Analyzing Student Feedback to Enhance Teaching Quality Using Explainable AI

Sultanul Arifeen Hamim1, a), Md.Mazid-Ul-Haque1, b), Md.Saef Ullah Miah1, c), Abhijit Bhomik1, d) and Liew Tze Hui2, e)

1*Department of Computer Science, American International University - Bangladesh (AIUB), Dhaka, Bangladesh*

2*Faculty of Information Science and Technology, Multimedia University, Jalan Ayer Keroh Lama,*

*75450, Bukit Beruang, Melaka.*

*e) Corresponding author: thliew@mmu.edu.my*

*a)hamim@aiub.edu*

*b)mazid@aiub.edu*

*c)saef@aiub.edu*

*d)abhijit@aiub.edu*

**Abstract.** Analyzing student feedback is important in evaluating teaching performance and enhancing educational quality. This research examines the use of direct machine learning and deep learning models including, Random Forest, Decision tree, Naive Bayes, LSTM, BiLSTM, and BERT to conduct a sentiment analysis of students’ comments. Using explainable AI techniques such as SHAP allowed the current study to provide interpretability and insight into how input features are leveraged to create model predictions. The results show that more sophisticated models such as BiLSTM and BERT had a great success in capturing the fine-grained context of student feedback and became more contextually aware and accurate than conventional methods. Furthermore, the most salient predictive features--student effort, teacher engagement, and accessibility--identified as being critical to feedback sentiment; as well, while several promising accuracies were achieved across models, issues with class imbalance relating to the minority class also showed opportunities for further work. The contribution from this study is that it reveals connections between predictive modeling and downstream educational applications, by marrying accuracy and interpretability so that educators and stakeholders can use data based on student feedback to make informed decisions for improving pedagogical practices or create a more student-centered learning experience. The implications also demonstrate how AI systems can change the landscape in the field of educational feedback analysis and indicate future work possible in this space.

# **INTRODUCTION**

The quality of instructional delivery is one of the significant determinants of students' academic success and learning outcomes. To improve the quality of education, universities worldwide strive to systematically assess the effectiveness of their teachers [1]. Traditional teaching evaluations normally include both quantitative indicators, such as academic performance and graduation rates, and qualitative student feedback. Whereas quantitative metrics are easily quantifiable, qualitative feedback comprises complex, unstructured data that provides richer insight into students' attitudes. However, qualitative comments are time-intensive, biased, and not scalable to large-scale course or institution levels [2]. Sentiment analysis computerizes the task of text-based comment analysis, and this can help us measure the quality of instruction well by categorizing comments as belonging to a positive, negative, or neutral class [3-5]. The traditional methods like Decision Trees and Naive Bayes being simple, they cannot identify subtle expressions like sarcasm and inference in each situation [6]. Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM) are deep learning models that carry out a better formation of text sequence patterns and language dependencies compared to the traditional ones [1]. Advanced sentiment analysis models surpass traditional machine learning methods in accuracy and scalability which makes them better suited for extracting student feedback quality and supporting teaching method improvements. Neural networks, especially LSTM and transformer architectures like BERT, are superior in analyzing student feedback in that they detect salient contextual and sequential text features required for accurate sentiment interpretation. Although effective in sequential text modeling, LSTMs' computational overhead and limited interpretability motivated the integration of BERT's self-attention mechanism, which processes local and global linguistic correlations more efficiently. To provide transparency, BERT has been coupled with explainable artificial intelligence (XAI) methods, enabling teachers to understand and trust model decisions [8]–[11].

# This research closes the divide between advancements in artificial intelligence and actionable changes in education, by systematically highlighting both traditional approaches (Naive Bayes, Decision Trees) versus new approaches via deep learning (LSTM, BiLSTM, BERT). In doing so, it determines the best methods of determining types of feedback with an eye on actionable recommendations for improving markers of teaching quality, such as availability, engagement, and clarity. By accommodating interpretability in conjunction to empirical accuracy, this study offers a framework allowing schools to objectively assess feedback assessments with both demonstrable accuracy, and transparency, which is generally lacking in AI evaluations. These results give academic institutions the opportunity to develop automated feedback analysis, prioritize data-informed improvement of teaching, and sustain continuous improvement cycles grounded in student perspectives. In sum, this study illustrates that customized AI solutions adhering to explainable principles have the power to change educational evaluation and bolster student-minded pedagogical innovation.

# **RELATED WORKS**

In the past few years, the application of machine learning processes to educational environments has been a popular topic, especially in the context of evaluating students' feedback to improve teaching processes. Sentiment analysis, which is a subfield of natural language processing, has been used to analyze students' overall thoughts and learnings, and provide data that teachers can effectively use. The author's [12] comprehensive review provides an overview of different machine learning and deep learning methods that have been used to gather and analyze students' feedback, for a variety of platforms. The study emphasized the value of creative teaching strategies and authors conclude that sentiment analysis is necessary for grasping students' perspectives, to engage in meaningful practices to improve educational experience. The author of [13] also employed a variety of natural language processing and machine learning platforms to help university administrators interpret students' feedback. They noted a connection between course performance and sentiment classification, in their attempts to extract useful information from the feedback [13]. Current research shows incongruent performance of machine learning approaches in education sentiment analysis. Qaiser et al. reported that deep learning approaches attained a better accuracy rate of 96.41% with a sample size of 4,289, outperforming Naive Bayes (87.18%), SVM (82.05%), and Decision Trees (68.21%) but limited by sample size [14]. Likewise, Mabunda et al.'s research with a sample size of 185 for feedback samples showed bias towards neutral/negative classes for K-NN, neural networks, Naive Bayes, SVM and Random Forest, suggesting a requirement for larger sample size and the utilization of neural approaches to counteract bias [8]. The two studies stress that despite promising current content, scaling the data and tuning deeper architectures for learning is required to enhance accuracy and minimize bias in education sentiment analysis. Agreement between the studies highlights the requirement for larger sample size and sophisticated model techniques in the domain to yield accurate, unbiased results. Recent research illustrates the potential of explainable AI in education through large linguistic models for reliable feedback [7] as well as prescriptive analytics for at-risk students [15]. And, despite the mentioned potential, data paucity, class imbalance, and bias pose challenges. Transparent AI approaches are essential to enable pedagogically sound, trustful sentiment analysis. Future advances demand larger sample size, sophisticated neural networks, and model interpretability to optimize educational outcomes [16].

# **METHODOLOGY**

This section details the end-to-end methodology for student feedback analysis, covering dataset preparation, feature engineering, model training, and evaluation. Figure 1 provides a diagram of the overall workflow and the main steps of the process, which runs from data pre-processing to feature extraction to modeling and finally to using explainable AI aspects. The systematic methodology allows for careful, repeatable outcomes, while remaining open and understandable through interpretability techniques.

**Dataset Preparation**

The dataset used in this study consists of student feedback collected through institutional surveys, comprising X instances. Each instance represents a comment labeled as either positive (satisfactory teaching) or negative (indicating areas for improvement). An initial analysis revealed a class imbalance, with Y% positive feedback and Z% negative feedback. This imbalance was addressed during preprocessing to prevent bias in model training.

**Data Splitting**

The feedback dataset used in this study contained a total of 918 instances of student feedback labelled as either positive or negative in terms of their perception of teaching quality. The data was split into three sets, with the intention of modifying the data for analysis to ensure the models were trained to provide robust and generalizable predictions. The original data was arranged into the following training, validation, and test sets, as outlined in Table 1. In this case the training data is 80% of the dataset, while both the validation and test are the remaining 10% each. This division was chosen so there was enough data during training, and retained instances which could be used for evaluation and hyper-parameter tuning. The previous described stratification split kept the class distribution of the positive mentioned feedback and negative feedback consistent in each of the sets ensuring that the class imbalance was controlled, and no bias developed in the class's imbalance.

A diagram of a computer process

Description automatically generated with medium confidence

**FIGURE 1.** Methodology diagram

**TABLE 1.** Dataset splits for machine learning and deep learning model

|  |  |  |
| --- | --- | --- |
| **Dataset** | **Percentage** | **Instances** |
| Training Set | 80 | 734 |
| Validation Set | 10 | 92 |
| Test Set | 10 | 92 |

**Feature Engineering**

The study used separate extraction methods for different model architectures. For traditional machine learning methods (Naive Bayes, Decision Trees) we used TF-IDF vectorization to convert text into sparse numerical form. This extraction method accentuated discriminative terms and reduced the weight of common words. Using this extraction method provided us with feature matrices that were computationally efficient and did an adequate job of including the discriminative term importance when presented with classification tasks. For deep learning methods we used several methods of extracting features. The LSTM and BiLSTM methods completed the text processing using pre-trained GloVe embeddings. GloVe embeddings transformed words to dense vectors capturing semantic relationships (for example in GloVe the words "teacher" and "educator" are close in the embedding space). BiLSTM's bidirectional aspect enhanced this because LSTM's only provided a context of words looking backwards in the sequence. In feedback texts, this means the BiLSTM provided contextual dependency of both preceding and succeeding words to complete the context of the feedback. BERT methods required the most cognitive processing. BERT used WordPiece tokenization to break words into bytes for modeling in its feature extraction (for example in BERT, the word "educational" is separated into "educate" + "##ional"). For feature extraction in BERT, we added special classification tokens ([CLS]) at the start and separator tokens ([SEP]) are added at the turn shifts, as well as attention masks to inform BERT what parts of the texts were valuable content versus padding. BERT then, based on the transformer architecture, produced embeddings from self-attention mechanism that produces contextualized embeddings of the words with respect to both the local word relationships and the global context of the document. This comprehensive approach ensured optimal text representation across all model types while preserving each architecture's unique strengths in handling sequential, semantic, and contextual information.

# **MODEL ARCHITECTURES**

This section describes the architecture and configurations of the models used for classifying student feedback. The models span from traditional machine learning approaches to advanced deep learning frameworks, including transformer- based architectures. Each model was selected to explore its unique strengths and evaluate its effectiveness in analyzing textual feedback.

**Traditional Models**

This study employs two fundamental models for sentiment analysis: Naive Bayes and Decision Trees. The Naive Bayes classifier works on TF-IDF features and uses the probabilistic reasoning of its models by taking a model from reasoning which assumes independence of features which can make the model quick to apply although naive in its application. Thus, Naive Bayes is a substantial foundational model. On the other hand, Decision Trees are used to form hierarchical decision paths of Gini impurity/information gain to split the TF-IDF transformed corpus text into, providing the ability to capture relationships on features that Naive Bayes cannot. While both models applied in this research use the same TF-IDF representations, decision processing used by Naive Bayes is inherently going to be different than Decision Trees: Naive Bayes providing fast probability-like classifications and Decision Tree providing interpretable rule-like structures. Thus we would be able to interpret the two foundational models as offering complementary styles of traditional sentiment classification.

**Deep Learning Models**

Deep learning models have been found to make a very significant difference in text classification issues because of their capacity to model contextual information and subtle dependencies in sequential data. This research employed three varied deep models, specifically LSTM, BiLSTM, and BERT, to classify student comments as negative or positive. Each of the models was optimized with precise architectural components to function optimally in the task. The current research made use of both LSTM and its bidirectional variant (BiLSTM) to model the sentiment in student comments. The LSTM model solved RNNs vanishing gradient issue with special memory cells and gate functions (input gate, forget gate, output gate), 128 hidden units for sequential pattern capture, and 0.2 dropout for preventing overfitting. Despite the effectiveness in dealing with ordered sequences in words with GloVe embeddings and sigmoid classification, the unidirectional structure reduced overall contextual comprehension. The structure in the BiLSTM augmented this function by processing both directions with dual LSTM with 128 units each. The strategy retained similar core components - GloVe embeddings, 0.2 dropout, and sigmoid output - in processing words as both past and future context. The bidirectional processing was especially beneficial for sentiment comprehension where words in the context have a crucial role in deciding comprehension, augmenting the comprehension of subtle patterns in comments over regular LSTM. Both structures were effective in managing long-range sequence in text, but BiLSTM's bidirectional context comprehension offered higher quality in modeling subtle sentiment dependencies, demonstrating the strength in architectural development in augmenting NLP tasks without necessarily augmenting computational requirement.

**BERT (Bidirectional Encoder Representations from Transformers)**

TFBert (bert-base-uncased) was used for sequence classification in this study, which was fine-tuned to student feedback data. BERT has three basic steps in its architecture: 1) it uses sub-word tokenization with [CLS]/[SEP] markers; 2) it has a 12 transformer architecture with multi-head self-attention for contextualization; and 3) it uses a sigmoid classification head for the sentiment probabilities. The model was fine-tuned with the AdamW optimizer (lr=2e-5) using a batch size of 16 with 3 epochs, including early stopping to prevent overfitting. BERT uses bidirectional attention which is particularly useful in capturing subtle sentiment in feedback since it examines both the local word relationship, and the overall context of the text. This illustrates that transformer architectures represent a recent advance in NLP state-of-the-art for educational sentiment analysis tasks.

# **RESULTS**

Six machine learning models were examined in this study to classify students' feedback sentiment. Random Forest, Decision Tree, and BiLSTM were championed as the best performing models with 99% accuracy. Random Forest was able to use ensemble learning to be robust in its predictions, Decision Tree was attuned to using a more balanced dataset (where the class sizes were unbalanced), and BiLSTM was the superior model in being able to consider context from both sides. BERT performed well, reaching 97% accuracy, and demonstrating the ability to contextualize nuanced language patterns. LSTM said 95% out of 100%, but the reason it didn't get a higher score was due to only using one direction of the data to predict context. Naive Bayes was the baseline and reached an accuracy of 88% but was still useful for less complex processing needs as it will have limitations with complex text segments. Overall, this study indicates that the traditional models still perform quite well, but they lack context-sensitive understanding compared to advanced and sophisticated model architecture. Naive Bayes will still serve a purpose, especially when considering the limited resource aspect. The best model to use will greatly depend on the task information requirements, how complex the data is, and the resources at hand. The metrics for all models are found in Table 2.

**TABLE 2.** Summary of the performance metrics for all models

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy (%)** | **Precision (%)** | **Recall (%)** | **F1-Score (%)** |
| Random Forest | 99 | 98 | 97 | 97 |
| Decision Tree | 99 | 96 | 96 | 96 |
| Naive Bayes | 88 | 87 | 86 | 86 |
| LSTM | 95 | 94 | 93 | 93 |
| BiLSTM | 99 | 98 | 97 | 97 |
| BERT | 97 | 96 | 95 | 96 |

**Explainable Artificial Intelligence (XAI)**

In order to assist with model interpretability and provide useful information about the categorization of student feedback, this study implemented Shapley Additive Explanations (SHAP). SHAP values allow model predictions to be decomposed into each feature's individual contributions and provide a transparent way to visualize the effects of input features on the outcome. The use of SHAP is consistent with a goal of increasing accessibly/explanations for the non-technical stakeholder such as an educator or administrator for machine learning models.

* SHAP Summary Plot: The SHAP plot in Figure 2 illustrates that positive values determine the features that drive authors' feedback predictions. Student Effort was the strongest predictor - more self-perceived effort resulted in more positive evaluations. Other features related to the teaching dimensions, such as Effective Lecturer and Available/Helpful were equally impactful. This demonstrates the importance of participatory engagement on the part of the instructor. Presentation quality led to clear moderating but consistently impactful acknowledgement. The Knowledge Start/End showed that the eventual ordering of course delivery could affect how the learning experience is perceived by students. The color gradient on the SHAP plot corresponds to feature values, which were generally red (high values) associated with the positive SHAP values presented in Figure 2. Overall, this analysis confirms quantitatively that both student participatory engagement and teaching quality dimensions lead to positive feedback.
* SHAP Interaction Plot: The interaction plot identifies features of Teacher Effort and Teacher Availability as positively correlated feedback (refer Figure 3). This is, when efforts and availability are high (or low) possibilities to positively impact positive evaluations occur. When an instructor's high efforts is coupled with low availability even when accessible negatively impacts evaluation. Feature importance confirms (Student Effort) as the most important feature with an importance value of +0.85 (refer Figure 4). Only suggestions that could have more effect than not are Effective (0.37) and Available/Helpful (0.28). Preferred Target (Knowledge Start, +0.22) and Teacher Effort (0.20) are structured teaching characteristics that are consistently positively impactful. Presenting Clearly (+0.14) and Knowledge End (+0.15), demonstrate positive impact. Engagement section factors that presented as positive at times are Stimulates Interest and Effective Time Use (+0.09 for each). The numeric results indicate the multidimensionality of teaching effectiveness.

A graph with red and blue dots

AI-generated content may be incorrect.

**FIGURE. 2.** SHAP value impact on model output

|  |  |
| --- | --- |
|  |  |
| **FIGURE 3.** Teacher effort | **FIGURE 4.** SHAP feature importance plot |

SHAP values reveal key drivers of positive feedback: Student Effort, Effective Lecturers, and teacher accessibility. Clear communication (Clear Presentations, Stimulates Interest) also significantly impacts evaluations. These explainable AI insights enable educators to target specific improvements, bridging predictive analytics with practical teaching enhancements while ensuring transparent, trustworthy AI applications for feedback analysis.

# **DISCUSSION**

The results from this study provide a thorough analysis of machine learning models for student feedback with good and bad attributes taken into consideration. Advanced models such as the Bi-LSTM and BERT were found statistically superior in reasonably capturing the subtle context of the text, compared to traditional methods like Decision Trees and Random Forest-based classification models that build on structured learning and struggled to capture complex language patterns. This comparison pinpoints how crucial it is to select a model when intending to align it with the type of feedback data and the intended educational purpose. The societal ramifications of this research are vast, especially in higher education. By taking a more systematic approach to stakeholder feedback, universities will be able to shift away from subjective assessments based on perceptions toward fact-finding-based decision making, thus facilitating improved teaching practices for the betterment of students. The study identifies factors like crypto-currency student engagement experiences, clarity of the instruction, and accessibility of learning material that provided positive feedback. Furthermore, institutions can quickly assess both strengths and weaknesses using sentiment analysis and develop a culture that embraces pedagogical improvement! This approach also better ensures or protects student voice through deeper level analysis of the feedback they provided, with the end goal of providing ever-better learning experiences. The study incorporated analytical models of student feedback while using SHAP explanations to develop trust in AI determined "insights" by faculty. The issue of current class imbalance continues to be addressed through further data correlation; however, this methodology is encouraging for potentials beyond education like customer services and health care. By not relying solely in AI analysis in lieu of instructional planning and decisions, the study moves towards developing interpretable and actionable models for institutional feedback that results in improved practices from data-driven methods.

# **CONCLUSION**

This study successfully applied machine learning and explainable AI to analyze student feedback, with BiLSTM and BERT excelling in capturing contextual nuances. While achieving high accuracy and interpretability through SHAP, challenges like category imbalance persist, suggesting needs for improved preprocessing and model refinement. The work demonstrates significant progress in automating feedback analysis while highlighting opportunities for enhancement through additional data and architectural improvements. These findings advance educational technology by merging AI innovation with practical teaching applications, establishing a foundation for developing more sophisticated, student-centered evaluation systems. Future research should address current limitations to further bridge the gap between predictive analytics and educational improvement.

# **REFERENCES**

1. P. Ren, L. Yang, and F. Luo, “Automatic scoring of student feedback for teaching evaluation based on aspect-level sentiment analysis,” *Education and Information Technologies* **28**(1), 797–814 (2023).
2. G.S. Chauhan, P. Agrawal, and Y.K. Meena, “Aspect-based sentiment analysis of students’ feedback to improve teaching–learning process,” in *Information and Communication Technology for Intelligent Systems: Proceedings of ICTIS 2018, Volume 2*, pp. 259–266 (Springer, 2019).
3. S. Seemab, M.S. Zaigham, A. Bhatti, H. Khan, and U.A. Mughal, “Sentiment analysis of student feedback in higher education using natural language processing (NLP),” *Contemporary Journal of Social Science Review* **2**(4), 1741–1750 (2024).
4. P. Shaik, T.X.L.Y.D.C.M.J.R., and L. Galligan, “A review of the trends and challenges in adopting natural language processing methods for education feedback analysis,” *IEEE Access* **10**, 56720–56739 (2022).
5. Z.H.S. Nasim and Q. Rajput, “Sentiment analysis of student feedback using machine learning and lexicon based approaches,” in *2017 International Conference on Research and Innovation in Information Systems (ICRIIS)*, pp. 1–6 (IEEE, 2017).
6. N. Nikolić, O. Grljević, and A. Kovačević, “Aspect-based sentiment analysis of reviews in the domain of higher education,” *The Electronic Library* **38**(1), 44–64 (2020).
7. H. Li and A. Botelho, “Developing explainable AI systems to support feedback for students,” in *Proceedings of the 17th International Conference on Educational Data Mining*, pp. 998–1002 (2024).
8. J.G.K.R. Mabunda and A. Jadhav, “Sentiment analysis of student textual feedback to improve teaching,” in *Interdisciplinary Research in Technology and Management*, pp. 643–651 (CRC Press, 2021).
9. R.A. Nawaz, S.Q.S.M.K.G.A.N.V., and S. Hassan, “Leveraging AI and machine learning for national student survey: actionable insights from textual feedback to enhance quality of teaching and learning in UK’s higher education,” *Applied Sciences* **12**(1), 514 (2022).
10. S. Sadriu, K.P. Nuci, A.S. Imran, I. Uddin, and M. Sajjad, “An automated approach for analysing students’ feedback using sentiment analysis techniques,” in *Mediterranean Conference on Pattern Recognition and Artificial Intelligence*, pp. 228–239 (Springer, 2021).
11. K. Diwali, K. Saeedi, K. Dashtipour, M. Gogate, E. Cambria, and A. Hussain, “Sentiment analysis meets explainable artificial intelligence: A survey on explainable sentiment analysis,” *IEEE Transactions on Affective Computing* (2023).
12. R. Bhagat and S. Dhande-Dandge, *A Literature Review on Sentiment Analysis Using Machine Learning in Education Domain* (Springer Nature Singapore, 2023).
13. M. Lazrig and S.L. Humpherys, “Using machine learning sentiment analysis to evaluate learning impact,” *Information Systems Education Journal* **20**(1), 13–21 (2022).
14. S.R.M. Qaiser, N. Yusoff, and H.K.A. Alwi, “A comparison of machine learning techniques for sentiment analysis,” *Turkish Journal of Computer and Mathematics Education* (2021).
15. T. Susnjak, “A prescriptive learning analytics framework: Beyond predictive modelling and onto explainable AI with prescriptive analytics and ChatGPT,” *arXiv preprint* arXiv:2208.14582 (2022).
16. B.P. Saue, O.N. Chukuigwe, and I.S. Bassey, “Development and Validation of Autotronic Training Module for Automobile Technology Students in Polytechnics in Southern Nigeria,” Journal of Informatics and Web Engineering **3**(2), 188–198 (2024).