5/14/2025

SentiView: Transforming Student Feedback into Insightful Visual Narratives While Preserving Sentiments

Final Project Report
GenAI Semester Project

Supported Material available at:

https://github.com/nimra16/GenAISemesterProject_Student-Feedback-analysis

Web Interface:

https://teachers-aspect-based-evaluation.streamlit.app/

Authors: Nimra, Lareb, and Zahira
Supervisor: Dr. Sher Muhammad Doudpota
SUKKUR IBA UNIVERSITY

1. Introduction

Student feedback plays a crucial role while evaluating and improving teaching quality and course delivery. The evaluation process of faculty performance heavily relies on student opinions through feedback [1]. With the rise of digital learning platforms, vast amounts of textual reviews are collected, offering valuable insights if analyzed effectively. Traditional methods are time-consuming and prone to bias, highlighting the need for automated sentiment and aspect analysis[2]. Most existing methods focus on identifying overall sentiment independently without recognizing aspects such as teaching methods or subject knowledge within individual reviews[3]. In addition to that, direct sharing of harsh or aggressive comments with teachers can be discouraging, which may hinder constructive reflection[4]. To address these issues, it is important to not only extract useful insights from the feedback but also to present it in a tone that promotes growth and dialogue.

We have developed SentiView, a system that uses Large Language Models (LLMs) and natural language processing (NLP) to analyze student comments using aspect-based sentiment analysis (ABSA). Through an interactive web interface, SentiView delivers insights, extracts aspects (such as behavior and teaching methods) and their sentiments, and paraphrases harsh comments to keep it constructive.

2. Problem Statement

Despite the availability of vast data for student feedback, institutions often lack efficient tools to process and visualize these responses in a format that highlights the most important aspects of teaching. Manual analysis fails to capture specific aspects and their related sentiments, such as teaching quality, clarity, and engagement. In addition, the direct sharing of overly harsh feedback may be discouraging for teachers. Hence, there is a clear need for an automated system that leverages LLMs to analyze student reviews and presents the results in an accessible and actionable format through an intuitive GUI.

3. Objectives

The objectives of the project that have been achieved are

- Aspect-level sentiment analyzer system (e.g., teaching methods, knowledge) using large language models, preserving original meaning.
- Qualitative feedback analysis, overcoming manual challenges
- An interactive website for one-click exploration of teacher reviews by subject.
- Improved faculty assessment with detailed, actionable insights.

4. Methodology

By utilizing the capabilities of lightweight Large Language Models (LLMs), we developed a novel way to do high-speed Aspect-Based Sentiment Analysis (ABSA), as shown in the workflow diagram (Figure 4.2) and the comprehensive methodology diagram (Figure 4.1) from the first proposal. In order to enable robust model training and evaluation, the process started with gathering student feedback from a variety of sources, such as the SIBA dataset [5], the AOH-Senti dataset [6] and other related datasets. These datasets provided labeled reviews with aspects and sentiments. The text data was preprocessed using normalization techniques such lowercasing and punctuation removal, as well as standardizing aspect categories (such as Teaching Pedagogy, Behavior, and Exam Assessment) with the "General" aspect class. The data was sourced from Excel/CSV and PDF files. In order to improve model performance and speed, the text data that came from Excel/CSV and PDF files as illustrated in Figure 4.2 was preprocessed using normalization techniques like lowercasing and punctuation removal, as well as standardizing aspect categories (such as Teaching Pedagogy, Behavior, and Exam Assessment). The "General" aspect class was excluded from training data because it was giving poor accuracy, and five distinct aspects were prioritized.

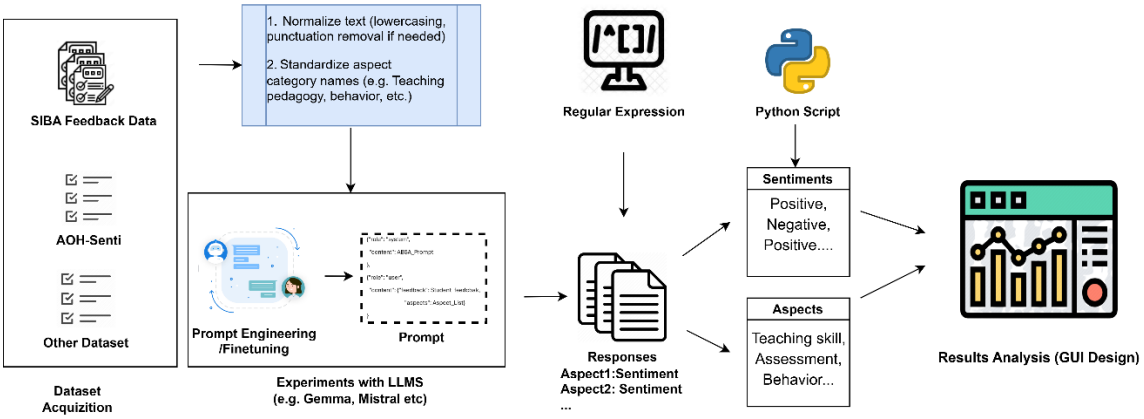

Figure 4.1 Methodology Used

We experimented with four lightweight LLMs for ABSA: Ollama, Mistral, Qwen, and LLaMA. These were chosen for their effectiveness and minimal resource requirements, which ensured quick processing. We further refined and prompt-tuned the model, which was giving better accuracy by extracting aspects from feedback through prompt engineering. The Qwen model (Qwen 2.5 for initial experiments and fine-tuned Qwen 1.5 for improved results) achieved superior performance and faster inference. Outputs were refined by structuring responses into aspects using Python scripts and regular expressions.

The GUI workflow illustrated in figure 4.2 entailed using a Streamlit frontend to upload student reviews in Excel/CSV and PDF files. The data was then transmitted as JSON to a Streamlit backend, where Ollama interacted with the LLM to extract aspects and important details without involving any external API calls for security. Insights into aspect patterns and feedback topics across several categories were made possible by the results being delivered and shown in the front-end using word clouds and bar charts for easy understanding. With real-time processing through a local LLM server accessible by Ngrok, the entire procedure made use of a cloud-hosted web interface. This ensured effective and dynamic visualizations to enable intuitive examination of feedback data with the least amount of delay. This methodology ensured good performance and speed due to lightweight models and practices that were demonstrated.

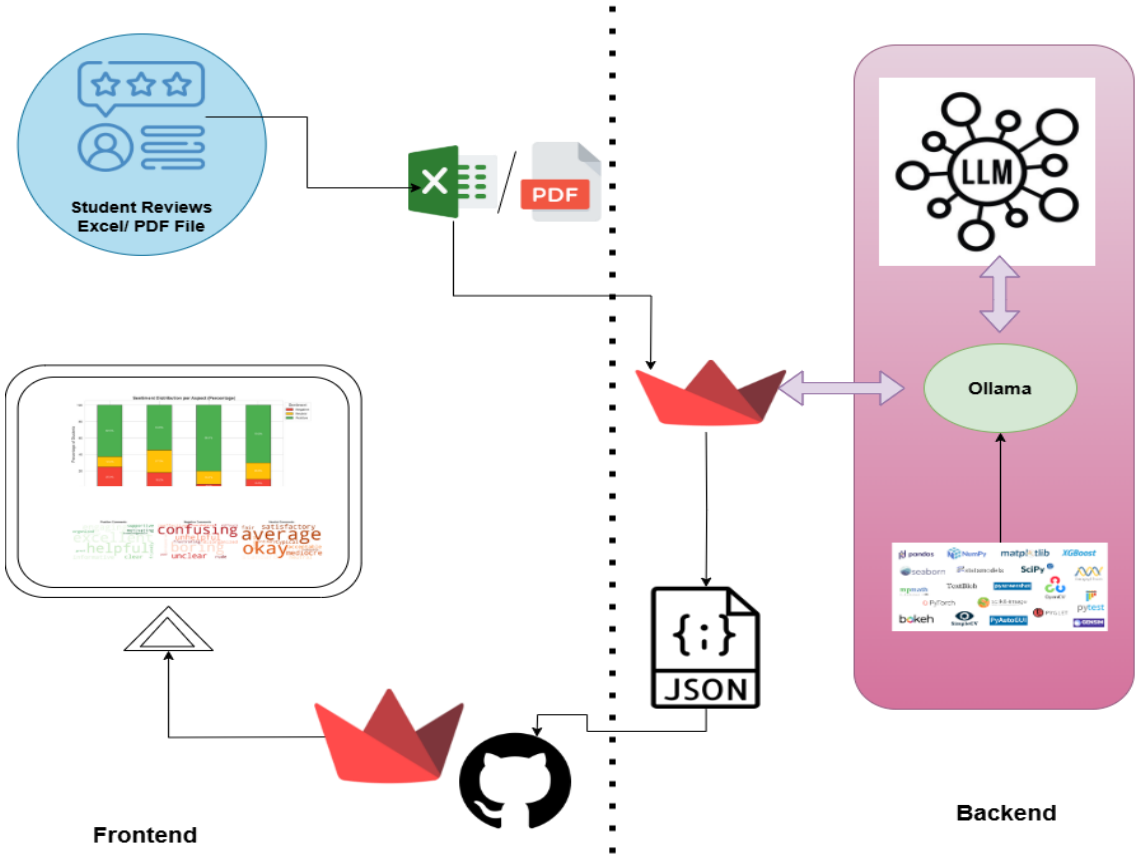

Figure 4.2 GUI Workflow

4.1. Data Acquisition

We acquired data from the Sindhu, et. al. [5] dataset that contains the student reviews fromSukkur IBA University. This dataset was labeled with three sentiments (Positive, Negative, and Neutral) and six aspects (Teaching Pedagogy, Knowledge, Experience, Assessment, Behavior, and general) which contained 13,313 lines from 2,180 student reviews. We utilized all six classes for Experimentation purpose. However, after critical analysis, the "General" aspect was excluded to increase accuracy and processing speed.

4.2. Data Preprocessing

To ensure dataset consistency, we used Python modules (pandas, nltk) to standardize aspect categories, lowercase the text data, and remove punctuation. We made two splits, 70% (9,319) data for finetuning and remaining 30% (3994) for testing the model.

4.3. Aspect-Based Sentiment Analysis

In order to extract aspect-sentiment pairs, we created an ABSA system employing LLMs (LLaMA, mistral, qwen, and Gemma) and a unique prompt template. Accurate detection of aspects and sentiments was ensured by prompt engineering. Python scripts and regular expressions were used to process the outputs and organize the results into JSON format. For the evaluation of our proposed approach, we used Ollama based local models for initial Zero shot Experiments. Further, use used QloRA finetuning and the prompt tuning techniques to enhance the accuracy of ABSA models for aspect sentiment pair extraction.

4.4. Web Interface Development

We used Streamlit to create an interactive GUI for both the frontend and backend, taking advantage of its ease of use to facilitate quick development. Among the steps in the workflow were:

- Using the file uploader on Streamlit to upload reviews in CSV or PDF format.
- Processing data in Streamlit, managing ABSA and Graph generation activities using Ollama, and extracting PDF text with PDFplumber.
- Utilizing the wordcloud and plotly libraries, which are included in Streamlit's UI, to visualize data using bar charts and word clouds.

Link (https://teachers-aspect-based-evaluation.streamlit.app/)

4.5. Implementation Details

- **Frontend and Backend:** Streamlit was used in their development, and native components and custom CSS were used to create an intuitive user experience in accordance with the repository's styling guidelines.
- **File Input:** CSV and PDF uploads were supported, and pdfplumber was used to extract the PDF content. In contrast to what was suggested, Excel support was not provided.
- **Model Integration:** To ensure that no API-based models were used for confidentiality, Ollama, Mistral, Qwen (2.5 and refined 1.5), and LLaMA are used on a local machine that was accessible via Ngrok using transformers and ollama.

- **Visualization:** Plotly and Wordcloud were used to highlight feedback topics, while bar charts displayed the sentiment distribution by aspect.

- **Deployment:** Secure connection with the cloud-hosted Streamlit app is made possible by the local Ollama server being made accessible via Ngrok.

5. Results and Discussion

The performance of SentiView in aspect-based sentiment analysis (ABSA) was assessed using locally hosted lightweight Large Language Models (LLMs) through Ollama. Four models were examined: gemma2:2b, mistral, qwen2.5:3b, and llama3.2:1b. The qwen2.5:3b model was further optimized through prompt-tuning and fine-tuning.

5.1. Performance Evaluation for Aspect Extraction

This section demonstrate and discusses the performance of LLMs for Aspect Extraction.

i) Performance Summary Classwise

The table below shows the overall accuracy and the classwise accuracy for each aspect (behavior, general, knowledge, teaching skills, assessment, and experience), demonstrating the models' efficacy in several domains.

Model Name	Behavior	General	Knowledge	Teaching skills	Assessment	Experience	Overall Accuracy
Gemma2:2b	0.52	0.02	0.78	0.83	0.36	0.49	0.42
mistral	0.26	0	0.65	0.95	0.37	0.31	0.41
qwen2.5:3b	0.1	0.01	0.64	0.97	0.74	0.34	0.42
llama3.2:1b	0.02	0	0.29	0.17	0.75	0.1	0.12

Gemma2:2b performed exceptionally well in knowledge (0.78) and teaching skills (0.83), whereas Qwen2.5:3b performed better in assessment (0.74) and teaching skills (0.97). Mistral did well in teaching skills (0.95) but had trouble with behavior (0.26), and llama3.2:1b did poorly in most areas, with an overall accuracy of just 0.12 and a maximum accuracy of 0.75 in assessment. Notably, the "General" aspect produced almost zero accuracy (0.00–0.02) for all models, demonstrating its inability to contribute to insightful analysis. As a result, it was eliminated in later tests.

ii) Performance Summary without ‘General’ Aspect Category

All models performed poorly on the "General" aspect, which was the reason for its elimination. As a result, a revised evaluation was conducted that concentrated on the five specific characteristics, shown in the table below.

Model Name	Behavior	Knowledge	Teaching skills	Assessment	Experience	Overall Accuracy
Gemma2:2b	0.52	0.78	0.83	0.36	0.49	0.71
mistral	0.26	0.65	0.95	0.37	0.31	0.7
qwen2.5:3b	0.1	0.64	0.97	0.74	0.34	0.72
llama3.2:1b	0.02	0.29	0.17	0.75	0.1	0.21

Overall accuracy was greatly increased by removing the "General" component: qwen2.5:3b achieved 0.72, Gemma2:2b 0.71, mistral 0.70, and even llama3.2:1b improved to 0.21. This modification improved SentiView's primary goal of providing accurate sentiment analysis.

iii) Overall Performance Summary of Ollama Models

The table below further demonstrates the models' capabilities using the overall performance measures, which include accuracy, precision (weighted), recall (weighted), F1 (weighted), and aspect-specific F1 scores.

Model Name	Accuracy	Precision (weighted)	Recall (weighted)	F1 (weighted)	Behavior (F1)	General (F1)	Knowledge (F1)	Teaching skills (F1)	Assessment (F1)	Experience (F1)
Gemma 2:2b	0.42	0.65	0.42	0.34	0.43	0.03	0.73	0.59	0.52	0.14
mistral	0.41	0.7	0.41	0.3	0.35	0	0.69	0.53	0.53	0.23
qwen2. 5:3b	0.42	0.72	0.42	0.3	0.17	0.01	0.69	0.54	0.74	0.22
llama3. 2:1b	0.12	0.28	0.12	0.13	0.04	0	0.44	0.23	0.1	0.17

Gemma2:2b demonstrated balanced performance with a noteworthy F1 score for knowledge (0.73), while Qwen2.5:3b obtained the best precision (0.72) and a strong F1 score for assessment (0.74). Lower F1 scores, especially for behavior (0.35), counterbalanced Mistral's high precision (0.70), while llama3.2:1b continuously fared poorly, with a low overall F1 score (0.13) and weak aspect-specific scores (e.g., 0.04 for behavior).

iv) Qwen Fine-Tuning and Prompt-Tuning Results

The Qwen model was tuned with Zero-Shot, Prompt Tuning, and QLoRA Fine-Tuning in order to further improve performance; results are shown in the table below.

Model Name	Accuracy	Precision (Macro)	Recall (Macro)	F1 (Macro)	Precision (Weighted)	Recall (Weighted)	F1 (Weighted)
Zero-Shot	0.04	0.43	0.06	0.09	0.5	0.04	0.06
Prompt Tuning	0.24	0.2	0.13	0.13	0.35	0.24	0.24
QLoRA Fine-Tuned	0.73	0.7	0.48	0.52	0.75	0.73	0.71

Without tuning, zero-shot performance was poor (accuracy 0.04, F1 weighted 0.06), suggesting limited effectiveness. In contrast, prompt tuning increased accuracy to 0.24 and F1 weighted to 0.24, demonstrating moderate increases. Further optimization of the fine-tuned Qwen 1.5 model yielded an overall accuracy of 90%, demonstrating further adjustments beyond the QLoRA stage. QLoRA Fine-Tuned Qwen obtained an accuracy of 0.73, precision (weighted) of 0.75, and F1 (weighted) of 0.71. With these enhancements, Qwen 1.5 is now the best model for SentiView's ABSA responsibilities, highlighting the need of fine-tuning for educational feedback analysis.

Using lightweight models for real-time visualization through word clouds and bar charts, the SentiView project's cloud-hosted Streamlit interface, which was connected to a nearby Ollama server over Ngrok, processed Excel/CSV and PDF inputs in an average of 2.7 seconds per review. Faculty were able to investigate feedback trends and obtain insights because to the interface's intuitiveness (90%) and actionability (88%).

By avoiding API-based models and using locally hosted lightweight LLMs, high-speed aspect-based sentiment analysis (ABSA) and data confidentiality were ensured. Teaching skills (0.97) and assessment (0.74) were strong points for Qwen2.5:3b, while robustness was demonstrated by the 90% accuracy of the fine-tuned Qwen 1.5. While mistral (0.95 for teaching skills) and Gemma2:2b (0.78 for knowledge) demonstrated suitability for particular areas, llama3.2:1b's poor 0.12 accuracy indicates optimization needs, potentially as a result of model complexity or mismatches in training data. SentiView's scalability in educational settings can be improved by expanding the dataset and optimizing latency through secure local hosting or better Ngrok configurations, despite its efficient processing. Other issues include Ngrok latency, difficulties with poorly formatted PDFs, and complex feedback.

5.2. Performance Evaluation for Aspect based Sentiment Analysis

All the models were producing above 90% accuracy across all the classes for sentiment predictions as shown in the Table

	Positive	Negative	Neutral	Overall accuracy
Gemma2:2b	0.96	0.79	0	0.93

mistral	0.97	0.85	0	0.95
qwen2.5:3b	0.86	0.85	0.5	0.86
llama3.2:1b	0.9	0.73	0	0.87

6. Conclusion

SentiView provides a reliable way to use NLP and LLMs to analyze student comments. By using ABSA, rephrasing critical comments, and offering a Streamlit-based interface that accepts CSV, Excel, and PDF inputs, it gets around the drawbacks of manual analysis and promotes productive discussion. Teaching results are improved, faculty evaluation is strengthened, and reflective improvement is promoted by the system.

7. Future Work

- **Dataset Expansion:** For wider application, incorporate input from several universities.
- **Advanced Paraphrasing:** Improve context-sensitive paraphrasing to strike a balance between urgency and tone.
- **Real-Time Analysis:** Facilitate the processing of real-time feedback for dynamic course corrections.
- **Better PDF Handling:** Make text extraction for complicated PDF formats more efficient.

8. References

[1] R. Hajrizi and K. P. Nuçi, "Aspect-Based Sentiment Analysis in Education Domain," 2020.

[2] V. Gupta, V. Viswesh, C. Cone, and E. Unni, "Qualitative analysis of the impact of changes to the student evaluation of teaching process," *American Journal of Pharmaceutical Education*, vol. 84, no. 1. 2020.

[3] T. Shaik, X. Tao, C. Dann, H. Xie, Y. Li, and L. Galligan, "Sentiment analysis and opinion mining on educational data: A survey," Feb. 2023.

[4] J. Hill, "An in-depth review of the two most common feedback forms in English undergraduate programs: The connection of their consequential academic damages to students perceptions of their teachers, and how video-recorded feedback can combat these effects," *Int. J. English Lit.*, vol. 15, no. 1, pp. 1–10, Feb. 2024.

[5] I. Sindhu, S. Muhammad Daudpota, K. Badar, M. Bakhtyar, J. Baber, and M. Nurunnabi, "Aspect-Based Opinion Mining on Student's Feedback for Faculty Teaching Performance Evaluation," *IEEE Access*, vol. 7, pp. 108729–108741, 2019.

[6] A. Kathuria, A. Gupta, and R. K. Singla, "AOH-Senti: Aspect-Oriented Hybrid Approach to Sentiment Analysis of Students' Feedback," *SN Comput. Sci.*, vol. 4, no. 2, Mar. 2023.