Journal Pre-Proofs

Cutting-edge Technologies for Analyzing Student Feedback to Inform Institutional Decision-making in Higher Education

Authors: Sabur Butt, Sandra Dennis Núñez Daruich, Joanna Alvarado-Uribe, Hector G. Ceballos

To appear in: *Foresight and STI Governance*

Received Date: 28.03.2025
Accepted Date: 15.08.2025

Please cite this article as: Butt S., Núñez Daruich S.D., Alvarado-Uribe J., Ceballos H.G. (2025) Cutting-edge Technologies for Analyzing Student Feedback to Inform Institutional Decision-making in Higher Education. *Foresight and STI Governance*, vol. 19, no 4 (forthcoming).

This article is a part of the special issue **"Transforming Education in the 21st Century: Foresight and Sustainable Development"**

Edited by Asad Abbas (Writing Lab, Institute for the Future of Education, Tecnologico de Monterrey, Monterrey, Mexico), Ahsan Ali (Zhejiang Sci-Tech University, Hangzhou, China), Jose Luis Martin-Nuñez (Instituto de Ciencias de la Educación, Universidad Politécnica de Madrid, Madrid, Spain), Mehul Mahrishi (Swami Keshvanand Institute of Technology, Management & Gramothan, Jaipur, India)

Cutting-edge Technologies for Analyzing Student Feedback to Inform Institutional Decision-making in Higher Education

Sabur Butt[1],
Sandra Dennis Núñez Daruich[2],
Joanna Alvarado-Uribe[1,3],
Hector G. Ceballos[1]

[1]Institute for the Future of Education, Tecnológico de Monterrey, Monterrey, Mexico
[2]Vicerectory of Academic Affairs, Tecnológico de Monterrey, Monterrey, Mexico
[3]School of Engineering and Sciences, Tecnológico de Monterrey, Monterrey, Mexico

Author's E-mail address: saburb@tec.mx, snunezda@tec.mx, joanna.alvarado@tec.mx, ceballos@tec.mx

ABSTRACT

Aspect-Based Sentiment Analysis (ABSA) has emerged as a powerful tool for deriving actionable insights from qualitative feedback in education. This study presents a multitask learning framework to analyze student evaluations of teaching (SET) by extracting and classifying opinions on specific aspects of teaching performance. Leveraging a novel and first open-sourced dataset of 6,025 Spanish-language comments, the proposed framework integrates opinion segmentation and multi-label classification to capture nuanced feedback on nine predefined aspects, such as "Teaching Quality" and "Classroom Atmosphere." Applications of this approach extend beyond SET analysis, offering valuable insights for course improvement, faculty assessment, and institutional decision-making in higher education. The paper compares the performance of fine-tuned transformers (BERT and RoBERTa) with large language models (LLMs), including GPT-4o, GPT4o-mini, and LLama-3.1-8B, using both fine-tuned and Few-shot Chain of Thought (CoT) methodologies. Evaluation results reveal that fine-tuned GPT-4o outperformed all other models, achieving a weighted F1-score of 0.69 for positive aspects and 0.79 for negative aspects, while Few-shot CoT approaches demonstrated competitive performance with greater scalability and interpretability. Our findings demonstrate the framework's potential to transform unstructured feedback into structured insights, aiding educators and institutions in enhancing teaching quality and student engagement.

Key Words: Aspect-Based Sentiment Analysis (ABSA); Student Evaluations of Teaching (SET); Opinion Segmentation; Multi-label Classification; Large Language Models (LLMs); Few-shot Chain of Thought (CoT)

1. INTRODUCTION

Uncovering and interpreting opinions from user-generated content has become crucial across various domains, particularly in education, where analyzing different forms of feedback can significantly enhance the overall experience of education. Student evaluations of teaching (SET) (Marsh and Roche, 1997; Nowell et al., 2010) offer valuable insights that help educators refine their methods and improve student engagement. However, given the substantial volume of feedback collected each semester, manually analyzing these responses to extract actionable insights is neither practical nor scalable. According to the Carnegie Classification[1], small colleges, typically private institutions like Colgate, Grinnell, and Reed, have fewer than 5,000 students, collecting approximately 25,000 SETs per semester if each student takes 5 courses. In contrast, large universities with more than 15,000 students, such as UCLA or the University of Texas at Austin, would gather at least 75,000 SETs per semester, with some huge public universities like the University of Michigan and UC Berkeley collecting over 150,000 SETs given their student populations exceeding 30,000. This challenge underscores the need for automated frameworks capable of processing sentiments embedded within unstructured text. As a result, sentiment analysis and opinion mining techniques have become essential tools for efficiently interpreting qualitative data in educational contexts (Zhou and Ye, 2023).

In Natural Language Processing (NLP), Aspect-Based Sentiment Analysis (ABSA) (Gallardo et al., 2023; Nazir et al., 2020) has emerged as a powerful framework to address the limitations of conventional sentiment analysis, which typically focuses on predicting sentiments at the sentence or document level. ABSA generally involves two key phases: Aspect Extraction (AE) and Aspect Sentiment Analysis (ASA). AE focuses on identifying aspects explicitly mentioned in the text or implied through context, such as specific teaching attributes or Opinion Target Expressions (OTEs). Once identified, ASA classifies the sentiment polarity associated with these aspects. Conventional methods often assume that a single sentiment is conveyed toward a unified topic within the text. However, in real-world scenarios like SET, multiple sentiments may coexist within a single response, each targeting different aspects of the teaching experience. ABSA shifts the focus from holistic sentiment evaluation to a more granular analysis of specific entities and their associated attributes. Hence, the primary objective of ABSA in this domain is to construct a detailed opinion summary at the aspect level, providing nuanced and actionable insights into teaching performance.

[1] https://carnegieclassifications.acenet.edu/

This paper focuses on the crucial problem of aspect and opinion extraction in SET. Specifically, the study addresses two subtasks: opinion segmentation and multi-label classification of aspects. Opinion segmentation aims to separate student feedback into positive and negative sentiments, while multi-label classification identifies multiple aspects related to these sentiments. The significance of this work lies in its ability to transform unstructured feedback into structured, actionable insights. Traditional sentiment analysis methods, which classify feedback as merely positive, negative, or neutral, fail to capture the diversity of opinions students express about different aspects of teaching. For example, a student might praise a teacher's subject knowledge while criticizing their classroom management within the same response. By identifying and categorizing these specific aspects, our approach offers targeted insights that help educators improve their teaching practices and enable institutions to make informed decisions about faculty performance.

However, this problem presents several significant challenges (Butt et al., 2023; Nazir et al., 2020). Student feedback often contains complex and subjective language, including idiomatic expressions, colloquialisms, and implicit sentiments. For instance, the Spanish phrase "Es un crack," which literally means "He's a crack," is a high compliment in the educational context but could be misinterpreted as negative in other settings. The model must not only interpret such nuances but also accurately associate them with the relevant teaching aspects. Additionally, the morphological richness of the Spanish language introduces further complexity. Words can take on various forms depending on grammatical context, complicating the task of extracting consistent aspect-opinion pairs. Moreover, feedback is often unstructured, with multiple sentiments directed at different aspects within a single sentence, making segmentation and classification even more challenging. Another difficulty arises from feedback that includes unrelated or ambiguous content. Students sometimes provide general advice for future course takers or discuss factors unrelated to teaching performance. Our model must distinguish and exclude such instances, marking them as "null/not-mentioned" to ensure the analysis remains focused on actionable teaching insights.

This paper seeks to address these challenges by presenting fine-tuned transformer-based models like BERT and RoBERTa alongside state-of-the-art LLMs, such as GPT-4o and LLama-3.1-8B, the framework captures the complexities of multilingual and multi-faceted feedback. This paper makes the following contributions to the field of Aspect-Based Sentiment Analysis (ABSA) in Student Evaluations of Teaching (SET):

1. Dataset Expansion and Annotation: The study expands an existing dataset (2026 comments) (Butt et al., 2023) to include 6,025 Spanish-language student comments, ensuring balanced

representation across multiple campuses, courses, and domains. A rigorous multi-stage annotation process guarantees high-quality data suitable for training advanced models. (See Section 3: Dataset)

2. A Multitask Learning Framework: We introduce a novel multitask learning framework that integrates opinion segmentation and multi-label classification to extract detailed, actionable insights from unstructured student feedback. This framework bridges the gap between conventional sentiment analysis methods and the nuanced demands of SET analysis. (See Section 4: Methodology)
3. Error Analysis and Future Recommendations: The paper identifies key challenges, problems, and limitations of the current state-of-the-art algorithms, such as GPT-4o, GPT4o-mini, and Llama-3. Detailed error analyses are provided for the models and recommendations for improvement are included. (See Section 5.4: Error Analysis)

2. LITERATURE REVIEW

The increasing use of AI in education research exemplifies the convergence of computational linguistics, pedagogy, and behavioral science. By analyzing student feedback through ABSA and LLMs, this study contributes to the interdisciplinary discourse on how AI-driven insights can reshape educational strategies and improve teaching effectiveness. To limit the scope of this literature review, we will concentrate on studies specifically addressing Aspect-Based Sentiment Analysis (ABSA) and Aspect-Based Opinion Mining (ABOM) within the education domain.

2.1 Pre-LLM era

Early research in ABSA primarily concentrated on polarity assessments, employing binary or multi-class settings to demonstrate sentiment polarities for education-related aspects. For instance, (Chauhan et al., 2019) demonstrated the polarity-based sentiment analyses based on several aspects in improving the teaching-learning process through a detailed examination of student feedback. They proposed a hybrid approach that leveraged machine learning and lexicon-based methods to extract and analyze opinions on aspects of education, such as teaching methods and course content. They employed part-of-speech tagging and dependency parsing to identify relevant aspect-opinion pairs. Sentiments were then classified into positive or negative using Naïve Bayes and sentiment lexicon-based classifiers. They got an F1 score of 0.80 for teaching aspects and 0.81 for course aspects.

In another study, (Dehbozorgi and Mohandoss, 2021) explored the application of Aspect-Based Emotion Analysis (ABEA) on student speech in collaborative learning environments to predict performance. Their study focused on identifying correlations between emotions expressed during teamwork and students' academic performance in an introductory programming course. The authors implemented a K-Nearest Neighbor (KNN) regression model, using the combination of emotion classes(happiness, anger, sadness, surprise, and fear) and aspect patterns as feature vectors to predict performance scores. Similarly, (Chen et al., 2024) focused on the application of Aspect-Based Opinion Mining (ABOM) to analyze course reviews in online learning environments. Given the rapid growth of online education, coupled with persistently high dropout rates and suboptimal learning outcomes, student feedback provided a crucial resource for improvement. Learners frequently posted reviews containing fine-grained aspects and opinions, which offered valuable insights for educators to enhance course quality and learning performance. The authors proposed a multi-step approach for extracting and analyzing aspects and opinions from course reviews. First, they applied Latent Dirichlet Allocation (LDA) to identify topics and associated keywords, which served as core aspects of the course. Next, using dependency syntax analysis, they designed a set of rules and algorithms to extract fine-grained aspects and opinions, generating an aspect-based opinion candidate set. Finally, the sentiment polarity of each opinion was determined using a combination of dictionary-based methods and Pointwise Mutual Information (PMI), which provided a sentiment score for the associated aspects. These approaches, however, heavily rely on predefined sentiment dictionaries, which may not adequately capture domain-specific expressions, idiomatic phrases, or contextual nuances in educational feedback. This can lead to misclassification of sentiments, especially for ambiguous or mixed-opinion sentences. Additionally, manual or semi-automated methods, such as those involving manual annotation or static lexicons, limit the scalability of existing approaches (Chauhan et al., 2019). Our framework is designed for large-scale automation, making it suitable for analyzing extensive datasets typical in educational institutions.

2.2 Post-LLM era

With the advent of Large Language Models (LLMs), the ABSA landscape witnessed significant advancements. (Mughal et al., 2024) conducted a comparative analysis of ABSA models, highlighting the evolution from traditional machine learning to modern deep learning architectures and transformer-based models like BERT. The introduction of LLMs, such as GPT-3.5 and PaLM, marked a transformative era, with these models demonstrating superior generalization capabilities and exceptional performance in handling implicit sentiments across various ABSA tasks. Studies emphasized the ability of LLMs to match or surpass fine-tuned transformer models in zero-shot and few-shot settings (Hadi et

al., 2024; Zhang et al., 2024; Zhong et al., 2023). (Shaikh et al., 2023) explored ChatGPT's performance in sentiment classification on a richly annotated dataset of student feedback. Their results revealed an impressive F1-score of 88%, outperforming both traditional deep learning and transformer-based models. The study underscored the importance of sentiment analysis in academic decision-making, providing actionable insights to enhance teaching quality. Similarly, (Qiao et al., 2023) demonstrated ChatGPT's efficacy in automatically labeling MOOC datasets, achieving higher accuracy and consistency compared to manual annotation. This approach reduced reliance on manual labor, highlighting LLMs' potential for scalable and reliable dataset labeling.

Building on early LLMs, (Butt et al., 2023) introduced MF-SET, a multi-task learning framework for ABSA in Student Evaluations of Teaching (SET). The framework addressed gaps in non-English educational feedback, particularly Spanish, by incorporating opinion segmentation, multi-class classification, and multi-label classification. Utilizing BETO (a Spanish BERT model) and GPT-3 Ada, the framework captured both explicit and implicit sentiments across nine aspects, including Teacher Quality, Teaching Content, Teaching Attitude, Teaching Ability, Teaching Effectiveness, Teacher Image, Teaching Method, Teacher-Student Relationship, and Classroom Atmosphere. Notably, Task 2 (multi-label classification and opinion segmentation) achieved micro F1-scores of 0.68 for positive aspects and 0.72 for negative aspects, demonstrating the model's robustness in handling multi-faceted and multilingual feedback.

Despite its advancements, MF-SET faced challenges related to dataset size and imbalance, which limited its applicability for certain models and tasks. To overcome these limitations, our study expands on the dataset by introducing additional annotated instances, ensuring a more diverse and realistic collection of student feedback. Additionally, we explore a broader array of cutting-edge techniques, enhancing the model's ability to generalize and deliver actionable insights across various educational contexts.

3. DATASET

This study focused on multi-label aspect detection and opinion segmentation, originally outlined in (Butt et al., 2023), which involves extracting aspects and opinions from student feedback. While the original experiment was conducted on 2,026 comments, we recognized the need to expand the dataset to enhance its suitability for deep learning methods and smaller variants of Large Language Models (3B, 7B, and 13B parameters). The aspects that can be found in the student feedback are defined below:

- Teacher quality: Referred to teachers' stable personal characteristics, including personality, temper, and morality.
- Teaching content: Teaching content refers to teachers' subject knowledge or life experience passed on to students.
- Teaching attitude: Referred to teachers' attitudes towards teaching and students.
- Teaching ability: Referred to teachers' knowledge level and classroom management.
- Teaching effectiveness: Referred to students' achievement, interest, and self-confidence due to the teaching.
- Teacher image: Referred to teachers' physical characteristics, including dress and posture.
- Teaching method: Referred to teachers' methods of instruction.
- Teacher-student relationship: Refers to the relationship between teachers and students formed in the education process and their attitudes towards one another.
- Classroom atmosphere: Referred to the overall attitudes and emotions within the teaching environment.
- Null: The absence of an aspect or opinion related to the class was marked as null.

Our expanded dataset comprises survey responses from various campuses and departments of Tecnológico de Monterrey, Mexico. The original data before filtering contained 26 unique campuses, and 781 unique courses, and around 9406 course identifiers. The data was collected during the February to June 2023 semester. Initially, 123,444 comments were gathered, with only those in Spanish retained through a language detection filter. Given the challenges and costs associated with human annotation at this scale, we applied further filtering to ensure an even distribution of comments across different campuses, domains, and courses. To achieve balanced sampling, we limited the selection to a maximum of two comments per course identifier, ensuring fair representation across all courses without favoring those with disproportionately high feedback volumes. Table 1 shows some examples from the dataset. Our dataset is open-sourced and publicly available for research[2]. The metadata description of the dataset is as follows:

@data{FK2/QSKJRM_2025, author = {Butt, Sabur and Sandra Dennis Núñez Daruich and Alvarado-Uribe,Joanna and Ceballos Héctor G}, publisher = {Tecnológico de Monterrey}, title = {{Multi-label Classification Dataset for Student Evaluation of Teaching (SET)}}, UNF = {UNF:6:MBsfls9pietCgLbZ/6kEpQ==}, year = {2025}, version = {V1}, doi = {10.57687/FK2/QSKJRM}, url = {https://doi.org/10.57687/FK2/QSKJRM} }

[2] https://datahub.tec.mx/dataset.xhtml?persistentId=doi:10.57687/FK2/QSKJRM

Text	Positive aspects	Positive opinions	Negative aspects	Negative opinions
Es estricto en la dinámica de la clase y en la ponderación del examen, conoce mucho sobre los temas que imparte y siempre se empeña en dar la explicación a leyes, teoremas que no siempre es común saber. (He is strict in the dynamics of the class and in the weighting of the exam, he knows a lot about the subjects he teaches and always insists on giving an explanation to laws, theorems that are not always common to know.)	Teacher attitude, Teaching ability	conoce mucho sobre los temas que imparte y siempre se empeña en dar la explicación a leyes, teoremas que no siempre es común saber. (He knows a lot about the subjects he teaches and always strives to explain laws, theorems that it is not always common to know.)	Teaching method	Es estricto en la dinámica de la clase y en la ponderación del examen (He is strict in the dynamics of the class and in the weighting of the exam)
Es una maestra que procura explicar paso a paso las metodologías a realizar y ejemplificar con ejercicios (She is a teacher who tries to explain step by step the methodologies to be carried out and exemplify with exercises)	Teaching content, Teaching method	procura explicar paso a paso las metodologías a realizar y ejemplificar con ejercicios (Tries to explain step by step the methodologies to be carried out and exemplify with exercises)	NULL	NULL

Table 1. The table shows the examples from the dataset

3.1 Data annotation and guidelines

The annotation process was carried out in multiple stages to ensure accuracy and consistency. Initially, two annotators—one female and one male—were hired to label the data. Recognizing the need for enhanced consistency and conflict resolution in the final labeling, three additional experts (two males and one female) were brought in, bringing the total to five annotators. These experts were in their 3rd or 4th years of college, fluent in Spanish, and citizens of Spanish-speaking countries. Their tasks included resolving annotation conflicts, managing sensitive information, and verifying the overall consistency of annotations. All personal names in the dataset were replaced with the placeholder tag *[name]*, and institutional references were masked with *[institute]*. Additionally, checks were performed to address any errors related to string consistency and formatting to maintain a high standard of data quality. Each opinion, whether positive or negative, was required to contain exactly two aspects. In cases where more than two aspects were present, the annotators were instructed to select the two most significant ones. They were expected to memorize the aspect definitions to ensure their choices were well-informed and consistent. Annotations had to follow a strict format, such as *[`x', `y']* or *[`not-mentioned']*, with any sensitive data being highlighted for redaction. Special attention was given to cases where the two initial annotators disagreed, with these cases reviewed by the conflict-resolution experts to ensure uniformity.

Given the natural overlap between some of the aspects—such as ‘Teaching Content’ and ‘Teaching Method’—we adopted a multi-label classification approach in our machine learning framework. This decision was motivated by the fact that students’ responses often reference multiple interrelated aspects, and forcing a single-label classification would not accurately reflect the complexity of their feedback. Allowing multi-label annotations enabled us to capture the richness of these opinions more effectively, and also aligned with the guidelines followed during manual annotation.

This multi-layered annotation process, with its rigorous guidelines and multiple stages of review, ensured the creation of a reliable, high-quality dataset. We achieved an inter-annotator agreement score of Cohen's Kappa for Positive Aspects: 0.52 and Cohen's Kappa for Negative Aspects: 0.75 between two initial annotators. More dissent in the positive labels is due to different aspects being prioritized in the comments, and also wrong aspects in some cases, which were corrected in the final annotations.

3.2 Dataset Statistics

Figure 1 shows the label distribution between negative and positive aspects. We can see that positive and negative aspects have different trends between the aspects. Although the negative labels are fewer in number, they identify the problems with teaching methodology the most, and whether the teaching was effective in their opinion, whereas teacher quality was the most identified aspect among positive aspects. After filtering, the final dataset consists of 6,025 comments. On average, each comment contains 1.30 sentences, with an overall average word count of 21.25 words per comment. Additionally, the average word count per sentence is 16.30 words. The wordclouds for all positive and negative aspects can be seen in Appendix (Section 9.1 and 9.2).

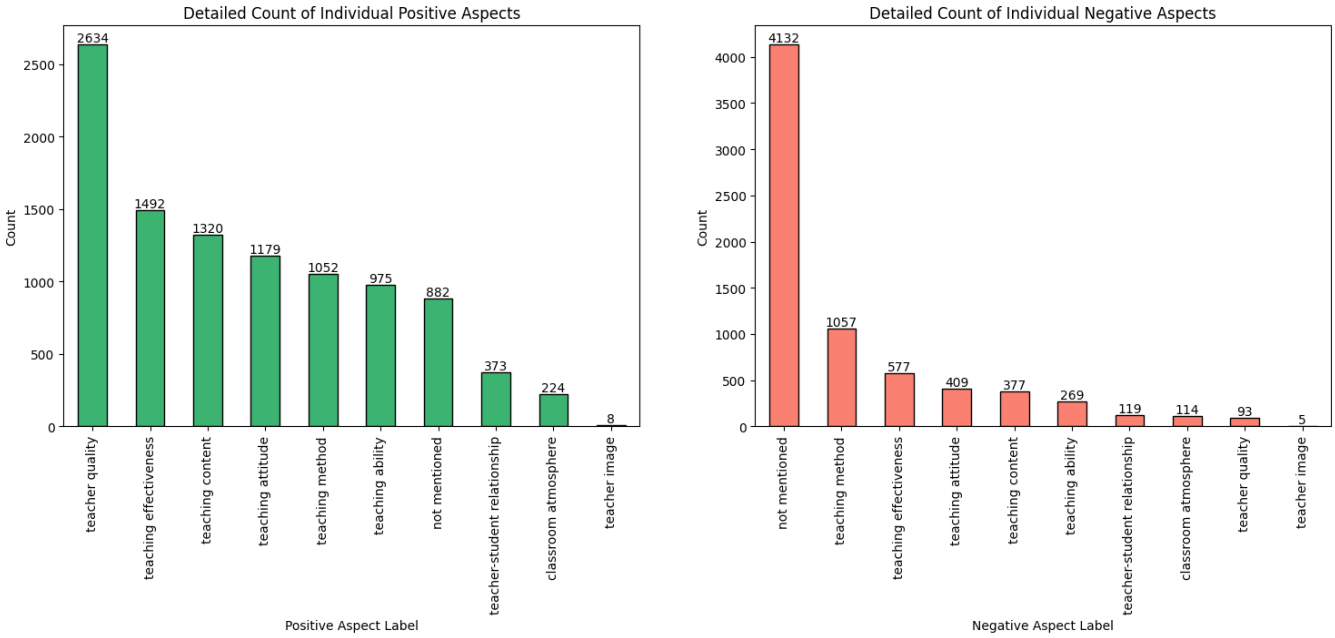


Fig. 1. The figure shows the class distribution between positive and negative aspects

4. METHODOLOGY

4.1 Finetuned Transformers

Although we consider advanced large language models (LLMs) to be the future of multitask setups, particularly for complex multi-label classification, we initiated our experiments by gauging the performance of transformers. By utilizing BERT (Kenton and Toutanova, 2019) and RoBERTa (Liu, 2019), pre-trained models within the transformer family, we aimed to capture both linguistic and contextual nuances within each feedback entry. We applied Optuna to optimize our hyperparameters, focusing on factors like the number of training epochs (6 to 10), learning rate (between $1*10^{-5}$ and 1*

10^{-3}), and batch sizes (8 and 16) for both training and evaluation. The experiments were conducted on base variants of both transformers. The transformer experiments were conducted using 5-fold cross-validation.

4.2 Large Language Models

4.2.1 Finetuned Models

Fine-tuned models leverage the pre-trained capabilities of LLMs by tailoring them to specific datasets and tasks. This approach enables the models to capture domain-specific nuances and achieve higher performance on specialized tasks like Aspect-Based Sentiment Analysis (ABSA). For this study, we focused on fine-tuning advanced LLMs such as GPT-4o, and GPT4o-mini. These models were selected based on their demonstrated efficiency in handling multilingual data, contextual embeddings, and complex multitask frameworks.

The fine-tuning process required the dataset to be formatted as structured prompt-completion pairs. Each prompt provided a clear task description, definitions of all possible aspects (e.g., "Teacher quality," "Teaching method"), and specific rules for annotation. Additionally, five examples were included to illustrate how the model should extract positive and negative aspects and opinions from the text. These examples demonstrated the reasoning process necessary to associate feedback with relevant aspects.

The completion format contained extracted aspects, opinions, and reasoning in a structured JSON schema. Each completion followed a strict guideline to ensure consistency, with exactly two aspects (positive and/or negative) selected for each instance. The reasoning provided a concise justification for the selection of aspects and sentiments. An example of the prompt format included:

- A clear explanation of the task, including definitions of aspects, rules for annotation, and expected output format.
- Five examples showing the association between feedback, aspects, and opinions.
- The student feedback text is to be analyzed.

In this experiment, the dataset was divided into two halves to maximize representation across all instances during training and evaluation. Each half was used alternately for fine-tuning and testing, ensuring robust performance across diverse feedback samples. Specifically, we trained two fine-tuned models for each LLM (e.g., GPT-4o, GPT4o-mini, and Meta-Llama-3.1-8B). The first model was trained using the first half of the dataset and evaluated on the second half, while the second model was trained on the second half and evaluated on the first. This strategy ensured that all data instances contributed to both training and evaluation without overlap, providing a fair assessment of the model's

generalization capabilities. We gave the combined results of each model by averaging all evaluation metrics.

4.2.2 *Few-shot Chain of Thought (CoT)*

Few-shot learning with Chain of Thought (CoT) reasoning enables large language models (LLMs) to generate step-by-step explanations for their predictions, enhancing interpretability and reducing reliance on extensive fine-tuning. We used models such as GPT-4o, GPT4o-mini, and Meta-Llama-3.1-8B-Instruct-Turbo with few shot CoT setup because of their ability to respond with json schemas, which makes evaluations scalable, less hallucinable, and less human-intensive for multilabel tasks. Similar to Section 4.2.1, the prompt for Few-shot CoT consisted of three primary components:

- A clear explanation of the task, including definitions of aspects, rules for annotation, and expected output format.
- Five examples demonstrating how to associate student feedback with predefined aspects and sentiments. Each example included extracted aspects, corresponding opinions, and a reasoning process.
- The student feedback text to be analyzed, provided at the end of the prompt, where the model was expected to generate structured outputs in a predefined format.

4.3 Evaluation

Given the multi-label nature of the task and the significant imbalance in class distributions, the evaluation is performed using standard classification metrics: precision, recall, F1-score, and their corresponding macro, micro, weighted, and sample averages. These metrics provide a comprehensive understanding of the models' capabilities across individual labels and overall performance. Per-class precision, recall, and F1-scores are crucial for understanding how well the models perform on each label. This is particularly important for this task because certain labels, such as "Not Mentioned" and "Teaching Quality," dominate the dataset in terms of support, while others, like "Teacher Image" and "Classroom Atmosphere," are rare. Precision evaluates the relevance of predicted instances, which is critical for avoiding false positives, especially for rare labels. For example, incorrectly labeling a text as "Teacher Image" when it does not apply may distort insights from the analysis. Recall ensures that the models capture the maximum number of true instances for each label, which is particularly relevant in educational settings where missing relevant feedback could lead to incomplete conclusions. The F1-score provides a harmonic mean of precision and recall, striking a balance between these two metrics.

The macro average treats all labels equally, irrespective of their frequency. This metric is critical for this task because it ensures fair representation of underrepresented labels like "Teacher Image" and "Classroom Atmosphere," which might otherwise be overshadowed by dominant labels. Macro F1-score highlights the models' ability to perform consistently across all classes, making it ideal for a task where even rare feedback categories hold significant importance. The weighted average accounts for class imbalance by weighting the per-class metrics by the number of instances in each class. This metric provides a performance measure that reflects the dataset's inherent class distribution. It is particularly relevant for this task because it ensures that the overall evaluation is not skewed by poor performance on rare labels, while still reflecting the importance of correctly predicting prevalent classes like "Teaching Quality" and "Teaching Content." Weighted F1-score is often more reliable when the real-world importance of classes correlates with their frequency in the dataset. And hence, in this case, we will compare different models using the weighted F1 scores. The micro average provides a single performance measure across all classes. While this metric is often suitable for tasks with balanced data, its heavy bias towards dominant classes makes it less ideal for tasks like this one. However, we are indicating these scores as a secondary indicator of overall performance.

For the purposes of this study, the macro F1-score is chosen as the primary metric for model evaluation. This decision is driven by the need to fairly evaluate the performance of the models across all classes, irrespective of their frequency. The weighted F1-score is reported as a complementary metric to account for the influence of class distribution in real-world scenarios. Per-class metrics are also analyzed to identify the models' strengths and weaknesses on individual labels.

5. Results and Discussion

The results of this study demonstrate the efficacy of various models and methodologies applied to Aspect-Based Sentiment Analysis (ABSA) on student feedback. This section highlights key insights derived from the results presented in Tables 2 to 7.

5.1 Performance of Fine-Tuned Transformer Models

Table 2 and Table 3 present the performance of fine-tuned transformer models on positive and negative aspects, respectively. Among the models, RoBERTa consistently outperformed BERT in terms of precision, recall, and F1-score across both positive and negative aspect categories. For positive aspects, RoBERTa achieved a weighted F1-score of 0.640 compared to BERT's 0.600. Similarly, for negative aspects, RoBERTa obtained a weighted F1-score of 0.678, surpassing BERT's 0.670. We observed that

the overall performance of transformer models was limited by their inability to address rare aspect categories effectively. For example, "Teacher Image" (TI) and "Classroom Atmosphere" (CA) consistently exhibited lower precision and recall due to their sparse representation in the dataset. This underscores the need for advanced models capable of handling data sparsity and class imbalance.

Model	Metric/Labels	CA	NM	TSR	TI	TQ	TA	TAT	TC	TEF	TM	Micro	Macro	Weighted	Samples
BERT (base)	Precision	0.62	0.8	0.37	0	0.75	0.48	0.6	0.65	0.48	0.52	0.62	0.53	0.62	0.63
	Recall	0.47	0.72	0.26	0	0.77	0.4	0.57	0.65	0.45	0.49	0.59	0.48	0.59	0.6
	F1-score	0.54	0.76	0.3	0	0.76	0.44	0.58	0.65	0.46	0.5	0.61	0.5	0.6	0.6
RoBERTa (base)	Precision	0.61	0.83	0.5	0	0.78	0.52	0.61	0.65	0.55	0.58	0.66	0.56	0.66	0.67
	Recall	0.5	0.78	0.38	0	0.78	0.45	0.62	0.66	0.51	0.58	0.63	0.52	0.63	0.64
	F1-score	0.55	0.8	0.43	0	0.78	0.48	0.61	0.65	0.53	0.58	0.64	0.54	0.64	0.64

Table 2. Average Classification report (5-folds) for Positive Aspects. The table presents metrics for different finetuned transfomer models (BERT, RoBERTa) across multiple labels: CA (Classroom Atmosphere), NM (Not Mentioned), TSR (Teacher-Student Relationship), TI (Teacher Image), TQ (Teacher Quality), TA (Teaching Ability), TAT (Teaching Attitude), TC (Teaching Content), TEF (Teaching Effectiveness), and TM (Teaching Method). Each row includes precision, recall, and F1-score for each label, followed by overall Micro, Macro, Weighted, and Samples averages for the respective models.

Model	Metric/Labels	CA	NM	TSR	TI	TQ	TA	TAT	TC	TEF	TM	Micro	Macro	Weighted	Samples
BERT (base)	Precision	0.16	0.92	0.3	0	0	0.28	0.43	0.19	0.36	0.56	0.77	0.32	0.7	0.74
	Recall	0.09	0.9	0.03	0	0	0.15	0.35	0.16	0.2	0.56	0.66	0.24	0.66	0.72
	F1-score	0.12	0.91	0.06	0	0	0.19	0.39	0.16	0.26	0.55	0.7	0.26	0.67	0.73
RoBERTa (base)	Precision	0.28	0.94	0.0	0	0	0.28	0.45	0.13	0.38	0.53	0.77	0.3	0.7	0.76
	Recall	0.04	0.9	0.0	0	0	0.12	0.36	0.01	0.34	0.66	0.67	0.24	0.67	0.73
	F1-score	0.06	0.92	0.0	0	0	0.15	0.39	0.02	0.36	0.59	0.71	0.25	0.68	0.74

Table 3. Average Classification report (5-folds) for Negative Aspects. The table presents metrics for different finetuned transfomer models (BERT, RoBERTa) across multiple labels: CA (Classroom Atmosphere), NM (Not Mentioned), TSR (Teacher-Student Relationship), TI (Teacher Image), TQ (Teacher Quality), TA (Teaching Ability), TAT (Teaching Attitude), TC (Teaching Content), TEF (Teaching Effectiveness), and TM (Teaching Method). Each row includes precision, recall, and F1-score for each label, followed by overall Micro, Macro, Weighted, and Samples averages for the respective models

5.2 Performance of Few-shot Chain-of-Thought (CoT)

Tables 4 and 5 evaluate the Few-shot Chain of Thought (CoT) methodology applied to LLMs for positive and negative aspects. GPT-4o demonstrated superior performance across most metrics, achieving a macro F1-score of 0.56, weighted F1-score of 0.65 for positive aspects and 0.52 (macro F1-score), 0.78 (weighted F1-score) for negative aspects. This indicates the model's strength in balancing precision and recall across all aspect categories. Notably, GPT4o-mini also performed well, achieving comparable scores with a lighter computational footprint, making it a practical choice for resource-constrained settings. Table 4 shows that for positive aspects GPT4o-mini outperforms 4o and llama-3.1-8b by identifying instances of teaching image and giving a significant difference in precision. Similarly, teaching ability was better scored in GPT4o-mini. A similar pattern is observed in Table 5 where teaching ability was scored higher in GPT4o-mini. However, for negative aspects, the teacher image was better identified with GPT4o. LLama-3.1-8B on the other hand showed a mixed performance, excelling in some categories like "Teaching Method" (TM) but struggling with others, particularly rare aspects such as "Teacher Image" (TI). Its macro F1-score of 0.48 (0.57 weighted-F1) for positive aspects and 0.48 (0.73 weighted F1) for negative aspects was lower compared to GPT-4o and GPT4o-mini, indicating room for improvement in handling diverse feedback categories.

Model	**Metric/Labels**	**CA**	**NM**	**TSR**	**TI**	**TQ**	**TA**	**TAT**	**TC**	**TEF**	**TM**	**Micro**	**Macro**	**Weighted**	**Samples**
GPT-4o	Precision	0.75	0.9	0.47	0.0	0.78	0.5	0.71	0.67	0.67	0.49	0.68	0.59	0.68	0.69
	Recall	0.54	0.86	0.58	0.0	0.81	0.18	0.44	0.8	0.52	0.82	0.65	0.55	0.65	0.67
	F1-score	0.62	0.88	0.52	0.0	0.79	0.27	0.54	0.73	0.59	0.61	0.66	0.56	0.65	0.68
GPT4o-mini	Precision	0.76	0.81	0.61	0.5	0.71	0.53	0.69	0.63	0.47	0.56	0.63	0.63	0.64	0.64
	Recall	0.5	0.8	0.24	0.12	0.88	0.2	0.56	0.82	0.81	0.62	0.69	0.56	0.69	0.66
	F1-score	0.61	0.8	0.35	0.2	0.79	0.29	0.62	0.71	0.59	0.59	0.66	0.56	0.64	0.66
LLama-3.1-8b	Precision	0.73	0.91	0.47	0.15	0.69	0.33	0.47	0.55	0.61	0.52	0.58	0.54	0.59	0.59
	Recall	0.24	0.68	0.23	0.25	0.73	0.35	0.52	0.77	0.32	0.63	0.57	0.47	0.57	0.58
	F1-score	0.36	0.78	0.31	0.19	0.71	0.34	0.49	0.64	0.42	0.57	0.57	0.48	0.57	0.58

Table 4. Classification report for Positive Aspects. The table presents metrics for different models on a Few shot CoT setting (GPT4o, GPT4o-mini, LLama-8b) across multiple labels: CA (Classroom Atmosphere), NM (Not Mentioned), TSR (Teacher-Student Relationship), TI (Teacher Image), TQ (Teacher Quality), TA (Teaching Ability), TAT (Teaching Attitude), TC (Teaching Content), TEF (Teaching Effectiveness), and TM (Teaching Method). Each row includes precision, recall, and F1-score for each label, followed by overall Micro, Macro, Weighted, and Samples averages for the respective models.

Model	Metric/Labels	CA	NM	TSR	TI	TQ	TA	TAT	TC	TEF	TM	Micro	Macro	Weighted	Samples
GPT-4o	Precision	0.45	0.98	0.53	0.18	0.26	0.47	0.71	0.56	0.53	0.67	0.82	0.53	0.82	0.84
	Recall	0.4	0.91	0.53	0.4	0.26	0.27	0.62	0.57	0.31	0.77	0.76	0.51	0.76	0.81
	F1-score	0.42	0.95	0.53	0.25	0.26	0.34	0.66	0.57	0.39	0.72	0.79	0.52	0.78	0.83
GPT4o-mini	Precision	0.36	0.92	0.75	0.0	0.22	0.45	0.68	0.61	0.39	0.74	0.77	0.51	0.78	0.81
	Recall	0.46	0.95	0.34	0.0	0.24	0.4	0.63	0.44	0.62	0.43	0.75	0.45	0.75	0.8
	F1-score	0.41	0.93	0.47	0.0	0.23	0.42	0.65	0.51	0.48	0.55	0.76	0.47	0.76	0.8
LLama-3.1-8b	Precision	0.16	0.97	0.33	0.17	0.14	0.23	0.55	0.48	0.43	0.75	0.74	0.42	0.78	0.79
	Recall	0.23	0.92	0.45	0.8	0.39	0.48	0.62	0.54	0.2	0.43	0.71	0.47	0.71	0.77
	F1-score	0.19	0.95	0.38	0.29	0.21	0.31	0.58	0.5	0.27	0.55	0.73	0.48	0.73	0.77

Table 5. Classification report for Negative Aspects. The table presents metrics for different models on a Few shot CoT setting (GPT4o, GPT4o-mini, LLama-8b) across multiple labels: CA (Classroom Atmosphere), NM (Not Mentioned), TSR (Teacher-Student Relationship), TI (Teacher Image), TQ (Teacher Quality), TA (Teaching Ability), TAT (Teaching Attitude), TC (Teaching Content), TEF (Teaching Effectiveness), and TM (Teaching Method). Each row includes precision, recall, and F1-score for each label, followed by overall Micro, Macro, Weighted, and Samples averages for the respective models

5.3 Performance of Fine-tuned Large Language Models (LLMs)

Tables 6 and 7 present the classification results for fine-tuned GPT-4o and GPT24o-mini models on positive and negative aspects, respectively. These tables highlight the performance of these models after task-specific fine-tuning, demonstrating their ability to capture nuanced sentiments and handle multi-label classification effectively. Compared to the baseline transformer results, the fine-tuning improved the macro-F1 score for positive aspects by 6% and negative aspects by 25%.

For positive aspects (Table 6), GPT-4o achieved a macro F1-score of 0.60 (0.69 weighted F1-score), while GPT4o-mini achieved a slightly lower macro F1-score of 0.59 (0.66 weighted F1-score). This indicates that GPT-4o excelled at balancing precision and recall across all positive aspect categories, with notable strengths in frequently occurring labels like "Teaching Quality" (TQ) and "Teaching Content" (TC). GPT4o-mini, although slightly less performance, still demonstrated strong capability in capturing positive feedback, making it a lighter but effective alternative for resource-constrained environments.

For negative aspects (Table 7), GPT-4o maintained its superior performance with a macro F1-score of 0.51 (0.79 weighted F1-score), while GPT4o-mini achieved a macro F1-score of 0.50 (0.78 weighted F1-score). Both models performed well in identifying dominant negative categories, such as "Not Mentioned" (NM) and "Teaching Attitude" (TAT), though challenges persisted with rare categories like "Teacher Image" (TI). Despite these challenges, GPT-4o and GPT4o-mini consistently demonstrated robustness in handling multi-label classification tasks for negative sentiments.

Model	**Metric/Labels**	**CA**	**NM**	**TSR**	**TI**	**TQ**	**TA**	**TAT**	**TC**	**TEF**	**TM**	**Micro**	**Macro**	**Weighted**	**Samples**
GPT4o-mini	Precision	0.73	0.83	0.59	0.25	0.78	0.48	0.67	0.71	0.59	0.64	0.68	0.62	0.68	0.69
	Recall	0.61	0.86	0.47	0.09	0.72	0.44	0.66	0.68	0.58	0.61	0.65	0.57	0.65	0.67
	F1-score	0.66	0.84	0.52	0.13	0.75	0.45	0.66	0.7	0.58	0.62	0.66	0.59	0.66	0.67
GPT-4o	Precision	0.63	0.86	0.58	0.0	0.8	0.51	0.67	0.715	0.64	0.62	0.7	0.6	0.7	0.71
	Recall	0.71	0.85	0.63	0.0	0.78	0.42	0.7	0.72	0.57	0.69	0.69	0.61	0.69	0.7
	F1-score	0.67	0.86	0.6	0.0	0.79	0.46	0.68	0.72	0.6	0.66	0.69	0.6	0.69	0.7

Table 6. Classification report for Positive Aspects. The table presents metrics for finetuned GPT-4o and GPT4o-mini on a Few shot CoT finetuned setting across multiple labels: CA (Classroom Atmosphere), NM (Not Mentioned), TSR (Teacher-Student Relationship), TI (Teacher Image), TQ (Teacher Quality), TA (Teaching Ability), TAT (Teaching Attitude), TC (Teaching Content), TEF (Teaching Effectiveness), and TM (Teaching Method). Each row includes precision, recall, and F1-score for each label, followed by overall Micro, Macro, Weighted, and Samples averages for the respective models.

Model	**Metric/Labels**	**CA**	**NM**	**TSR**	**TI**	**TQ**	**TA**	**TAT**	**TC**	**TEF**	**TM**	**Micro**	**Macro**	**Weighted**	**Samples**
GPT4o-mini	Precision	0.54	0.95	0.59	0.17	0.27	0.45	0.68	0.53	0.45	0.71	0.8	0.54	0.8	0.83
	Recall	0.43	0.93	0.55	0.1	0.23	0.41	0.7	0.44	0.49	0.67	0.77	0.5	0.77	0.82
	F1-score	0.43	0.94	0.57	0.13	0.24	0.43	0.68	0.48	0.47	0.69	0.79	0.5	0.78	0.82
GPT-4o	Precision	0.52	0.96	0.59	0.5	0.32	0.47	0.65	0.57	0.5	0.68	0.81	0.58	0.81	0.84
	Recall	0.44	0.94	0.57	0.1	0.17	0.36	0.71	0.46	0.43	0.77	0.78	0.5	0.78	0.82
	F1-score	0.43	0.94	0.57	0.17	0.22	0.41	0.67	0.5	0.46	0.72	0.79	0.51	0.79	0.83

Table 7. Classification report for Negative Aspects. The table presents metrics for finetuned GPT-4o and GPT4o-mini on a Few shot CoT setting across multiple labels: CA (Classroom Atmosphere), NM (Not Mentioned), TSR (Teacher-Student Relationship), TI (Teacher Image), TQ (Teacher Quality), TA (Teaching Ability), TAT (Teaching Attitude), TC (Teaching Content), TEF (Teaching Effectiveness), and TM (Teaching Method). Each row includes precision, recall, and F1-score for each label, followed by overall Micro, Macro, Weighted, and Samples averages for the respective models.

5.4 Error Analyses

To better understand the behavior of pre-trained large language models (LLMs) in handling student feedback, we analyzed the results exclusively using Chain of Thought (CoT) reasoning. The analysis focused on both positive and negative aspects across three models: LLaMA 8B, GPT4o-mini, and GPT-4o. Confusion matrices provided insights into common confusions, performance trends, and limitations. We provide all the confusion matrices in the Appendix (Section 9.3).

The confusion matrices revealed several key trends. For negative aspects, "Not Mentioned" consistently showed the highest number of correctly predicted instances, reflecting the models' strong capability to detect its absence in feedback. However, this dominance may also stem from dataset imbalances, as "Not Mentioned" is highly frequent. Significant confusions were observed between semantically similar aspects, such as "Teaching Effectiveness" and "Teaching Method," and between "Teaching Attitude" and "Teaching Quality," indicating the models' difficulty in distinguishing nuanced feedback. Rare aspects like "Teacher Image" and "Classroom Atmosphere" were underrepresented and often misclassified, showing the impact of dataset imbalance on model performance. Among the models, GPT-4o exhibited the most balanced performance with fewer misclassifications, outperforming LLaMA 8B and GPT4o-mini in handling both frequent and rare categories.

For positive aspects, similar trends were noted. The models performed well on frequent categories such as "Teacher Quality" and "Teaching Ability," but confusions between related aspects like "Teaching Content" and "Teaching Method" persisted. Rare categories like "Teacher Image" again showed low true positive counts, underscoring the models' reliance on frequency distributions in the training data. GPT4o-mini demonstrated superior performance on "Teaching Effectiveness," whereas GPT-4o excelled in most other categories, highlighting its robustness. Across both positive and negative aspects, semantic overlap between categories and dataset imbalance emerged as the primary sources of error.

6. COMPARATIVE ANALYSIS AND INSIGHTS

For positive aspects, as shown in Tables 4 and 5, fine-tuned GPT-4o outperformed its Few-shot CoT counterpart, achieving a weighted F1-score of 0.69 compared to 0.65 in the Few-shot setup. Similarly, GPT4o-mini achieved a weighted F1-score of 0.66 in the fine-tuned setting, slightly surpassing its Few-shot performance of 0.64. These results underscore the effectiveness of task-specific fine-tuning in leveraging annotated datasets to capture subtle nuances in positive feedback. However, despite its lower scores, the Few-shot CoT approach demonstrated competitive performance while requiring significantly fewer annotated examples. Its step-by-step reasoning mechanism further enhanced interpretability, offering a transparent prediction process, which is particularly beneficial in low-resource scenarios.

For negative aspects, as illustrated in Tables 5 and 7, fine-tuned GPT-4o achieved a weighted F1-score of 0.79, outperforming its Few-shot CoT result of 0.78. Similarly, fine-tuned GPT4o-mini attained a weighted F1-score of 0.78, exceeding its Few-shot CoT counterpart, which scored 0.76. This consistent improvement highlights the advantages of fine-tuning in capturing complex, multi-faceted negative sentiments, especially in rare aspect categories. Despite the superior performance of fine-tuning, the Few-shot CoT approach remains valuable for scenarios prioritizing scalability and rapid deployment. Notably, Few-shot CoT models performed well on dominant categories such as "Not Mentioned" (NM) and "Teaching Quality" (TQ) but struggled with rare categories like "Teacher Image" (TI).

Beyond the comparison of Few-shot CoT and fine-tuning, an analysis of transformer-based models versus large language models (LLMs) reveals further distinctions in performance. While transformers exhibited robustness in capturing linguistic and contextual nuances—particularly for well-represented aspects like "Teaching Quality"—they struggled with rare categories such as "Teacher Image" and "Classroom Atmosphere." For instance, RoBERTa achieved a weighted F1-score of 0.64 for positive aspects, which, while commendable, remained lower than the fine-tuned LLMs such as GPT-4o (0.69). This discrepancy highlights the advanced generalization capabilities of LLMs, which are better suited for handling dataset imbalances and semantic complexities.

It should be noted that compared to the previously published version of this dataset (Butt et al., 2023), each class representation, its inter-annotator agreement and the scores have significantly increased. The average macro-F1 score using our method increased the results from 0.38 to 0.56, while the average weighted-F1 score improved from 0.70 to 0.74. However, the dataset in its real-life representation poses notable challenges, particularly in terms of class imbalance and the presence of underrepresented aspect categories. For example, "Not Mentioned" (NM) and "Teaching Quality" (TQ) dominated the dataset, skewing performance metrics. Although GPT-4o effectively managed these imbalances—achieving high recall for these frequent categories—rare aspects like "Teacher Image" (TI) continued to yield lower scores. Addressing these challenges requires targeted strategies such as data augmentation for rare categories, refining aspect definitions to reduce overlaps, and adopting hierarchical classification to better structure related aspects. Future research should focus on these improvements to enhance model performance across all aspect categories. Additionally, we recognize that incorporating a more structured format in future questionnaires could enhance the consistency and computational tractability of the data. Aligning the questionnaire with the nine aspects identified in our analysis—by explicitly prompting students to provide feedback on each—may result in more balanced and aspect-specific responses. At the same time, to preserve the richness and nuance of student perspectives, we propose a

hybrid approach that combines structured prompts with optional free-text justifications. This design would not only improve the robustness of data collection but also facilitate semi-automated annotation by leveraging the students' aspect-aligned expressions, thereby reducing the need for extensive manual labeling.

Moreover, through our results we acknowledge the importance of considering potential negative impacts that student feedback could have on educators. Although student surveys offer valuable insights for improving educational quality, consistently negative or critical evaluations—if not presented constructively—can negatively affect teachers' morale, enthusiasm, and motivation. In extreme cases, educators might even disengage from their teaching responsibilities. We recommend that university management strategically incorporate supportive structures alongside student feedback mechanisms. These structures could include personalized mentorship, professional development programs, or workshops that help educators interpret and contextualize feedback constructively, thereby fostering an environment of continuous improvement rather than discouragement.

On a broader scale, transformative trends in educational policy influenced by student assessments are becoming increasingly evident. Universities are likely to adopt data-driven educational strategies, utilizing computational methods—such as those demonstrated in our study—to systematically analyze and interpret student input. These practices not only enhance curriculum design and pedagogical effectiveness but also reinforce student-centered approaches in education policy. However, the success of these transformative policies will largely depend on institutional commitments to robust teacher support frameworks. Aligning educational improvement goals with sustained teacher engagement and professional satisfaction will thus be essential to achieving balanced and effective educational outcomes.

7. CONCLUSION

This study presents a novel dataset in Spanish and a multitask learning framework for Aspect-Based Sentiment Analysis (ABSA) in Student Evaluations of Teaching (SET), demonstrating the potential of leveraging advanced machine learning techniques to extract actionable insights from qualitative feedback. By integrating opinion segmentation and multi-label classification, the framework effectively captures nuanced feedback on nine predefined aspects of teaching performance. Through an extensive evaluation of fine-tuned transformer models (BERT, RoBERTa) and Large Language Models (LLMs), such as GPT-4o, GPT4o-mini, and LLaMA-3.1-8B, we demonstrate the effectiveness of task-specific fine-tuning and the scalability of Few-shot Chain of Thought (CoT) reasoning. The results indicate that

fine-tuned GPT-4o consistently outperformed other models, achieving the highest macro F1-scores for both positive and negative aspects. Few-shot CoT, while slightly less accurate, provided competitive performance with the added benefits of interpretability and reduced reliance on annotated data. However, the analysis also uncovered challenges, including semantic overlap between similar aspects, imbalances in dataset distribution, and underperformance in rare categories, such as "Teacher Image" and "Classroom Atmosphere." To address these challenges, future work should focus on balancing datasets through augmentation, refining aspect definitions to reduce ambiguity, and exploring hierarchical classification approaches for related aspects. The findings of this study underscore the potential of combining ABSA and advanced LLMs to transform unstructured student feedback into structured, actionable insights. This framework has significant implications for improving teaching quality, fostering student engagement, and guiding institutional decision-making in education.

8. REFERENCES

Butt, S., Mejía-Almada, P., Alvarado-Uribe, J., Ceballos, H.G., Sidorov, G., Gelbukh, A., 2023. MF-SET: A Multitask Learning Framework for Student Evaluation of Teaching, in: Proceedings of the Future Technologies Conference. Springer, pp. 254–270.

Chauhan, G.S., Agrawal, P., Meena, Y.K., 2019. Aspect-based sentiment analysis of students' feedback to improve teaching–learning process, in: Information and Communication Technology for Intelligent Systems: Proceedings of ICTIS 2018, Volume 2. Springer, pp. 259–266.

Chen, J., Wang, R., Fang, B., Zuo, C., 2024. Fine-grained aspect-based opinion mining on online course reviews for feedback analysis. Interact. Learn. Environ. 32, 4380–4395.

Dehbozorgi, N., Mohandoss, D.P., 2021. Aspect-based emotion analysis on speech for predicting performance in collaborative learning, in: 2021 IEEE Frontiers in Education Conference (FIE). IEEE, pp. 1–7.

Gallardo, K., Butt, S., Ceballos, H., 2023. Improvement of Teaching Competencies Training in Higher Education Faculty Based on Student Evaluations of Teaching and AI Systems, in: International Conference in Information Technology and Education. Springer, pp. 555–563.

Hadi, M.U., Al Tashi, Q., Shah, A., Qureshi, R., Muneer, A., Irfan, M., Zafar, A., Shaikh, M.B., Akhtar, N., Wu, J., others, 2024. Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects. Authorea Prepr.

Kenton, J.D.M.-W.C., Toutanova, L.K., 2019. Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of naacL-HLT. Minneapolis, Minnesota, p. 2.

Liu, Y., 2019. Roberta: A robustly optimized bert pretraining approach. ArXiv Prepr. ArXiv190711692 364.

Marsh, H.W., Roche, L.A., 1997. Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. Am. Psychol. 52, 1187.

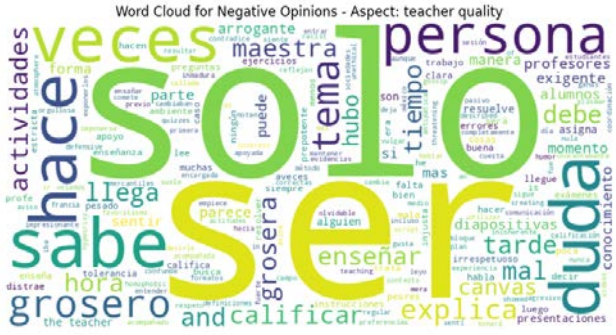
Mughal, N., Mujtaba, G., Kumar, A., Daudpota, S.M., 2024. Comparative Analysis of Deep Natural Networks and Large Language Models for Aspect-Based Sentiment Analysis. IEEE Access.

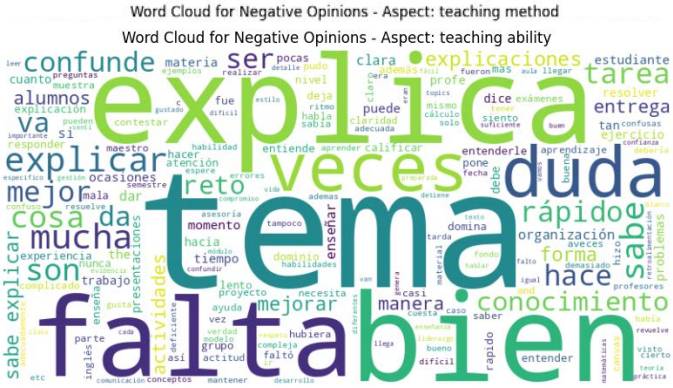
Nazir, A., Rao, Y., Wu, L., Sun, L., 2020. Issues and challenges of aspect-based sentiment analysis: A comprehensive survey. IEEE Trans. Affect. Comput. 13, 845–863.

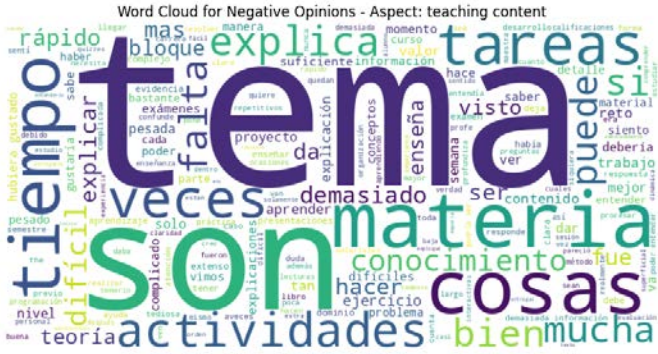
Nowell, C., Gale, L.R., Handley, B., 2010. Assessing faculty performance using student evaluations of teaching in an uncontrolled setting. Assess. Eval. High. Educ. 35, 463–475.
Qiao, K., Li, G., Zeng, X., Li, W., 2023. Utilizing Large Language Models for the Generation of Aspect-Based Sentiment Analysis Datasets, in: 2023 4th International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE). IEEE, pp. 121–124.
Shaikh, S., Daudpota, S.M., Yayilgan, S.Y., Sindhu, S., 2023. Exploring the potential of large-language models (LLMs) for student feedback sentiment analysis, in: 2023 International Conference on Frontiers of Information Technology (FIT). IEEE, pp. 214–219.
Zhang, W., Deng, Y., Liu, B., Pan, S., Bing, L., 2024. Sentiment Analysis in the Era of Large Language Models: A Reality Check, in: Findings of the Association for Computational Linguistics: NAACL 2024. pp. 3881–3906.
Zhong, Q., Ding, L., Liu, J., Du, B., Tao, D., 2023. Can ChatGPT Understand Too? A Comparative Study on ChatGPT and Fine-tuned BERT. ArXiv E-Prints arXiv-2302.
Zhou, J., Ye, J., 2023. Sentiment analysis in education research: a review of journal publications. Interact. Learn. Environ. 31, 1252–1264.

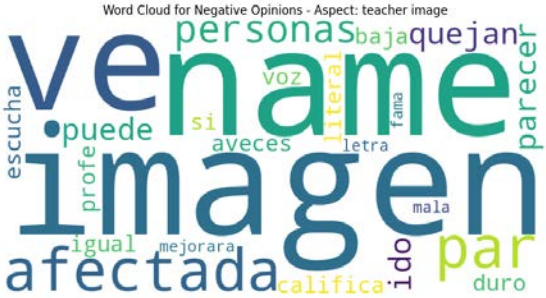
9. Appendix

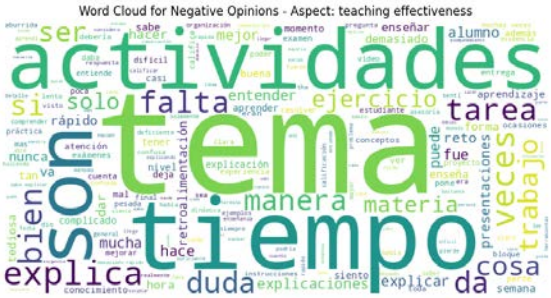
9.1 Negative Aspects











9.2 Positive Aspects

Word Cloud for Positive Opinions - Aspect: classroom atmosphere

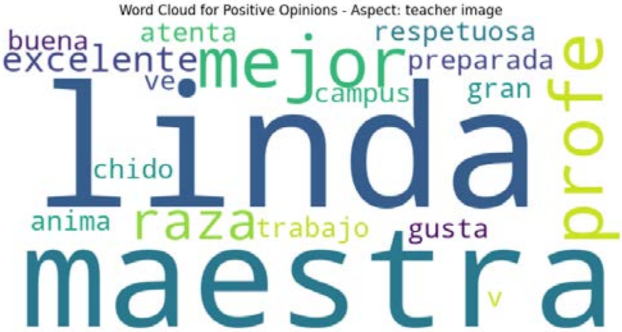
Word Cloud for Positive Opinions - Aspect: teaching attitude

Word Cloud for Positive Opinions - Aspect: teaching ability

Word Cloud for Positive Opinions - Aspect: teaching content

Word Cloud for Positive Opinions - Aspect: teaching effectiveness

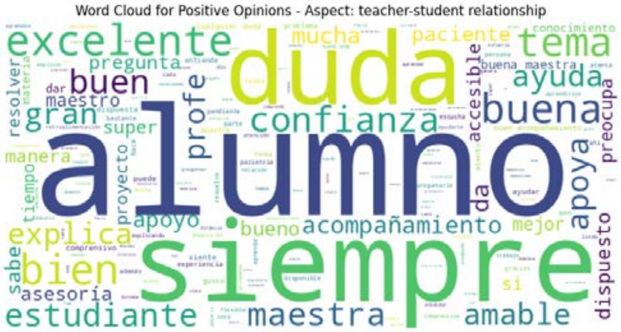
Word Cloud for Positive Opinions - Aspect: teacher image



Word Cloud for Positive Opinions - Aspect: teaching method

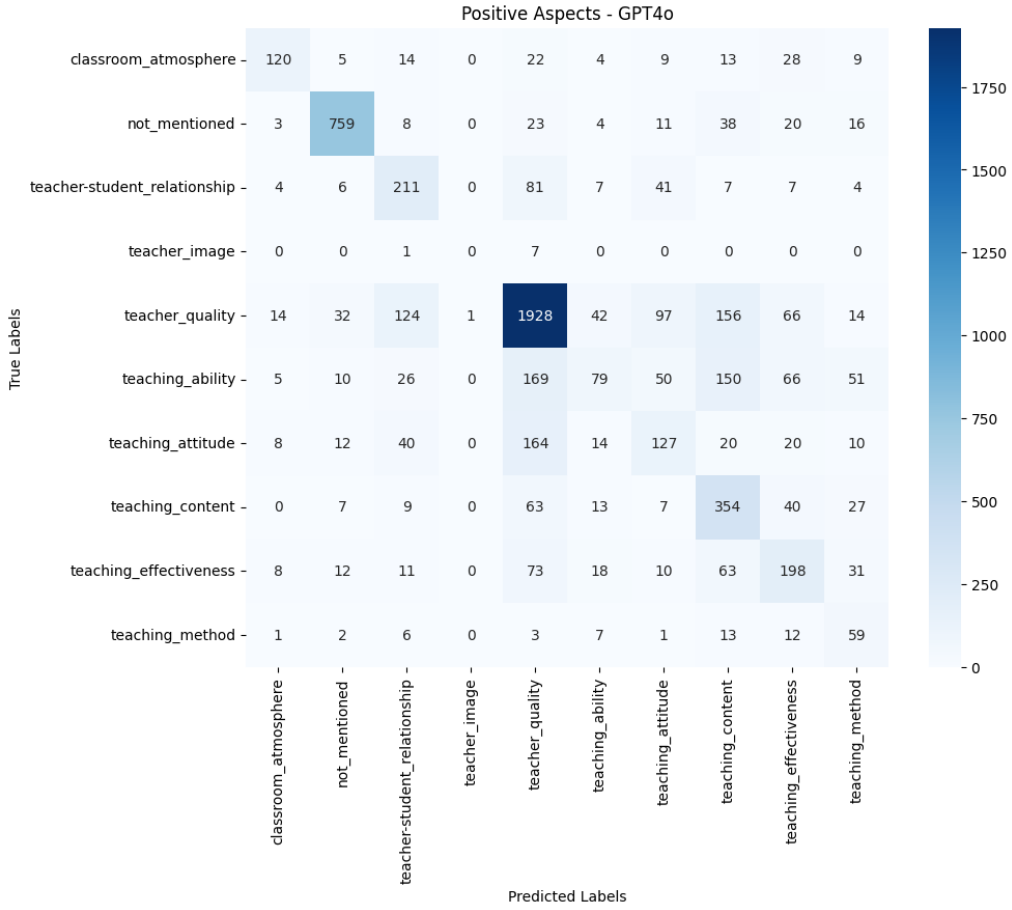
Word Cloud for Positive Opinions - Aspect: teacher quality

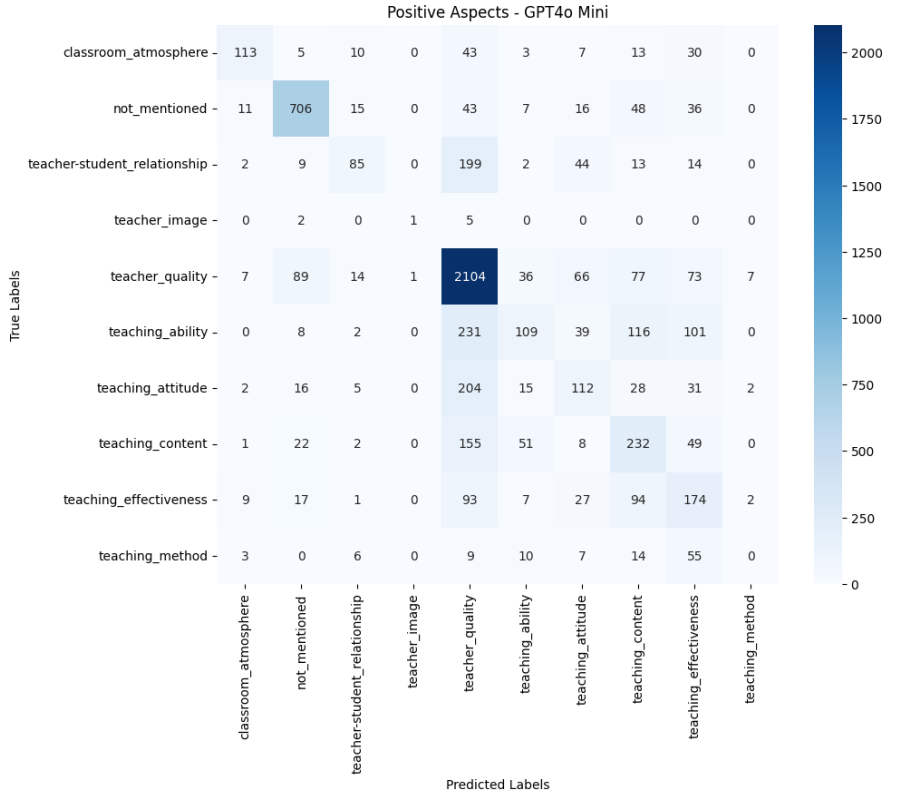
Word Cloud for Positive Opinions - Aspect: teacher-student relationship

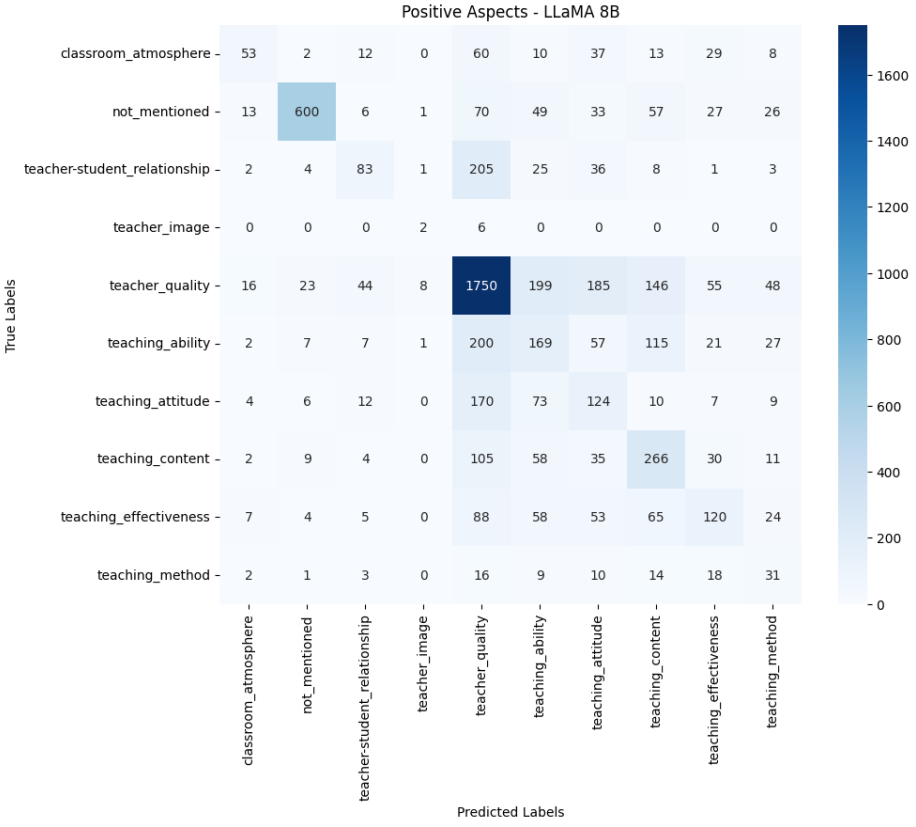


9.3 *Confusion Matrices*

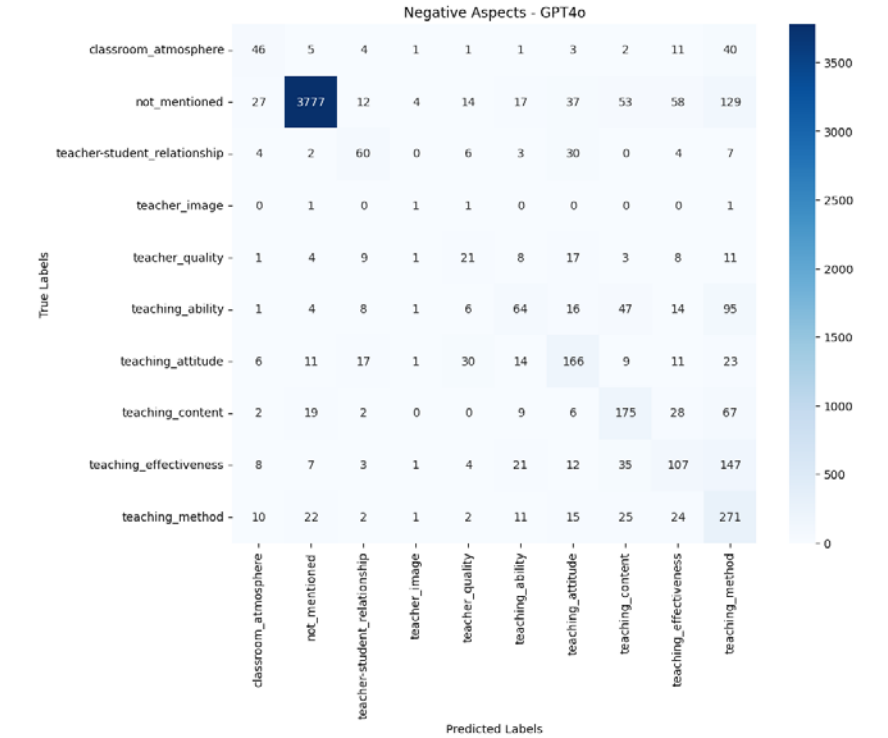
9.3.1 *Positive aspects for CoT experiment*

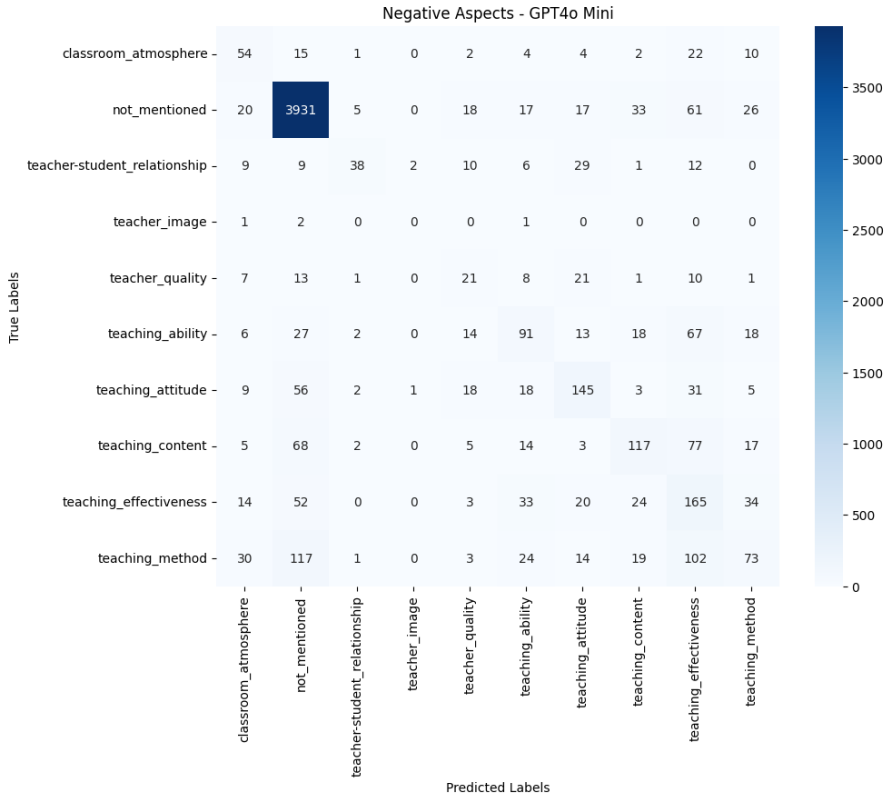


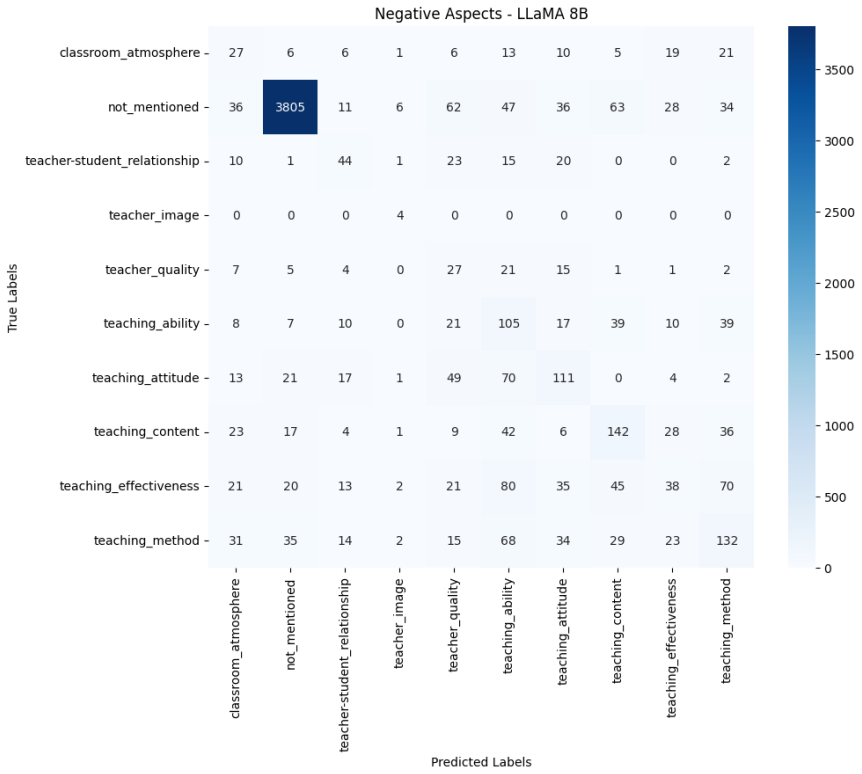




9.3.2 *Negative aspects for CoT experiment*


Negative Aspects - GPT4o
classroom_atmosphere 46 5 4 1 1 1 3 2 11 40
not_mentioned 27 3777 12 4 14 17 37 53 58 129
teacher-student_relationship 4 2 60 0 6 3 30 0 4 7
teacher_image 0 1 0 1 1 0 0 0 0 1
teacher_quality 1 4 9 1 21 8 17 3 8 11
teaching_ability 1 4 8 1 6 64 16 47 14 95
teaching_attitude 6 11 17 1 30 14 166 9 11 23
teaching_content 2 19 2 0 0 9 6 175 28 67
teaching_effectiveness 8 7 3 1 4 21 12 35 107 147
teaching_method 10 22 2 1 2 11 15 25 24 271
True Labels
classroom_atmosphere
not_mentioned
teacher-student_relationship
teacher_image
teacher_quality
teaching_ability
teaching_attitude
teaching_content
teaching_effectiveness
teaching_method
Predicted Labels
3500
3000
2500
2000
1500
1000
500
0


Negative Aspects - GPT4o Mini
classroom_atmosphere 54 15 1 0 2 4 4 2 22 10
not_mentioned 20 3931 5 0 18 17 17 33 61 26
teacher-student_relationship 9 9 38 2 10 6 29 1 12 0
teacher_image 1 2 0 0 0 1 0 0 0 0
teacher_quality 7 13 1 0 21 8 21 1 10 1
teaching_ability 6 27 2 0 14 91 13 18 67 18
teaching_attitude 9 56 2 1 18 18 145 3 31 5
teaching_content 5 68 2 0 5 14 3 117 77 17
teaching_effectiveness 14 52 0 0 3 33 20 24 165 34
teaching_method 30 117 1 0 3 24 14 19 102 73
True Labels
classroom_atmosphere
not_mentioned
teacher-student_relationship
teacher_image
teacher_quality
teaching_ability
teaching_attitude
teaching_content
teaching_effectiveness
teaching_method
Predicted Labels
3500
3000
2500
2000
1500
1000
500
0


Negative Aspects - LLaMA 8B
True Labels
Predicted Labels
classroom_atmosphere
not_mentioned
teacher-student_relationship
teacher_image
teacher_quality
teaching_ability
teaching_attitude
teaching_content
teaching_effectiveness
teaching_method
3500
3000
2500
2000
1500
1000
500
0