided to automatically omit all semantic entities that consist four or less characters in length.

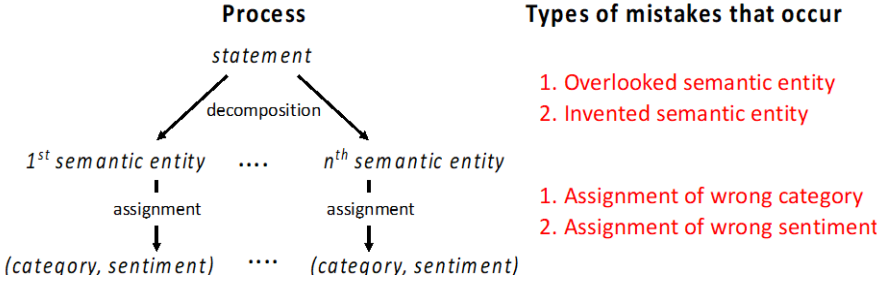
Process
Types of mistakes that occur
statement
decomposition
1st semantic entity
....
nth semantic entity
assignment
assignment
(category, sentiment)
....
(category, sentiment)
1. Overlooked semantic entity
2. Invented semantic entity
1. Assignment of wrong category
2. Assignment of wrong sentiment

Figure 1. The process of the analysis of answers, and possible errors that are generated by the GenAI-tools we used.

Over all, we analyzed a total of 107 answers to the question *"What is the main reason why you would or would not recommend your degree program?"*, by a human expert and by two GenAI-tools. Table 1 summarizes the results of detecting entities. The manual analysis detected 217 semantic entities. The zeros in the Human Expert column should be read with caution, as even humans can make mistakes and have to make decisions in case of ambiguities!

Table 1. Numerical results of the process of breaking down 107 statements into semantic entities.

	Human Expert	OpenAI	Llama
Semantic entities detected	217	221	241
Overlooked semantic entitites	0	6	1
Additional semantic entitites	0	10	25

One interesting observation was that, unlike Llama, OpenAI did some minor rewriting of the entities it detected, which made it difficult to automatically compare the results of the two GenAI-tools. For example the statement *"I think computer science is a very interesting course of study"* was shortened by OpenAI to *"Computer science is a very interesting course of study"*. Both models, however, rated this as a positive personal opinion.

Comparing the results generated in terms of categories and sentiments, we found that sentiments were assigned identically by both models in 98% of the cases. Table 2 shows the correlation between the two models in terms of categorization. We have 210 comparable units, resulting from an initial total of 217 (as detected by the human expert), of which 7 were overlooked by the GenAI-tools. For roughly 84% of the comparable items, both GenAI-tools agreed in their notion of the expressed sentiment.

Note that some statements are inherently ambiguous for human raters as well, e. g.: *"professionally very demanding"*. Both models categorized this statement as Content, OpenAI rated it as positive and Llama as neutral. We use this as hint to redesign question 1 for the next iteration of the survey by splitting it up into two separate questions, in order to explicitely ask for positive and for negative feedback, respectively.

Table 2. Number of matching and non-matching categorizations of the two models.

	Categorization by Llama				
Categorization by OpenAI	**Content**	**Qualityof Teaching**	**Study Conditions**	**Pre-requisites**	**Personal Opinion**
Content	**66**	5	5	1	4
Qualityof Teaching	2	**30**	1	0	1
Study Conditions	1	2	**59**	2	0
Pre-requisites	0	0	2	**4**	0
Personal Opinion	4	1	1	1	**14**

Finally, we have made some interesting observations that are worth mentioning:

1. Both models coped unexpectedly well with typos and spelling mistakes.
2. For four entities, the OpenAI API invented the additional category "level of difficulty". We believe that it is still inherent in current models for these types of errors to occur.

Altogether, the results are informative and useful. They could be forwarded to teaching and organizational staff as sorted lists or in condensed form as shown in Figure 2, which can easily be generated automatically from the results. On this basis, the faculty's decision-makers then have a starting point for deciding where it is worth taking a closer look (study conditions in this case).

4.2. Results of Automated Summaries of Student Answers

Furthermore, we used both models to summarize a total of 52 answers to the question *"In what aspect of your studies would you need more flexibility?"*. OpenAI identified six areas in which the most, and most urgent, need for change was expressed, whereas Llama detected seven. Surprisingly, the categories found by OpenAI, were much more commonplace and less concrete than those detected by Llama, as shown in Table 3.

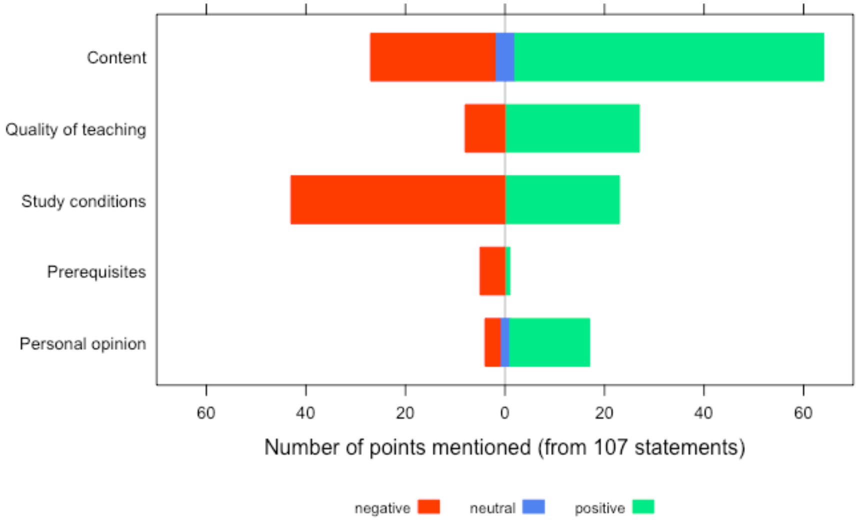

Figure 2: Automatically generated Likert-like plot from the data taken from 107 free text statements and processed via OpenAI. Includes 217 out of 221 detected semantic units.

Table 3. Comparison of the summaries that the two language models created from 52 answers. Items that are considered to be too commonplace are marked with an asterisk "*".

OpenAI	Llama
Flexibility in choosing and designing courses	More options for electives
Expansion of online learning opportunities	More online offers and hybrid lectures
Practical lectures and internships	Adjustment of the internship semester
Reduce pressure and strain	
Course organisation*	More flexibility in the selection of professors and courses
Organization of study*	More flexible timetables
	More support for commuters
	Longer examination periods and more flexible examination schedules

5. Conclusions and Future Work

Our analysis shows that a self-hosted Llama model leads to results whose quality is comparable to the commercial version of OpenAI's API. We recommend creating a collection of several hundred statements from evaluations that can be agreed on as being relevant, unambiguous and well understood by humans (see e. g. Zyout and Zyout 2022). This set of statements can then be used as test data to calibrate a GenAI-based analysis software. Furthermore, errors identified in the GenAI-tools' classification might indicate weaknesses in the survey, and thus might help to improve the survey's clarity. Automatically cleaning, organizing, and judging qualitative feedback offers a huge potential for qualitative analysis, while reducing human workload at the same time.

Lessons learned from the analysis of students' answers to the questionnaire provided many hints for improvement for the next cycle of this survey. Thus, surveys can be adapted in a way that make them easier to interpret for AI-based sentiment analysis. Having accomplished this, hundreds of replies to open questions can be analyzed and summarized quickly and with little human effort, thus offering great potential for a continuous improvement process.

A non-AI based finding is that we need to increase students' participation rate in the future, as only 8.6% of our students completed the survey. This participation rate could be increased by asking students to complete the survey form during class time.

References

Bashiri, H., & Naderi, H. (2024). Comprehensive review and comparative analysis of transformer models in sentiment analysis. Knowledge and Information Systems, 66(12), 7305–7361. https://doi.org/10.1007/s10115-024-02214-3

Dake, D.K., & Gyimah, E. (2023). Using sentiment analysis to evaluate qualitative students' responses. Educ Inf Technol 28, 4629–4647. https://doi.org/10.1007/s10639-022-11349-1

Farrelly, T., & Baker, N. (2023). Generative Artificial Intelligence: Implications and Considerations for Higher Education Practice. Education Sciences 13(11), 1109. https://doi.org/10.3390/educsci13111109.

Gray, D. (2023). Doing Research in the Real World, SAGE Publications Ltd, London, 5th ed.

Jin, Y., Cheng, K., Wang, X., & Cai, L. (2023). A Review of Text Sentiment Analysis Methods and Applications. Frontiers in Business, Economics and Management. 10. 58-64. https://doi.org/10.54097/fbem.v10i1.10171.

Schulhoff, S., et al. (2024). The Prompt Report: A Systematic Survey of Prompting Techniques (Version 3). arXiv. https://doi.org/10.48550/ARXIV.2406.06608

Shaik, T., Tao, X., Dann, C., Xie, H., Li, Y., & Galligan, L. (2023). Sentiment analysis and opinion mining on educational data: A survey. Natural Language Processing Journal, 2, 100003. https://doi.org/10.1016/j.nlp.2022.100003

Sunar, A. S., & Khalid, M. S. (2024). Natural Language Processing of Student's Feedback to Instructors: A Systematic Review. IEEE Transactions on Learning Technologies, 17, 741–753. https://doi.org/10.1109/TLT.2023.3330531

Zyout, I., & Zyout, M. (2024). Sentiment analysis of student feedback using attention-based RNN and transformer embedding. IAES International Journal of Artificial Intelligence (IJ-AI), 13(2), 2173. https://doi.org/10.11591/ijai.v13.i2.pp2173-2184