SPECTRUM OF ENGINEERING SCIENCES
ISSN (E) 3007-3138 (P) 3007-312X VOLUME 4, ISSUE 3, 2026

ASPECT BASED SENTIMENT ANALYSIS EDUCATION FEEDBACK FOR IMPROVING EDUCATION QUALITY AND PERFORMANCE USING MACHINE LEARNING AND DEEP LEARNING

Fayyaz Ali[*1], Sajida Karim[2], Ali Orangzeb Panhwar[3], Muhammad Azeem Junejo[4]

[*1]*Software Engineering Department Sir Syed University of Engineering and Technology*
[2]*Assistant Professor, The University of Mirpurkhas, Pakistan*
[3]*Department of Computer Science SZABIST University Gharo, Sindh.*
[4]*Data Analyst, The University of Mirpurkhas, Pakistan*

[*1]fayyaz.ali@ssuet.edu.pk,[2]sajidakarim@umpk.edu.pk,[3]ali.orangzeb@ghr.szabist.edu.pk,
[4]azeem.junejo@umpk.edu.pk

DOI: https://doi.org/10.5281/zenodo.21062496

Keywords
Student Feedback, Education Data mining, Aspect Extraction, ABSA, Machine Learning and LSTM and BiLSTM

Article History
Received: 17 January 2026
Accepted: 07 March 2026
Published: 20 March 2026

Corresponding Author: *
Fayyaz Ali

Abstract
The education domain generates a substantial volume of textual feedback through student reviews, online learning platforms and course evaluation, providing valuable information for evaluating the teaching quality, course material, course design and institutional services. The Aspect based sentiment analysis (ABSA) has gained noteworthy attention as useful technique for getting fine grained options regarding the education domain. In this research study, we have implemented five machine learning methods such as Support vector Machine (SVM), Random Forest (RF), and Logistic Regression (LR), Naïve Base and XGBoost and two deep learning architectures (LSTM and BiLSTM). The models are trained and evaluated on 13,070 annotated student review instances spanning four polarity classes. Due to data diversity, experimental results indicate that XGBoost achieved the highest accuracy (87%) among the seven models capturing contextual and semantic relationships with educational texts. Misclassification analysis reveals that all models struggle most with neutral polarity due to severe class imbalance. With BiLSTM showing the most balanced tradeoff between majority class accuracy and minority class sensitivity, these findings provide actionable insights for education data mining systems aiming to automate the analysis of student feedback at scale.

1. INTRODUCTION

Education domain has witnessed rapid increase in the generation of text data through student feedback and course evaluation, discussion forms, online learning platforms and surveys. The feedback contains important information about the course details, course material, teaching methods, teaching skills and student perceptions and satisfaction level of the different learning [1][3-5]. Besides this, higher education institutions also used SA to enhance the quality of teaching, curriculum design and decision making of the institution. But, as universities move to large, open-ended feedback mechanisms, the volume and heterogeneity of student comments go beyond the practical limits of manual analysis, hindering institutions' capacity to gain timely and actionable insights.

However, manual analyzing data such large scale of the unstructured time consuming, subjective and impractical on a scale. Therefore, automated ABSA is important for extracting meaningful insights from the textual data in education domain. Sentiment classification has been demonstrated to be useful in examining the faculty performance and student's perception of courses. Sentiment Analysis is subfield of natural language processing (NLP) offers an automated means of extracting opinion polarity from text and has been widely applied on social media, product reviews and education. While prior studies sentiment analysis operates at the document of sentence level, assigned a single polarity label to entire review. While useful for general satisfaction monitoring, documents level analysis fails to capture the multi-faceted nature of student feedback [3][4-10][12-14]. In which a single comment can express positive sentiment toward one aspect, for example difficulty of teaching methods while expressing the negative sentiment toward another aspect. For example, difficulty with the quality of material and examination. If a student says: "The lectures are inspiring, but the exams are impossibly hard, and the classroom is freezing" a standard system is confused. Clash between positive, negative and neutral elements result in a useless "Mixed/Neutral" score. So, Aspect based sentiment analysis overcomes these restrictions by linking the polarity of sentiment with specific categories revealed in the text and thus provides fine grained and actionable insights for curriculum designers, instructors and academic administrators. While great strides have been taken, challenges remain in terms of accurately capturing implicit sentiments and managing imbalanced datasets in educational text analysis. In this study, we aim to tackle these by comparing machine learning and deep learning for aspect-based sentiment classification and education domain. So, in this research study, we implemented seven supervised learning methods, five from machine learning and rest of two from deep learning methods for aspect level polarity classification student feedback in education. The experimental pipeline combines course, aspect and comment/review into the unified textual representation. Which is then transformed using TF-IDF feature vectors for five classical machine learning models LR, RF, NB, SVM and XGBoost and tokenized and embedded for two deep learning architectures LSTM, BiLSTM.

The contributions of this are summarized as follows:

- To identify and extract essential education aspects such as course content, teaching quality, learning resources and assessment methods form the student feedback.
- To develop and apply machine learning and deep learning methods for analysis education aspects and polarity
- To compare the performance of machine learning and deep learning methods comparison accuracy, precision, recall and F1 score and confusion matrices.
- Delivers actionable insights that can help the instructors, educators and institutions in enhancing curriculum design, student engagement and teaching strategies.

The rest of this paper is organized as follows: section 2 discuss the aspect-based sentiment analysis machine learning and deep learning approaches in the education sector. Section 3 explains the proposed methodology steps such as data collection, data processing, feature engineering and aspect extraction with machine learning and deep learning methods. Section 4 results and discussion, and the details of experimental step, hyperparameters and evaluation matrices. And provides comparative analysis and discussion and highlights the strengths and weaknesses and finally section 5 concludes the study and outline future research directions for improving ABSA in education text analysis.

2. Related Work

A. Sentiment Analysis in Educational Data Mining

Sentiment analysis has become an established tool within educational data mining for processing student feedback, course evaluations, and online discussion forum posts. Early approaches relied on lexicon-based methods and bag-of-words representations combined

with classical classifiers such as Naive Bayes and Support Vector Machines. More recent surveys highlight that artificial intelligence methodologies including machine learning, deep learning, and transformer-based architecture, are now capable of learning nuanced student opinions and classifying or predicting emotions even for unlabeled feedback corpora [1][4-6]. A persistent challenge identified in this body of work is the handling of negation and multi-polarity statements, where a single comment expresses both positive and negative sentiment toward different aspects of the same course [1][2].

B. Aspect-Based Sentiment Analysis

ABSA extends conventional sentiment analysis by identifying the specific entities or attributes (aspects) toward which an opinion is directed and assigning sentiment polarity independently to each aspect [2]. In the educational domain, common aspects include instructor performance, course content, assessment quality, learning materials, and technological infrastructure [2], [3]. A review of the trends and challenges in applying NLP to educational feedback analysis emphasizes that ABSA remains underexplored relative to document-level sentiment classification, and that aspect extraction and aspect-level sentiment assignment present distinct sub-problems, each with explicit and implicit variants [4].

Kastrati et al. proposed an aspect-level sentiment analysis framework for Massive Open Online Courses (MOOCs) that used weak supervision to reduce reliance on manually annotated data, employing convolutional neural networks for aspect category identification and sentiment classification, achieving F1-scores of 86.13% and 82.10% respectively for the two sub-tasks [5], [6]. Sindhu et al. applied a two-layer LSTM architecture in which the first layer predicted aspects from student feedback and the second layer predicted sentiment polarity, evaluating the approach on both university-collected data and the SemEval-2014 benchmark [6].

C. Machine Learning Approaches with TF-IDF Features

Tradional machine learning combined with TF-IDF for feature extraction extensively have been used sentiment analysis and text classification due to their computational competence and strong baseline performance. TF-IDF converts the textual data into numerical numbers/vectors by highlighting informative terms while reducing the influence of frequency occurring words. Al-Dwish and Aljohani joint the TF-IDF with various machine learning algorithms incorporating logistic regression, K-Nearest Neighbor, SVM and XGBoost for sentiment analysis of online course reviews. And the study proved that the traditional machine classifiers achieved outstanding performance with integration of the TF-IDF representation on logistic regression [7-11]. At the same time, Nasim et al combined lexicon-based features with TF-IDF for sentiment analysis, the proposed model combined RF demonstrated state-of-art performance with comparison to other classifiers for classification of sentiment. Previous studies successfully applied TF-IDF on education data analysis. For instance, kandhro et al employed the TF-IDF for feature extraction in conjunction machine learning and analyzed the meaningful patterns from the textual data [19]. Kabir et al evaluated SVM, Gradient booting, LR and NB on lexical features extracted from course teaching evaluation [7]. Despite these advantages the TF-IDF have some limitations, the contextual dependencies and semantic relationships between the words, the TF-IDF deal the terms individual and cannot capture polysemy, and contextual subtleties present in student feedback [7][18-20].

D. Deep Learning for Sentiment Classification

Deep Learning (DL) have significantly advanced sentiment classification by automatically learning complex feature representation of data. The DL methods have state-of-art on NLP problems. The performance of models for prediction of sentiment analysis and text classification could be knowingly enhanced on

Recurrent neural networks, mostly LSTM and extended version BiLSTM shown the great performance on sequential classification because they capture long range dependencies and contextual word order that frequency-based features cannot represent. Goularas and Kamis combined BiLSTM with multiple convolutional layers using embeddings GloVe with this, the model achieved the outstanding performance CNN and RNN [8]. Addition to this, another comparative study on Indonesian mobile review app, integrated two-layer BiLSTM achieved a weighted F1-score of 0.80% on a held-out test set. Well performed on SVM, DT, MNB built on TF-IDF with five-fold cross validation [9]. Recent study kandhro et al has demonstrated the effectiveness of deep learning methods, containing LSTM and hybrid methods for classification of text [19]. Another research study comparing TF-IDF with SVM against Bi LSTM with word embedding and paragraph vectors reported that Bi LSTM combined with word embedding and vector paragraph achieved outstanding performance with respect to precision and F1 score (0.93%, 0.93%). The finding suggest that bidirectional recurrent architecture provides great performance and representation of contextual sentiment cues and compares unidirectional models and static frequency-based features for text comparison and negation and mix-aspect options [9-10][19-23].

F. Class Imbalance and Statistical Evaluation

Class imbalance is a widespread challenge in sentiment classification, where sentiment categories are significantly underrepresented compared to others imbalanced datasets. a recurring issue in sentiment classification of real-world feedback datasets is imbalanced. Where positive and neutral sentiment control while minority classes such as mixed or conflicting opinions are underrepresented and result in reduced classification performance. The traditional machine learning methods have been applied on data level and code level to handle class imbalanced class issues, Synthetic Minority Over-sampling Technique (SMOTE), under sampling and oversampling which focus to balance classes distributions for model training. Studies addressing food delivery and product review have applied techniques such as random oversampling to mitigate this imbalance prior to training hybrid BiLSTM attention architectures [9]. In sentiment classification, addressing class imbalanced and conducting rigorous statistical evaluation are essential for developing generalized models [10-15].

3. Proposed Methodology

A. Dataset Description

The dataset used in this research study is collected from different public and private universities from academic learning and course evaluation, and institutional surveys. It contains 13,070 instances of student feedback collected across multiple courses. The dataset contains positive, negative and neutral feedback related to courses and teaching, assessment methods, and overall learning experience. Common examples of the feedback include the comments includes the response of instructor, difficulty level of assignment, course delivery, and availability of learning resources. Aspect which denotes the specific feedback dimension being discussed (teaching quality, assessment, course content and infrastructure) and polarity represents sentiment label associated with the comment aspect pair.

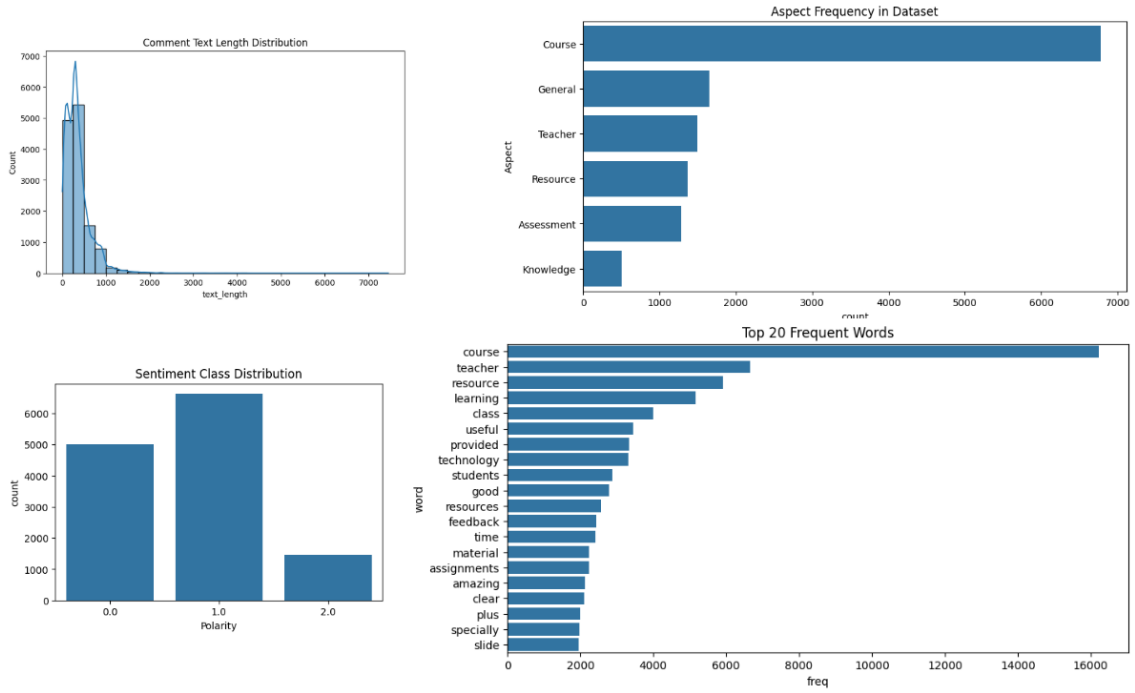

Figure .1 Data distribution and text analysis on education domain.

B. Data Preprocessing

Data preprocessing is an important step in the sentiment classification to transform raw data into structured suitable format for machine learning, initially data is collected unstructured or noisy format. Several steps should be taken to improve the model's quality and performance as shown in figure 1. Prior to feature extraction, the dataset underwent the processing steps. First, records containing null or empty polarity values were removed and data frame index was reset to ensure contiguous indexing. Second step to convert the polarity numerical form assign labels such as (0,1,2) with help of encoder. And the unified representation of text related columns in single form. This concatenation method embeds course level and aspect level context directly into the input representation allowing both machine leaning and deep learning models via the embedding layer to implicitly lean associations between course, aspects and polarity. The text constitution splits training and testing partitions ensuring the relative properties of the three polarity classes were maintained in both partitions.

C. Feature Extraction for Machine Learning Models

Feature Extraction is fundamental stage in machine learning sentiment classification, as converts the textual data into the TF-IDF vectors for machine learning and deep learning algorithms. The TF-IDF vectorizer was configured with maximum vocabulary size of 5000 features and an n gram range of (1, 2) capturing unigram and bi gram lexical patterns. This allows the model to capture the individual words but also short contextual phrases that might carry the polarity, TF-IDF is combing term frequency which measure how frequently words term appears in the document. The vectorizer fitted on training partition and same also used the testing section so preventing the leakages between the two sets. TF-IDF is simple but a powerful feature representation to effectively analyze the student feedback.

D. Machine Learning Classifiers

Five classical machine learning algorithms were trained on the TF-IDF feature representations as discussed in

Tabe 1:

TABLE 1. Classical Machine Learning Models and Configurations

Model	Configuration
LR	A linear discriminative classifier configured with a maximum of 1,000 iterations, used as a baseline due to its interpretability and computational efficiency for high-dimensional sparse text features.
SVM	A margin-maximizing classifier with default radial basis function kernel parameters, selected for its strong performance on sparse, high-dimensional TF-IDF representations.
RF	An ensemble of 200 decision trees (n_estimators = 200, random_state = 42), capturing non-linear interactions between TF-IDF features through bootstrap aggregation.
NB	A probabilistic classifier based on the multinomial event model, well-suited for term-frequency representations and computationally efficient on large vocabularies.
XGB	A gradient-boosted decision tree ensemble configured with 200 estimators, a learning rate of 0.1, and a maximum tree depth of 6 (random_state = 42), representing a state-of-the-art ensemble method for structured and sparse feature classification.

E Deep Learning Architectures

For the deep learning models, the composite text field was tokenized using a Keras Tokenizer restricted to the top 10,000 most frequent words (MAX_WORDS = 10,000), with out-of-vocabulary tokens mapped to a dedicated <OOV> token. Each text sequence was padded or truncated to a fixed length of 100 tokens (MAX_LEN = 100) using post-padding. The resulting padded sequences were split into training and testing partitions using the same 80:20 stratified split strategy applied to the machine learning pipeline.

Two recurrent neural network architectures were implemented:

1) LSTM Model: The input sequence is first passed through a trainable Embedding layer (input dimension = 10,000, output dimension = 128), followed by a single LSTM layer with 128 hidden units, a Dropout layer with a rate of 0.5 to mitigate overfitting, and a final Dense output layer with softmax activation producing probabilities across the four polarity classes.

2) BiLSTM Model: The architecture mirrors the LSTM model but replaces the unidirectional LSTM layer with a Bidirectional LSTM layer of 128 hidden units per direction (256 effective hidden units), enabling the model to learn contextual dependencies from both preceding and following tokens within each comment. This bidirectional context is particularly valuable for student feedback, where sentiment-bearing words (e.g., negations or contrastive conjunctions such as "but" and "however") may appear after the aspect term they modify.

Both models were compiled using the Adam optimizer, sparse categorical cross-entropy loss (appropriate for integer-encoded multi-class labels), and accuracy as the monitored metric. Each model was trained for 10 epochs with a batch size of 32 and a 10% validation split drawn from the training partition, enabling monitoring of training and validation accuracy/loss curves throughout training.

4. Results and Discussion

All experiments were implemented in Python using the scikit-learn library for classical machine learning models and preprocessing utilities, the XGBoost library for gradient boosting, and TensorFlow/Keras for the LSTM and BiLSTM deep learning models. Data manipulation was performed using pandas and NumPy, and

visualizations (distribution plots, confusion matrices, learning curves, and heatmaps) were generated using Matplotlib and Seaborn. Experiments were executed in a cloud-based notebook environment with GPU acceleration available for deep learning model training.

TABLE 2. Hyperparameter Configuration for All Seven Models

Model	Key Hyperparameters
Logistic Regression	max_iter = 1000; solver = default (lbfgs)
SVM	kernel = RBF (default); C = 1.0 (default)
Random Forest	n_estimators = 200; random_state = 42
Naive Bayes	Multinomial distribution; default smoothing (α = 1.0)
XGBoost	n_estimators = 200; learning_rate = 0.1; max_depth = 6; random_state = 42
LSTM	Embedding (10000, 128); LSTM (128); Dropout (0.5); Dense(4, softmax); epochs = 10; batch = 32
BiLSTM	Embedding(10000, 128); BiLSTM(128 × 2); Dropout(0.5); Dense (4, softmax); epochs = 10; batch = 32

The dataset (N = 13,079) was partitioned using stratified random sampling with an 80:20 train-test ratio (random_state = 42), yielding 10,463 training instances and 2,616 testing instances. Stratification ensured that the proportional representation of each of the four polarity classes was preserved across both partitions. For the deep learning models, an additional 10% of the training partition was reserved for validation during the 10-epoch training process, allowing monitoring of generalization performance at each epoch. The TF-IDF vectorizer was configured with max_features = 5000 and ngram_range = (1, 2), producing sparse feature matrices of dimensionality 10,463 × 5000 for the training set and 2,616 × 5000 for the test set. The vectorizer was fit exclusively on the training partition to avoid data leakage. For the LSTM and BiLSTM models, the Keras Tokenizer was configured with num_words = 10,000 and an out-of-vocabulary token (<OOV>). Sequences were padded to a fixed length of MAX_LEN = 100 using post-padding. The resulting input tensors had shape (N, 100), where N is the number of instances in the respective partition.

TABLE 3. Test-Set Performance of Classical Machine Learning Models (TF-IDF Features)

Model	Accuracy	Macro Precision	Macro Recall	Macro F1	Weighted F1
LR	0.846	0.81	0.74	0.75	0.83
SVM	0.846	0.86	0.71	0.71	0.82
RF	0.830	0.81	0.70	0.70	0.80
NB	0.766	0.77	0.62	0.61	0.73
XGBoost	0.870	0.83	0.76	0.77	0.85

Table 3 and Figure 3 summarizes the test-set performance (accuracy, macro-averaged precision, recall, and F1-score, and weighted-averaged F1-score) of the five classical machine learning models trained on TF-IDF features. Among the five classical models, XGBoost achieved the highest overall accuracy (0.870) and the highest macro-averaged recall (0.76) and F1-score (0.77), indicating that its ensemble of gradient-boosted decision trees was most effective at capturing non-linear interactions among TF-IDF unigram and bigram features. Logistic Regression and SVM both achieved an accuracy of 0.846, with SVM exhibiting a slightly higher macro-precision (0.86) but a lower macro-recall (0.71), reflecting a more conservative decision boundary that favors

precision over recall for the minority classes. Random Forest achieved a moderate accuracy of 0.830, while Multinomial Naive Bayes recorded the lowest accuracy (0.766) among all seven models, consistent with its strong independence assumptions limiting its ability to capture the contextual word combinations present in bigram TF-IDF features.

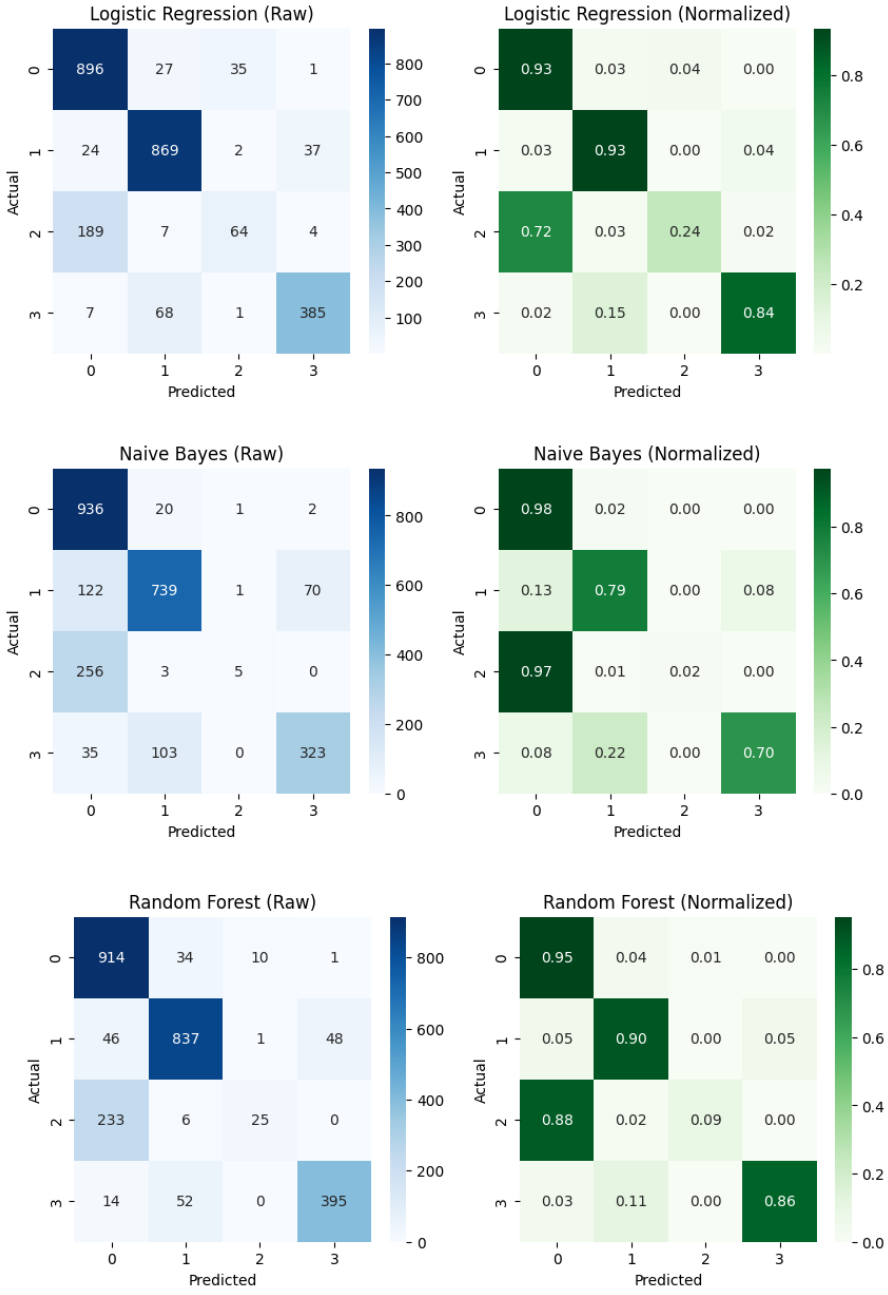

Figure 2 Aspect based Classification using LR, NB and RF on education domain.

Classifications reports shows consistent patterns across all machine learning models, Classes (1-postive, 0-negative and 2-neutral) are classified with relatively high precision and recall. Where class 2 conflict is classified with substantially lower recall across all models, the observed performance trend is primarily driven by significant class imbalance associated with class 2, which contains 26,16 test samples, moreover, mixed sentiment instance are inherently difficult to classify because

they simultaneously convey the negative and positive opinions regarding the different aspects of the same text.

TABLE 4. Test-Set Evaluation of LSTM and BiLSTM Models

Model	Test Accuracy	Test Loss	Training Accuracy	Validation Accuracy
LSTM	0.844	0.482	0.843	0.851
BiLSTM	0.823	0.660	0.959	0.838

Based on the test results, LSTM model achieved 0.84%, slightly better results as compared to BiLSTM model, The BiLSTM model achieved 0.82% accuracy. However, analysis of the training behavior reveals that BiLSTM model performs better as compared to LSTM. On sequence input, bidirectional LSTM architecture was able to capture contextual information. The combination of high accuracy and increased test loss suggest that the model learns patterns specific to the training samples but is unable to generalize on unseen samples. This type of behavior normally shows BiLSTM due to their large number of trainable parameters and representational capacity as summarized in Table 2. Furthermore, Bidirectional architecture enabled the preceding and succeeding an aspect term and this capability only student feedback data. Where sentiment are often expressed though complex sentence structure and contrastive for instance "lecture is interesting, but assignment very difficult" by incorporating information both directions, in this condition BiLSTM model perform well it captured the sentient relationships, but LSTM overlook such kind review or feedback.

Figure 1 demonstrated and visualized the different model performance, where XGBoost achieved state-of-art results and narrow margin separating six models that shows that choice of model does not rest on accuracy alone. The BiLSTM encode richer sequential of the input text with embedding and recurrent layers, which shows that superior handling the minority ambiguous polarity. This distinction between class sensitive performance and aggregate accuracy and despite XGBoost marginal edge.

Table 4 showed learning behavior of the LSTM and BiLSTM on 10 epochs. LSTM exhibited stable and steady convergence with training accuracy (0.49% to 0.84%) and validation accuracy reaching peak 0.85% at 9 epoch. The close alignment between training and validation which shows minimal overfitting on a first epoch learning. Furthermore, validation reached a peak of two to three epochs. It gradually declined thereafter while validation loss increased. These observations suggest that early stopping is based on validation loss monitoring for suppose stopping training three to four epochs would likely yield a BiLSTM checkpoint

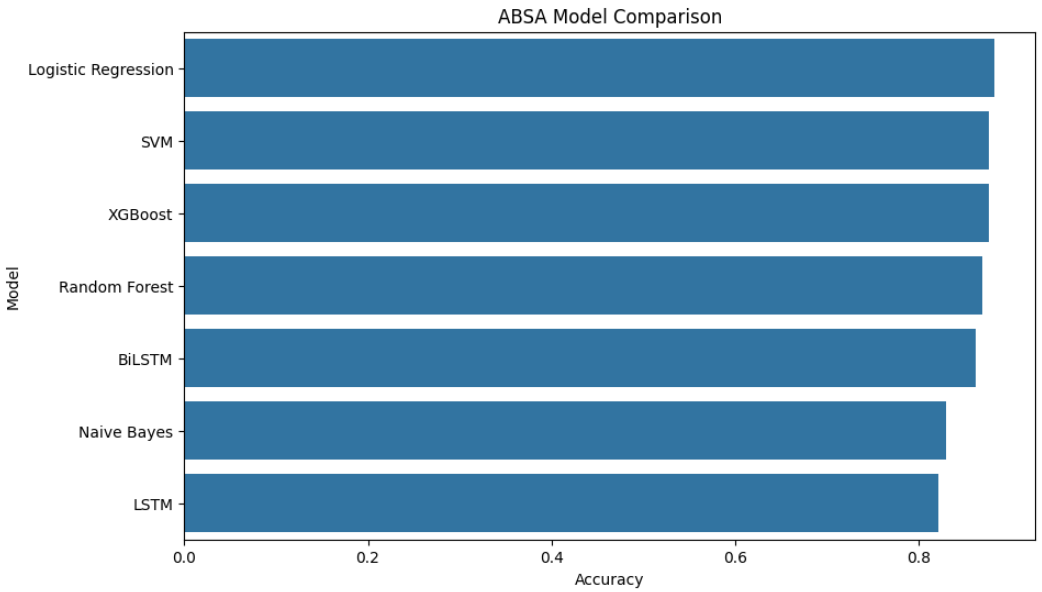

Figure 3 ABSA Performance comparison of different models on education domain

Figure 2 and Figure 4 showed Raw and row-normalized confusion matrices were generated for all seven models (Logistic Regression, SVM, Random Forest, Naive Bayes, XGBoost, LSTM, and BiLSTM) across the four polarity classes (0: negative, 1: positive, 2: conflict, 3: neutral), following the classification report support values of 959, 932, 264, and 461 test instances, respectively. Across all seven confusion matrices, a consistent diagonal-dominant pattern is observed for classes 0, 1, and 3, indicating that most negative, positive, and neutral instances are correctly classified, with normalized diagonal values typically exceeding 0.80. In sharp contrast, the normalized confusion matrices for class 2 (conflict) show that most misclassified conflict instances are predicted as either class 0 (negative) or class 1 (positive), rather than class 3 (neutral). This pattern is consistent with the linguistic nature of conflict-polarity comments, which typically contain both a positive sentiment-bearing segment and a negative sentiment-bearing segment; classifiers without explicit aspect-segmentation capability tend to be "pulled" toward whichever polarity dominates the overall lexical content of the comment. The BiLSTM model's normalized confusion matrix exhibits the most balanced off-diagonal distribution for class 2 among all seven models, with a comparatively higher proportion of conflict instances correctly retained in class 2 or redistributed toward class 3 (neutral) rather than being fully absorbed into classes 0 or 1. This is attributable to the bidirectional recurrent layer's ability to maintain representations of both halves of a contrastive sentence (e.g., the clauses before and after a conjunction such as "but"), enabling a more nuanced final representation that does not collapse entirely toward one polarity extreme. The unidirectional LSTM, by contrast, processes the sequence only in forward order, and its final hidden state is therefore disproportionately influenced by the sentiment expressed in the final clause of the comment—a structural limitation that the BiLSTM's backward pass directly addresses.

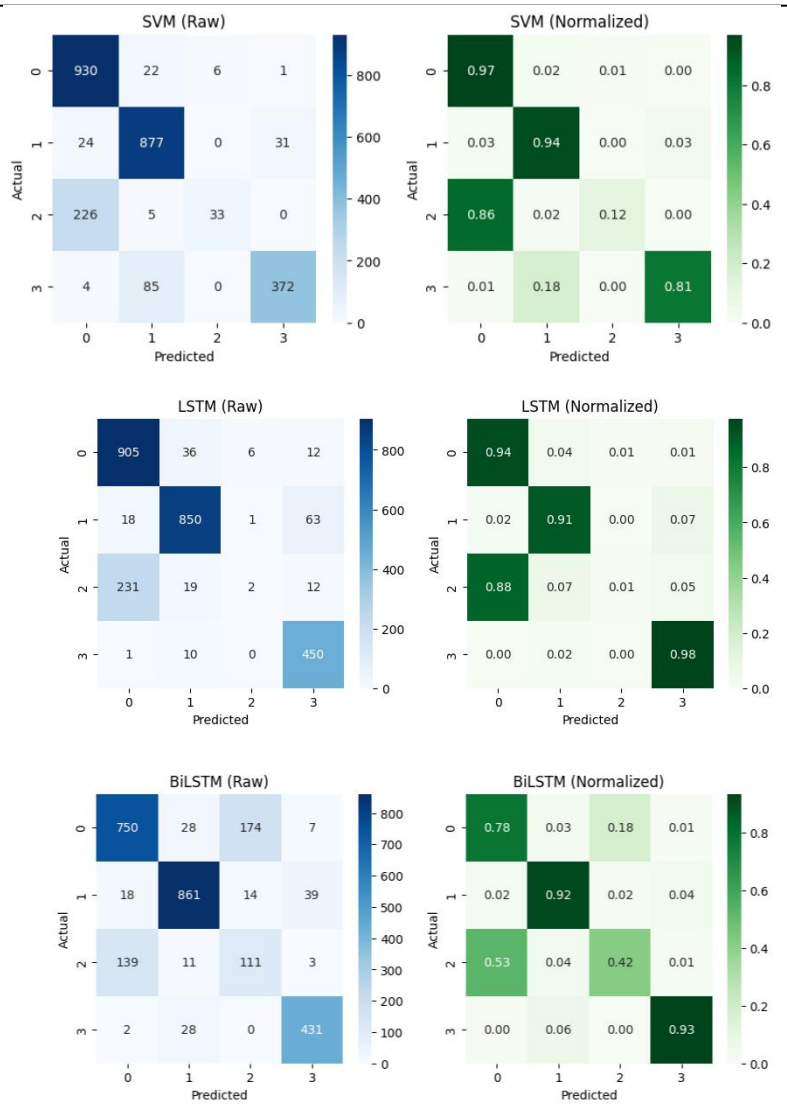

Figure 4 Aspect based Classification using SVM, LSTM and BiLSTM on education domain.

For the classical machine learning models, Naive Bayes exhibits the most extreme misclassification pattern for class 2, with a normalized recall of only 0.02 (i.e., 98% of conflict instances are misclassified), while Logistic Regression and XGBoost achieve the highest class-2 recall among the classical models at 0.24 and 0.20, respectively. This further supports the conclusion that models capable of capturing higher-order feature interactions (gradient boosting) or learned sequential context (BiLSTM) are better equipped to detect the linguistic markers of mixed-sentiment feedback than models relying on independent term frequencies (Naive Bayes) or purely linear decision boundaries over sparse TF-IDF vectors.

The aspect-versus-polarity heatmap (cross-tabulation of the Aspect and Polarity fields) provides a granular view of how sentiment is distributed across different feedback dimensions. Aspects related to instructor engagement and teaching delivery show a predominance of positive-polarity instances, suggesting that students are generally satisfied with direct pedagogical interaction. Conversely, aspects related to assessment workload, grading transparency, and course difficulty show a higher relative proportion of negative and conflict-polarity instances, indicating that these dimensions represent the primary sources of student dissatisfaction or ambivalence.

Notably, the conflict-polarity class (class 2) is not uniformly distributed across aspects: it occurs disproportionately within aspects that combine evaluative judgments about both content quality and workload (e.g., "the course material was excellent, but the workload was excessive"). This observation has direct practical implications for educational administrators: aspects with elevated conflict-polarity proportions represent areas where students perceive genuine trade-offs, and these aspects may warrant qualitative follow-up (e.g., focus groups) in addition to automated sentiment monitoring, since a high conflict-polarity rate cannot be resolved by simply tallying positive versus negative counts.

From a modeling perspective, the aspect-wise analysis reinforces the importance of the BiLSTM's contextual modeling: because conflict-polarity instances are concentrated in specific aspects with characteristic linguistic patterns (contrastive conjunctions, workload-related vocabulary co-occurring with content-quality vocabulary), a model capable of learning aspect-conditioned contextual representations such as BiLSTM, whose input already encodes the Aspect token within the composite text field is structurally better positioned to learn these aspect-specific sentiment patterns than frequency-based TF-IDF models, which treat the aspect token as just another unigram/bigram feature without positional or sequential context.

5. Conclusion and Future Work

This paper presented a comprehensive aspect-based sentiment analysis pipeline for educational feedback data, evaluating five classical machine learning models trained on TF-IDF features and two deep learning models trained on tokenized and embedded sequence representations. Using a dataset of 13,079 student feedback, the study conducted exploratory data analysis, feature extraction, model training, multi-metric evaluation, confusion matrix analysis, learning curve analysis, aspect-wise sentiment breakdown, and Friedman-test-based statistical validation.

The results demonstrate that while XGBoost achieved the highest raw test accuracy 87%, the BiLSTM model emerged as the most contextually robust architecture, achieving the fastest convergence among the deep learning models, the highest training accuracy 0.95%, and the most balanced confusion matrix with respect to the challenging minority conflict-polarity class. Error analysis revealed that the conflict-polarity class remains the principal source of misclassification across all models due to severe class imbalance and the inherent linguistic complexity of mixed-sentiment text, with BiLSTM's bidirectional contextual encoding providing the most effective though incomplete mitigation of this challenge among the seven evaluated models. As shown figure 3 Overall, these findings position BiLSTM as the preferred architecture for deployment in educational feedback analysis systems were capturing nuanced, contextually dependent, and mixed-sentiment feedback is of practical importance, complementing its slightly lower raw accuracy with superior representational depth and faster convergence. Future work includes pre-trained embeddings (GloVe, FastText) or transformer models (BERT, RoBERTa) can be explored to enhance generalization and minority-class performance. Additionally, future studies should incorporate automatic aspect extraction (e.g., BiLSTM-CRF) and apply stronger regularization techniques such as early stopping and learning rate scheduling.

REFERENCES

Shaik, Thanveer, et al. "Sentiment analysis and opinion mining on educational data: A survey." Natural Language Processing Journal 2 (2023): 100003.

Hajrizi, R., & Nuçi, K. P. (2020). Aspect-based sentiment analysis in education domain. arXiv preprint arXiv:2010.01429.

Ngwira, B., Gobin-Rahimbux, B., & Sahib, N. G. (2023). A Deep-learning framework for analysing students' review in higher education. Computational Intelligence and Neuroscience, 2023(1), 8462575.

Shaik, T., Tao, X., Li, Y., Dann, C., McDonald, J., Redmond, P., & Galligan, L. (2022). A review of the trends and challenges in adopting natural language processing methods for education feedback

analysis. Ieee Access, 10, 56720-56739.

DAHIYA, S., ARORA, A., BHARDWAJ, A., KUMAR, M., & RAY, M. (2024). Comparative analysis of machine learning and deep learning models for aspect-based sentiment analysis in education. JOURNAL OF SCIENTIFIC RESEARCH, 30(12), 567-576.

Koufakou, A. (2024). Deep learning for opinion mining and topic classification of course reviews. Education and Information Technologies, 29(3), 2973-2997.

Deshpande, S. B., Tangod, K. K., Srinivasaiah, S. H., Alahmadi, A. A., Alwetaishi, M., Ong Michael, G. K., & Rajendran, S. (2025). Elevating educational insights: sentiment analysis of faculty feedback using advanced machine learning models. Advances in Continuous and Discrete Models, 2025(1), 89.

Kamruzzaman, M., & Kim, G. L. (2023, November). Efficient sentiment analysis: a resource-aware evaluation of feature extraction techniques, ensembling, and deep learning models. In Proceedings of the 11th international workshop on natural language processing for social media (pp. 9-20).

Purba, U. W., Parhusip, A. H. R., Maulana, S., Muthoharoh, L., Satria, A., & Manullang, M. C. (2026). Sentiment Analysis of Indonesian Spotify Reviews Using Machine Learning and BiLSTM. arXiv preprint arXiv:2605.03443.

Purwarianti, A., & Crisdayanti, I. A. P. A. (2019, September). Improving bi-lstm performance for indonesian sentiment analysis using paragraph vector. In 2019 International Conference of Advanced Informatics: Concepts, Theory and Applications (ICAICTA) (pp. 1-5). IEEE.

Alshaikh, K. A., Almatrafi, O. A., & Abushark, Y. B. (2023). BERT-based model for aspect-based sentiment analysis for analyzing Arabic open-ended survey responses: A case study. IEEE Access, 12, 2288-2302.

DAHIYA, S., ARORA, A., BHARDWAJ, A., KUMAR, M., & RAY, M. (2024). Comparative analysis of machine learning and deep learning models for aspect-based sentiment analysis in education. JOURNAL OF SCIENTIFIC RESEARCH, 30(12), 567-576.

Li, R., Patel, T., Wang, Q., & Du, X. (2024). Mlr-copilot: Autonomous machine learning research based on large language models agents. arXiv preprint arXiv:2408.14033.

Koufakou, A. (2024). Deep learning for opinion mining and topic classification of course reviews. Education and Information Technologies, 29(3), 2973-2997.

Edalati, M. (2020). The potential of machine learning and NLP for handling students' feedback (A short survey). arXiv preprint arXiv:2011.05806.

Ranjan, S., & Mishra, S. (2020, July). Comparative sentiment analysis of app reviews. In 2020 11th international conference on computing, communication and networking technologies (ICCCNT) (pp. 1-7). IEEE.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.

Kandhro, I. A., Wasi, S., Kumar, K., Rind, M., & Ameen, M. (2019). Sentiment analysis of students' comment using long-short term model. Indian Journal of Science and Technology, 12(8), 1-16.

Kandhro, I. A., Chhajro, M. A., Kumar, K., Lashari, H. N., & Khan, U. (2019). Student feedback sentiment analysis model using various machine learning schemes: A review. Indian Journal of Science and Technology, 12(14), 1-9.

Kandhro, I. A., Ali, F., Uddin, M., Kehar, A., & Manickam, S. (2024). Exploring aspect-based sentiment analysis: an in-depth review of current methods and prospects for advancement. Knowledge and Information Systems, 66(7), 3639-3669.

Kandhro, I. A., Jumani, S. Z., Ali, F., Shaikh, Z. U., Arain, M. A., & Shaikh, A. A. (2020). Performance analysis of hyperparameters on a sentiment analysis model. Engineering, Technology & Applied Science Research, 10(4), 6016-6020.