Published in AJSE, Vol:22, Issue: 3
Received on 22nd September 2023
Revised on 3rd December 2023
Accepted on 3rd December 2023

AIUB JOURNAL OF SCIENCE AND ENGINEERING
ISSN: 1608 – 3679 (print) 2520 – 4890 (Online)

Aspect-based Sentiment Analysis Model for Evaluating Teachers' Performance from Students' Feedback

Abhijit Bhowmik, Noorhozaimi Mohd. Nur, M. Saef Ullah Miah, Debajyoti Karmekar

***Abstract*—Evaluating teachers' performance is a fundamental pillar of educational enhancement, guiding the evolution of pedagogical practices and fostering enriched learning environments. This study pioneers an innovative approach by harnessing sentiment analysis within an aspect-based framework to decipher the intricate emotional nuances embedded within students' feedback. By categorizing sentiments as positive, negative, and neutral, we delve into the diverse perceptions of teaching aspects, offering a multifaceted portrait of educators' contributions. Through meticulous data collection, preprocessing, and a deep learning sentiment analysis model, we dissected student comments into distinct teaching aspects. The subsequent sentiment analysis unearthed positive, negative, and neutral sentiments. Positive sentiments highlighted strengths and effective communication, while negative sentiments illuminated areas for growth. Neutral sentiments provided contextual equilibrium, forming a holistic tapestry of teachers' performance. The proposed model achieved 86% F1 score for classifying sentiments into three classes.**

***Index Terms*—Teachers' performance evaluation, BiLSTM, Deep Learning, GRU, CNN, Sentiment Analysis**

I. Introduction

IN today's rapidly evolving educational landscape, the pursuit of effective teaching methodologies and pedagogical strategies stands as a paramount goal [1]. As educational institutions strive to provide high-quality learning experiences, assessing teachers' performance is critical [2], [3]. Student feedback, often encapsulating diverse perspectives on teaching methods, classroom dynamics, and instructional efficacy, is invaluable for evaluating educators' contributions and shaping future teaching practices [4].

Traditionally, teacher evaluations have relied on quantitative metrics and standardized assessments, offering a quantitative glimpse into instructional effectiveness [5]. However, the richness of students' opinions, insights, and emotions – embedded within their comments and feedback – remains largely untapped [6], [7]. With the advent of natural language processing (NLP) and sentiment analysis techniques, a transformative opportunity emerges to harness the power of textual data and gain nuanced insights from students' expressions [8], [9], [10].

This article presents an innovative approach to teacher performance evaluation, leveraging aspect-based sentiment analysis to delve deep into the multifaceted dimensions of students' feedback. By dissecting student comments into distinct aspects encompassing various facets of teaching, such as communication skills, subject knowledge, and engagement, our methodology offers a comprehensive understanding of the strengths and areas for improvement in educators' practices.

Sentiment analysis, a subfield of NLP, is key to extracting sentiments, emotions, and attitudes expressed within textual content [11]. In the context of teacher evaluations, sentiment analysis enables us to discern the overall tone of student comments and the sentiments associated with specific aspects of teaching. By incorporating sentiment analysis into the evaluation process, we aim to bridge the gap between quantitative assessment and qualitative perception, creating a holistic framework that empowers educators with actionable insights derived from the sentiments conveyed by their students.

In the subsequent sections of this article, we delve into the intricate details of our approach. We describe the methodology employed to collect and preprocess student feedback, the techniques used for aspect identification, and the sentiment analysis model designed to classify sentiments within each aspect. Through a rigorous analysis of real-world student comments, we illustrate the applicability and efficacy of our approach in shedding light on the multifaceted nature of teacher performance.

By combining the power of sentiment analysis with a granular examination of teaching aspects, our research contributes to advancing teacher evaluation methods. Moreover, this study underscores the potential for sentiment analysis and NLP to revolutionize educational assessments, enhancing the feedback loop between students and educators and fostering a culture of continuous improvement in teaching practices. As we explore sentiment-laden insights from student comments, we pave the way for a more informed, personalized, and effective approach to evaluating and enhancing teachers' performance.

Evaluating teachers' performance through student feedback has long been recognized as a valuable component of educational quality assurance and professional development. Traditionally, teacher evaluations have relied on quantitative

Abhijit Bhowmik is with the Faculty of Computing, College of Computing and Applied Sciences, Universiti Malaysia Pahang Al-Sultan Abdullah (UMPSA), 26600, Pekan, Malaysia, And Faculty of Science and Technology, Department of Computer Science, American International University-Bangladesh (AIUB), 1229, Dhaka, Bangladesh (e-mail: ovi775@gmail.com).
Noorhuzaimi Mohd Noor is with the Faculty of Computing, College of Computing and Applied Sciences, Universiti Malaysia Pahang Al-Sultan Abdullah (UMPSA), 26600, Pekan, Malaysia (e-mail: nhuzaimi@umpsa.edu.my).
M. Saef Ullah Miah is with the Faculty of Science and Technology, Department of Computer Science, American International University-Bangladesh (AIUB), 1229, Dhaka, Bangladesh (e-mail: md.saefullah@gmail.com).
Debajyoti Karmaker is with the Faculty of Science and Technology, Department of Computer Science, American International University-Bangladesh (AIUB), Research Associate, The University of New South Wales (UNSW). (e-mail: d.karmaker@aiub.edu).

instruments, such as Likert-scale surveys and standardized test scores, to gauge instructional effectiveness [12]. While these methods provide structured insights, they often overlook the nuanced perspectives and sentiments that students convey through their qualitative comments [13].

Integrating sentiment analysis techniques into educational assessment represents a significant stride toward unlocking the latent potential of textual data [14], [15]. Sentiment analysis, a subfield of natural language processing (NLP), empowers researchers and educators to automatically detect and classify emotions, attitudes, and opinions embedded within the text. Applying sentiment analysis in the educational context offers a novel lens through which to interpret student feedback, enabling a more holistic understanding of teaching dynamics.

A growing body of research has explored the intersection of sentiment analysis and teacher evaluations, focusing on enhancing the precision and depth of assessment. Okoye et al. [16] conducted a study where sentiment analysis was employed to categorize student comments according to emotional valence, revealing insights into the affective impact of teacher behaviors. Similarly, Ezzameli et al. [17] leveraged sentiment analysis to quantify the emotional tone of student feedback, providing educators with a nuanced understanding of the emotional states elicited by their teaching methods.

To move beyond the one-size-fits-all sentiment analysis approach, recent studies have embraced the concept of aspect-based sentiment analysis. This approach involves segmenting textual content into distinct aspects or categories, allowing for sentiment classification specific to each aspect. Ishaq et al. [18] demonstrated the efficacy of aspect-based sentiment analysis in evaluating online course instructors, emphasizing the importance of considering diverse aspects such as content delivery, responsiveness, and course organization.

Moreover, within the realm of teacher evaluations, aspect-based sentiment analysis has the potential to unlock intricate insights into the multifaceted nature of teaching. Anwar et al. [19] showcased how aspect-based sentiment analysis could provide educators with a detailed breakdown of student sentiments across teaching dimensions, enabling targeted improvement strategies. Similarly, Melba and Suguna [20] employed aspect-based sentiment analysis to categorize student feedback into teaching aspects, facilitating a more comprehensive evaluation of instructional performance.

While sentiment and aspect-based sentiment analyses offer promising avenues for teacher evaluation, challenges persist. Ensuring the accuracy of sentiment classification, addressing linguistic nuances, and accounting for potential bias in student comments remain areas of ongoing research. Additionally, integrating sentiment analysis into practical teaching evaluation frameworks necessitates considering scalability and interpretability.

The literature reflects a growing consensus on the potential of sentiment and aspect-based sentiment analysis to enrich teacher evaluations by tapping into the qualitative insights encapsulated within student feedback. Researchers and educators are poised to unlock a deeper understanding of teaching dynamics by combining NLP techniques with pedagogical assessment, shaping a more holistic and data-driven approach to enhancing teachers' performance. The contributions of this study are as follows,

1) Innovative Aspect-Based Sentiment Analysis Model: This study introduces a novel sentiment analysis model based on LSTM, focusing on evaluating teachers' performance through diverse aspects.
2) Customized Teacher Evaluation: The model utilizes deep learning techniques to analyze student sentiments, offering tailored insights into teachers' strengths and areas for improvement.
3) Enhanced Educational Practices: By providing a student-centric perspective and promoting data-driven decision-making, this research significantly improves educational practices, enabling institutions to implement targeted teacher training programs based on specific sentiment analysis outcomes.

II. Methodology

This section outlines the methodology adopted to conduct aspect-based sentiment analysis to evaluate teachers' performance through student feedback. The process involves data collection, preprocessing, dataset description, and developing a deep learning model for sentiment analysis. The architecture of the deep learning model encompasses the input layer, embedding layer, dropout layer, bidirectional LSTM (BiLSTM) layer, dense layer, and the subsequent evaluation metrics. The experimental setup is also detailed to provide context for the model's implementation and performance assessment.

A. Data Collection

The dataset used in this study consists of a diverse range of student comments collected from educational institutions. Comments from various courses, disciplines, and academic levels were sourced to ensure inclusivity and representativeness. Comments encompassed students' perceptions of teaching aspects, allowing for an aspect-based sentiment analysis approach. The dataset is available at https://data.mendeley.com/datasets/b2yhc95rnx/1, and the details of the dataset can be found in another study by the authors [21].

B. Data Preprocessing

The collected text data underwent several preprocessing steps to ensure the effectiveness of the sentiment analysis model. This included text cleaning to remove irrelevant characters, symbols, and formatting artefacts. Tokenization was employed to break down comments into individual words or subword units. Stop-word removal and lemmatization were performed to enhance the model's ability to extract meaningful features from the text.

C. Sentiment Analysis Deep Learning Model Development

The sentiment analysis model uses a deep learning architecture to capture sentiment patterns in textual data. The architecture consists of the following components:

Input Layer: The input layer accepts tokenized and padded sequences of words as input, where each word is represented as a numerical index.

Embedding Layer: An Embedding layer is the initial layer in the neural network architecture. Its primary purpose is to convert tokenized words from the input text into dense numerical vectors. Tokenization is the process of breaking down text into individual words or tokens. In this case, it is assumed that there is a predefined vocabulary of 10,000 words. The Embedding layer has an input dimension of 10,000, corresponding to the vocabulary size, and an output dimension of 32. This means each word in the vocabulary is represented as a 32-dimensional dense vector. These dense vectors capture semantic relationships between words, enabling the model to understand the contextual meaning of words in the text.

Bidirectional LSTM Layer: Following the Embedding layer is the Bidirectional Long Short-Term Memory (LSTM) layer. LSTM is a recurrent neural network (RNN) for sequential data processing. LSTM units are essential for handling text sequences because they can effectively capture and remember long-range dependencies and sequential patterns. The term ”Bidirectional” indicates that this layer processes input sequences in both forward and backward directions. This bidirectional aspect allows the model to consider the preceding words and subsequent words when making predictions. The LSTM layer plays a crucial role in understanding the temporal dependencies and nuances in natural language, which are critical for accurate sentiment analysis.

Dense Layer with ReLU Activation: Following the Bidirectional LSTM layer, there is a Dense layer with 64 units and a Rectified Linear Unit (ReLU) activation function. Dense layers are fully connected layers that introduce non-linearity to the model. They enable the model to learn complex relationships between features in the data. ReLU activation, which stands for Rectified Linear Unit, is a common activation function used in neural networks. It introduces non-linearity by converting all negative values to zero and leaving positive values unchanged. This layer enhances the model’s ability to capture intricate patterns in the data, further improving its predictive power.

Dropout Layer: The Dropout layer is a regularization technique to prevent overfitting. Overfitting occurs when a model learns the training data too well and performs poorly on unseen data. The Dropout layer randomly deactivates a fraction of neurons during training, effectively introducing noise and reducing the model’s reliance on any specific neuron or feature. By applying dropout, the model becomes more robust and less likely to memorize the training data, which enhances its generalization ability to make accurate predictions on new, unseen text samples.

Final Dense Layer with Softmax Activation: The last layer in the model architecture is the output layer, referred to as the ”Final Dense” layer. This layer consists of three units, each corresponding to one of the sentiment categories: ’negative,’ ’neutral,’ and ’positive.’ In other words, it has three output neurons, one for each sentiment class. The activation function used in this layer is softmax. Softmax transforms the raw model outputs into probability scores, each representing the likelihood of the input text belonging to a specific sentiment category. After passing through the softmax layer, the model’s output is a probability distribution over the sentiment categories. The category with the highest probability is the predicted sentiment label for the input text.

Figure 1 presents the architecture graphically. This figure shows the architecture of the proposed model, including all the layers and the connection between them. Algorithm 1 shows the architecture of the proposed model with all employed values. Table I shows the layer-wise configuration of the model.

Algorithm 1 Aspect-based sentiment analysis model

```
Require: Load necessary libraries and dataset
Import necessary libraries: numpy, tensorflow, pandas, train_test_split, Tokenizer, pad_sequences, Sequential, Embedding, LSTM, Dense, Bidirectional, Dropout, classification_report, roc_auc_score, roc_curve, auc, matplotlib, drive, EarlyStopping.
Mount Google Drive
Check GPU availability: print("Num GPUs Available:", len(tf.config.list_physical_devices('GPU')))
Configure GPU growth to save memory.
Load dataset from TSV file
Create dataset: dataset = df.drop('Unnamed: 0', axis=1).copy()
Take a small sample of the dataset: percentage = 0.1 (Take 10% of the total data)
Display the sampled data: print(len(dataset))
Extract texts, aspect labels, and sentiment labels.
Perform train-test split: train_test_split() for texts, aspect labels, and sentiment labels.
Tokenization and padding: Initialize Tokenizer and pad_sequences.
Define aspect and sentiment label mappings.
Convert aspect and sentiment labels to one-hot encoding.
Build the LSTM model.
Compile the model: model.compile()
Define early stopping: early_stopping.
Train the model: model.fit() with training data, labels, and callbacks.
```

TABLE I: Layer-wise configuration of the proposed model

Layer	Configuration
Embedding	input_dim: 10000 output_dim: 32
BiLSTM	units: 128 activation: tanh recurrent_activation: sigmoid
Dense	unit: 64 activation: relu
Dropout	rate: 0.2
Dense	unit: 3 activation: softmax

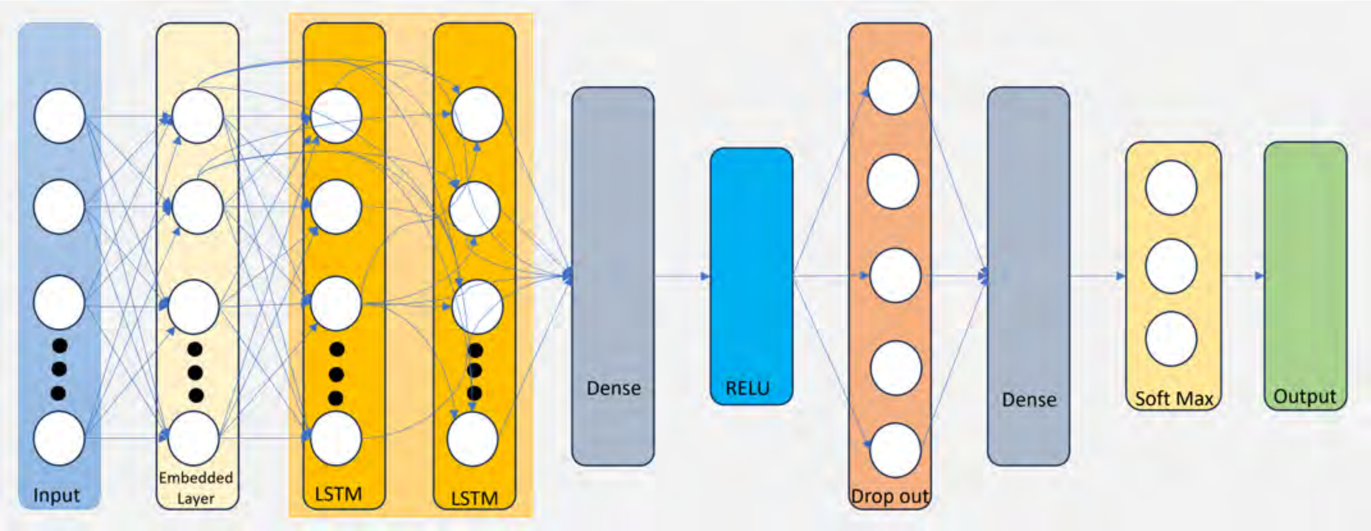

Fig. 1: Architecture of the proposed aspect-based sentiment analysis deep learning model

D. Evaluation Metrics

The performance of the sentiment analysis model is evaluated using standard metrics such as precision, recall, and F1-score. These metrics provide insights into the model's ability to classify sentiment labels for each teaching aspect correctly. The evaluation metrics are discussed as follows.

Precision: Precision is a metric that quantifies the model's ability to make accurate positive predictions. In sentiment analysis, it measures the proportion of correctly predicted instances of a specific sentiment label (e.g., "Positive") out of all instances predicted as that label.

$$Precision = \frac{TP}{TP + FP}$$

Here, TP (True Positives) are the instances correctly classified as a particular sentiment label (e.g.,"Positive"). FP (False Positives) are the instances incorrectly classified as belonging to a particular sentiment label when they do not. Precision is valuable for assessing the model's ability to avoid false positives, which occur when it mistakenly assigns a sentiment label to a text that does not belong to that category.

Recall (Sensitivity): Recall, also known as Sensitivity or True Positive Rate, measures the model's capability to correctly identify instances of a specific sentiment label out of all that genuinely belong to that label.

$$Recall = \frac{TP}{TP + FN}$$

Here, FN (False Negatives) are the instances incorrectly classified as not belonging to a particular sentiment label when they do. Recall is important for evaluating the model's ability to avoid false negatives, which are instances of the target sentiment the model misses.

F1-Score: The F1-Score is a metric that balances precision and recall. It provides a single value that considers both false positives and false negatives, making it useful when the cost of these errors is significant.

$$F1\text{-}Score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

The F1-Score is particularly valuable when dealing with imbalanced datasets or when an uneven cost is associated with false positives and false negatives. It combines precision and recall into a single metric, providing a balanced assessment of the model's performance.

E. Experimental Setup

The experiment was conducted using Google's collaborative IDE, Google Colab. Colab was accessed through the Microsoft Edge web browser on a personal computer powered by an Intel Core i5 processor. The computer boasts 16 GB of RAM and operates on the Windows 11 operating system. To execute the experiment, the Colab environment utilized a T4 Tensor Processing Unit (TPU) with 12 GB of RAM.

The subsequent section presents and discusses the results of our aspect-based sentiment analysis, shedding light on the intricate sentiments that shape the landscape of teacher evaluations.

III. Results and Discussion

This section presents the outcomes of our aspect-based sentiment analysis, focusing on the sentiments expressed by students in their feedback on teachers' performance, specifically in the realms of positive, negative, and neutral teaching aspects. Initially, the experimental results of the proposed models are discussed in comparison with other models, and then a discussion of their implications for teacher evaluation and professional development is provided.

The proposed model is compared with two popular deep-learning models: CNN and GRU. The results are compared on precision, recall, and F1 metrics. Table II shows the experimental results.

Table II summarizes the performance metrics of three different models, namely "Proposed Model", "CNN", and

TABLE II: Experimental results

Model	Precision	Recall	F1	Support
Proposed Model	0.85	0.88	0.86	40155
CNN	0.8	0.87	0.83	40155
GRU	0.83	0.88	0.85	40155

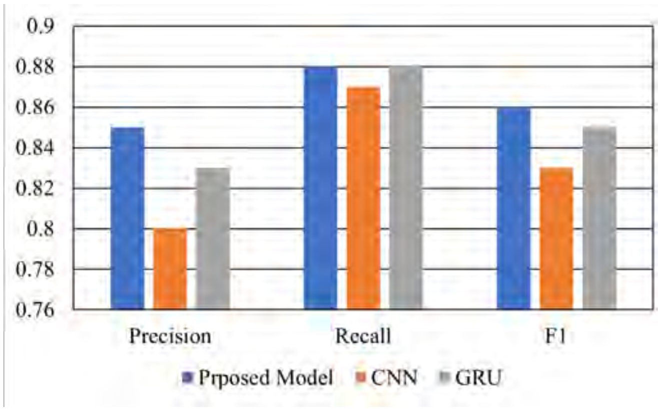

Fig. 2: Comparison of evaluation metrics between different models compared in this study

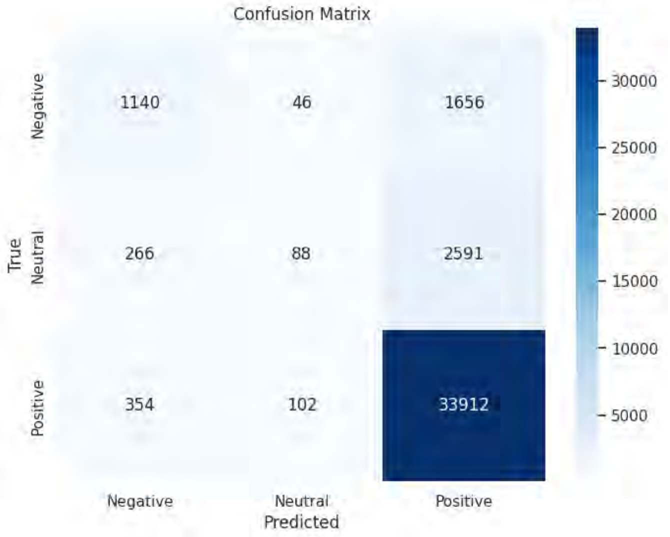

Fig. 3: Confusion matrix of the proposed model

“GRU”, in the context of a sentiment analysis task. These models have been evaluated using standard evaluation metrics, including precision, recall, F1-score, and support, to assess their effectiveness in sentiment classification.

The proposed model demonstrates strong performance, achieving a precision of 0.85, accurately identifying positive sentiment instances. The model also exhibits a high recall of 0.88, suggesting its ability to capture the majority of positive sentiment cases correctly. Consequently, the F1-score for this model stands at 0.86, reflecting a well-balanced trade-off between precision and recall. Furthermore, the model has been evaluated on a substantial dataset with a support count 40155. The CNN model, while slightly lower in precision at 0.80, maintains a commendable recall of 0.87. This indicates its capability to effectively identify positive sentiment instances, albeit with a slightly higher false positive rate than the proposed model. The F1-score for the CNN model is 0.83, indicating a strong overall performance. Like the proposed model, the CNN model has been evaluated on a dataset with a support count 40155. The GRU model exhibits a precision of 0.83 and a recall of 0.88, aligning closely with the proposed model’s performance. This suggests the GRU model’s proficiency in correctly classifying positive sentiment instances. The F1-score for the GRU model is 0.85, indicating a robust balance between precision and recall. Like the other models, the GRU model was evaluated on the same dataset with a support count of 40155. Figure 2, shows the comparison visually.

A confusion matrix provides a visual representation of a model’s performance by displaying the count of correctly and incorrectly classified instances. This allows for a clear assessment of its accuracy, precision, recall, and overall effectiveness in classification tasks. The confusion matrix of the proposed model is shown in Figure 3.

In scientific parlance, this matrix elucidates the performance of the classification model in differentiating between the various sentiment categories. Specifically, the matrix delineates the following metrics:

- True Negatives (TN): 1140 instances were correctly classified as “Negative”, avoiding false identification.
- False Positives (FP): 46 instances that were originally “Neutral” were inaccurately classified as “Negative”.
- False Negatives (FN): 1656 instances that were indeed “Positive” were misclassified as “Negative”.
- True Neutrals (NN): 88 instances were identified as “Neutral”, representing accurate predictions.
- False Positives (FP): 266 instances that were initially “Negative” were wrongly classified as “Neutral”.
- False Negatives (FN): 2591 instances that were truly “Positive” were mistakenly categorized as “Neutral”.
- True Positives (TP): 33912 instances were accurately predicted as “Positive”, reflecting sound classification.
- False Positives (FP): 102 instances that were originally “Negative” were falsely identified as “Positive”.
- False Negatives (FN): 354 instances that were indeed “Neutral” were erroneously classified as “Positive”.

This comprehensive representation facilitates the assessment of the model’s efficacy in assigning sentiment labels to text data, offering insights into the strengths and shortcomings of its performance for each sentiment category.

Receiver Operating Characteristic (ROC) is a graphical representation used for evaluating the performance of classification models. It illustrates the trade-off between the model’s true positive rate (sensitivity) and the false positive rate (1-specificity) for various threshold settings. The ROC plot for the proposed model is shown in Figure 4.

From the ROC curve, it can be seen that it has an AUC of 0.95. The AUC of 0.95 suggests that the classification model can correctly classify positive instances while keeping false positives to a minimum. This characteristic is good for many classification tasks, indicating a robust and accurate model.

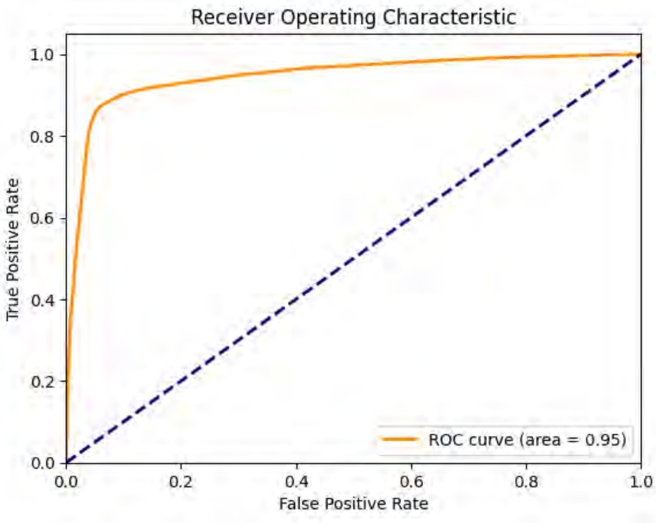

Fig. 4: Receiver Operating Characteristic plot for the proposed model

TABLE III: Sentiment-wise precision, recall, and f1 scores

Sentiment	Precision	Recall	F1-Score	Support
Negative	0.70	0.36	0.48	2842
Neutral	0.54	0.02	0.04	2945
Positive	0.88	0.99	0.94	34368

Table III shows the sentiment-wise precision, recall, and F1 scores achieved from the experiment. “Positive” sentiment achieves an impressive F1-Score of 0.94, demonstrating a robust balance between precision and recall. In contrast, “Neutral” sentiment exhibits challenges with a recall of just 0.02. Figure 5 graphically shows the precision, recall, and f1 curves. This figure represents the training and testing scores for each of the sentiments.

The sentiment analysis revealed a prevalent occurrence of positive sentiments within various teaching aspects. Students’ comments often conveyed appreciation, recognition, and commendation for specific teaching attributes. This abundance of positive sentiments signifies an overall favorable perception of teachers’ performance and underscores the significance of these attributes in creating an effective learning environment.

While positive sentiments dominated the feedback, a modest yet notable presence of negative sentiments was also observed within the teaching aspects. These negative sentiments conveyed constructive criticism, concerns, or areas for improvement. This nuanced feedback offers educators valuable insights into aspects of their teaching that may warrant attention and enhancement.

Within the neutral teaching aspects, sentiments were relatively balanced, indicating a lack of strong emotional valence. Neutral sentiments often reflected factual observations, acknowledgements, or statements without overtly positive or negative connotations. This suggests that certain aspects of teaching are perceived without strong affective bias, serving as neutral elements of the teaching experience.

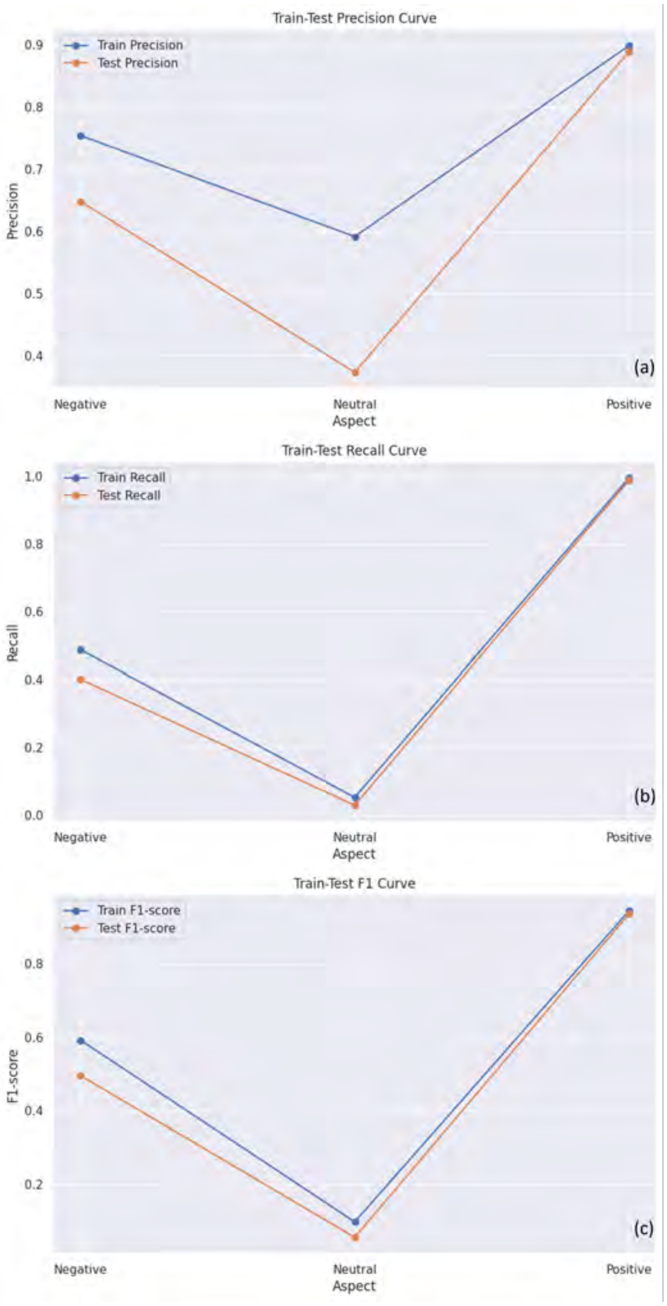

Fig. 5: Sentiment-wise Train and Test curves for (a) Precision, (b) Recall, and (c) F1 for the proposed model

A. *Implications and Discussion*

The prevalence of positive sentiments underscores the importance of reinforcing and nurturing the highlighted teaching attributes, as they contribute significantly to positive student experiences. Recognizing and leveraging these positive aspects can serve as a foundation for amplifying teaching effectiveness and sustaining student engagement.

The presence of negative sentiments presents an opportunity for growth and refinement. Educators can utilize this feedback to address specific concerns, adapt instructional strategies, and cultivate a learning environment responsive to student needs and preferences.

The balanced distribution of sentiments within the neutral teaching aspects signifies a baseline understanding and recognition of these aspects. While neutral sentiments may not drive strong emotional reactions, they form an essential backdrop supporting the teaching experience.

Future research endeavors might involve a comparative analysis between sentiment-based evaluation and traditional assessment methods. This could shed light on the degree of alignment between sentiment-driven insights and established evaluation metrics, further enhancing the validity and effectiveness of the sentiment analysis approach.

Continued development of sentiment analysis models, expansion of the dataset to encompass diverse educational contexts, and exploration of sentiment-driven feedback mechanisms are among the potential avenues for future investigations.

In a nutshell, aspect-based sentiment analysis provides a nuanced understanding of students' sentiments within the positive, negative, and neutral teaching aspects. The implications of these sentiments extend to shaping teacher evaluation, fostering professional growth, and nurturing a holistic approach to educational enhancement. By harnessing the power of sentiment analysis, educators can create a more responsive and impactful teaching environment that resonates with the multifaceted perceptions of their students.

IV. Limitations and Future Work

While our aspect-based sentiment analysis offers valuable insights into teachers' performance through the lens of student sentiments, several limitations warrant consideration:

1) Limited Aspect Coverage: Our study focuses on a subset of teaching aspects, potentially omitting other crucial dimensions of teachers' performance that might impact student experiences. Expanding the aspect framework could provide a more comprehensive evaluation.
2) Contextual Nuances: Sentiment analysis might struggle with capturing context-dependent sentiments, sarcasm, or nuanced expressions. The model's inability to grasp intricate linguistic subtleties could lead to misclassifications.

Our study lays the groundwork for further advancements in sentiment-based teacher evaluation. Several avenues for future research and development emerge:

1) Aspect Expansion and Refinement: Incorporating additional teaching aspects and refining the existing aspect framework could provide a more detailed and accurate assessment of teachers' performance.
2) Fine-Tuned Sentiment Analysis: Improving sentiment analysis models better to understand context, idiomatic expressions, and emotional subtleties would enhance accuracy in sentiment classification.
3) Multimodal Analysis: Exploring the integration of multiple data modalities, such as audio or video, along with text, could offer a richer understanding of student sentiments.
4) Human-AI Collaborative Frameworks: Developing hybrid approaches that combine human expertise with AI-powered sentiment analysis can mitigate subjectivity and improve sentiment classification.

V. Conclusion

In the ever-evolving education landscape, assessing teachers' performance is pivotal in fostering effective pedagogical practices and ensuring enriched learning experiences. This study embarked on a journey to harness the power of sentiment analysis within an aspect-based framework, uncovering the intricate emotional tapestry woven within students' feedback. Our findings shed light on the sentiments expressed across positive, negative, and neutral teaching aspects, offering a dynamic portrait of teachers' performance from the perspective of their students. Through sentiment analysis, we unveiled the significance of effective communication, subject expertise, student engagement, and classroom management. Positive sentiments underscored strengths, negative sentiments pointed to opportunities for growth, and neutral sentiments provided contextual balance. This comprehensive evaluation enriched the understanding of teachers' multifaceted contributions and provided a foundation for targeted improvements. While our study holds promise, it is not without limitations. The constrained aspect coverage and contextual nuances underscore the evolving nature of sentiment-based evaluation. These limitations inspire us to delve deeper, refine methodologies, and explore uncharted territories. In the broader context of educational enhancement, our study contributes to a paradigm shift in teacher evaluation, emphasizing the value of student sentiments as a potent tool for continuous improvement. This approach resonates with the evolving dynamics of education, where personalized and holistic insights drive meaningful progress.

Acknowledgment

The work has been supported by UMPSA RDU GRANT RDU230352, titled ”A fraud detection model for instructor's evaluation based on semantic keyword extraction using machine learning”.

References

[1] N. Mirra and A. Garcia, “In search of the meaning and purpose of 21st-century literacy learning: a critical review of research and practice,” *Reading research quarterly*, vol. 56, no. 3, pp. 463–496, 2021.

[2] R. Stiggins, “Assessment through the student's eyes,” *Educational leadership*, vol. 64, no. 8, p. 22, 2007.

[3] J. DeMonte, “High-quality professional development for teachers: Supporting teacher training to improve student learning.” *Center for American Progress*, 2013.

[4] L. Bardach, R. M. Klassen, and N. E. Perry, “Teachers' psychological characteristics: Do they matter for teacher effectiveness, teachers' well-being, retention, and interpersonal relations? an integrative review,” *Educational Psychology Review*, vol. 34, no. 1, pp. 259–300, 2022.

[5] S. L. Hood, M. E. Dilworth, and C. A. Lindsay, “Landscape of teacher preparation program evaluation policies and progress. evaluating and improving teacher preparation programs.” *National Academy of Education*, 2022.

[6] L. X. Jensen, M. Bearman, D. Boud, and F. Konradsen, “Digital ethnography in higher education teaching and learning—a methodological review,” *Higher Education*, vol. 84, no. 5, pp. 1143–1162, 2022.

[7] P. Sikström, C. Valentini, A. Sivunen, and T. Kärkkäinen, “How pedagogical agents communicate with students: A two-phase systematic review,” *Computers & Education*, vol. 188, p. 104564, 2022.

[8] M. R. Khatri, “Integration of natural language processing, self-service platforms, predictive maintenance, and prescriptive analytics for cost reduction, personalization, and real-time insights customer service and operational efficiency,” *International Journal of Information and Cybersecurity*, vol. 7, no. 9, pp. 1–30, 2023.

[9] M. S. U. Miah, J. Sulaiman, T. B. Sarwar, N. Ibrahim, M. Masuduzzaman, and R. Jose, “An automated materials and processes identification tool for material informatics using deep learning approach,” *Heliyon*, p. e20003, 2023.

[10] M. S. U. Miah, J. Sulaiman, T. B. Sarwar, K. Z. Zamli, and R. Jose, “Study of keyword extraction techniques for electric double-layer capacitor domain using text similarity indexes: an experimental analysis,” *Complexity*, vol. 2021, pp. 1–12, 2021.

[11] K. L. Tan, C. P. Lee, and K. M. Lim, “A survey of sentiment analysis: Approaches, datasets, and future research,” *Applied Sciences*, vol. 13, no. 7, p. 4550, 2023.

[12] A. M. Ponsiglione, F. Amato, S. Cozzolino, G. Russo, M. Romano, and G. Improta, “A hybrid analytic hierarchy process and likert scale approach for the quality assessment of medical education programs,” *Mathematics*, vol. 10, no. 9, p. 1426, 2022.

[13] L. Lohman, “Evaluation of university teaching as sound performance appraisal,” *Studies in Educational Evaluation*, vol. 70, p. 101008, 2021.

[14] T. B. Chiyangwa, J. Van Biljon, and K. Renaud, “Natural language processing techniques to reveal human-computer interaction for development research topics,” in *Proceedings of the International Conference on Artificial Intelligence and its Applications*, 2021, pp. 1–7.

[15] M. S. U. Miah, J. Sulaiman, T. B. Sarwar, N. Ibrahim, M. Masuduzzaman, and R. Jose, “An automated materials and processes identification tool for material informatics using deep learning approach,” *Heliyon*, vol. 9, no. 9, p. e20003, sep 2023. [Online]. Available: https://doi.org/10.1016%2Fj.heliyon.2023.e20003

[16] K. Okoye, A. Arrona-Palacios, C. Camacho-Zuñiga, N. Hammout, E. L. Nakamura, J. Escamilla, and S. Hosseini, “Impact of students evaluation of teaching: A text analysis of the teachers qualities by gender,” *International Journal of Educational Technology in Higher Education*, vol. 17, no. 1, pp. 1–27, 2020.

[17] K. Ezzameli and H. Mahersia, “Emotion recognition from unimodal to multimodal analysis: A review,” *Information Fusion*, p. 101847, 2023.

[18] A. Ishaq, S. Asghar, and S. A. Gillani, “Aspect-based sentiment analysis using a hybridized approach based on cnn and ga,” *IEEE Access*, vol. 8, pp. 135 499–135 512, 2020.

[19] A. Anwar, I. U. Rehman, M. M. Nasralla, S. B. A. Khattak, and N. Khilji, “Emotions matter: A systematic review and meta-analysis of the detection and classification of students’ emotions in stem during online learning,” *Education Sciences*, vol. 13, no. 9, p. 914, 2023.

[20] J. Melba Rosalind and S. Suguna, “Predicting students’ satisfaction towards online courses using aspect-based sentiment analysis,” in *International Conference on Computer, Communication, and Signal Processing*. Springer, 2022, pp. 20–35.

[21] A. Bhowmik, N. M. Noor, M. S. U. Miah, M. Mazid-Ul-Haque, and D. Karmaker, “A comprehensive dataset for aspect-based sentiment analysis in evaluating teacher performance,” *AIUB Journal of Science and Engineering (AJSE)*, vol. 22, no. 2, pp. 200–213, 2023.

Abhijit Bhowmik completed his B.Sc. in Computer Science & Engineering in 2009 and M.Sc. in Computer Science in 2011 from the American International University– Bangladesh (AIUB). He is pursuing his PhD from University Malaysia Pahang Al-Sultan Abdullah in NLP and Machine Learning. He is an Associate Professor and Special Assistant Office of Student Affairs (OSA) in the Department of Computer Science, AIUB. His research interests include NLP, Machine Learning, software engineering, mobile & multimedia communication, and data mining. Mr. Bhowmik can be contacted at ovi775@gmail.com.

Noorhuzaimi Mohd Noor has been in the academic, research & consultancy field since 2003. She is Head of the Program (Entrepreneurship) at the Centre of Creative Entrepreneur Development and Senior Lecturer at The Universiti Malaysia Pahang Al-Sultan Abdullah, Malaysia. She received her B.Sc. in Computer Science from the Universiti Putra Malaysia, Malaysia, in 1999, followed by a Master’s in Science from the same university in 2003. She received her PhD in Computer Sciences from Universiti Kebangsaan Malaysia, Malaysia 2016. She is the author of more than 20 research articles. Her research interests include natural language processing, expert system, and computer security. She is also a Reviewer for the Journal of Information and Communication Technology (JICT) and editor for the International Journal of Software Engineering and Computer Systems (IJSECS). Dr. Noorhuzaimi is also a certified Professional Technologist from the Malaysia Board of Technologists (MBOT), where he is actively involved as Assessor Panel for Technology and Technical Academic Programs Accreditation.

M. Saef Ullah Miah is working as an assistant professor in the Department of Computer Science at American International University-Bangladesh (AIUB). He is currently engaged in research and teaching activities and has practical experience in software development and project management. He obtained his PhD from Universiti Malaysia Pahang and earned his Master of Science and Bachelor of Science degrees from AIUB. In addition to his professional activities, he is passionate about working on various open-source projects. His main research interests are data and text mining, natural language processing, machine learning, material informatics, and blockchain applications.

Debajyoti Karmaker is working as an Associate professor in the Department of computer science at American International University-Bangladesh. He worked as Postdoctoral Research Fellow at Australian National University (ANU), and Stanford University. Before joining ANU, he completed PhD from The University of Queensland (UQ). His research interests are Deep Learning, Computer Vision, & Machine Learning. He is particularly interested in image classification, object detection, segmentation, bio-inspired collision avoidance strategies, and Robust Decision-making and Learning. Before starting his PhD, he worked as a Lecturer at the American International University-Bangladesh (AIUB) - in the Department of Computer Science. He also worked as a software engineer at Infra Blue Technology (IBT Games).