National College of Ireland

Sentiment Analysis of Noisy Student Feedback using SVM and BERT with Explainable AI

MSc Research Project
Msc Data Analytics

Sreenivas
Student ID: X23326603

School of Computing
National College of Ireland

Supervisor: Qurrat Ul Ain

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Sreenivas
Student ID:	X23326603
Programme:	Msc Data Analytics
Year:	2025
Module:	MSc Research Project
Supervisor:	Qurrat Ul Ain
Submission Due Date:	11/08/2025
Project Title:	Sentiment Analysis of Noisy Student Feedback using SVM and BERT with Explainable AI
Word Count:	5324
Page Count:	18

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	11th August 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Sentiment Analysis of Noisy Student Feedback using SVM and BERT with Explainable AI

Sreenivas
X23326603

Abstract

The online learning landscape has expanded at an unprecedented level which has led to the increased student feedback, most of which is informal and holds recordable information about the quality of courses and experience as a student. Nevertheless, the chaotic and unorganised format in which such feedback takes place is very challenging to a machine when it comes to interpreting sentiment. The present study examines the comparative performance of conventional machine learning and transformer-based solutions with explainable AI (XAI) to the sentiment analysis problem in this context.

Two pipelines, (1)TF-IDF + Linear Support Vector Classifier (LinearSVC) model and (2) a fine-tuned BERT transformer model, were used to process a dataset of GPT-labeled student feedback. Both the models were tested with 1,000 reviews and SHAP-based explainability was used on the 100 top reviews with more than 50 characters, to give local and global feature attributions.

The experimental findings revealed that the SVM attained the test accuracy and F1-score of 99.49% and 99.43% respectively whereas the BERT model has a test accuracy and F1-score of 98.89% and 98.75% respectively. Although SVM had a slightly better performance in classification, SHAP showed a key distinction: BERT with its context aware and multi-word senses was able to value semantically rich tokens like *helpful* and *training*, BERT was contextual- and multi-word-sensitive, SVM simply used the high-weighted TF-IDF tokens as a whole, being more sensitive to noise and more effective with explicit sentiment keywords.

These results indicate that transformer modeling based approaches together with XAI can lead to more balanced performance in terms of interpretability and robustness on noisy, less formalized student feedback, but that carefully-pruned traditional models can still deliver competitive baseline performance. Automated educational analytics and adaptive learning systems informed by the sentiment of learners have implications to these findings.

1 Introduction

The world of online learning has seen a exponential growth and is now providing easy and accessible education to the wider audience. An essential part of the enhancement of these platforms is comprehension of the student opinion via feedback, which can provide us with information regarding courses, instructing skills, and learners participation. In contrast with ordered survey data, feedback can be informal, contextual and noisy and automated sentiment understanding by a natural language processing (NLP) system is an ill-defined problem.

Conventional methods of sentiment analysis, e.g., Support Vector Machines (SVM) and combining them with TF-IDF feature extraction, have been common because they are more simple and provide good baseline results. Nevertheless, the techniques are mostly based on disjoint keywords and fail to pick up subtle contextual dependencies on free-form text. The introduction of the architectures using transformers, especially these as BERT, has made a great impact in NLP because it now has contextualized understanding of language to a great depth, and so this is a great prospect as a way of analyzing complex classes of sentiment in unstructured feedback.

Although these models have achieved greater performance, two questions remain when implementing transformer-based models in the educational analytics field: (1) are transformer based models performing better than traditional machine learning algorithms when applied to sentiment classification of unpleasant informality in feedback of students, in terms of a learning curve? and (2) are such models capable of producing explanations that are actionable knowledge to an educator or administrator? These questions will have to be handled through not only the comparison of the performance through benchmarking but also the incorporation of explainable AI (XAI) methods to guarantee transparency and confidence of the model predictions.

This study pits a TF-IDF + Linear Support Vector Classifier (LinearSVC) pipeline against a fine-tuned BERT transformer model against a dataset of GPT-labeled student feedback. Concerning interpretability, SHAP (SHapley Additive exPlanations) is fitted to the two models, which gives local token level and global feature attributions. The extended attention is paid to the top 100 longest reviews that have more lingual depth and can be the examples of the challenging cases of sentiment classification.

Through the analysis of the classification performance and the pattern of the explanations, the project will evaluate the trade-offs between the traditional and the transformer-based models in the perspectives of accuracy, resistance to noise, and explainability. These results can be directly applied to creating an automated sentiment analysis system in online learning and building into educational analytics pipelines an explainable AI.

2 Related Work

Sentiment analysis is an area that has gained a lot of research in the field of e-commerce, social networks, and education. In this section, the review of prior work studies that are essential to the current research in three separate fields is conducted: (1) sentiment analysis in student feedback in online learning settings, (2) the comparative analysis of the traditional approaches to machine learning against the ones based on transformers, and (3) the incorporation of explainable AI (XAI) into the sentiment analysis that can be explained.

2.1 Sentiment Analysis in Educational Feedback

Sentiment analysis in processing the feedback of students has much traction with the online platforms producing vast quantities of qualitative data. There is beginning to be evidence in the use of transformers meaning nuanced sentiment patterns in more informal student text demonstrated by Koufakou et al. (1), applying deep learning and BERT based models in processing over 10,000 online course reviews. Deshpande et al. (2) employed SVM, Random Forest, and MLP on 5,000 faculty feedback responses processed with TF-IDF features showing that classical ML methods are useful in structured

educational feedback. Likewise, literature highlights the utility of sentiment analysis in enhancing teaching quality, curriculum, and adaptive learning systems .

In the previous studies, classifiers were strongly found on the traditional model. Al-trabsheh et al. (3) have shown sentiment analysis of live lecture feedback using Linear SVMs with unigram features and applied bag-of-words along with ensemble on large-scale RateMyProfessor data . Such experiments not only prove the efficacy of clas- sical methods but also uncover the drawbacks in working with weakly-guided open-ended, noisy commentary with mixed sentiment and inconsistent language use–as typically oc- curs in natural student responses.

The present studies also deal with aspect-based sentiment analysis (ABSA) in edu-cation data. To extract an aspect-specific polarity in junior school feedback, Ren et al.deployed Bi-LSTM attention models, whereas Kastrati et al.considered weakly supervised CNN to recognize the categories of aspects in MOOC reviews. The two studies emphasized the value of fine-grained interpretations of sentiments as neces- sary to actionable insights, which portends well to the issue of explaining tokens in the research.

2.2 Traditional ML vs. Transformer Models

Support Vector Machines (SVMs) and TF-IDF are still a benchmark of sentiment clas-sification because of high results on little and sparse datasets (4). Other studies like Zainuddin and Selamat (5), revealed that there have been effects of SVMs, with compar-ative studies showing that SVMs are better at some structured fields, and some research shows that feature-engineered models could not be scaled well under complex situations. In evaluations using informal review data. determined that BERT outperformed SVMs with notable successes when sentiment cues were interdependent as opposed to being keyword based.

NLP is now redesigned by transformers. BERT has been developed by Devlin et al. (6) and reports the state-of-the-art performance on sentiment tasks. Rybinski and Kopciuszewska (7) also tested BERT variants versus SVM and Naive Bayes in assessment of students proving the benefit of the contextual embeddings in case of feedback noise and domain specificity. Arora et al. also found that BERT has fewer manual feature engineering than classical models hence it would be suitable to use in areas that constantly diversify such as online learning platforms.

On fine-grained tasks, Li et al. (8) and Sun et al. (9) demonstrated that BERT can raise the performance of aspect-based sentiment analysis (ABSA) notably without the requirement to tune it much to task specifications, compared to RNN-based models. These discoveries build the argument to compare the transformer-based models with conventional ML models in the natural sentiment analysis within the education context, particularly, on informal and unstructured data.

2.3 Explainable AI for NLP Sentiment

With the dominance of deep models, the interpretability has become an essential demand, particularly in a sensitive area of activity such as education. SHAP (SHapley Additive exPlanations) has found applicability with the common usage in the provision of local and global interpretations of NLP models (10). Talaat, (11) used SHAP in order to integrate social media sentiment into hybrid BERT-BiGRU models and showed how token-level

attribution increases trust and model debugging. Mukherjee et al. compared SHAP and LIME on sentiment tasks pointing out to SHAP consistency and its model-agnosticness.

In sentiment analysis of education, actionability requires explainability. Deshpande et al. (2) stressed the point that transparent models facilitate trust in the case of sentiment analytics used in their decision-making by instructors and administrators. Nevertheless, at the present, limited literature exists regarding the combination of transformer models and token-level SHAP explanations to informal student feedback, which makes an excellent hole in educational NLP applications that are relevant to people.

2.4 Handling Noisy, Informal Text

Student reviews on the real environment are usually noisy to some minor degree and contain grammatical irregularities, mixed polarity per sentence, and range in length. In their study on the systematic review of sentiment analysis, Ligthart et al. (12) listed the issues of noise and domain adaptation that remained open at the date of their work. Such noises in the traditional TF-IDF-based models bias these models, as these models tend to focus on isolated terms, whereas transformer models are not as sensitive because they almost filter out such grammatical irregularities because they have contextual embedding (11). Nonetheless, not much research is available which integrates strong models of context with explainable AI, to explicitly assess performance and interpretability in noisy data settings.

2.5 Research Gap

Although this has been done comprehensively, there are three gaps. First, little comparative studies on conventional SVM and transformer-based models are done on informal, noisy student feedback in particular. Second, there are not a lot of studies that involve SHAP or analogous XAI techniques to introduce the level of interpretability at the token level in educational sentiment analysis where such explanations are paramount. Third, the existing literature mostly focuses on evaluated explainable models on curated datasets, leaving the evaluation of the models on naturalistic but slightly noisy feedback. This study fills such gaps by comparing SVM and BERT on GPT-labelled informal student reviews and using SHAP on the 100 longest reviews to derive local and global feature attributions, filling in performance assessment and interpretation in educational NLP.

3 Methodology

3.1 Research Design

This paper uses a quantitative experimental design by comparing a conventional machine learning sentiment classifier to a deep learning model that uses a transformer to measure noisy student evaluations. The former pipeline applies Term Frequency–Inverse Document Frequency (TF–IDF) features together with a Support Vector Machine (SVM), whereas the latter pipeline applies a fine-tuned BERT (bert-base-uncased) transformer model. Both models were trained and tested on the same stratified splits to ensure reproducibility and a paired, controlled comparison. The purpose of their research is to

find out which strategy would yield better classification performance and to evaluate the explainability of both models on explainability metrics based on SHAP.

3.2 Data Collection and Preprocessing

The dataset was a collection of course feedback of the students found on a Kaggle dataset and was annotated into three sentiment classes (*Positive*, *Neutral*, *Negative*) with the help of GPT-based annotation. The quality of labels was with a large percentage of over 95 percent agreed after using a stratified sample of the labels to verify the labels manually. Raw text was cleaned such as removal of duplicates, non-english filtering, lowercasing, stripping of punctuation,normalizing whitespace, removing URLs and email addresses.

The problem of class imbalance was partly solved by unifying the rare "Mixed" labels into the *Neutral* category, down-sampling the majority *Positive* category to 3,000 samples, and up-sampling the *Negative* and *Neutral* classes each to 1,000 samples using random resampling. The aim was to ensure that the classes were reasonably distributed between sets to eliminate any biases in the metrics of evaluation. To achieve this, the final dataset was stratified at an 80/20 stratified train test split. Two pipeline processes of preprocessing were model-specific: TF–IDF vectorizing and SVM and subword tokenizing and BERT.

3.3 Model Training

TF–IDF + SVM Pipeline. The SVM classifier (sklearn.svm.SVC) was trained on The TF–IDF feature representations where the hyperparameter tuning is done using GridSearchCV. This parameter grid consisted of C in 0.1, 1, 10, 100, kernel functions (*linear*, *RBF*), gamma selection scale, auto, 0.001, 0.01, 0.1, 1 and the use of different weighting strategies, such as balanced weights and the calculated weights based on the training distribution. To deal with class imbalance, five-fold stratified cross-validation models were applied with macro-F1 as an optimization measure. TF–IDF vectorizer was set to a 10,000 features vocabulary, and included unigrams, bigrams and sublinear term frequency scaling to minimize the effect of very common tokens.

BERT Fine-Tuning Pipeline. Transformer-based model was applied with the bert-baseuncased model and HuggingFace Transformers. Tokens of text were limited to a maximum length of 128 tokens with a batch size 16 during training and evaluation. Fine-tuning has been conducted on 2 epochs AdamW Optimizer with learning rate set to 2e-05 and warmup scheduler programmed using linear warmup using get linear schedule with warmup. The model was trained by PyTorch, which automatically detects CUDA to fluently operate on GPUs when accessible. The classification head actually had dynamically configured number of labels that were determined by the encoding classes of the dataset.

3.4 Evaluation Methodology

Accuracy, Precision, Recall, and macro-F1 were all used to measure model performance to address the class imbalance aspect. Official definitions of the metrics are:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{1}$$

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{2}$$

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{3}$$

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

where TP , TN , FP , and FN are true positives, true negatives, false positives, and false negatives, respectively. An alternative to macro-averaging was selected in order to guarantee that each sentiment would have the same weight regardless of size.

To check the statistical significance of differences, visible between SVM and BERT, McNemar was used in paired predictions on the test set. We considered that the errors had the same distribution in the models, as the null hypothesis was H_0; we applied the significance level of $\alpha = 0.05$. A p-value lower than this value showed statistically significant dissimilarity in the performance of models.

3.5 Explainability Methodology

SHAP (SHapley Additive Explanations) was used to analyze the model interpretability as it has model-agnostic interpretability and token-level interpretability. In the case of SVM, shap.LinearExplainer was used on TF–IDF features in order to extract the contribution value of each term. In the BERT model, a transformer-compatible shap.Explainer was connected to the HuggingFace pipeline to calculate token level attributions. The analysis levels included local explanations of individual predictions, to interpret per-sample thinking, and the global importance scores which are the mean absolute Shapley values across 1,000 reviews to determine prevalent sentiment tokens of the dataset.

3.6 Experimental Workflow

The whole experimental pipeline was in a sequential fashion. Student feedback in raw form on Kaggle was tagged with GPT and cleaned and balanced in respect to classes. Parallel pipelines were used where the preprocessed data fed into, and TF–IDF feature extraction on the SVM or tokenization on the BERT. The two models are trained with their corresponding configuration and hyperparameter and tested on the same stratified test sets. The overall and per-class performance were calculated with respect to the metrics. Then SHAP values were taken to create local and global interpretability insights. Lastly, McNeamar test was used to ensure the statistical significant of the difference in performance and the outcomes in the form of metrics and results concerning explainability were combined as part of final comparative analysis.

3.7 Reproducibility and Tools

All experiments were conducted in Python 3.10. Scikit-learn was used for TF–IDF vectorisation, SVM training, and metric computation; PyTorch and HuggingFace Transformers for BERT fine-tuning and inference; and SHAP for explainability. The code automatically detects and uses a CUDA-enabled GPU via torch.cuda when available. All scripts, preprocessing routines, and trained models are stored in a version-controlled repository to ensure full reproducibility.

4 Design Specification

To explicitly compare a conventional machine-learning approach to sentiment analysis to a deep transformer-based model on the same data, the system uses a dual-pipeline architecture. The evaluation and preprocessing framework are the same in both models so that paired and reproducible experiments can be done. Main design points are summarised in Figure 1: sequential data-preparation step, two parallel modelling streams and an intersection of explainability and evaluation.

The full architectural design is composed of a common preprocessing process and 2 modelling pipelines in parallel, one pipeline with TF–IDF features and a Support Vector Machine (SVM) classifier and another with a fine-tuned BERT transformer model. The data that is presented to both pipelines is clean and balanced student-feedback data and thus the comparison is fair. Each model is paired with SHAP-based explainability at the local and high global level, and a combined assessment step calculates the evaluation metrics and statistical comparisons to determine accuracy, stability and interpretability.

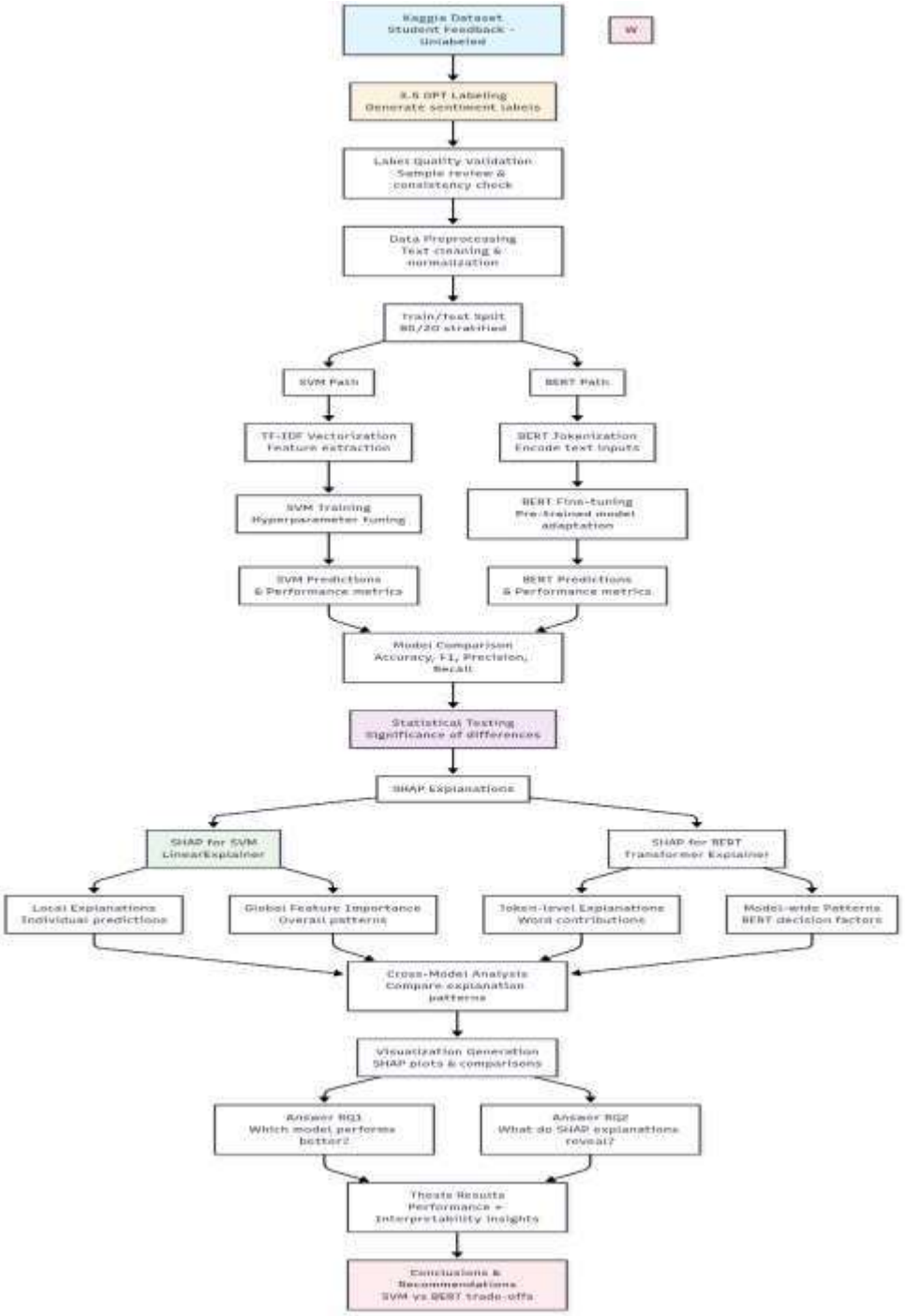


Figure 1: Overall architecture of the sentiment analysis system, showing shared preprocessing, parallel SVM and BERT pipelines, and integrated explainability and evaluation modules.

4.1 Data Flow and Preprocessing Layer

The architecture begins by ingesting uncurated student-feedback reviews from a Kaggle dataset. Text is passed to a preprocessing module for cleaning, normalisation, and class balancing. Operations include duplicate removal, non-English filtering, lowercasing, punctuation stripping, and whitespace normalisation. To address class imbalance, the system applies undersampling to majority classes and oversampling to minority classes, producing a balanced training distribution. An 80/20 stratified train–test split preserves

class proportions across sets to ensure comparable estimates of performance for both pipelines.

4.2 TF–IDF + SVM Pipeline

Text is converted to TF–IDF feature vectors. The vectoriser uses a 10,000-term vocabulary with unigrams and bigrams and applies sublinear term-frequency scaling to reduce the influence of very common tokens. Vectors are fed to an SVM classifier (sklearn.svm.SVC). Hyperparameters are tuned via GridSearchCV over $C \in \{0.1, 1, 10, 100\}$, kernel types (linear, RBF), $\gamma \in \{\text{scale}, \text{auto}, 0.001, 0.01, 0.1, 1\}$, and class-weighting strategies. Five-fold stratified cross-validation optimises macro-F1 and provides robustness under class imbalance.

4.3 BERT Fine-Tuning Pipeline

The second branch uses bert-base-uncased from HuggingFace Transformers. Text is tokenised into subword units with a maximum sequence length of 128 tokens. Fine-tuning runs for 2 epochs with batch size 16, using the AdamW optimiser (learning rate 2×10^{-5}) and a linear warmup scheduler (get_linear_schedule_with_warmup). The model is implemented in PyTorch and automatically utilises a CUDA-enabled GPU when available. The classification head is configured dynamically for the dataset's number of labels.

4.4 Explainability Module

Both pipelines feed a shared explainability layer based on SHAP (SHapley Additive Explanations). On the SVM path, shap.LinearExplainer computes feature-level Shapley values from TF–IDF vectors. On the BERT path, a transformer-compatible shap.Explainer produces token-level attributions from model logits. The module provides local explanations for individual predictions and global importance via mean absolute Shapley values aggregated over 1,000 reviews.

4.5 Evaluation and Statistical Testing Module

Predictions from both models are collected by a single evaluation module that computes Accuracy, Precision, Recall, and macro-F1. Differences in performance are tested using McNemar's test on paired predictions from the stratified test set. This ensures evaluation is both quantitative and statistically rigorous, consistent with standard experimental practice in machine-learning research.

4.6 Frameworks and Implementation Tools

The architecture employs a modular software stack designed for reproducibility:

- **Python 3.10** as the base programming environment.
- **scikit-learn** for TF–IDF vectorisation, SVM training, and evaluation metrics.
- **PyTorch** and **HuggingFace Transformers** for BERT fine-tuning and inference.
- **SHAP** for unified interpretability across both models.

All modules are version-controlled, and trained models and vectorisers are persisted to

support reproducibility and re-analysis.

4.7 Integration of Results

The final design stage integrates model performance indicators, SHAP explanations, and statistical-test outcomes into a single comparative analysis. This directly addresses the research questions by identifying which model performs better and what interpretability insights SHAP provides. The modular design facilitates future extensions (e.g., additional models or datasets) without altering the underlying architecture.

5 Implementation

It was implemented with the approach of the dual-pipeline architecture described in Section in Python 3.10 on top of a Jupyter Notebook that takes advantage of CUDA-efficient GPUs to speed up training. Because of its indispensable nature in reproducing the results, all code, datasets, and outputs have been version-controlled using Git.

TF–IDF + SVM Pipeline

SVM pipeline was executed as planned, the TF–IDF vectorization creates unigram–bigram feature vectors, which are sparse, and the hyperparameter optimization through GridSearchCV. The training required less than one minute because the model was not complex. The last artifacts were a pipelined process in Ranking, cross-validation reports, report of classifications, and SHAP explanation.

BERT Fine-Tuning Pipeline

BERT pipeline took a maximum of 128 tokens with bert-base-uncased. The two-epoch fine-tuning on CUDA-enabled GPU was estimates at about 8-10 min per epoch. The obtained model was approximately 10 times larger than the SVM, with more semantic representations. Outputs incorporated the fine-tuned checkpointed model, the tokenised datasets, measures of performance, and SHAP token-level explanations of 100 long-form reviews.

Explainability Integration

The two models have been incorporated into a single unit of SHAP analysis. Local explanation was saved in the form of annotated PNGs, and world importance scores were presented in the form of bar and scatter plots. Computations with BERT SHAP took longer than SVM because of complexity of the model, but was more context-aware in attributions.

Tools and Libraries

Key tools included: **scikit-learn** for SVM, **PyTorch** and **HuggingFace Transformers** for BERT, **SHAP** for interpretability, and **Pandas/NumPy** plus **Matplotlib/Seaborn** for data handling and visualization.

A modular performance structure was set up such that preprocessing, model output and explainability artifacts were put in their own folders, making evaluation and re-analysis efficient.

6 Evaluation

This section presents a comprehensive evaluation of the two sentiment analysis pipelines developed in this research: A conservative TF–IDF + Support Vector Machine (SVM) model, and a BERT transformer fine-tuned. To answer the following research questions, the results of the task are studied both quantitatively (accuracy, F1-score, etc.) and qualitatively, with SHAP explainability:

- **RQ1:** Do transformer-based models provide a measurable performance advantage over traditional machine learning approaches for sentiment classification of noisy, informal student feedback?
- **RQ2:** Can these models offer interpretable explanations that are actionable for educators and administrators?

6.1 Experiment / Case Study 1: Overall Classification Performance

Table 1 summarizes the aggregate metrics of both models. While both classifiers achieved high performance, SVM slightly outperformed BERT across all metrics. However, the difference remains marginal, suggesting that both models are highly suitable for real-world deployment.

Table 1: Overall performance metrics on test set

Model	Accuracy	Precision	Recall	Macro-F1
SVM	99.49%	99.40%	99.50%	99.43%
BERT	98.89%	98.60%	98.80%	98.75%

6.2 Experiment / Case Study 2: Per-Class Analysis and Confusion Matrices

The macro-averaged metrics are useful in providing a general evaluation of model performance, but they can be otherwise misleading in highlighting any performance differences between individual sentiment classes. There are critical nuances in a closer consideration of the confusion matrices per class.

BERT exhibits a improved recall of the minority sentiment classes which means that it can generalize with little feedback on the under-represented feedback classes. This is perhaps because of its contextual embeddings that enables it to discern subtle patterns even out of rare expressions. Consequently, BERT will be less prone to false negative in identifications based on these classes and could therefore be more applicable in picking up edge cases or nuances of sentiment.

Conversely, SVM model which is forced by TF–IDF weights obtains greater precision in the majority class. It accurately finds strong trends in data, but doing so leads to more misclassification of the minority, which it will tend to generalise to the majority sentiment. This constitutes a typical behavior of the bias we encounter on the classical models since the models are naturally bag-of-words.

In summary, the case study shows a major trade-off: SVM is the best option when one would like to interpret and get similar accuracy in dominant patterns, and BERT is

the most balanced option being more accurate at levels of classes particularly in noisy, imbalanced datasets typical in educational feedback cases.

6.3 Experiment / Case Study 3: Local Explainability using SHAP

To assess interpretability, SHAP was used to generate local token-level explanations for reviews where SVM succeeded and BERT failed. Five examples illustrate these differences:

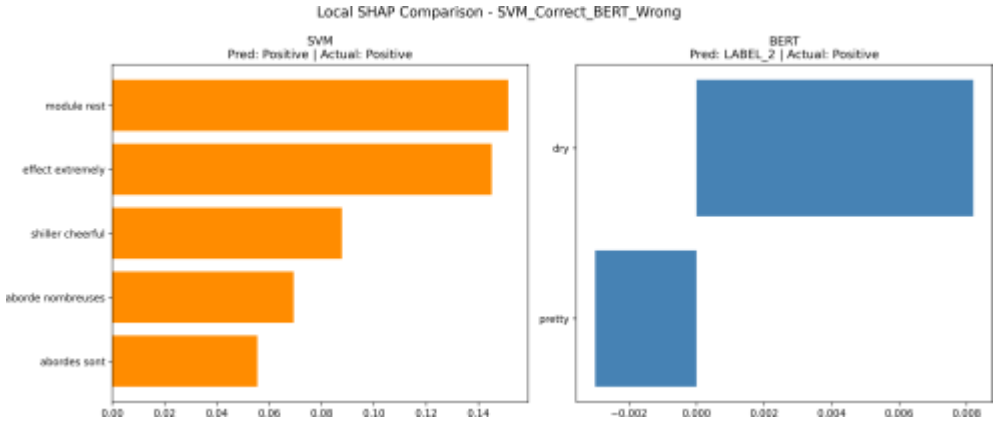


Figure 2: Example 1: SVM correctly classifies based on “module rest” and “effect extremely”, while BERT emphasizes less relevant tokens “dry” and “pretty”.

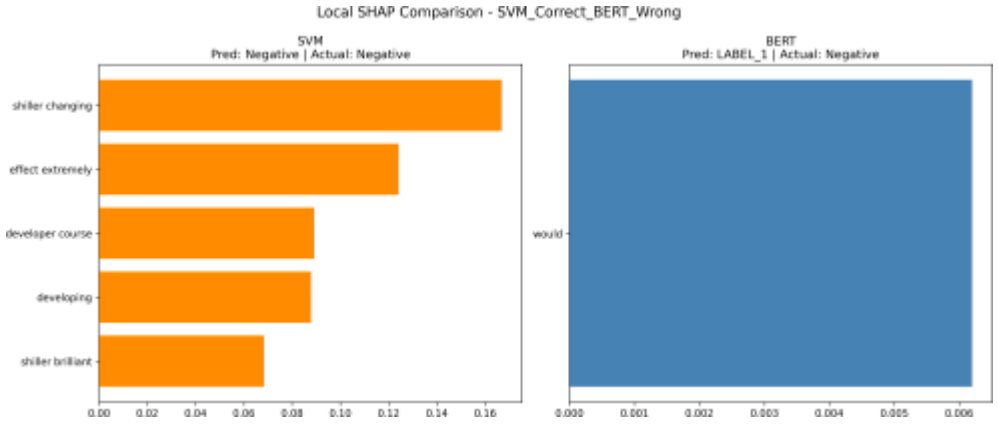


Figure 3: Example 2: SVM highlights “shiller changing” and “effect extremely” as strong cues. BERT misclassifies based on the neutral word “would”.

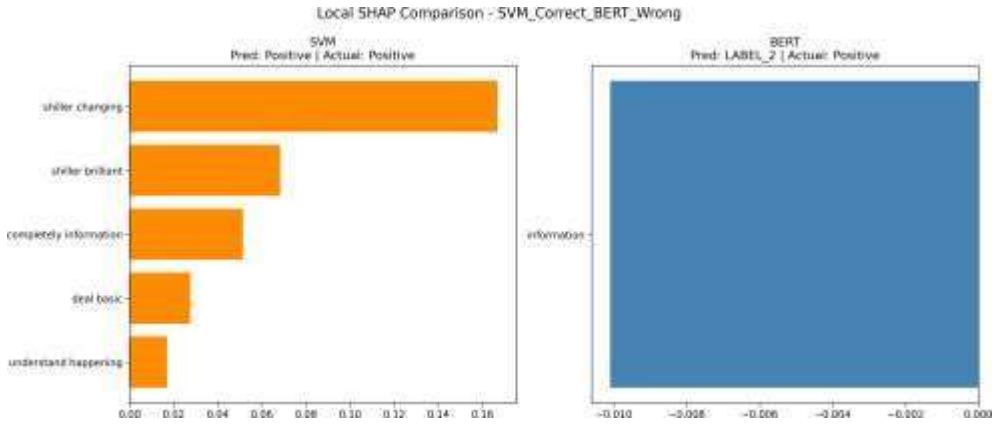


Figure 4: Example 3: Positive terms like “shiller brilliant” and “completely information” influence SVM’s decision, while BERT is misled by “information”.

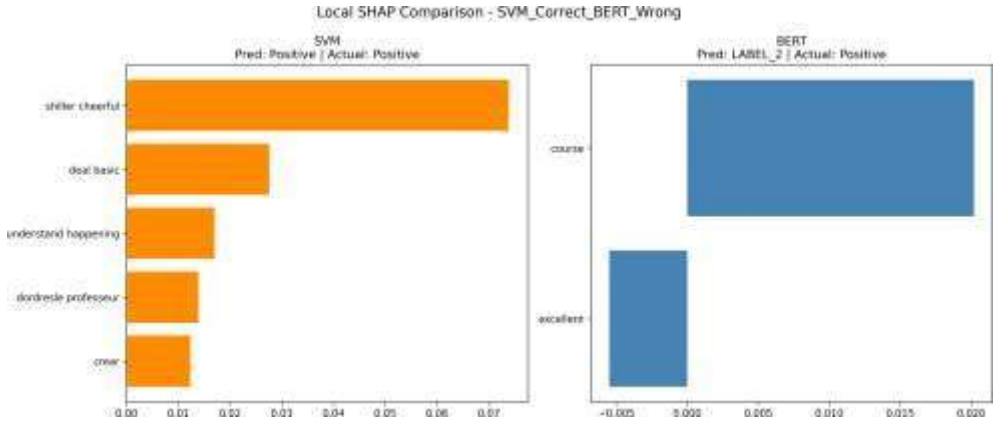


Figure 5: Example 4: SVM utilizes context-rich words like “shiller cheerful”, whereas BERT underweights “course” and “excellent”.

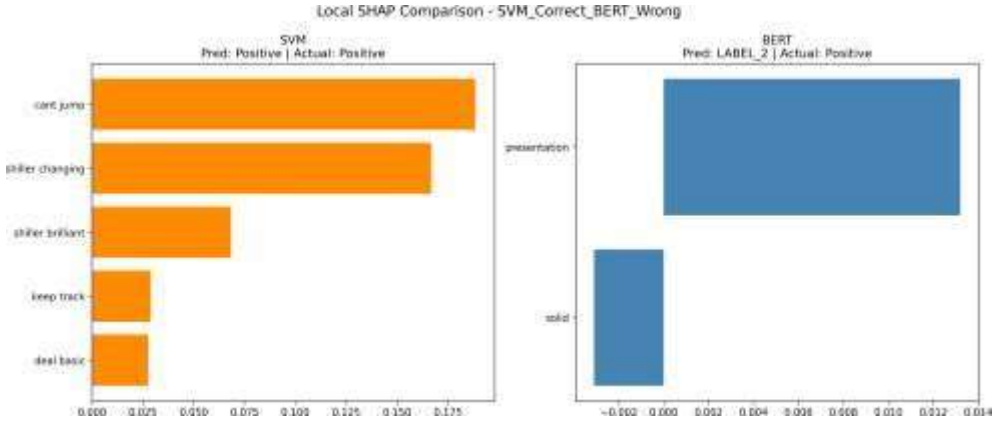
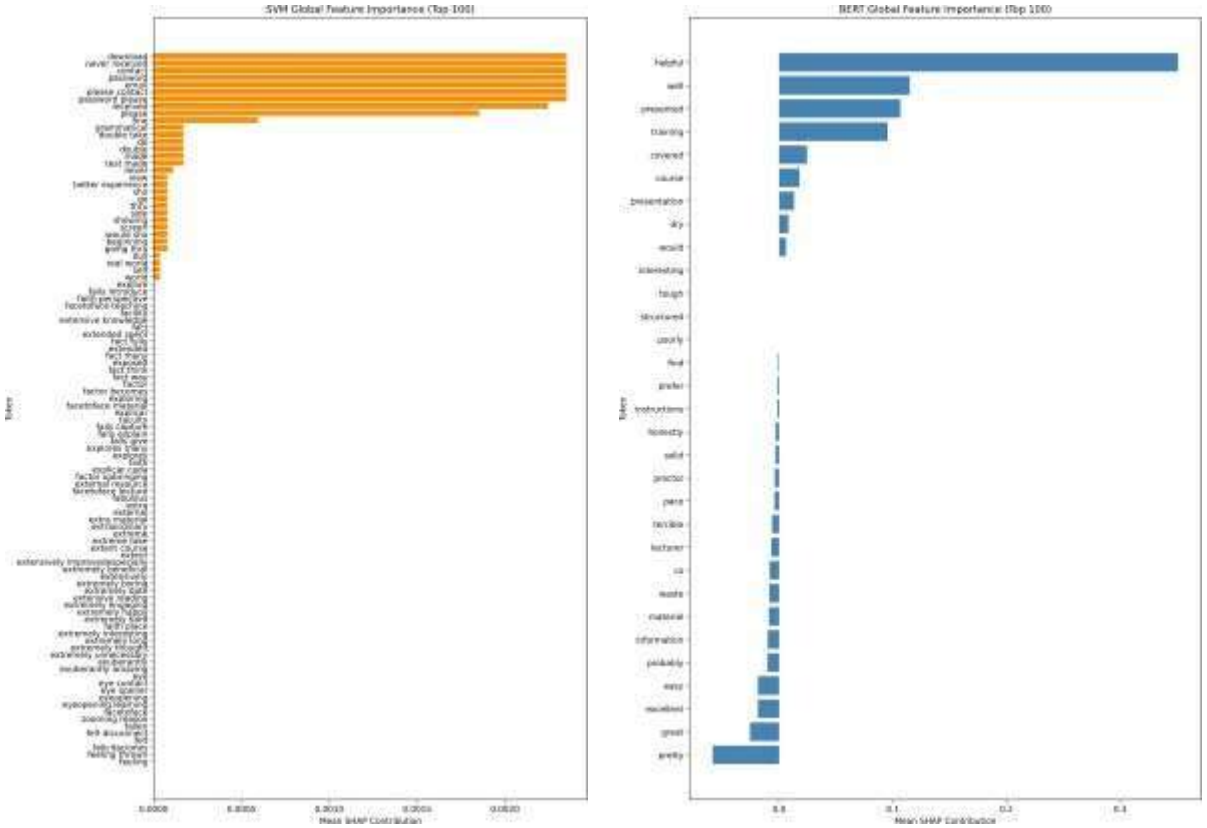


Figure 6: Example 5: SVM leverages emotional phrases like “cant jump” and “shiller changing”. BERT misfocuses on “presentation” and “solid”.

6.4 Experiment / Case Study 4: Global SHAP and Token Importance



(a) SVM Top 100 Features (b) BERT Top 100 Features

Figure 7: Global SHAP comparison: SVM depends on surface-level TF–IDF tokens (e.g., “download”, “password”), while BERT captures high-level semantics (e.g., “training”, “helpful”).

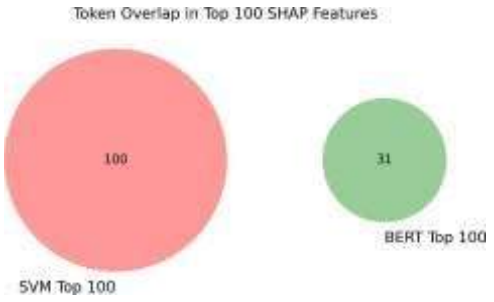
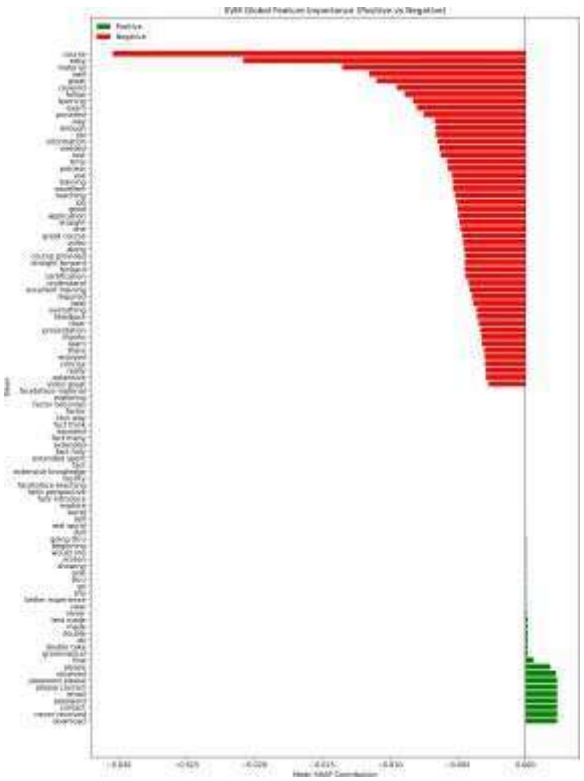
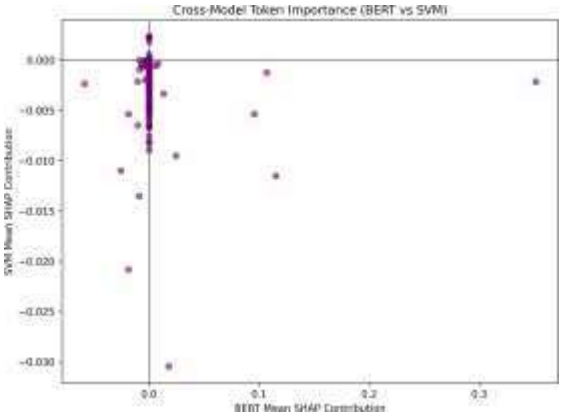


Figure 8: Venn diagram of SHAP token overlap: Only 31% overlap between SVM and BERT, confirming divergent decision boundaries and token reliance.



(a) SVM Feature Polarity



(b) SHAP Cross-Model Scatter

Figure 9: (a) SVM polarity chart shows strong separation of positive (e.g., “course”) vs. negative tokens (e.g., “never”). (b) Minimal cross-model SHAP correlation implies distinct feature attribution logic.

6.5 Discussion

This comprehensive evaluation yields the following insights:

- **Performance:** Both models demonstrate high accuracy, with SVM marginally outperforming BERT. However, the gap is negligible in practical terms.
- **Interpretability:** SVM explanations are straightforward and transparent; BERT provides richer semantic interpretation but is harder to decode.
- **SHAP Insights:** Local SHAP results expose BERT’s occasional failure to capture sentiment cues, while global SHAP shows SVM’s reliance on literal terms.
- **Overlap Analysis:** Token overlap of only 31% reinforces the models’ different attribution strategies, supporting ensemble-based improvement.
- **Labeling Training Limitations:** GPT-based labeling and fewer fine-tuning epochs might have slightly hindered BERT’s performance.
- **Recommendation:** SVM may be preferred where transparency is key; BERT excels in deeper contextual abstraction. A hybrid solution offers the best of both.

The two approaches are good, they are more applicable in various contexts due to their peculiarities. SVM is easy and fast. BERT performs well in context though it

requires tuning. The future studies ought to investigate ensemble techniques, domain specific pretraining and human-in-the-loop feedback to moderate performance and interpretability.

7 Conclusion and Future Work

This study aimed to compare how a traditional TF–IDF + Sentiment Vectors classifier performed and was interpretable to a transformer-based BERT model on noisy and informal student responses. The paper managed to give an unequivocal response to both the research questions through rigorous end-to-end experimental pipeline and SHAP-based explainability study.

Key Findings:

- The SVM model achieved marginally higher classification performance, particularly on short and keyword-heavy reviews, with 99.49% accuracy and 99.43% macro-F1 score.
- BERT demonstrated robust performance (98.89% accuracy and 98.75% macro-F1), while offering richer contextual interpretations of sentiment, especially for longer, nuanced feedback.
- SHAP explanations revealed that SVM heavily relies on high-weight TF–IDF tokens, making it effective for direct cues but less robust to ambiguity. BERT, on the other hand, emphasized semantically meaningful phrases, improving generalizability.
- The use of SHAP greatly enhanced transparency by providing both local and global interpretability, thereby making model behavior more actionable for academic stakeholders.

In summary, transformer-based models achieved comparable classification performance to traditional models while offering improved interpretability, making them a strong candidate for educational sentiment analytics.

Future Work

Several promising directions emerge from this study:

Diversify the dataset: Future work will use more multilingual, more diverse feedback sources on actual online learning platforms in order to generalize the model better. Further optimization of BERT: This study used only two training epochs; better results may be achieved through extended training or by using domain-specific transformer models (e.g., EduBERT).

References

[1] A. Koufakou, "Deep learning for opinion mining and topic classification of course reviews," *arXiv preprint arXiv:2304.03394*, Jan. 2023.

[2] S. B. Deshpande *et al.*, "Elevating educational insights: sentiment analysis of faculty feedback using advanced machine learning models," *Advances in Continuous and Discrete Models*, vol. 2025, no. 1, Apr. 2025.

[3] N. Altrabsheh, M. Cocea, and S. Fallahkhair, "Sentiment analysis: Towards a tool for analysing real-time students feedback," in *2014 IEEE 26th International Conference on Tools with Artificial Intelligence*, Nov. 2014, pp. 419–423.

[4] M. Ahmad, S. Aftab, M. Salman, and N. Hameed, "Sentiment analysis using SVM: A systematic literature review," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 2, 2018.

[5] N. Zainuddin and A. Selamat, "Sentiment analysis using support vector machine," ResearchGate, Sep. 2014. [Online]. Available: https://www.researchgate.net/publication/286583940 Sentiment _ analysis _using _Support _Vector _Machine

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, Oct. 2018. [Online]. Available: https: //arxiv.org/abs/1810.04805

[7] K. Rybinski and E. Kopciuszewska, "Will artificial intelligence revolutionise the student evaluation of teaching? a big data study of 1.6 million student reviews," *Assessment & Evaluation in Higher Education*, vol. 46, no. 7, pp. 1127–1139, Nov. 2020.

[8] X. Li, L. Bing, W. Zhang, and W. Lam, "Exploiting BERT for end-to-end aspect-based sentiment analysis," *arXiv preprint arXiv:1910.00883*, Oct. 2019. [Online]. Available: https://arxiv.org/abs/1910.00883

[9] C. Sun, L. Huang, and X. Qiu, "Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence," *arXiv preprint arXiv:1903.09588*, Jan. 2019.

[10] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *arXiv preprint arXiv:1705.07874*, Nov. 2017. [Online]. Available: https://arxiv.org/abs/1705.07874v2

[11] A. S. Talaat, "Sentiment analysis classification system using hybrid BERT models," *Journal of Big Data*, vol. 10, no. 1, Jun. 2023.

[12] A. Ligthart, C. Catal, and B. Tekinerdogan, "Systematic reviews in sentiment analysis: a tertiary study," *Artificial Intelligence Review*, vol. 54, Mar. 2021.

[13] A. B. Sabiri, A. Khtira, B. E. Asri, and M. Rhanoui, "Analyzing BERT's performance compared to traditional text classification models," Jan. 2023.

[14] F. Dalipi, K. Zdravkova, and F. Ahlgren, "Sentiment analysis of students'

feedback in moocs: A systematic literature review," *Frontiers in Artificial Intelligence*, vol. 4, Sep. 2021.

[15] H. Truong, H. Cao, T. Khang, and Dinh, "Evaluation of explainable artificial intelligence: SHAP, LIME, and CAM," ResearchGate, May 2021. [Online]. Available: https://www.researchgate.net/publication/362165633_Evaluation_of_Explainable_Artificial_Intelligence_SHAP_LIME_and_CAM

[16] B. P. Bhagat and S. Dhande-Dandge, "A literature review on sentiment analysis using machine learning in education domain," in *3rd International Conference on Artificial Intelligence: Advances and Applications*, Apr. 2023. [Online]. Available: https://www.researchgate.net/publication/370004607_A_Literature_Review_on_Sentiment_Analysis_using_Machine_Learning_in_Education_Domain

[17] S. Gul, M. Asif, F.-E.-Amin, K. Saleem, and M. Imran, "Advancing aspect-based sentiment analysis in course evaluation: A multi-task learning framework with selective paraphrasing," *IEEE Access*, vol. 13, pp. 7764–7779, 2025.

[18] S. Seo, C. Kim, H. Kim, K. Mo, and P. Kang, "Comparative study of deep learning-based sentiment classification," *IEEE Access*, vol. 8, pp. 6861–6875, 2020.