Improving Student Engagement Using NLP Techniques
Final Report

Rimika Dhara & Akansha Kamineni
CSCI 5541 NLP: Rimika & Akansha
dhara015@umn.edu, kamin143@umn.edu

1 Introduction

Student feedback is a cornerstone of evaluating teaching effectiveness, course design, and academic resources, offering actionable insights to improve the educational experience. However, in high-enrollment courses, the sheer volume of feedback—often expressed as nuanced, open-ended text—presents significant challenges for traditional analysis methods, which rely heavily on manual processing or basic quantitative techniques. These methods struggle to capture the contextual and emotional depth of the feedback, leaving many insights unexplored.

Our project addresses this gap by introducing a structured framework to collect, analyze, and derive actionable insights from student feedback, using advanced Natural Language Processing (NLP) techniques. The objective is to empower educators with data-driven recommendations that improve teaching practices and enhance student engagement. By focusing on Computer Science students at the University of Minnesota, our objective is to analyze feedback from introductory-level courses, where foundational teaching plays a critical role in shaping long-term learning outcomes. Traditional feedback analysis methods fail to address the nuances in student sentiments, especially when dealing with large datasets and diverse linguistic expressions. Our objective is to overcome these challenges using scalable and sophisticated NLP models that can classify sentiments, detect emotions, and generate meaningful feedback summaries.

Our results align closely with human benchmarks, demonstrating accuracy and reliability across varied setups by adjusting for factors like course difficulty and employing robust preprocessing. This project provides actionable insights that improve student engagement and satisfaction, making it highly relevant for educators and administrators. Beyond education, the methodology extends to feedback analysis in domains like corporate training. By bridging NLP advancements with real-world applications, this scalable framework transforms how institutions interpret and respond to feedback, enhancing teaching and learning outcomes.

Ultimately, this project introduces a replicable, scalable, and innovative framework for analyzing student feedback, bridging the gap between theoretical advancements in NLP and their practical applications in education. By addressing both the technical challenges and the real-world needs of educators, our work represents a significant step forward in leveraging NLP to enhance teaching and learning outcomes.

2 Approach

Our project aims to address the challenges in analyzing large volumes of nuanced student feedback to derive actionable insights. This section outlines the methodology, the hypothesis behind our approach, the challenges encountered, and the novel aspects of our work.

2.1 Methodology

Step 1: Data Collection. We utilized a web scraper to extract student feedback from *RateMyProfessors (RMP)*, focusing on Computer Science professors at the University of Minnesota. The extracted data included both numerical ratings and open-ended textual comments. However, this dataset definitely has its limitations because it isn't as robust as student feedback should be. Additionally, there is a higher chance of the data being biased and clouded by smaller pool of student judgement. It is crucial to note that our approach was initially designed fro the SRT data that is collected by the University. However, we were not able to gain access to the data collected by the University.

Step 2: NLP Analysis.

We employed VADER Sentiment Analysis to classify comments as positive, neutral, or negative, providing a baseline for understanding the general tone of the feedback. To capture more nuanced emotional expressions, we utilized a pre-trained DistilRoBERTa model for emotion classification, identifying key emotions such as joy, anger, fear, and surprise within the textual feedback. Additionally, we leveraged NLTK and Gensim tools to extract key phrases and classify recurring tags for each professor. This approach highlighted themes such as "clear grading criteria," "engaging lectures," and "challenging coursework," offering a granular understanding of the feedback.

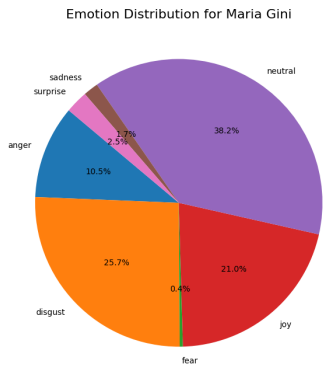

Figure 1: Robust emotion analysis visualized.

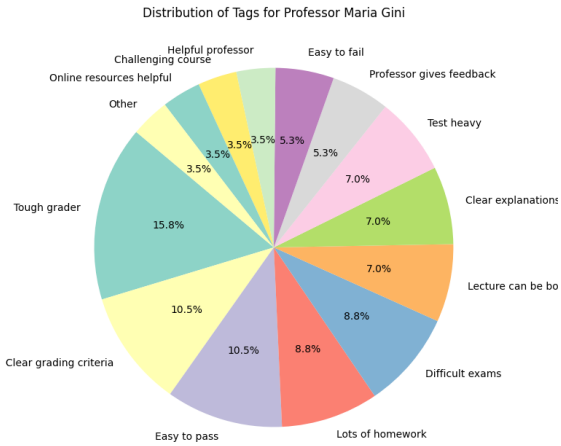

Figure 2: Key word tag occurrences visualized.

Step 3: Feedback Summarization. The processed data was used to prompt the Google Gemini model to generate structured feedback summaries and actionable recommendations for instructors. This step transformed raw insights into practical guidance.

Step 4: Evaluation. Human evaluations were conducted to compare model-generated outputs against manually curated benchmarks. Accuracy, precision, recall, and F1 scores were computed to validate the effectiveness of the sentiment and keyword analyses. Additionally, we assessed the quality and relevance of Gemini's feedback summaries through qualitative metrics.

2.2 Hypothesis and Rationale

We hypothesized that combining sentiment analysis with emotion and keyword classification would provide a comprehensive understanding of student feedback. By leveraging advanced NLP models like DistilRoBERTa and integrating contextual factors such as course difficulty, our approach aimed to reduce biases and enhance the reliability of results. We believed this framework would successfully address limitations in existing methods, offering both clarity and actionable insights.

2.3 Challenges and Limitations

Data Accessibility: A major challenge was the inability to access the Student Rating of Teaching (SRT) data due to privacy and security policies. As a result, we relied on publicly available RateMyProfessors (RMP) data, which provided only a partial view of student feedback.

Data Bias: RMP data is often biased, reflecting feedback from a self-selected group of students, which may not represent the broader population. Extreme opinions, whether overly positive or negative, further limited the reliability of our analysis.

Limitations of Existing Approaches: Traditional machine learning models like Support Vector Machines (SVMs) and Random Forests, commonly used in past studies, lack the contextual understanding of transformer-based models. While our approach addressed this gap, its effectiveness was constrained by the quality of the input data.

Generalization Challenges: Although we incorporated contextual adjustments, such as accounting for course difficulty, the absence of comprehensive datasets limited the depth and generalizability of our insights.

Future Directions: To overcome these limitations, future work should focus on collaborating with institutions to access anonymized SRT data

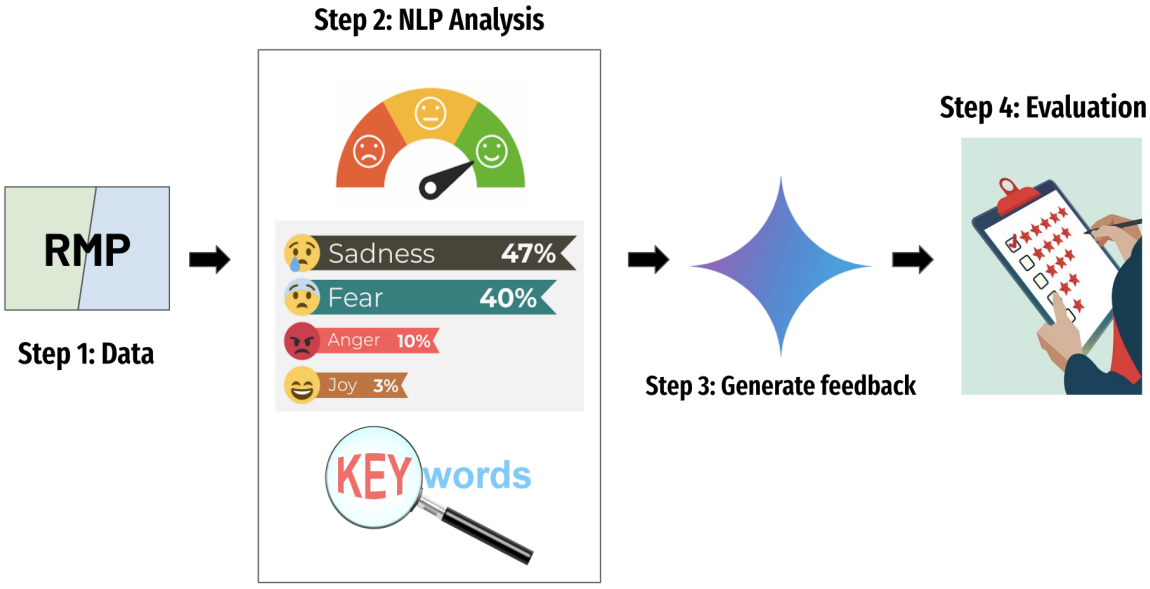

Figure 3: Overview of the methodology used, illustrating key stages in data processing, sentiment analysis, and keyword extraction for actionable insights.

and expanding datasets to reduce biases. Enhanced annotation methods and diverse feedback sources will improve the robustness and applicability of our framework.

2.4 Novelty

Our approach distinguishes itself through the use of state-of-the-art transformer models like BERT, which leverage bidirectional context representation to capture nuanced language features. We enhance this with an emotional weight system inspired by Rani and Kumar's wheel of emotions and integrate student crowdsourcing techniques to reduce biases during data preprocessing. Furthermore, unlike previous research, which typically focuses on static visualizations like word clouds, we provide instructors with actionable, personalized feedback based on the analysis, ensuring a real-world application that is both practical and innovative. Our approach stands out as a novel initiative because it seeks to integrate advanced NLP tools into the domain of student feedback analysis and academia. By doing so, we aim to enhance the quality of education by providing actionable insights that directly address student needs. This project bridges the gap between cutting-edge natural language processing techniques and their practical application in improving teaching practices and course design. Our work emphasizes creating a more data-driven and student-centric approach to academic feedback, fostering an environment where educational institutions can better align with the expectations and requirements of their students.

3 Experiment & Results

The main goal of our project shifted a few times as we worked through it. Originally, we wanted to explore student sentiment on professors to see how we could apply the NLP techniques we learned to education, an area where technology often isn't used to its full potential. As we got deeper into the project and looked at the literature, we noticed a gap—not just in how student sentiment was being analyzed, but also in how the results were being used. A lot of the research we saw relied on traditional machine learning models and stopped at identifying sentiment without much analysis on the implications of that sentiment. We realized we could focus on improving these processes in two ways: using specialized pre-trained models instead of traditional ones and exploring how the results could be applied in more actionable ways.

3.1 Measuring Success

We defined success in terms of both the accuracy of our models and the usefulness of their outputs. For the sentiment and emotion analysis, we measured accuracy by comparing the model's predictions with those of two human evaluators (Rimika

and Akansha). The evaluators read the comments and marked the emotions they believed were expressed. Their responses were compared to the model's predictions to calculate how often they matched. This analysis was limited to three professors, and we combined results across all three to minimize human effort.

For the keyword analysis, we took a slightly different approach. Instead of human evaluations, we compared the tags generated by our model with those listed on RateMyProfessors (RMP). To evaluate performance, we calculated metrics like accuracy, precision, recall, and F1 scores. Since RMP uses a limited set of predefined tags chosen by students, we anticipated some mismatch but felt this would still provide a solid benchmark for comparison.

Another key aspect of measuring success was ensuring the outputs were usable and meaningful for professors. To evaluate this, we reviewed the feedback generated by the Gemini model. We visually inspected the summaries and action items to ensure they aligned with the input tags and comments. We also shared the application with a small group of educators to gather qualitative feedback on its usefulness and readability.

3.2 Overall Results

For sentiment/emotion analysis, the results showed notable variations between emotions. Some, like "fear" and "sadness," were relatively easy to classify and achieved perfect accuracy. On the other hand, "joy" was not classified correctly at all, which we found surprising given that it was one of the only positive emotions on our list. "Anger" performed moderately well with an accuracy of 0.66, while "disgust," "neutral," and "surprise" each had an accuracy of 0.33. Overall, the model achieved an accuracy of 0.66 for this task. While this is below the commonly used threshold of 70%, we still considered it relatively successful for a few reasons. First, this project wasn't focused on high-stakes data where perfect accuracy is essential. Second, the small sample size (only two human evaluators and three professors) likely influenced the results, creating some variability.

For keyword analysis, the results were consistent across most professors. Using Maria Gini as an example, we achieved an accuracy score of 20%, while precision, recall, and F1 scores were all at 74.04%. The low accuracy score could be attributed to differences between our model's tags

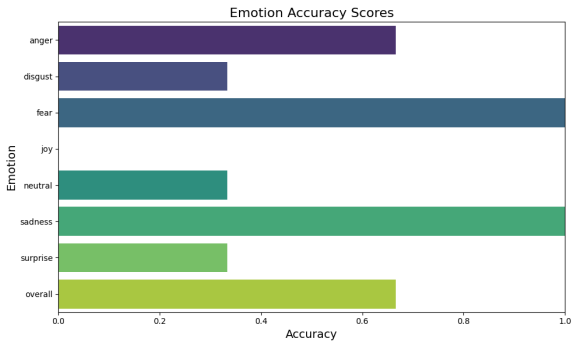

Figure 4: Accuracy of emotion classification across different emotions.

and RMP's limited, predefined set. Our model generated a broader range of tags, often capturing nuances in the comments that RMP's system might overlook. While this mismatch affected accuracy, we saw it as a strength, as our model provided more detailed insights.

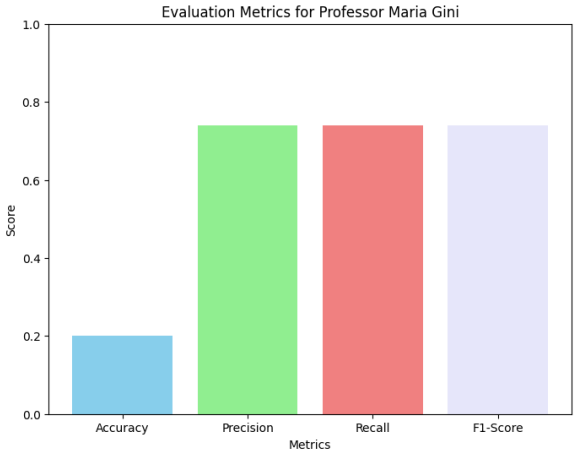

Figure 5: Accuracy of key words for a single professor.

Finally, for Gemini's feedback generation, we didn't use specific numerical metrics but conducted a qualitative sanity check. We reviewed the tags, summaries, and action items generated for the three professors we analyzed, as well as a small sample of others, to ensure consistency and readability. The feedback generally matched the input tags and comments, and we received positive responses from the educators who tested the application. Based on these evaluations, we considered the feedback generation process successful.

3.3 Failure Cases

While the project was successful in several areas, we identified a few challenges and areas for improvement. One notable issue was the model's

inability to classify "joy." This may have been due to the small sample size or a lack of positive emotions in the dataset. Since "joy" was one of the only positive emotions in the set, the model may not have had enough examples to learn from effectively. Addressing this would require expanding the dataset to include more examples of positive emotions.

Another challenge was developing clear evaluation metrics for the final outputs. Unlike traditional research projects, our work included a creative element—generating actionable feedback for professors—which made it difficult to use standard evaluation metrics. While we conducted qualitative evaluations and sanity checks, a more robust scoring system that integrates all inputs (like sentiments, keywords, and feedback) would improve the evaluation process and help optimize the model further.

To address these challenges, we plan to expand the dataset to include more diverse examples of emotions, especially positive ones, to improve classification accuracy. Additionally, we aim to develop a more comprehensive evaluation system that combines quantitative and qualitative metrics. This would allow us to better assess the alignment between input data, generated tags, and feedback. By addressing these areas, we hope to further refine the model and its outputs, making them even more useful for professors.

4 Discussion

Based on the results, we can conclude that our model performs relatively well in meeting the objectives of this project. The outputs provided the type of data necessary to build a web application that enables instructors to derive actionable insights from student feedback. A thorough review of the results, including comparisons with the initial dataset, revealed no significant outliers or inconsistencies, which supports the robustness of the model and its ability to produce meaningful analyses. While the models are not flawless, they successfully identify key strengths and areas of improvement, making them a valuable tool for instructors seeking to enhance teaching effectiveness and student engagement.

In terms of replicability, the methods and models used in this project were implemented using publicly available tools and frameworks, including VADER for sentiment analysis and pre-trained DistilRoBERTa for emotion classification. These tools, combined with detailed documentation of our pipeline and pre-processing methods, make it feasible for others to reproduce and extend our results. Additionally, our web application provides an interactive platform for deploying the models, further demonstrating the replicability and scalability of the project.

Our choice of dataset, primarily collected from RateMyProfessors, presents both opportunities and limitations. While the dataset provided valuable insights into student feedback, it does not encompass more formal or institution-specific feedback mechanisms such as Student Rating of Teaching (SRT) data. However, the methodology developed in this project could inspire other researchers to explore similar applications in the education domain or extend the analysis to alternative feedback datasets.

From an ethical perspective, there is potential harm if the insights derived from our models are misinterpreted or used punitively against instructors without appropriate context. To mitigate this risk, our web application emphasizes the importance of interpreting results as guidelines rather than definitive judgments. Furthermore, the ethical handling of sensitive data remains a priority, and the implementation of secure data handling and anonymization measures ensures that student feedback is used responsibly and confidentially.

Despite its strengths, the project has several limitations. The lack of access to comprehensive and institution-specific datasets, such as SRT data, limited the scope and granularity of our analysis. Additionally, the models may struggle with the inherent subjectivity and ambiguity in natural language feedback, leading to occasional misclassifications or oversimplifications of nuanced sentiments. Future work could address these limitations by incorporating more diverse and granular datasets, refining the models to handle context-specific language, and exploring hybrid approaches that combine machine learning with expert validation.

To extend this work, we propose formalizing instructor interviews to collect detailed feedback on improving the models' usability and visualization methods. Expanding the web application to dynamically integrate new feedback data and developing stronger security measures would also make the platform more robust and self-sufficient. Additionally, resolving privacy and security concerns around accessing SRT data could significantly en-

hance the project's impact by enabling deeper and more reliable analyses. By addressing these challenges, our work could evolve into a transformative tool for improving educational practices and student engagement.

5 Contributions

Rimika's Contributions: The work for this project was evenly divided. My contributions focused on key aspects from data collection to application development. I started by developing and running the web scraper to collect data from RMP, ensuring proper pre-processing of structured and unstructured data to create a clean dataset for analysis. During the literature survey, I researched advanced sentiment and emotion analysis techniques alongside Akansha, identifying gaps in existing methods and exploring the use of cutting-edge NLP models for extracting meaningful insights from nuanced student feedback. I implemented Sentiment Analysis using VADER to classify feedback into positive, neutral, or negative tones and extended this with Emotion Analysis using a pre-trained DistilRoBERTa model to capture finer emotional nuances like joy, anger, and fear. Additionally, I set up and refined the integration of Google Gemini to summarize feedback and generate actionable advice for instructors, iterating on prompt engineering to ensure outputs were accurate, efficient, and aligned with the project's objectives. My primary technical contribution was the end-to-end development of a static web application that integrates the analysis pipeline with an intuitive interface for instructors, enabling real-time interaction with feedback insights. This involved ensuring the application was scalable, user-friendly, and secure. Throughout the project, I gained invaluable experience in NLP, prompt engineering, web development, and practical problem-solving, making this an enriching learning opportunity that showcased the potential of combining technical innovation with user-centered design.

Akansha's Contributions: My contributions focused on various other key aspects of the project itself, as well as some of the deliverables. I focused on some of the initial project pitches and idea generation/idea gathering, which helped us frame the overall goals of the project as well as the work that needed to be completed. Although these did not end up getting used in our final project, I created the initial data collection forms for student feedback. I soon realized that we would not get the amount of data that would be needed from the surveys, and was able to propose that we pivot to the RateMyProfessors data. I also found various articles that contributed to the literature survey, defined the initial project goals, and created the overall project plan. My research focused on understanding what had already been done and how we could adapt and contribute to what has already been done by the community. I proposed that we make a usable product rather than focusing on the research aspect, since some of the initial research had been done and mainly needed optimization. For the project itself, I was able to implement the identification of key words using Gensim and NTLK, which I then used to create tags for each professor and converted into usable visual data for our web application. I was also able to create the initial framework for the Gemini API and integration, which Rimika was able to build off of for prompting and generating actionable feedback. While Rimika was setting up the web application, I was able to focus on the project deliverables required for the class. I created the main project poster and report webpage, which included a detailed diagram of how our code worked, our main objectives, results, and so on.

References

[1] Baker, R. S., Ocumpaugh, J., Andres, J. M. A. L., & Marino, M. T. (2017). Neurodiversity in education: Consequences for learning and teaching. ACII 2017. Retrieved from https://learninganalytics.upenn.edu/ryanbaker/ACII2017_183.pdf

[2] Brar, R. (n.d.). *Student Feedback Dataset* [Data set]. Kaggle. Retrieved from https://www.kaggle.com/datasets/brarajit18/student-feedback-dataset

[3] Dake, D.K., & Gyimah, E. (2023). Using sentiment analysis to evaluate qualitative students' responses. *Education and Information Technologies, Vol. 28*, 4629–4647. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9581765/

[4] Kasumba, R., & Neumman, M. (2024). Practical Sentiment Analysis for Education: The Power of Student Crowdsourcing. *Proceedings of the AAAI Conference on Artificial Intelligence, 38*(21), 23110-23118. https://doi.org/10.1609/aaai.v38i21.30356

[5] Rani, S., & Kumar, V. (2017). A sentiment analysis system to improve teaching and learning. *International Journal of Computer Applications, 174*(24), 28-33. Retrieved from https://ieeexplore.ieee.org/document/7924253

[6] Shaik, T., Tao, X., Dann, C., Xie, H., Li, Y., & Galligan, L. (2022). Sentiment analysis and opinion mining on educational data: A survey. *Data-Centric AI*. Retrieved from https://www.sciencedirect.com/science/article/pii/S2949719122000036