WIP: AI-based Sentiment Analysis and Grader Enhancements

Mr. Bobby F Hodgkinson, University of Colorado Boulder

Bobby Hodgkinson is an Associate Teaching Professor in the Smead Aerospace Engineering Sciences Department (AES) and co-manages the educational electronics and instrumentation shop. He assists students and researchers in the department for sensor and data acquisition needs as well as manages several lab courses and experiments. He is a member of the Professional Advisory Board for the senior capstone projects course. Prior to joining Smead Aerospace department in 2012, he was the lab manager at the Institute of Networked Autonomous Systems at the University of Florida, Gainesville where he focused on the research and development of small, autonomous aerial and underwater vehicles, sensors and actuators. He received a BS and MS degree from the Aerospace Engineering Sciences department at CU Boulder in 2010 and 2011 respectively.

Nathan Eric Whittenburg, University of Colorado Boulder

Nathan Whittenburg is currently pursuing a degree in Aerospace Engineering with a minor in Computer Science at the University of Colorado Boulder. He serves as a Lab Assistant in the Aerospace department, where his responsibilities include employing Large Language Models and Natural Language Processing to enhance educational outcomes, managing and maintaining lab equipment, and assisting students with conducting experimental procedures.

WIP: AI-based sentiment analysis and grader enhancements

Summary

In the realm of higher education, peer feedback for group activities gives instructors a valuable tool into the inner workings of a group. Peer assessment can also provide useful, constructive feedback to the individual participants. In experiential learning environments, particularly in disciplines like Aerospace Engineering, group work plays a valuable role to prepare students for a career in collaborative environments and feedback on an individual's performance can be a useful pedagogical tool. To enhance the peer review process, this study implements sentiment analysis, specifically using a roBERTa sentiment analysis model [1], to provide a quantitative assessment of reviews received by individual students. Additionally, the work quickly evolved to include AI-based constructive criticism paraphrasing to allow for timely individualized feedback in a large-enrollment setting. This work also explored the capabilities of an AI-based suite to aid report graders in order to improve the efficiency of the grading and feedback process for large scale laboratory classes. The motivation for this work is to investigate the utility of Artificial Intelligence as a way to increase the efficiency of the instructional team in large enrollment lab courses with enrollments on the order of 250+ students.

Background and Pedagogical Context

This work in progress aims to explore the correlation between sentiment analysis scores and numerical peer evaluations in an Aerospace Engineering sophomore experiential learning course. In recent years, sentiment analysis has emerged as a powerful tool for interpreting online comments and reviews, with applications ranging from product feedback to academic peer reviews. Basiri *et al.* [2] developed a novel approach for sentiment analysis by exploiting the comment histories of reviewers, demonstrating that considering the historical context of a reviewer's comments can enhance the accuracy of sentiment evaluations. Pankaj *et al.* [3] applied sentiment analysis to customer feedback data on Amazon, categorizing opinions into positive, negative, and neutral sentiments, showcasing its utility in understanding customer perceptions. In the context of academic peer reviews, Kim and Calvo [4] introduced a method for summarizing feedback in academic essay writing, employing sentiment score-based techniques to analyze reviews written by engineering students, highlighting the application of sentiment analysis in educational settings. Finally, Wang and Wan [5] focused on sentiment analysis of peer review texts for scholarly papers, proposing a multiple instance learning network with an abstract-based memory mechanism to predict overall recommendations and identify sentiment polarities in peer review texts, thereby demonstrating the potential of sentiment analysis in scholarly

communication. These studies collectively illustrate the diverse applications of sentiment analysis, from e-commerce to academic peer reviews, underscoring its significance in extracting valuable insights from textual data. Kastrati *et al.* [6] developed an automatic sentiment analysis framework for student reviews in MOOCs. The framework utilizes aspect-level sentiment analysis to identify opinions expressed towards specific aspects of a MOOC, reducing the need for manual data annotation and providing efficient sentiment categorization in large-scale online education settings. One study [7] proposes a system to analyze the group emotions of students in a classroom setting using multimodal sentiment prediction. It combines audio, video, and text data using deep learning models to assess and predict the overall group sentiments during lectures, demonstrating the potential of AI in understanding classroom dynamics.

Methodology

Using natural language processing and machine learning, this work delves into the sentiment expressed in peer reviews in a project based course. The hypothesis posits that sentiment analysis can offer a nuanced view of the peer reviews, highlighting elements like positivity and constructiveness, particularly in critical feedback.

This investigation also leverages a large language model (LLM) to transform peer feedback into more constructive input, especially for lower-performing students. Additionally, this approach enables automated monitoring of the professionalism in student feedback, a task that is challenging for the teaching team to manage manually due to the sheer volume of anonymous submissions in large scale environments. An AI-based tool can efficiently perform this task with minimal strain on the instructional staff.

The course serving as our proving ground emphasizes group work, where students assess their peers' contributions through comment-based reviews. They provide both a summary of contributions and areas for improvement, and a numerical score on a scale of 1-10. Historically, the teaching team has relied primarily on these numerical scores to determine individual grades due to the impracticality of thoroughly analyzing approximately 1500 peer review submissions per lab activity. This method, however, is limited by potential subjectivity and bias, issues that sentiment analysis could potentially mitigate.

Peer evaluations, both quantitative and qualitative, are useful in large enrollment lab settings for assessing individual performances within groups. Studies confirm that these evaluations not only predict individual performance but also align well with other effectiveness metrics within team-based projects in engineering education [8–10]. In our investigatory course, students rate each teammate and provide a summary, which this study aims to correlate with the numerical scores. It is hypothesized that students may tend to assign higher numerical scores than what the qualitative feedback suggests, especially for lower-performing peers. The large-scale nature of the course makes it impractical for instructors to individually assess hundreds of qualitative statements for grade assignment.

This study involves approximately 250 sophomore Aerospace Engineering students in their first semester course. The development of automated tools for this research facilitates potential application in various courses and educational levels. As this work is still in progress, this paper

will primarily discuss the methodology and tool development, providing only a preliminary summary of the findings.

The sentiment analysis project not only demonstrated the potential of AI-based techniques in extracting insights from student peer evaluations but also marked a successful integration with our learning management system. These achievements paved the way for the initial development of an AI-based grading assistant. Mindful of the ethical considerations associated with a fully automated grader, we focused on creating a tool to assist, rather than replace, human graders.

This AI assistant streamlines the evaluation of group lab reports, traditionally a time-intensive task in large-scale courses. By uploading nameless lab reports to a LLM through an API interface, the system efficiently identifies and highlights segments that align closely with specific rubric items. This process is designed to isolate the most relevant sections of each report, providing a preliminary guide for human graders. The aim is to enhance grading efficiency and consistency while maintaining the crucial human element in evaluation and feedback.

In its current form, the tool offers a glimpse into the future of grading assistance, emphasizing precision and time-savings in the grading process. Although still in its early stages, the AI assistant promises to evolve into a more sophisticated aid, complementing the nuanced judgement of human graders. This approach, born from the insights of sentiment analysis, underscores our ongoing desire to harness AI technologies to improve instuctional efficiences.

The current literature on AI-assisted graders is more limited than sentiment analysis studies. Marchiori [11] introduces a command-line interface tool designed to assist in managing student work in computer science lab sections. This tool streamlines grading tasks and provides prompt, consistent feedback. It demonstrates the efficiency of automated tools in managing lab work and the potential for similar applications in lab report grading. Weinthal *et al.* [12] discuss the implementation of technology to ensure academic integrity in engineering labs. The study highlights various methods, such as the use of security features on grading sheets, mandatory lab image uploads, and metadata tracking of lab reports. These measures aim to authenticate lab grades and student work, maintaining integrity in the grading process.

Methodology - Sentiment Analysis

In the investigatory course, students are assigned to groups of four to six for lab sessions. Upon submitting their final report, they evaluate their lab mates by providing feedback on their performance and rating it on a scale of one to ten. To draw a quantitative comparison between the comments and scores, we perform sentiment analysis on the comments. Sentiment analysis is the process of assigning a number to how positive/negative a piece of text is. This number then allows us to directly compare the positively/negativity of the comments a student gives to a lab mate to the score they give.

First, data collection is done by assigning students a Google Form, shown in Figure 1, in which they are asked to select the names, provide comments, and provide scores for each of their lab mates. This Google Form is then exported to a comma-separated values (.csv) file that is later read by a Python script.

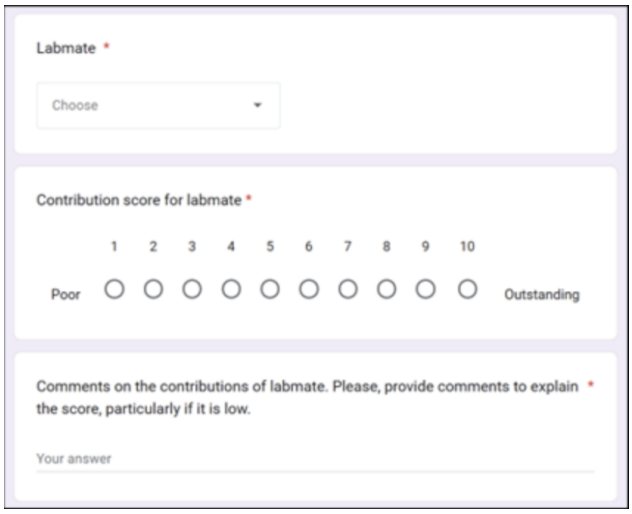


Figure 1: A screenshot of the Google Form assigned to students

Once data is collected, the Python script performs sentiment analysis on all of the comments for each student. The process of performing sentiment analysis involves the Hugging Face "Twitter-roBERTa-base for Sentiment Analysis". This is a language model that is trained on "approximately 58 million tweets and fine-tuned for sentiment analysis". The model was selected for its user-friendly interface. It is accessed via the Hugging Face API within a Python script. Each student comment is transmitted to the language model through an HTTP request via the API, which then returns a sentiment score. This sentiment score is composed of positive, negative, and neutral scores expressed as a percentage. Figure 2 shows an example of the sentiment score returned to the script, where "LABEL2" is positive, "LABEL1" is neutral, and "LABEL0" is negative.

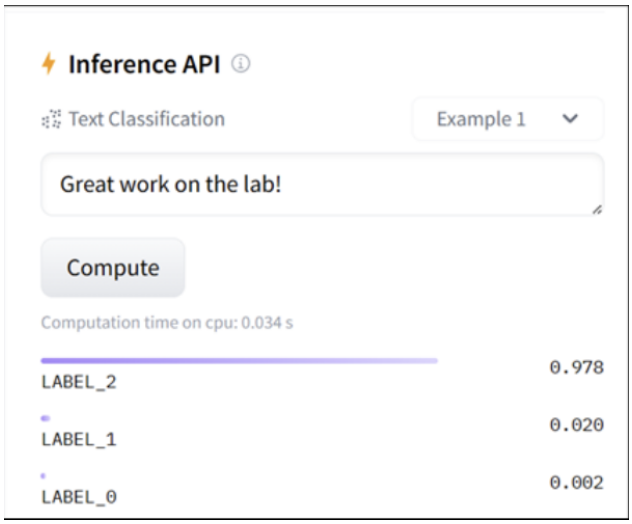


Figure 2: Example of the response sent back from the Hugging Face API

After we send each of the comments to the API and receive a sentiment score for every student, a Python dictionary (associative array data structure) is compiled with the student's name as the

key, and their student provided scores, comments, and sentiment scores as the values as shown in Figure 3. Simply, every student now has these values associated with their name, making it fast and easy to access this data in our Python script.

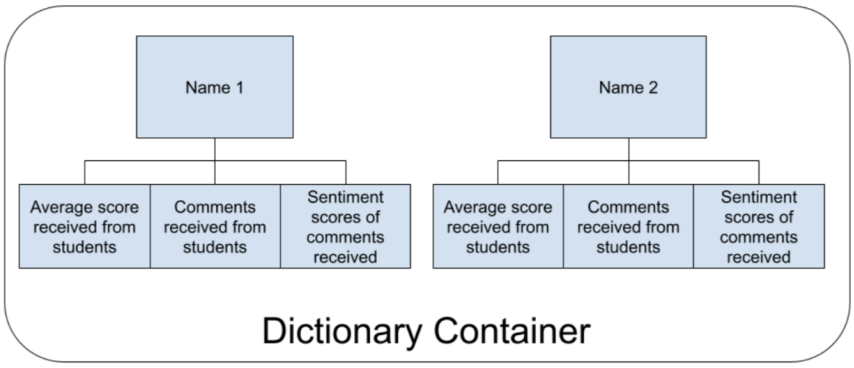


Figure 3: Visualization of the Python Dictionary

Historically, we've observed that students receiving low scores often receive feedback that lacks constructiveness and can be very critical. To address this, we sought a method to transform such comments into positive, constructive criticism. By integrating the OpenAI API into our Python script, we automatically paraphrase comments for students who revieve an average peer-provided score of 3 or lower. (It is worth noting that FERPA compliance is maintained by redacting any identifying information before sending to an LLM). This process uses OpenAI's 'gpt-3.5-turbo' language model, selected for its speed and affordability. We send the original feedback to the OpenAI LLM with a specific prompt: 'Give the student constructive feedback based on this comment. Be nice, but make sure your comment informs the student how they should improve if the comment implies they need to.' Our experience and adjustments have refined this prompt to ensure the paraphrased comment is encouraging yet informative, helping students understand where and how to improve. Once we receive a paraphrased comment from the API, we append it to the aforementioned dictionary container under the name of the student who's comment was paraphrased.

Finally, we've streamlined the process of returning sentiment analysis and peer feedback results to every student with several tools in our Python script. This includes the ReportLab module, which lets us create PDFs directly in the script, and the Instructure Canvas API for distributing these PDFs to Canvas. Each student receives a detailed PDF that outlines the peer feedback they received. This includes the comments made about them, their average peer score, and a breakdown of sentiment scores—minimum, maximum, and average—based on sentiment analysis. Students who receive paraphrased feedback receive the paraphrased feedback in lieu of the original feedback. We upload these PDFs directly to Canvas using the CanvasAPI module which uses HTTP requests to communicate with Canvas, similar to how we access the HuggingFace and OpenAI APIs.

Methodology - AI Grading Assistant

Building on the insights from our sentiment analysis study, we explored the use of generative AI to enhance the grading process. Mindful of ethical concerns around diminishing human involvement in grading, we developed a tool that identifies specific parts of lab reports corresponding to a grading rubric, while human graders continue to evaluate the content. This tool again uses the OpenAI API. Moreover, best practices and techniques from the field of 'prompt engineering' played a crucial role in achieving consistent results from the LLM. To optimize our interactions with the LLM, we provide four key types of information: instruction/role assignment, context, input data, and the desired response format.

First, the instruction/role assignment we provide for this task is as follows:

"Please find in the text where each of the following prompts are addressed. You must find the specific sentence where each prompt is addressed. The prompts whose answer is to be found are as follows:"

This instruction accurately and concisely tells the AI what we want it to do with our input data. Furthermore, it provides a transition to the next section of our prompt, the context. Context specifies to the AI how it should execute its instruction. The context we provide consists of each of the sections we want the AI to find. We provide these sections as a numerical list because it gives us a shorthand to specify sections in a PDF by assigning numbers to each section. Below is a snippet of the context from our prompt:

"Section 5. Explain the whiffletree design approach, discretization of loads (rectangular/trapezoidal/ triangular), and determination of bar lengths. How was the code checked for accuracy?

Section 8. Discuss the process and include Cp vs x/c plots for various angles of attack to go along with your discussion. Compare your results to the NACA data.

Section 9. Provide expressions for internal bending moment, stress and strain, and deflection due to rectangular load."

Some rubric section numbers (such as 6 and 7 from above) are excluded from our prompt. These sections are excluded because they ask the student to provide plots and/or images. In our script, we provide the AI with raw text data from the lab report, thus the AI is unable to find any plots/images. Leaving these sections in the context prompt would only confuse the AI.

Next, we provide the input data to the AI, which in this case is the lab submission from the student. We take the submitted PDF and use the PyMuPDF module in Python to extract all of the text from the document into a string. This string is then appended to the prompt that will be sent to the AI. (It should be noted that FERPA compliance is maintained by redacting any identifying information before sending the text to the LLM.)

Finally, we tell the AI how it should respond. To use the response from the AI in any meaningful way without tediously parsing the text in its response, we found that asking the AI to return a response in JavaScript Object Notation (JSON) is best. Simply, JSON is a text file format that allows for easy conversion from a human-readable text to a computer-readable object. We ask the

AI to associate 3-4 words that indicate the start of a section with the corresponding section number so that we can later identify the start location of each section in our script. The exact response format given in our prompt is shown below.

Please format your output the following way below. Just enter 3-4 words that identify the start of a section into the JSON object. Sections of the text should NOT overlap. If a specific section is not discussed in the text, you may add "MISSING" to its start attribute in the JSON object:
{"label": "Section 1", "start": "3-4 words that indicate where Prompt 1 is addressed"}

Note that JSON objects are associative containers. Each line contained in curly braces is an object with properties "label" and "start". Thus, we are essentially associating a section label with a string of words that start the section.

Now that we've obtained the response in JSON format, we can highlight sections in the lab report PDF that correspond to each prompt. Using the PyMuPDF module in our Python script, we can edit the PDF with highlighted annotations. This module allows us to search for specific text strings within the PDF, highlight these segments, and insert a text box indicating the section number that corresponds to the addressed prompt. Thus, we iterate through each entry in the JSON file, search for the "3-4 words that indicate where Prompt X is addressed", highlight those words, and add a text label to denote the section number ("label" in the JSON object).

In short, we preserve the original lab report while adding section labels that indicate to the graders which prompt the report addresses in each section of text. It is worth noting that this approach may be overly complicated with the recent deployment of ChatGPT 4.0, where PDFs can be uploaded and modified directly by the LLM. Nevertheless, the course under investigation has upward of 40 submissions which could quickly reach any ChatGPT data limits. Additionally, this approach is mostly automated so dozens to even hundreds of group reports could be analyzed and highlighted with minimal user interaction.

Results - Sentiment Analysis

We present findings from the application of our sentiment analysis technique on two lab activities each lasting about seven weeks. Despite the limited scope, the participation was substantial, with over 250 students providing peer evaluations for their four to five teammates, culminating in more than 1,000 peer reviews for processing and analysis per lab activity. Figure 4 illustrates the relationship between the peer-evaluated scores (ranging from 1, indicating poor performance, to 10, signifying an ideal teammate) and the positive sentiment analysis scores for Lab 1. Figure 5 illustrates the relationship between the peer-evaluated scores and the negative sentiment analysis scores for Lab 1.

Table 1 presents the Pearson correlation coefficients for the two lab activities. This statistical measure, varying from -1 to 1, is indicative of the strength and direction of a linear relationship between two variables. A coefficient of 1 implies a perfect positive correlation, -1 denotes a perfect negative correlation, and a value near 0 suggests no linear correlation.

As expected, the results reveal a strong correlation between the student-provided scores and both positive and negative sentiment scores. Notably, the negative sentiment score show a slightly stronger negative correlation with the student-provided score than the positive sentiment score.

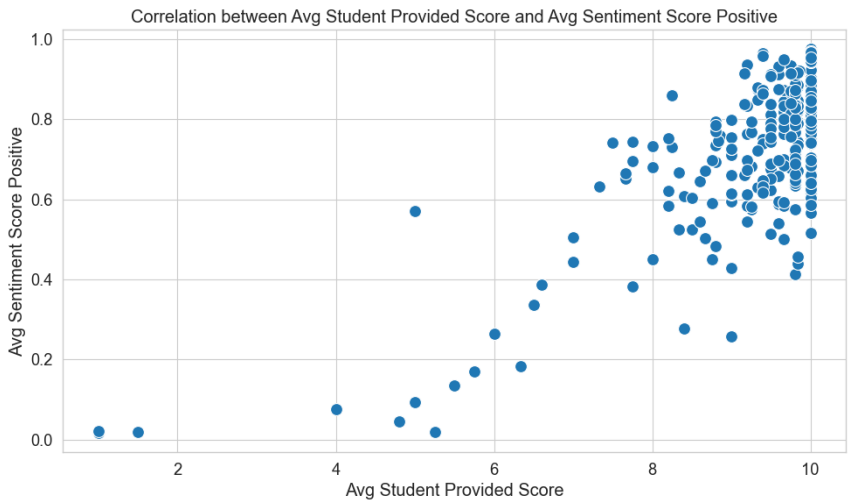


Figure 4: Positive sentiment score vs average student provided score

Table 1: Correlation between Student Provided Score and Sentiment Scores

Comparison	**Lab 1**	**Lab 2**
Student Provided Score and Sentiment Score Positive	0.569	0.737
Student Provided Score and Sentiment Score Neutral	0.17	-0.3731
Student Provided Score and Sentiment Score Negative	-0.879	-0.867

This could suggest that the sentiment analysis tool may be more adept at identifying lack of negative sentiment than at recognizing positive sentiment. Another possibility is that students' feedback for high performing scores are not as positive as the negative feedback students give for lower performing group mates. However, given the limited data set and the nascent stage of the methodology, these findings should be approached with caution. They are not definitive but indicate a trend worth exploring further.

What stands out is the relatively strong grouping of sentiment scores with the scores provided by students for the few lower-performing individuals. This correlation hints at the potential of sentiment analysis to provide meaningful insights into the small population of students who seem to perform well below the expectations of their group mates. We make an attempt to refine the analysis with the sentiment score where we consolidate positive, neutral, and negative feedback into a unified sentiment score. This is achieved by applying specific weights to each category: +1 for positive, 0 for neutral, and -1 for negative feedback. By summing these weighted values, we generate a composite sentiment score that effectively encapsulates the three distinct sentiment categories into a singular, insightful measure. We can then set various thresholds, depending on preference, to determine which students should be consulted about their performance in an attempt to improve their efforts in a group setting in the future. These results are shown in Figure 6. Using this approach we could elect to take a closer look at any student who has a weighted sentiment score less than 0, indicating that their peers' overall feedback sentiment was more negative than positive.

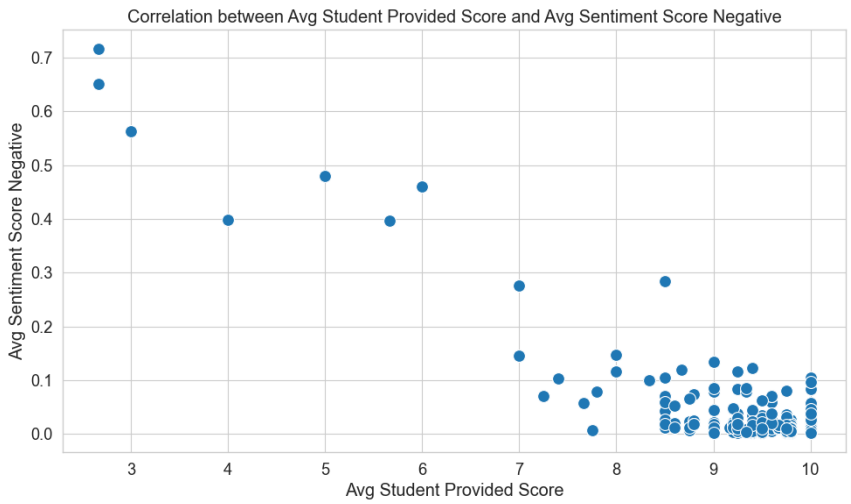


Figure 5: Negative sentiment score vs average student provided score

Results - Peer Feedback Paraphrasing

The sentiment analysis approach allows a relatively efficient way to pinpoint specific students who may benefit from constructive critiscm paraphrased peer feedback. Using the threshold technique from the previous section, we present a few examples using actual peer feedback from our study. These examples showcase the original peer review and the corresponding reinterpreted version by the LLM. The contrast between the two reviews highlight not only the ability of the LLM to retain the essence of the feedback but also its potential in rephrasing it in a manner that is more encouraging and less likely to be perceived as discouraging or overly critical.

The methodology involved feeding the original, less favorable peer reviews into an LLM. The model was then tasked with reinterpreting these reviews, with a focus on maintaining the core feedback while transforming the tone and content into a more positive and constructive format. This process aimed to preserve the original intent and substance of the feedback but present it in a manner that could be more readily received and acted upon by the student. Example peer feedback and the paraphrased result are shown in Table 2.

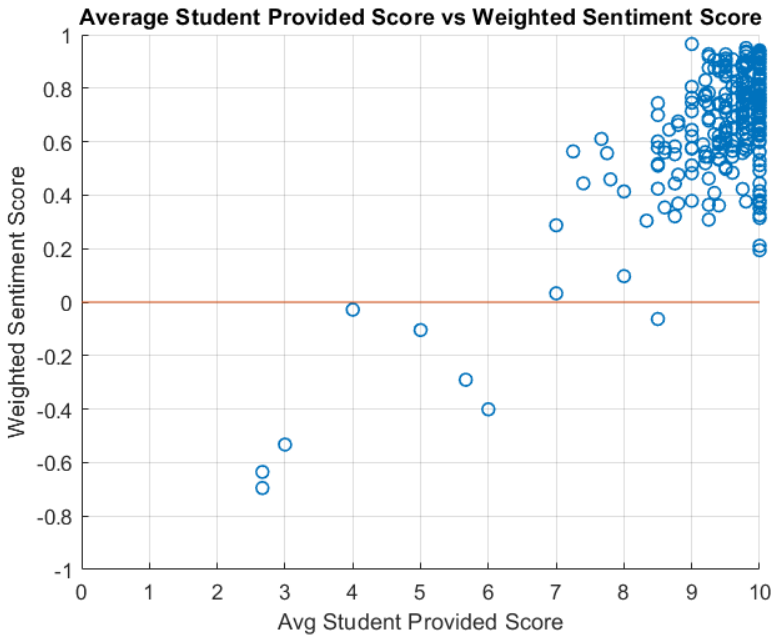


Figure 6: Negative sentiment score vs average student provided score

Original Comment	Paraphrased Comment
Didn't come to any of the lab periods, or do any of the work.	Attending lab periods and actively participating is important for your learning. To improve, make an effort to attend all the lab periods and complete the assigned work. This will enhance your understanding of the subject and contribute to your growth as a student.
Communicated literally twice with us, never showed up, did nothing on the lab or lab report	It's important to actively participate in lab groups by attending meetings and contributing to the tasks assigned. Going forward, make an effort to improve your communication and engagement with the group to ensure a more productive collaboration.
[REDACTED] was barely involved in the duration of this lab. His absence resulted in a lot of slack being left that was unfairly forced onto other structures team members. It felt as if we had 3 people on our subteam, when we should have had four. He was at the first couple labs, and then we never really heard from him again. He consistently missed labs, and on the ones that he showed up to, he was not very involved. On top of not contributing to the weekly deliverables, he did not contribute at all to the lab report	To improve your performance in future labs, it is crucial that you attend consistently and actively engage with your team. By being present and actively participating, you can ensure fair distribution of tasks and contribute to the success of your subteam. Additionally, make an effort to contribute to the weekly deliverables and collaborate on the lab report.

Table 2: Original Comments vs. Comments Paraphrased by LLM

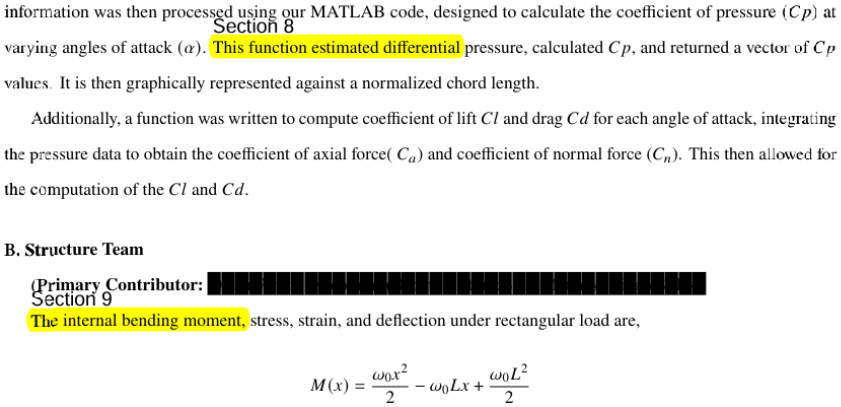
These examples serve as a testament to the capabilities of AI in educational settings, particularly in the context of peer reviews. By leveraging the power of an LLM to reinterpret feedback, we open up possibilities for a more nuanced and supportive educational environment where feedback, even when critical, is conveyed in a way that is constructive and empowering. This approach also underscores the potential of AI to augment human input, transforming it into a form that is potentially more effective for learning and personal development.

As with any AI-driven approach, this methodology is subject to continuous refinement and validation. The initial results, however, are promising and suggest significant potential for AI to enhance the quality and impact of peer feedback in educational settings.

Results - AI Grading Assistant

The concluding aspect of our study delved into evaluating the efficacy of a LLM as a grading assistant. An illustrative example of the tool's output is presented in Figure 7. This figure highlights the pertinent sections of a sample lab report, correlating them with specific criteria from the grading rubric. A statement from one of the graders encapsulates the utility of this approach: "The most challenging aspect of grading is pinpointing where students address each rubric item. The highlighted submissions significantly streamlined the grading process, enhancing efficiency."

Moreover, the course professor observed a notable reduction in the time needed to finalize grades post-submission closure. Historically, the turnaround time for grading this particular lab report for approximately 40 group submissions was about five calendar days. With the introduction of the highlighted grading assistant, this duration was reduced to under three calendar days. It's important to note that the grading personnel varied, making a direct comparison impossible. Nonetheless, this initial trial suggests a potential increase in grading efficiency and motivates further investigation.


information was then processed using our MATLAB code, designed to calculate the coefficient of pressure (Cp) at
Section 8
varying angles of attack (α). This function estimated differential pressure, calculated Cp, and returned a vector of Cp
values. It is then graphically represented against a normalized chord length.

Additionally, a function was written to compute coefficient of lift Cl and drag Cd for each angle of attack, integrating the pressure data to obtain the coefficient of axial force(C_a) and coefficient of normal force (C_n). This then allowed for the computation of the Cl and Cd.

B. Structure Team

(**Primary Contributor:** [redacted]
Section 9
The internal bending moment, stress, strain, and deflection under rectangular load are,

$$M(x) = \frac{\omega_0 x^2}{2} - \omega_0 L x + \frac{\omega_0 L^2}{2}$$

Figure 7: Sample student submission highlight with section titles. In the rubric, Section 8 asks the student to "discuss the process and include Cp vs x/c plots for various angles of attack to go along with your discussion." Section 9 asks the student to "provide expressions for internal bending moment, stress and strain, and deflection due to rectangular load.

Conclusion

Our investigation holds the potential to contribute to the ongoing discourse on peer assessment in educational settings. It aims to provide educators with insights into the advantages and challenges of integrating sentiment analysis and AI-based paraphrasing into the evaluation process. The goal of our study is to enhance the fairness and objectivity of peer assessments and help the students improve their performance from the constructive feedback of their peers, thereby improving the learning experience even in large-enrollment environments.

Given that this work is in the preliminary stages, there are several avenues for further, more detailed exploration to enhance our findings. There are also promising directions for future advancements. One key area of future research involves examining the discrepancies between student-provided reviews and the output of the sentiment analysis tool. Another aspect we intend to explore is the impact of modified peer feedback, rephrased for constructiveness by the AI tool, on student performance. Additionally, we are keen to quantify the efficiency gains achieved by graders using AI-highlighted reports compared to traditional grading methods without such assistance.

We also envision an extension of the grading assistant to offer groups a rapid feedback mechanism for their lab reports. In this proposed application, students could submit a 'final draft' of their report and receive AI-generated feedback on how well their submission aligns with specific rubric items. While we recognize the ethical considerations associated with employing AI for direct grade assignment, we believe that AI tools, when used appropriately, can significantly enhance the educational experience. Our aim is not to replace human grading but to augment it, facilitating a more efficient and effective learning and assessment environment regardless of enrollment numbers.

References

[1] C. NLP, "RoBERTa sentiment analysis model," https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment, Accessed: 2024.

[2] M. E. Basiri, N. Ghasem-Aghaee, and A. Naghsh-Nilchi, "Exploiting reviewers' comment histories for sentiment analysis," *Journal of Information Science*, vol. 40, pp. 313 – 328, 2014.

[3] Pankaj, P. Pandey, Muskan, and N. Soni, "Sentiment analysis on customer feedback data: Amazon product reviews," in *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, 2019, pp. 320–322.

[4] S. Kim and R. Calvo, "Sentiment-oriented summarisation of peer reviews," 2011.

[5] K. Wang and X. Wan, "Sentiment analysis of peer review texts for scholarly papers," in *The 41st International ACM SIGIR Conference on Research Development in Information Retrieval*, 2018.

[6] Z. Kastrati, A. S. Imran, and A. Kurti, "Weakly supervised framework for aspect-based sentiment analysis on students' reviews of moocs," *IEEE Access*, vol. 8, pp. 106 799–106 810, 2020.

[7] "Multimodal decision-level group sentiment prediction of students in classrooms," *International Journal of Innovative Technology and Exploring Engineering*, 2019.

[8] D. Gransberg, "Quantifying the impact of peer evaluations on student team project grading," *International Journal of Construction Education and Research*, vol. 6, pp. 17–32, 2010. [Online]. Available: https://consensus.app/papers/quantifying-impact-peer-evaluations-student-team-project-gransberg/00f09b7565c956f788cdd94c034bd473/?utm_source=chatgpt

[9] J. Wang and P. Imbrie, "Assessing team effectiveness: Comparing peer evaluations to a team effectiveness instrument," 2009. [Online]. Available: https://consensus.app/papers/assessing-team-effectiveness-comparing-peer-evaluations-wang/21471e575cbf52e284ea54aadeef6f60/?utm_source=chatgpt

[10] C. Wigal, "The use of peer evaluations to measure student performance and critical thinking ability," in *2007 37th Annual Frontiers In Education Conference - Global Engineering: Knowledge Without Borders, Opportunities Without Passports*, 2007, pp. S3B–7–S3B–12. [Online]. Available: https://consensus.app/papers/peer-evaluations-measure-student-performance-thinking-wigal/9b44579655a45376a9204ee5131629c6/?utm_source=chatgpt

[11] A. Marchiori, "Labtool: A command-line interface lab assistant and assessment tool," in *Proceedings of the 53rd ACM Technical Symposium on Computer Science Education - Volume 1*, 2022.

[12] C. P. Weinthal, M. M. Larrondo-Petrie, and L. F. Zapata-Rivera, “Academic integrity assurance methods and tools for laboratory settings,” in *2019 IEEE Frontiers in Education Conference (FIE)*, 2019, pp. 1–6.