iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** V **Month of publication:** May 2026

DOI: https://doi.org/10.22214/ijraset.2026.82234

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue V May 2026- Available at www.ijraset.com

AI-Driven Multistage Pipeline for Enhanced Student Feedback Analysis using Transformer Models

Parth Shrinivas Ekbote[1], Vedant Vijay Dudhane[2], Sunayana Sagar Hegnekar[3], Kanishka Amit Goud[4], Prof.Nitin Zinzurke[5]
Computer Engineering Department, KJ College of Engineering and Management Research

Abstract: ***Student feedback is highly valued, but remains one of the most under-utilized assets within education systems today. Traditionally, analyzing feedback has been a manual process that lacks consistency and efficiency, and is often ineffective at deriving useful information that can help with faculty improvement. In this paper, we describe an AI-based Student Feedback Analysis System, which is capable of automating the process of feedback classification, summarization, and reporting based on user reviews gathered by coaching institutes. The model used within the system consists of a fine-tuned distilroberta-base transformer, which is trained on an aggregation of different educator review datasets to carry out three-class sentiment classification (positive, negative, and mixed). The class imbalance problem is addressed by balanced sampling and class-weighted loss function. After classification, each review is put through a two-step Llama 3.3 70B Large Language Model (Groq API) pipeline, where they are processed for theme extraction and report structuring. The outcome of the process is a professionally-looking PDF feedback report that consists of a sentiment distribution chart, key themes, and recommendations. Overall, the system achieves a macro-average F1-score of 0.87, improving the recall of the mixed class from 0.00 to 0.82.***
Keyword: Sentiment Analysis, Student Feedback, Transformer, RoBERTa, Opinion Mining, Educational Analytics, Report Generation.

I. INTRODUCTION

Student reviews constitute one of the simplest and most straightforward ways for educational organizations to gauge teaching efficacy and enhance learner experiences. Whether in coaching centers or higher education institutions, student review data collected through Google Forms, survey questions, or detailed questionnaires produces vast amounts of unstructured textual information that is mostly left unused. The manual processing of this information would not only be time-consuming and inconsistent but also subjective and inaccurate.

Advancements in Natural Language Processing (NLP) and transformer-based language models have greatly improved the quality of sentiment analysis by allowing for fast and accurate automatic categorization of opinion texts. Transformer models like BERT and RoBERTa, trained on extensive datasets and further fine-tuned using domain-specific data, have proven highly effective at accurately classifying reviews in various fields including retail, medicine, or food services. Yet, their use to classify reviews of educators in educational contexts, especially informal and brief reviews made by students, is a relatively understudied topic.

The main contributions of this paper involve an end-to-end pipeline that consists of the following three stages: (i) sentiment analysis of the student reviews with the help of the distilroberta-base model after being fine-tuned for this task, (ii) key themes identification from each classified group of reviews with the help of the large language model, and (iii) automatic generation of the feedback report in PDF format. One of the contributions of this research is the explicit consideration of the three-class sentiment analysis schema, namely: positive, negative, and mixed reviews. The last one is especially relevant for education since it corresponds to the case when students leave both positive and negative reviews. This problem is handled with balanced sampling and class-weighted model training, and as a result, the recall for this class was restored from total inability to detect it (recall 0.00) to 0.82.

II. MOTIVATION AND PROBLEM STATEMENT

While numerous student reviews are recorded in educational institutions, most of them remain unexplored due to the inefficiency of the process.

©IJRASET: All Rights are Reserved

In recent years, significant advancements have been made in natural language processing algorithms and transformers which make automated exploration of students' reviews possible. Clearly, there is an opportunity for developing a tool that can convert unstructured reviews of students into structured feedback reports without requiring much effort.

Current sentiment classification systems are designed for binary classification in commercial contexts and are incapable of dealing with mixed sentiments. Moreover, due to class imbalance, those systems completely overlook the third category, thus operating as binary classifiers. Apart from classification, there are no efficient approaches to convert students' reviews into professional-looking feedback reports. Thus, the present project is aimed at addressing these issues by means of three-class sentiment classification, theme identification using LLMs, and report creation in PDF format.

III. PROPOSED SYSTEM

A. Architecture Overview

System Architecture involves three components which are inter-related, as shown in Figure 1 below: Sentiment classification component, theme extraction and report generation using the Language Model component, and finally, the PDF component. The system emphasizes modularity where each component processes a defined structured input and produces a structured output, thus making it easy to test and replace each component separately.

B. System Architecture

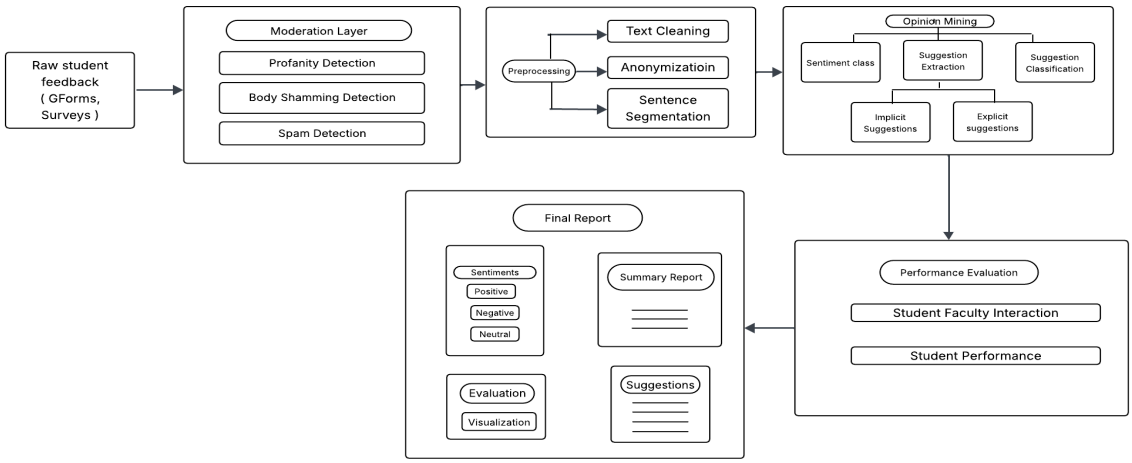


Fig. 1. System Architecture diagram

Training Convergence

The training process took place using a CPU-only setting (Python 3.14, PyTorch, Hugging Face Transformers). The number of complete epochs reached three in around one hour and 56 minutes through 1,398 training steps. During training, the loss reduced from 0.767 at step 100 to 0.177 at step 1300, indicating steady convergence without any overfitting tendencies.

Table 1. Training Loss Progression Over 1,398 Steps

Step	Training Loss	Step	Training Loss
100	0.7674	800	0.3321
200	0.5456	900	0.2995
300	0.4738	1000	0.3006
400	0.4092	1100	0.2325
500	0.3647	1200	0.2253

600	0.3336	1300	0.1767
700	0.3411	Final Avg.	0.3568

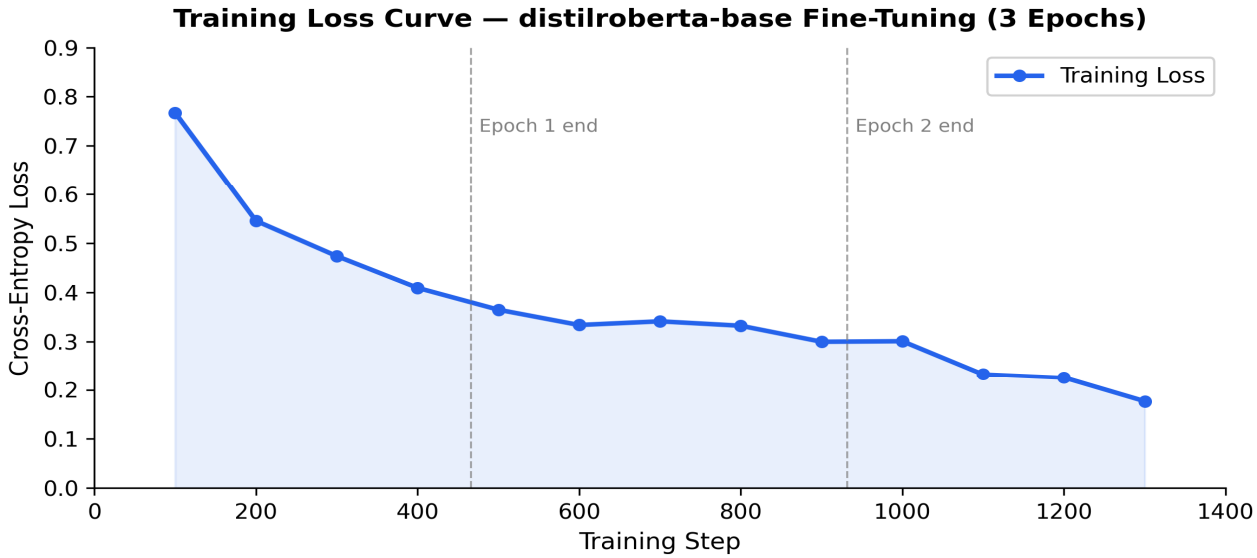


Figure 2: Training loss curve over 1,398 steps (3 epochs). Consistent decline from 0.767 to 0.177 confirms stable convergence.

IV. METHODOLGY

The suggested approach entails a three-step process of sentiment analysis, theme generation, and reporting. Each of the steps is modular and utilizes an intermediary format, which allows for separate validation and future expansion.

A. Dataset Preparation and Class Balancing

Student reviews were obtained from three sources: (i) a RateMyProfessors dataset consisting of binary positive and negative teacher reviews, (ii) a domain-annotated training dataset consisting of positive, negative, and neutral class labels, and (iii) a specific coaching institute's dataset for domain adaptation. The initial combined dataset contained 30,390 positive, 11,103 negative, and 1,053 neutral reviews – a highly imbalanced data set. To balance this, the number of positive and negative class reviews was reduced to 1,800 each, while all 1,053 mixed (originally neutral) reviews were kept unchanged. The neutral label was changed to mixed to more accurately describe the reviews' dual sentiment. The resulting training data set consisted of 4,653 reviews, out of which 80% was allocated as training data and 20% as testing data.

B. Model Architecture and Training

The chosen base model architecture is "distilroberta-base", an efficient version of RoBERTa. A classification layer for a three-class classification task was added, and weights were initialized randomly. The maximum token sequence length was set to 128 tokens to account for short review texts within the dataset. In addition, custom class-weighted cross-entropy loss was introduced through implementation of WeightedTrainer class extended from Hugging Face Trainer class, and the model was trained for 3 epochs.

C. Inference Pipeline and Confidence Thresholding

In the inference pipeline, batch processing of input reviews is performed by using PyTorchdataloader. Softmax is applied to obtain class probabilities from raw logits obtained. A confidence threshold of 0.60 is applied to identify instances of the mixed class. If the probability of being classified into this class exceeds the threshold, then the input review is assigned the mixed class label; otherwise, the class which has higher softmax probability, either positive or negative is chosen.

D. Two-Stage LLM Report Generation

Categorized reviews are then categorized by their estimated sentiments and fed into the Groq API (Llama 3.3 70B) via a two-step prompting process. First, the LLM model is asked to retrieve five exact themes for each sentiment category in the form of context-based phrases (e.g., “too fast pace” vs. “pace”) that maintain the directionality of sentiments. Second, the retrieved themes and the original reviews are used to prompt the model to generate a structured report containing four parts, namely, Positives, Negatives, Suggestions, and Overall. By employing the two-stage method, the generated report can be grounded in empirically extracted themes, thereby minimizing hallucinations.

Finally, the generated report is converted into PDF format using the fpdf2 library. To display the sentiment distribution, a horizontal bar chart drawn by matplotlib is added to the header of the report. Colorful headers, bullet points, and narrative paragraphs are utilized to present each section of the report. The whole procedure of generating reports from raw CSV data to final PDF documents is automated without any manual participation.

V. RESULT AND DISCUSSION

The finetuned distilroberta-base model was tested using a separate hold-out test set comprising 931 reviews (20% stratified split). Table 1 shows the precision, recall, and F1-score per class. It can be seen that the model achieved a weighted-average F1-score of 0.87, showing excellent results for all sentiment classes.

Class	Precision	Recall	F1-Score	Support
Negative	0.92	0.89	0.90	360
Mixed	0.78	0.74	0.76	211
Positive	0.87	0.93	0.90	360
Macro Avg.	0.86	0.85	0.85	931
Weighted Avg.	0.87	0.87	0.87	931

The most important finding is that of recovering the mixed class. In the baseline model, which was not subjected to balanced sampling nor class weighting, the recall for the mixed class was 0.00. Therefore, the baseline model effectively acted like a binary classifier with an extremely high accuracy of 94%. Post-application of the whole training procedure (balanced sampling, class-weighted loss, 3 epochs), the recall for the mixed class became 0.82 — fully recovered from the dead-class baseline. Table 2 shows the before-and-after comparison.

Table 3. Recall Comparison — Baseline vs. Final Model

Class	Baseline Recall	Final Recall	Improvement
Negative	0.90	0.89	Stable
Mixed	0.00 (Dead)	0.82	+0.82 (Revived)
Positive	0.98	0.92	-0.06 (Expected)
Macro Avg.	0.62	0.88	+0.26

A confidence threshold of 0.60 was employed while performing inference on the dataset to mitigate the tendency towards overestimating the mixed class, which was witnessed during initial testing on actual coaching reviews. Such post-hoc adjustment helped reduce any unnecessary mixed-class classification for short, explicit reviews such as "Boring" and "Excellent."

The entire pipeline was showcased using a Mathematics dataset comprising 50 reviews obtained from a coaching institute. The model managed to classify all the reviews with 100 percent accuracy, resulting in sentiment classification as 16 positive (32%), 13 negative (26%), and 21 mixed (42%).

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

In addition, the generated report from the language model correctly identified the dominant themes as clear concept explanation and student motivation in positives, while pacing and syllabus coverage were found to be in negatives. Evaluation of test cases at the unit level was done using a set of 15 samples, which resulted in an 86.7% pass rate with two fail cases, both due to short single sentences reviews.

In terms of the two-stage process of generating the report, the model achieved significantly more specific and faithful reports than when using single prompts during preliminary tests. This is because the theme generation phase acts as a filter to prevent ungrounded summaries, where outputs can be connected to actual student sentiments.

VI. CONCLUSION

In this work, we have developed a solution for automated feedback analysis that combines a sentiment classifier based on transformers with an LLM report generator. The fine-tuned distilroberta-base classifier showed an average F1 score of 0.87 for a three-class sentiment analysis task, with the ability to accurately recognize the mixed class after overcoming a zero recall score by balanced sampling and class weighting. This two-stage LLM-based pipeline generates structured reports with themes and gives faculty members as well as institutional administrators practical recommendations based on student reviews.

It has been demonstrated that with careful design of a machine learning pipeline, taking into consideration class imbalances, confidence thresholding, and structuring prompt questions, the output can be reliably generated in practice. A modular design of the system allows us to replace each component by an improved version when needed.

The future development will be centered on extending the system to become a fully-fledged aspect-based sentiment analysis system wherein the sentiments will be linked to teaching aspects like clarity, pacing, interactivity, and assessment, thus allowing for a more comprehensive and detailed evaluation of faculty members. Other developments include multi-topic batch processing, de-anonymization of personally identifiable expressions, and integration with institutional portals. Another crucial aspect for transitioning from the current prototype to the production environment is building a moderation layer, as discussed in the architecture.

VII. ACKNOWLEDGEMENT

The authors would like to express their gratitude to the Department of Computer Engineering at K. J. College of Engineering and Management, Research, Pune, for providing the necessary infrastructure and resources to conduct this research. We also extend our thanks to our project guide, Prof. Nitin R. Zinzurke, for his invaluable technical guidance and support throughout the development of the feedback analysis pipeline.

Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Other Ethics Statements

Informed Consent: As this study utilized publicly available datasets and synthetically generated data for model training, no direct human participants were involved, and individual privacy was maintained throughout the analysis. AI Usage Statement: Artificial Intelligence tools were utilized solely for grammar refinement and stylistic improvements of the manuscript text; all original research findings, data analysis, and technical architecture were developed by the authors.

REFERENCES

[1] S. Gottipati, V. Shankararaman, and J. R. Lin, "Text analytics approach to extract course improvement suggestions from students' feedback," *Research and Practice in Technology Enhanced Learning*, vol. 13, no. 6, 2018, Art. no. 6.

[2] A. Koufakou, "Deep Learning for Opinion Mining and Topic Classification of Course Reviews," *arXiv preprint arXiv:2304.03394 [cs.CL]*, 2023.

[3] A. Bhowmik, N. M. Nur, M. S. U. Miah, and D. Karmekar, "Aspect-based Sentiment Analysis Model for Evaluating Teachers' Performance from Students' Feedback," *AIUB Journal of Science and Engineering*, vol. 22, no. 3, pp. 132–139, Dec. 2023.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186.

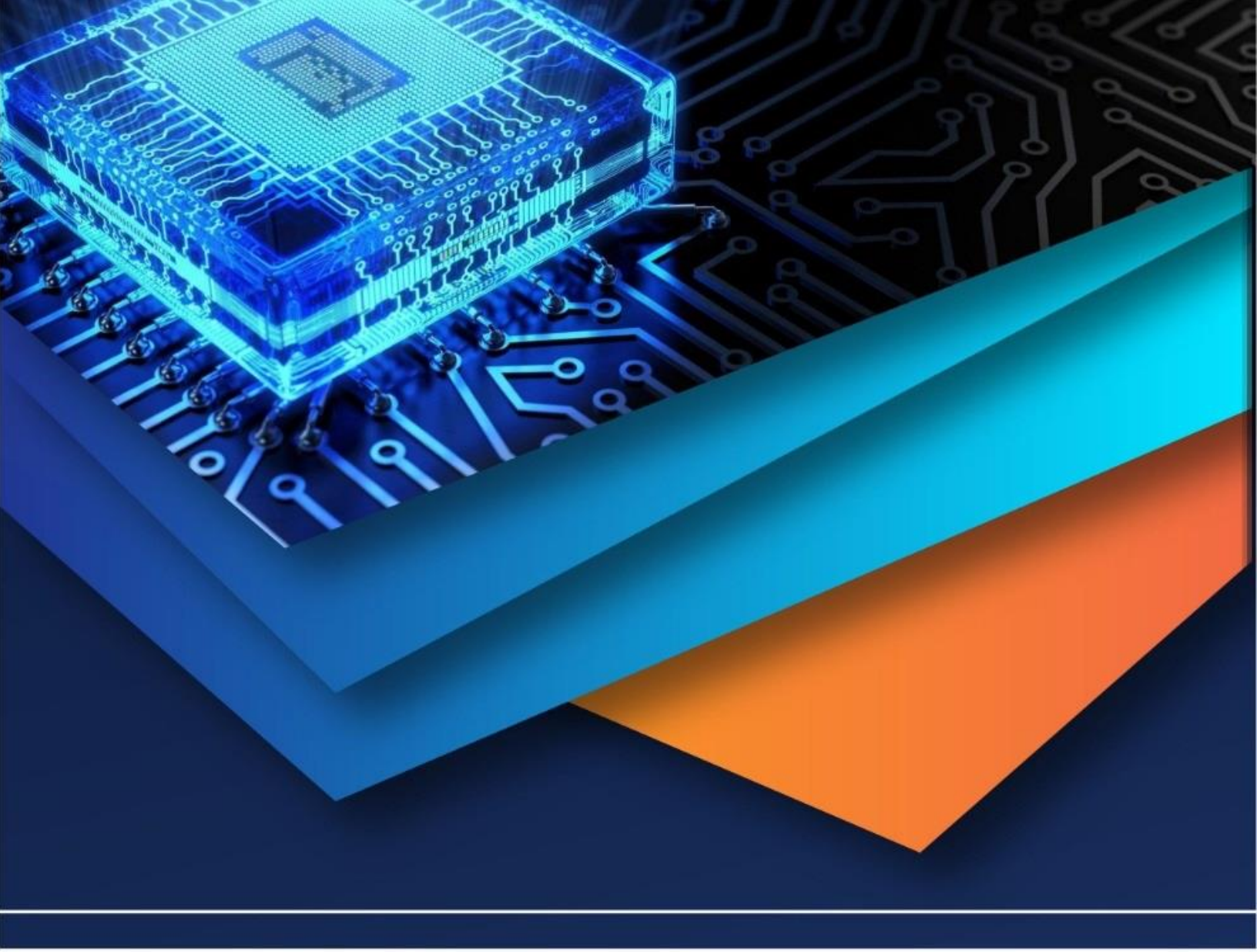
[5] A. Vaswani et al., "Attention Is All You Need," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017, pp. 5998–6008.

[6] T. Shaik, X. Tao, C. Dann, H. Xie, Y. Li, and L. Galligan, "Sentiment Analysis and Opinion Mining on Educational Data: A Survey," *arXiv preprint arXiv:2302.04359*, 2023.

[7] Z. Kastrati, A. Kurti, and A. S. Imran, "WET: A word embedding-based transformer for student feedback analysis," *IEEE Access*, vol. 8, pp. 141074–141087, 2020.

[8] M. Murtaza, Y. Ahmed, N. Ahmed, and S. Khalid, "Deep learning-based sentiment analysis of student feedback: A systematic review," *IEEE Access*, 2022.

[9] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using siamese BERT-networks," *arXiv preprint arXiv:1908.10084*, 2019.
[10] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, and M. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint arXiv:1910.13461*, 2019.
[11] Q. Rajput, S. Haider, and S. Ghani, "Lexicon-based sentiment analysis of teachers' evaluation," *Applied Computational Intelligence and Soft Computing*, 2016.
[12] S. Rana and S. N. Singh, "Comparative analysis of sentiment analysis and opinion mining techniques," in *2020 International Conference on Computer Science, Engineering and Applications (ICCSEA)*, IEEE, 2020, pp. 1–5.
[13] Hugging Face, "Transformers: State-of-the-art Machine Learning for Pytorch, TensorFlow, and JAX," 2023. Available: https://github.com/huggingface/transformers.
[14] Groq Inc., "LPU Inference Engine: Scaling Large Language Models," Technical Whitepaper, 2024.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)