Engineering Educators Bringing the World Together
2025 ASEE Annual Conference & Exposition
Palais des congrès de Montréal, Montréal, QC • June 22–25, 2025
ASEE
Paper ID #46985

BOARD #483: Student Feedback Analysis Using Natural Language Processing (NLP) and Sentiment Analysis

Ms. Sharmin Jahan Badhan, independent researcher

Sharmin Jahan Badhan is an independent researcher. She received an M.S. in Computer Science from United International University. In addition to her research interests in Artificial Intelligence, Deep Learning, and Natural Language Processing, she is actively engaged in exploring innovative applications of these technologies in construction site environments.

Dr. Rei Samsami, University of New Haven

Reihaneh Samsami (Ph.D., P.E.) has joined the University of New Haven (UNH), as a faculty in Construction Management, in Fall 2022. She has contributed to a new MS in Construction Management program development as the program director. She has also been involved in Entrepreneurial Mindset Learning by KEEN and Open Pedagogy at UNH. In addition to Engineering Education, she has 4+ years of experience in working with Departments of Transportation (DOTs) as a Graduate Research Assistant. Her research is positioned at the intersection of Automated Construction Inspection, Construction Information Modeling, and Data-Driven Decision-Making for project managers, contractors, inspectors, and other project stakeholders.

Dr. Goli Nossoni, University of New Haven

Dr. Goli Nossoni is currently an Associate Professor in the Department of Civil and Environmental Engineering at University of New Haven. She received her M.S. and Ph.D. from Michigan State University in civil engineering. In addition to her interest in

Student Feedback Analysis Using Natural Language Processing (NLP) and Sentiment Analysis

Abstract

This study explores the use of sentiment analysis to interpret qualitative student feedback from course evaluations, addressing the challenge of class imbalance, particularly the low number of negative comments. To balance the dataset, back-translation is applied, translating negative comments into Japanese and back into English, thus augmenting the dataset without altering the sentiment. A fine-tuned DistilRoBERTa model, based on the pre-trained RoBERTa architecture, is used for sentiment classification of 377 comments from 13 Civil Engineering courses. The model achieves a 90% testing accuracy.

Pearson Correlation analysis reveals a moderate positive correlation (r = 0.42) between sentiment polarity and quantitative review scores, suggesting that more positive sentiment generally aligns with higher ratings. Linear regression analysis further supports this relationship, with a coefficient of 0.24, indicating that a unit increase in sentiment polarity corresponds to a 0.24 unit increase in review scores. While sentiment analysis shows potential for predicting course effectiveness, the moderate strength of these correlations suggests that other factors influence student evaluations. This study demonstrates the value of sentiment analysis in understanding student feedback, while highlighting the complexity of capturing all dimensions of course effectiveness through sentiment alone.

Keywords: Student Feedback, Course Evaluation, Natural Language Processing (NLP), Sentiment Analysis, Qualitative Text Analysis

1. Introduction

Academic institutions regularly seek to assess and improve the quality of their educational offerings, often relying on student feedback through course evaluations. These evaluations typically involve both quantitative and qualitative components, with the former providing numerical assessments of course elements such as teaching effectiveness, course content, and overall satisfaction. The qualitative aspect, in the form of open-ended comments, offers valuable insights into students' personal experiences, highlighting elements of teaching and learning that are difficult to quantify, such as engagement, inclusivity, and the clarity of communication. While quantitative data is often straightforward to analyze and provides actionable insights, qualitative feedback tends to be underutilized due to its complexity and the labor-intensive nature of manual analysis.

Sentiment Analysis, a technique developed in recent years to evaluate emotions and opinions expressed in text, presents a promising solution to this challenge. By automating the process of analyzing the polarity of qualitative feedback, categorizing comments as positive, negative, or neutral, Sentiment Analysis can provide a more efficient and scalable way to interpret and utilize open-ended student responses. This paper explores the application of Sentiment Analysis to course evaluations, specifically using a pre-trained model named DistilRoBERTa to analyze student comments and derive meaningful insights about teaching effectiveness and course quality.

The goal of this research is to develop a methodology that allows for a quantitative assessment of the sentiment expressed in qualitative course evaluations, thereby complementing traditional quantitative ratings. By correlating sentiment with ratings from the quantitative sections, this approach aims to uncover deeper insights into student perceptions of specific aspects of the course, such as teaching style, course materials, and overall satisfaction. Additionally, it offers the potential to analyze feedback across diverse student populations and academic disciplines, providing institutions with a more comprehensive and nuanced understanding of teaching effectiveness.

Integrating Sentiment Analysis into course evaluation processes could transform the way academic institutions interpret and act on qualitative feedback. Rather than treating open-ended responses as secondary, this approach would enable a more data-driven and statistically rigorous analysis of the learning experience. Ultimately, combining both qualitative and quantitative data could lead to more informed decisions regarding course design, teaching improvements, and institutional policies, thereby enhancing the learning environment for students and supporting ongoing professional development for instructors.

2. Background

The approach of Machine Learning (ML) and Neural Network architecture has significantly advanced the field of Natural Language Processing (NLP), introducing methods that utilize large-scale datasets and complex models for sentiment analysis and related tasks. For instance, traditional supervised algorithms like Support Vector Machines (SVM) and Naïve Bayes (NB) have been foundational in classifying sentiments based on labeled data, with metrics such as accuracy, precision, and recall, commonly employed to assess their performance [1], [2], [3]. These algorithms often benefit from preprocessing techniques such as Term Frequency-Inverse Document Frequency (TF-IDF) to enhance feature representation and improve classification accuracy [4].

With the progression of computational capabilities, Neural Network architectures like Artificial Neural Networks (ANN) and Long Short-Term Memory (LSTM) models have emerged as powerful tools for sentiment analysis. ANNs, inspired by the structure of the human brain, consist of layers of interconnected neurons capable of learning complex relationships in high-dimensional data, for instance, ANNs have been demonstrated to outperform traditional methods such as Naïve Bayes and SVM in scenarios involving significant data, as they can capture complex relationships in high-dimensional spaces [2], [5]. In parallel, Long Short-Term Memory (LSTM) networks, an extension of Recurrent Neural Networks (RNNs).are designed to retain long-term dependencies within sequential data. LSTMs are particularly well-suited to textual analysis, as they capture the evolving structure and meaning of language. In educational applications, LSTMs have been employed to classify feedback based on instructional components such as pedagogy, behavior, and subject matter knowledge [6]. This has been further enhanced by incorporating domain-specific word embeddings, leading to more accurate and context-aware sentiment predictions.

Unsupervised learning models, which do not rely on labeled inputs, are also useful in sentiment analysis for detecting patterns and topics in unstructured text. Techniques have also found utility in sentiment analysis, particularly for tasks such as topic modeling and polarity detection. Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorization (NMF) have been employed to analyze unstructured student feedback, offering insights into themes and prevalent topics without the need for labeled data. These models identify latent structures in text without relying on labeled training data. Techniques such as LDA and NMF help discover themes within feedback, supporting further analysis and supervised modeling. [7], [8]. In the context of this study, unsupervised methods provide a foundation for understanding dominant themes and augment the sentiment classification process when labeled data are limited. Hybrid approaches that integrate unsupervised methods with supervised learning have shown promise, as evidenced by their improved performance in feature extraction and classification tasks [9],[8].

Furthermore, Lexicon-based methods, such as the Vader Sentiment Intensity Analyzer (VSIA), provide another dimension to sentiment analysis by employing predefined dictionaries of words and their associated sentiments. These rely on predefined dictionaries that assign sentiment scores to individual words. VADER, for instance, combines lexical knowledge with heuristics to determine sentence-level sentiment, making it interpretable and efficient. These methods have been particularly effective in sentence-level polarity detection, where additional features like capitalized words and emojis enhance the granularity of the analysis [10],[11]. These methods are lightweight and interpretable, making them useful for quick analysis. However, they often require normalization techniques to mitigate biases and ensure consistent representations of feedback sentiments [12],[13].

Lastly, the recent advancements in Deep Learning (DL) have introduced transformer-based models, such as Bidirectional Encoder Representations from Transformers (BERT), which leverage attention mechanisms and pre-trained embeddings to achieve state-of-the-art performance across various NLP tasks. These models, trained on billions of tokens in a self-supervised manner, excel in applications requiring minimal additional labeled data, making them particularly advantageous for educational sentiment analysis [14]. For example, Recursive Neural Tensor Networks (RNTNs) have been utilized to uncover patterns in feedback data, demonstrating their ability to capture the hierarchical structure of sentiments [4].

In the educational domain, NLP has been applied to analyze qualitative feedback and identify trends in student sentiments. Studies integrating supervised learning with numerical rating analysis have shown that combining qualitative and quantitative data yields more comprehensive insights [11], [13], [15]. Furthermore, techniques like K-means clustering have been employed to identify pivotal themes in student reflections, emphasizing the role of learning communities in fostering a sense of belonging [16]. Traditional methods such as rule-based systems remain prevalent in educational contexts, primarily due to their simplicity and interpretability[17].

Understanding the transformative potential of modern NLP techniques in sentiment analysis, particularly in educational settings, this study utilizes a version of BERT to derive deeper insights from student feedback, to ultimately enhance teaching methodologies and student experiences.

3. Methodology

The objective of this research is to design a sentiment analysis methodology to study student comments and their polarity (positive/negative/neutral) and determine if they are in agreement with students’ responses to the quantitative questions. Additionally, the study seeks to explore the relationship between qualitative feedback (sentiment polarity scores) and quantitative review ratings, providing insights into how sentiments in student comments correlate with numerical evaluations. The final goal is to offer a more comprehensive evaluation of course and teaching effectiveness, combining both quantitative and qualitative data. **Figure 1** illustrates the proposed model design for sentiment analysis. This methodology is explained in detail in the following four phases of (1) Data Collection and Pre-Processing, (2) Data Annotation, (3) Data Augmentation, and (4) Model Evaluation. Statistical Analysis including Correlation and Regression Analysis is also studies in phases 5 and 6, to further investigate the correlation between quantitative and qualitative feedback, and predict quantitative scores based on sentiment polarity scores.

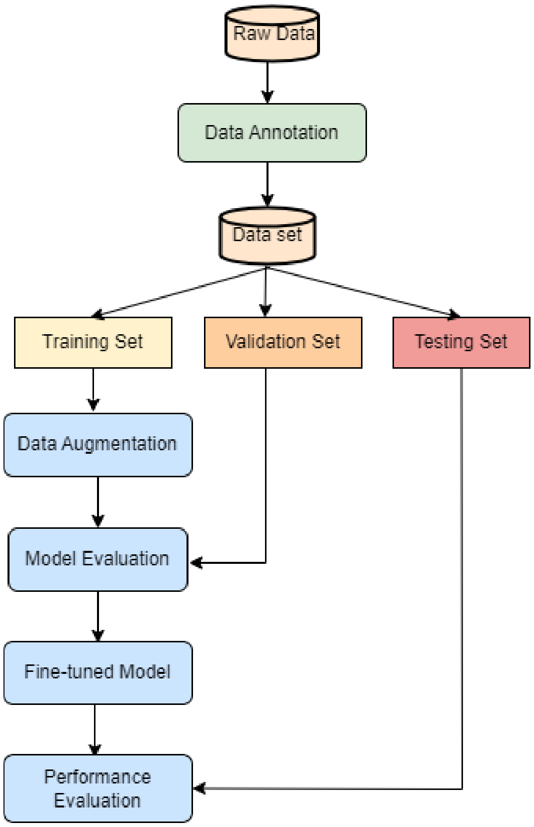

Figure 1 Proposed methodology for sentiment analysis

3.1. Data Collection and Pre-Processing

The dataset for this study consists of student evaluation feedback from the University of New Haven specifically focusing on Civil Engineering courses. The primary dataset comprises qualitative (unstructured text) data from 13 courses, where each student provides feedback to address the following three questions:

1. Which aspects of this course and/or of this instructor's teaching did you like most?
2. What suggestions would you make for improving this course or the instructor's teaching?
3. What advice would you give to another student who is considering taking this course?

The collected data is prepared for the next step, where sentiment polarity scores are computed and then compared with quantitative evaluation scores to identify trends and correlations.

3.2. Data Annotation

Sentiment annotation is performed using two methods of (1) Manual Annotation and (2) ML Annotation with DistilRoBERTa. In the manual annotation process, each comment is classified as positive, neutral, or negative based on its content. DistilRoBERTa is then used to generate sentiment predictions for the same dataset. Discrepancies between manual and machine-generated annotations are resolved through re-annotation to ensure consistency and accuracy.

3.3. Data Augmentation

A common challenge in sentiment analysis is the class imbalance, particularly due to the low number of negative comments. To address this issue, back-translation is used for data augmentation. Back-translation is a data augmentation technique where text is translated into another language and then translated back into the original language. This technique involves translating negative comments into a second language (Japanese in this study) and then back into English, generating additional comments while preserving their original meaning. This method is applied exclusively to negative comments, as they are significantly fewer in number compared to neutral and positive responses. Having too few negative samples can lead to a biased model that performs poorly in detecting critical or adverse feedback. Back translation is used to increase the number of negative comments, helping to balance the dataset. The counts for positive and neutral comments remain unchanged, resulting in a more balanced dataset for model training.

3.4. Model Fine-Tuning

During this phase, a version of the Bidirectional Encoder Representations from Transformers (BERT) model, named DistilRoBERTa is utilized for sentiment classification. BERT captures the contextual meaning of words by accounting for their relationships with surrounding words in a sentence. Unlike traditional approaches that process text sequentially, BERT model comprises 12 transformer encoders to process all input tokens simultaneously, enabling it to handle dependencies effectively. **Figure 2** presents BERT's architecture. It illustrates the multi-layered structure of BERT, starting from the input tokens (Token 1 to Token n) at the bottom, progressing through multiple transformer encoder layers, and finally integrating outputs via mechanisms like Multi-Head Attention, Add & Norm, and Feed Forward layers.

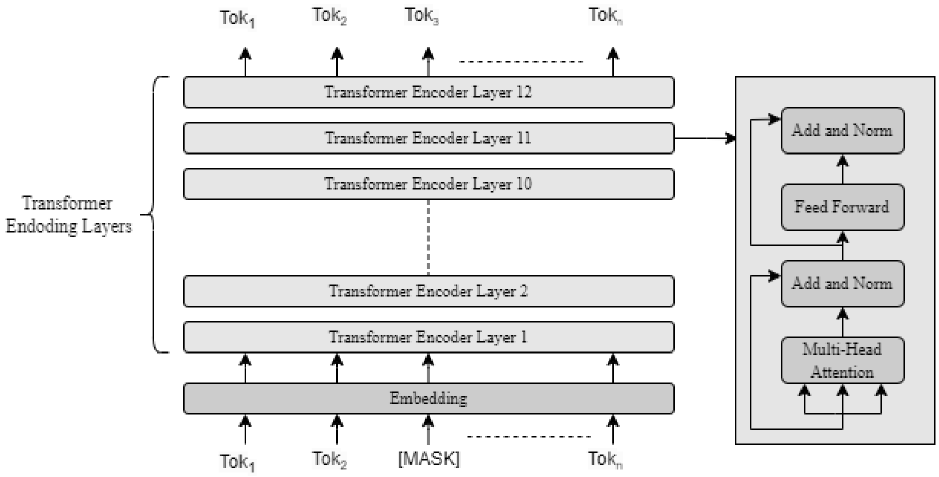
Tok_1
Tok_2
Tok_3
Tok_n
Transformer Encoder Layer 12
Transformer Encoder Layer 11
Transformer Encoder Layer 10
Transformer Endoding Layers
Transformer Encoder Layer 2
Transformer Encoder Layer 1
Embedding
Tok_1
Tok_2
[MASK]
Tok_n
Add and Norm
Feed Forward
Add and Norm
Multi-Head Attention

Figure 2 BERT model architecture

The DistilRoBERTa model is a distilled version of the RoBERTa model, which itself is an optimized version of BERT. It achieves similar performance with fewer computational resources by removing certain training components, such as the next-sentence prediction task, and optimizing training parameters like batch size and learning rate. These modifications make it more efficient while still retaining the core strengths of the original architecture.

Fine-tuning this model allows its pre-trained parameters to adapt to the specific requirements of sentiment classification, enhancing its performance. The steps for fine-tuning BERT for sentiment classification using the Hugging Face library and PyTorch are as follows:

a) Dividing the dataset into training, validation, and testing subsets.

b) Transforming the training data into PyTorch tensors for model compatibility.

c) Defining batch size and creating tensors and iterators for training.

d) Fine-tuning the BERT model with appropriate hyperparameters, monitoring loss, and validating performance.

e) Evaluating the model on the test set to assess its effectiveness.

These steps are applied to fine-tune the DistilRoBERTa for the sentiment analysis task. Test Accuracy is calculated to evaluate the model's performance.

Ultimately, the fine-tuned model is used to calculate sentiment polarity scores by summing weighted values of qualitative comments classified as positive (+1), neutral (0), or negative (-1). These scores are normalized and compared to quantitative review ratings in the next phase. This process ensures that the model can provide a reliable and interpretable analysis of student feedback, enabling a deeper understanding of course effectiveness.

3.5. Correlation Analysis

To evaluate the relationship between sentiment polarity scores and quantitative review scores, Pearson correlation Analysis is conducted. The Pearson correlation coefficient quantifies the linear correlation between quantitative review scores, as dependent variable, and sentiment polarity scores, as independent variable, and is defined as shown in **Equation 1**, where x_i and y_i are the individual data points for sentiment polarity and overall review, respectively, $\underline{x}$ and $\underline{y}$ are their mean values, and n is the number of observations.

$$r = \frac{\sum_{i=1}^{n} (x_i - \underline{x}) (y_i - \underline{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \underline{x})^2 \sum_{i=1}^{n} (y_i - \underline{y})^2}} \qquad \textbf{(Equation 1)}$$

This coefficient ranges from -1 to +1, where a value of 0 signifies no correlation between the variables. Positive values indicate a direct relationship, where an increase in one variable corresponds to an increase in the other. Negative values represent an inverse relationship, where an increase in one variable is associated with a decrease in the other. The closer the coefficient is to +1 or -1, the stronger the relationship, with the sign denoting whether the association is positive or negative.

3.6. Linear Regression

In the last phase of statistical analysis, a Linear Regression model is utilized to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data. The linear regression equation is given by **Equation 2**.

$$Y = \beta_0 + \beta_1 X + \epsilon \qquad \textbf{(Equation 2)}$$

In this equation, the polarity score X is the independent variable, and the overall review score Y is the dependent variable. The regression model aims to find the values of β_0 (intercept) and β_1 (slope) that minimize the sum of squared residuals, ϵ, between the observed review scores and those predicted by the model.

4. Results and Discussion

Building on the methodology developed in this study, this section provided details on the studied datasets and the results of the study. The first dataset comprised feedback collected from 13 Civil Engineering courses and 377 comments. Each comment in the dataset was annotated using Manual Annotation and Machine Learning Annotation with DistilRoBERTa. Discrepancies between manual and machine-generated annotations are resolved through re-annotation to ensure consistency and accuracy. This process also highlighted the inherent subjectivity and complexity involved in sentiment classification. For instance, the comment *"i would say do it but take thorough notes"* was initially labeled as positive in the manual annotation, likely due to the annotator's interpretation of the phrase "I would say do it" as a recommendation reflecting favorable sentiment. In contrast, both the machine-generated label and the final re-annotation classified the comment as neutral, based on the instructional nature of the statement "take thorough notes," which lacks explicit emotional content. This example underscored how subjective interpretations introduced inconsistencies in manual annotation, and how re-annotation supports a more standardized and objective approach to sentiment classification. **Table 1** illustrates a sample of 20 comments, where each comment is labelled by manual annotation, ML annotation, and finally re-annotated. **Figure 3** illustrates the results of this classification, for each category.

Table 1 Sample of 20 Comments from Feedback Dataset

Comment	Manual Annotation	DistilRoBERTa	Re-Annotation
she has a step by step process teaching systems is easy to follow.	Positive	Neutral	Positive
she is an amazing and knowledgeable professor. the many example problems in class really helped me to retain the material and to be able to solve more problems. this proved well for homework assignments and exams.	Positive	Positive	Positive
i really liked the way you explained the material and broke each problem into simple steps. as well as explain how you got each equation that you used. overall, i really enjoyed your class and wish i had room next semester to take advanced concrete with you.	Positive	Positive	Positive
i liked learning about concrete design because i believe that this is something i want to do in the future with my career.	Positive	Positive	Positive
very helpful and enthusiastic about the material being taught.	Positive	Positive	Positive
lectures were fairly clear but tricky	Neutral	Positive	Neutral
the personability of the topics we covered was good. as well as being challenged intellectually	Neutral	Positive	Positive
be ready to learn about leed. if you’re a competent engineering student, it’s not difficult.	Positive	Neutral	Positive
prepare very well for any topic in renewable energies for the final exam includes many calculation problems that are critical.	Neutral	Neutral	Negative
not applicable	Neutral	Neutral	Neutral
i would say do it but take thorough notes.	Positive	Neutral	Neutral

Table 1 Sample of 20 Comments from Feedback Dataset (Continued)

Comment	Manual Annotation	DistilRoBERTa	Re-Annotation
the advice i would give to another student would be to work on group projects early. there was an overlap on projects in the class and a lot was required of the students at that time.	Positive	Neutral	Neutral
know basic math	Neutral	Neutral	Neutral
ask questions if confused	Neutral	Neutral	Neutral
take this course online.	Neutral	Neutral	Neutral
if you have a problem with the homework, contact the professor asap.	Positive	Neutral	Neutral
do the assignments, the extra credit, and be active in the class.	Positive	Neutral	Neutral
come to class and pay attention. that way you can pick up the minimum amount of what is expected to learn.	Positive	Neutral	Neutral
take notes especially on the calculations part of the course.	Positive	Neutral	Neutral
good class and instructor	Positive	Neutral	Positive

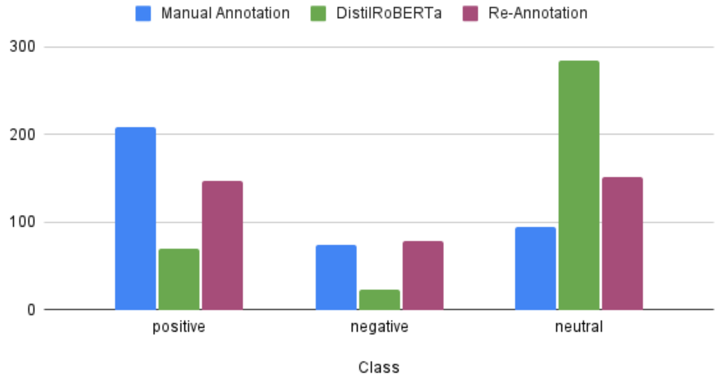

Figure 3 Comparison of sentiment class distribution across three annotation methods

As a result of re-annotation on 377 comments, 147 were labelled as positive, 78 as negative, and 152 as neutral. The negative comments were back translated, to augment more negative data, resulting in 132 negative comments (**Figure 4**).

This dataset was also used for training, validating, and testing the fine-tuned model. It was split into training, validation, and testing sets using a 70:15:15 ratio. This meant that 70% of the data was used for training the model, 15% for validation to fine-tune hyperparameters, and the remaining 15% for testing and evaluation. The split ensured a balanced representation across sentiment classes, facilitating effective training and evaluation.

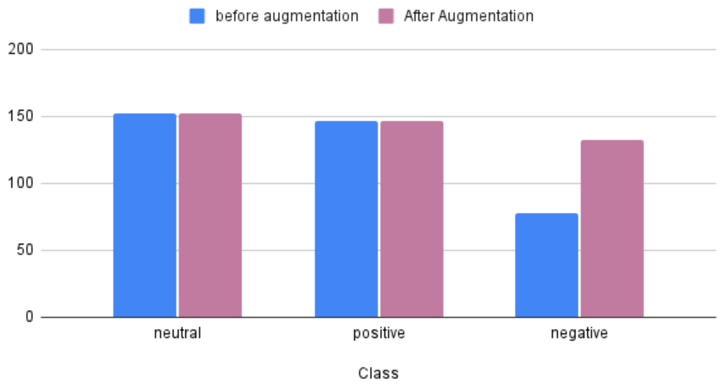

Figure 4 Sentiment class distribution before and after data augmentation

A pre-trained DistilRoBERTa model was fine-tuned for this sentiment classification on the student feedback dataset. To ensure optimal performance, hyperparameter tuning experiments were conducted to determine the most effective combination of learning rate and number of training epochs. These hyperparameters played a critical role in the model's capacity to learn meaningful patterns from the data while avoiding issues of overfitting or underfitting.. Throughout these experiments, the batch size was fixed at 32, while learning rates and epochs were varied to explore the model's performance under different conditions. The best performance was achieved with a learning rate of 5e-6 and 30 epochs, yielding a testing accuracy of 90% (**Table 2**).

Table 2 Performance of the Model with Different Learning Rates and Epoch (Experiment 4 provided the highest accuracy.)

No.	Learning Rate	epoch	Test Accuracy
1	5e-6	40	0.84
2	5e-5	40	0.79
3	5e-5	30	0.88
4	5e-6	30	0.90

The next phase after designing the model was correlation and regression analysis. To ensure that the model was evaluated on previously unseen data, a second dataset comprising feedback from seven additional Civil Engineering courses was selected. This dataset included both quantitative and qualitative components. For each course, students provided an overall review score through the quantitative evaluation section. These “overall review” scores represented aggregated responses to selected survey items, typically measured on a Likert scale (e.g., 1 to 5), reflecting students’ overall satisfaction with the course and the instructor. In parallel, a normalized sentiment polarity score was computed from the qualitative feedback using the sentiment classification model developed in earlier phases. The sentiment polarity score was calculated by summing weighted values of individual comments—classified as positive (+1), neutral (0), or negative (–1)—and normalizing the total for each course. This process enabled a direct comparison between qualitative sentiment and quantitative evaluations. . **Table 3** summarizes the results for dataset 2.

Table 3 Mean Overall Course Review Score and Polarity Score for Dataset 2

No.	Overall Review	Normalized Polarity Score
1	4.6	0.49
2	4.2	0.59
3	4.4	0.40
4	4.5	0.48
5	4.8	0.71
6	4.5	0.61
7	4.4	0.63

The Pearson correlation coefficient between sentiment polarity scores and quantitative review scores was calculated using **Equation 1** and resulted in 0.42, indicating a moderate positive correlation. This means there is a tendency for higher sentiment polarity scores to align with higher quantitative review scores, suggesting that as students' expressed sentiments about a course become more positive, their numerical ratings for the course also tend to be higher. However, while there is a relationship, it's not very strong, implying that other factors might also influence these scores, or that the sentiment scores do not capture all aspects that lead to higher quantitative ratings.

Ultimately, a linear regression model was constructed to predict quantitative scores based on sentiment polarity scores. The regression coefficient, using **Equation 2**, was 0.24, confirming a positive relationship between the two variables. This coefficient means that for each unit increase in sentiment polarity score, there is an expected increase of 0.24 units in the quantitative review score, holding all else constant. This suggests that while there is a positive impact of sentiment polarity on the review scores, the effect is moderate. Other factors not captured by the sentiment polarity scores might also play significant roles in determining the quantitative review scores. This could include aspects like course content, student expectations, or external influences not reflected in the sentiment analysis. The model confirms a relationship but also highlights the complexity of fully predicting student satisfaction and course ratings based on sentiment analysis alone.

To further explore this relationship, a scatterplot with a linear regression line was generated to visualize the distribution of the data. Each data point in the figure represents an individual course from the second dataset. The regression line models the predicted overall review scores based on sentiment polarity. Courses with data points above the line indicate higher-than-expected review scores, suggesting that factors beyond the sentiment expressed in comments may have positively influenced students' numerical evaluations. Conversely, data points below the line represent courses with lower-than-expected scores, which may reflect negative experiences or conditions not fully conveyed through qualitative feedback. The moderate spread of data points around the regression line illustrates the partial explanatory power of sentiment polarity in predicting review scores. **Figure 5** displays the linear regression line fitted to the relationship between normalized sentiment polarity scores (x-axis) and overall review scores (y-axis) for seven courses.

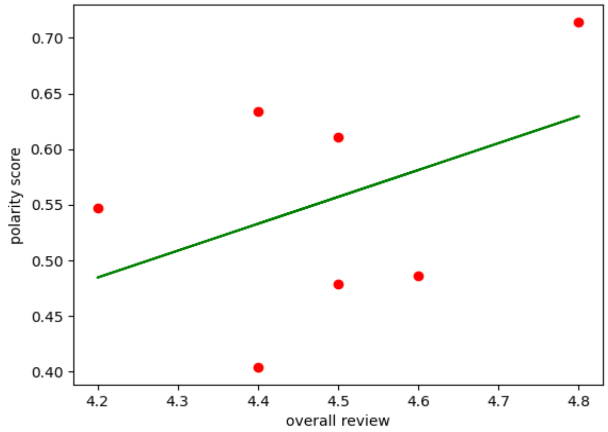

Figure 5 The relationship between overall review score and polarity score, linear regression

To summarize these results, the Pearson correlation coefficient of 0.42 suggested a moderate positive correlation, indicating that students who expressed more positive sentiments generally rated higher quantitative scores. On the other hand, the Linear Regression analysis yielded a coefficient of 0.24, quantitatively supporting the positive relationship between these variables. This indicated that an increase in the sentiment polarity score by one unit is associated with an increase in the quantitative score by 0.24 units. This numerical analysis confirmed the hypothesis that sentiment analysis can be a meaningful tool in understanding and predicting student satisfaction metrics, though the moderate strength of these correlations also suggests that other factors may influence student evaluations.

5. Conclusion

This study proposed a sentiment analysis methodology to interpret qualitative student feedback from engineering courses, integrating it with quantitative review scores to understand the relationship between these two. Fine-tuning a DistilRoBERTa model on 13 course evaluation data for sentiment classification, this study analysed student comments on 7 course evaluations, revealing a Pearson correlation coefficient of 0.42, demonstrating that students' sentiment in text comments generally aligns with their numerical evaluations, albeit other factors also play a role.

The Linear Regression model further substantiated this relationship with a regression coefficient of 0.24, indicating that sentiment analysis can be a predictive tool for assessing course effectiveness, though it should be used in conjunction with other evaluation methods to fully capture student feedback.

5.1 Practical Implications for Institutions and Faculty

While this study primarily focuses on the technical implementation of sentiment analysis in the context of course evaluations, its findings offer several important implications for educational institutions and individual faculty members. Sentiment analysis provides a scalable method for interpreting large volumes of open-ended student feedback, which is often overlooked due to the labor-intensive nature of manual review. By transforming qualitative comments into structured sentiment scores, institutions can begin to systematically analyze patterns in student perceptions over time.

At the institutional level, the aggregation of sentiment polarity scores across multiple courses and departments can serve as an early diagnostic tool for academic leadership. Courses or programs that consistently receive lower sentiment scores may indicate underlying issues that require intervention, such as instructional quality, resource availability, or student support services. Monitoring these scores over time can help administrators identify trends, assess the impact of policy changes, and allocate resources more effectively.

For faculty members, sentiment classification offers a concise overview of student attitudes that may otherwise be buried in lengthy qualitative responses. Instructors can use the sentiment breakdown to identify shifts in student sentiment across different course offerings or academic terms. Although these scores are not a replacement for reading full comments, they provide an initial lens through which faculty can determine whether more detailed feedback analysis is needed. For example, a trend toward increasing neutral or negative sentiment may prompt reflection on course materials, teaching methods, or classroom engagement.

Additionally, sentiment scores can support faculty in compiling teaching portfolios for annual evaluations, tenure, or promotion. By presenting a visual and data-driven summary of feedback trends, instructors can provide evidence of teaching effectiveness over time. This approach complements selected student comments and helps contextualize feedback within broader patterns.

5.2. Challenges

The initial dataset from 13 courses presented several challenges for analysing sentiment and categorizing feedback effectively. Its small size limited the ability to draw broad conclusions, and a lack of diverse sentiment expressions, particularly a shortage of negative comments, made training the analysis model difficult. To address this imbalance, synthetic data augmentation was used, which while helpful, introduces potential risks of bias and errors in the data.

Additionally, classifying sentiments from student feedback proved highly subjective, with some comments being ambiguous and hard to classify. The need to repeatedly check and correct these classifications not only added significant time to the project but also introduced the possibility of human error. These challenges emphasized the necessity for more refined methods to manage such data limitations, ensuring both accurate and reliable analysis results.

5.3. Future Work

Future research should focus on expanding the dataset to include feedback from diverse courses and institutions. Aspect-based sentiment analysis could provide more detailed insights by examining specific course elements, such as teaching style or content quality. Incorporating multimodal feedback, such as audio or video comments, could capture non-verbal cues and enhance the analysis. Real-time feedback systems could enable dynamic interventions, improving teaching practices. Cross-validation with other methods, such as interviews or focus groups, would further enrich the understanding of student satisfaction.

References

[1] V. Dhanalakshmi, D. Bino, and A. M. Saravanan, “Opinion mining from student feedback data using supervised learning algorithms,” in *2016 3rd MEC International Conference on Big Data and Smart City (ICBDSC)*, IEEE, Mar. 2016, pp. 1–5. doi: 10.1109/ICBDSC.2016.7460390.

[2] S. Katragadda, V. Ravi, P. Kumar, and G. J. Lakshmi, “Performance Analysis on Student Feedback using Machine Learning Algorithms,” in *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, IEEE, Mar. 2020, pp. 1161–1163. doi: 10.1109/ICACCS48705.2020.9074334.

[3] M. Bansal, S. Verma, K. Vig, and K. Kakran, “Opinion Mining from Student Feedback Data Using Supervised Learning Algorithms,” 2022, pp. 411–418. doi: 10.1007/978-3-031-12413-6_32.

[4] A. Koufakou, J. Gosselin, and D. Guo, “Using data mining to extract knowledge from student evaluation comments in undergraduate courses,” in *2016 International Joint Conference on Neural Networks (IJCNN)*, IEEE, Jul. 2016, pp. 3138–3142. doi: 10.1109/IJCNN.2016.7727599.

[5] V. S. Sadanand, K. R. R. Guruvyas, P. P. Patil, J. Janardhan Acharya, and S. Gunakimath Suryakanth, “An automated essay evaluation system using natural language processing and sentiment analysi,” *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 12, no. 6, p. 6585, Dec. 2022, doi: 10.11591/ijece.v12i6.pp6585-6593.

[6] 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, doi: 10.1109/ACCESS.2019.2928872.

[7] A. Bralin, J. Morphew, & R. C. M., and N. S. Rebello, “Analysis of student essays in an introductory physics course using natural language processing,” in *Physics Education Research Conference proceedings. 2023.*, Physics Education Research Conference proceedings. 2023., 2023.

[8] N. Kardam, D. Wilson, and S. Makhsous, “A Comparative Analysis of Natural Language Processing Techniques for Analyzing Student Feedback about TA Support,” in *2024*

ASEE Annual Conference & Exposition, 2024 ASEE Annual Conference & Exposition, 2024.

[9] N. Kardam and D. Wilson, “A Hybrid Approach to Natural Language Processing for Analyzing Student Feedback about Faculty Support,” in *2024 ASEE Annual Conference & Exposition Proceedings*, ASEE Conferences. doi: 10.18260/1-2--46447.

[10] M. Wook *et al.*, “Opinion mining technique for developing student feedback analysis system using lexicon-based approach (OMFeedback),” *Educ Inf Technol (Dordr)*, vol. 25, no. 4, pp. 2549–2560, Jul. 2020, doi: 10.1007/s10639-019-10073-7.

[11] T. Shaik, X. Tao, C. Dann, H. Xie, Y. Li, and L. Galligan, “Sentiment analysis and opinion mining on educational data: A survey,” *Natural Language Processing Journal*, vol. 2, p. 100003, Mar. 2023, doi: 10.1016/j.nlp.2022.100003.

[12] M. Misuraca, G. Scepi, and M. Spano, “Using Opinion Mining as an educational analytic: An integrated strategy for the analysis of students’ feedback,” *Studies in Educational Evaluation*, vol. 68, p. 100979, Mar. 2021, doi: 10.1016/j.stueduc.2021.100979.

[13] S. Crossley, J. Ocumpaugh, M. Labrum, F. Bradfield, M. Dascalu, and R. S. Baker, “Modeling Math Identity and Math Success through Sentiment Analysis and Linguistic Features,” in *International Educational Data Mining Society*, International Educational Data Mining Society, 2018.

[14] A. Katz, M. Norris, A. Alsharif, M. Klopfer, D. Knight, and J. Grohs, “Using Natural Language Processing to Facilitate Student Feedback Analysis,” in *2021 ASEE Virtual Annual Conference Content Access Proceedings*, ASEE Conferences. doi: 10.18260/1-2--37994.

[15] F. F. Balahadia, Ma. C. G. Fernando, and I. C. Juanatas, “Teacher’s performance evaluation tool using opinion mining with sentiment analysis,” in *2016 IEEE Region 10 Symposium (TENSYMP)*, IEEE, May 2016, pp. 95–98. doi: 10.1109/TENCONSpring.2016.7519384.

[16] A. Satyanarayana, K. Goodlad, J. Sears, P. Kreniske, M. F. Diaz, and S. Cheng, “Using Natural Language Processing Tools on Individual Stories from First Year Students to Summarize Emotions, Sentiments and Concerns of Transition from High School to College,” in *In 2019 ASEE Annual Conference & Exposition Proceedings*, In 2019 ASEE Annual Conference & Exposition Proceedings, 2019.

[17] A. Rashid, S. Asif, N. A. Butt, and I. Ashraf, “Feature Level Opinion Mining of Educational Student Feedback Data using Sequential Pattern Mining and Association Rule Mining,” *International Journal of Computer Applications 81.10 (2013).*, 2013.