Raj Kumar, 2026, 14:4
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

NLP-Based Sentiment Analysis Framework for Education Industry 4.0

Raj Kumar[1], Monti Saini[2], Shilpy Sharma[3]
Assistant Professor, Department of CSE, Quantum University, Roorkee, India

Abstract- Sentiment analysis is a powerful computational technique used to identify, extract, and interpret subjective information embedded within textual data. It focuses on recognizing and classifying opinions, attitudes, emotions, beliefs, and feelings expressed by individuals through written language. In the educational domain, sentiment analysis plays a crucial role in understanding student behaviour, engagement, motivation, and overall learning experiences. By analysing linguistic features such as word choice, sentence structure, contextual meaning, and sentiment lexicons, sentiment analysis systems can detect a wide range of emotional states, including positive, negative, neutral, and subtle affective expressions. Recent advancements in Natural Language Processing (NLP) have significantly improved sentiment analysis performance through the adoption of transformer-based architectures and Large Language Models (LLMs). These models are trained on massive datasets using deep neural networks, enabling them to capture contextual and semantic nuances more effectively than traditional machine learning approaches. Models such as BERT, RoBERTa, and ELECTRA have demonstrated remarkable success across various NLP tasks, including sentiment classification and emotion detection. However, the practical deployment of such models in educational settings often faces challenges related to limited labelled data, high computational costs, and resource constraints. To address these challenges, this project investigates the application of LLMs for student sentiment analysis within an Education 4.0 framework by examining three primary adaptation strategies: zero-shot learning, N-shot learning, and fine-tuning approaches. The proposed system analyses student feedback to assess sentiment polarity, emotional dimensions, and a composite Learning Quotient (LQ) that reflects engagement, comprehension, motivation, collaboration, and critical thinking. Experimental observations indicate that different adaptation strategies yield significantly varied performance outcomes, highlighting the importance of selecting appropriate modelling techniques based on available resources and application requirements. Overall, the results emphasize the strong potential of LLM-based sentiment analysis systems in enhancing data-driven decision-making, personalized learning, and adaptive educational environments despite existing resource limitations.

Keywords: Sentiment Analysis, Education, Student Behaviour, Natural Language Processing (NLP), Student Engagement.

I. INTRODUCTION

In recent years, sentiment analysis has become a major field of study due to its ability to extract and interpret emotions, opinions, and subjective information from text data [1], [2], and [3]. Sentiment analysis converts unstructured textual data into structured knowledge through the use of computational linguistics and natural language processing (NLP) techniques [4].Sentiment analysis research has grown significantly over the last ten years in the fields of social media analytics, business, education, and healthcare [3], [4], and [5]. Sentiment analysis is used in business applications to assess customer satisfaction and comprehend the behaviour of customers [3].

Sentiment analysis is used by researchers in education to gauge student feedback, gauge engagement, and assess the efficacy of instruction [31], [32], and [36].Analysing this data has become crucial for Education 4.0 because students are increasingly sharing their academic experiences on social media, digital platforms, and institutional surveys [40]. However, processing massive amounts of student-generated text is inefficient when using manual sentiment analysis. Lexicon-based and traditional machine learning techniques were largely used in early research [6], [7], and [11].Sentiment classification was enhanced by deep learning techniques like CNNs, RNNs, LSTMs, and GRUs [14], [17], and [18]. NLP performance has recently been greatly improved by transformer-based LLMs, such as BERT, GPT, RoBERTa, T5, and LLaMA [19], [20], [22], [23], and [25]. LLMs are excellent at few-shot and zero-shot learning, making tasks without big labelled datasets possible. [27, 28, 29].

LLMs have limitations despite their amazing capabilities:
Need a lot of processing power [20], [21]. rely on enormous datasets [19]demonstrate domain dependency problems in specialized domains, such as education [30].Limited interpretability (also known as "black box") [18] In educational research, acquiring large, high-quality labelled datasets is difficult due to the cost and subjectivity of annotation [31], [36], [38]. Model performance is further diminished by data imbalance [14]. In order to overcome these constraints, this study looks at how LLMs can be modified using scant domain-specific data by:

- Zero-shot learning [27]
- N-shot learning [28]
- Fine-tuning [23], [25]

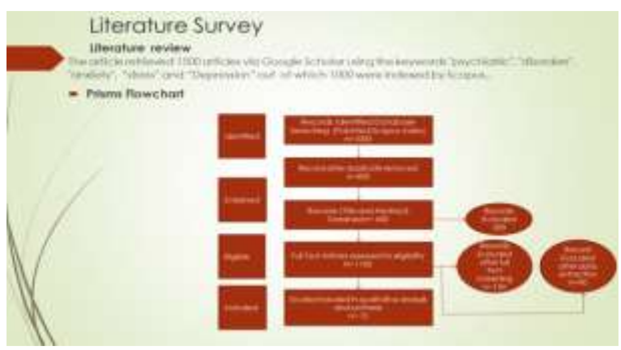
The goal is to develop sentiment analysis models for student feedback in higher education that are more precise and comprehensible. The impact of small, domain-specific datasets on LLM performance in educational sentiment analysis is examined in this study. It divides sentiment into positive, negative, and neutral categories using student comments from a Quantum University physics course. Aspect-based sentiment analysis is not included in the study, which only concentrates on overall polarity detection [15], [33]. In educational institutions with restricted finances, where it is impossible to train such massive models as GPT-3 or PaLM, this work will offer practical solutions to work with advanced LLMs [20], [21]. It provides institutions with examples of practical AI in understanding sentiments and features of adaptation such as N-shot learning and fine-tuning based on open-source and smaller models such as ELECTRA [26].

Research Questions
RQ1: To what extent can pre-trained general-purpose LLMs be successfully used to do zero-shot sentiment analysis on student-generated feedback? [27], [28].

RQ2: How can the student sentiment analysis be effectively trained using limited labelled data and limited computational budgets on LLMs? [23], [26].
The sample used is based on one course and one institution, which restricts its generalizability [31]. The sentiment is confined to polarity, but not intensity or particular emotions [7]. Subjectivity of annotation can cause bias [8], [38].

II. LITERATURE REVIEW

The chapter is a review and summary of the major concepts and basis of sentiment analysis in the field of Natural Language Processing (NLP) [1], [2]. It discusses different levels of analysis document, sentence, phrase, and aspect-level, their benefits, and restrictions [5], [7]. It also presents the latest sentiment analysis methods such as lexicon-based, machine-learning-based, hybrid and LLM-based methods [6], [11], [13], [25]. Lastly, it provides the analysis of the use of these methods in the sphere of education, particularly, the analysis of student sentiment [31], [32], [36].



Overview of Sentiment Analysis Research
Sentiment can be understood as the evaluative dimension of an individual's subjective experience reflecting positive, negative, or neutral attitudes [2], [20]. Sentiment analysis, also called opinion mining, identifies and classifies these sentiments in text [3], [5].

Research identifies three core aspects of sentiment analysis:

- Polarity classification (positive, negative, neutral) [3], [22]
- Intensity estimation (weakly or strongly expressive sentiment) [8]
- Subjectivity detection (factual vs. opinionated text) [8]

Sentiment analysis can be conducted at different text granularities:

- Document-level sentiment captures the overall sentiment of an entire text [2], [5]
- Sentence-level detects sentiment sentence by sentence [7], [28]
- Phrase-level analyses sentiment-bearing phrases [30]
- Aspect-level focuses on sentiment toward specific attributes or entities [15], [33]

Aspect-level sentiment analysis has become crucial for complex entities (e.g., courses, instructors), as different aspects may have different sentiments [34], [35].

Recent advancements extend beyond polarity classification, identifying complex emotions such as joy, anger, and frustration [18], [36], [37].

II. SENTIMENT ANALYSIS TECHNIQUES

Lexicon-based Approaches

Lexicon-based approaches are based on sentiment dictionaries like SentiWordNet [6], opinion lexicons [43] and domain specific resources [44]. They are efficient in computation and interpretable [45]. Nonetheless, they have problems with the issues of domain dependence, contextual polarity and vocabulary coverage [40], [48], [49].

Machine Learning (ML) Approaches.

Sentiment classifiers are trained by applying machine learning typologies, which include Naive Bayes, kNN, and SVM [11], [50], [51]. They need labeled datasets and manual feature engineering, which limits their performance in scenarios where text is domain-specific or when training data is limited [12].

Deep Learning Approaches

CNNs, LSTMs, GRUs, and RNNs are deep learning architectures that are effective in high-order linguistic regularities and long-term dependencies [13], [14], [17], [18]. They are more competitive compared to the traditional ML, yet they need huge annotated datasets and GPUs [55], [58]. Computational cost is also elevated by the use of GPUs [63], which are not favourable to institutions with limited resources.

Large Language Models (LLMs)

Transformer-based LLMs such as BERT [19], GPT-3 [20], GPT-4 [65], PaLM [21], and LLaMA [22] achieve state-of-the-art performance in NLP tasks. They are also good at zero-shot and few-shot learning and can do things without being trained on the tasks at hand [27], [28], [68].

BART is also auto-encodered to generate text and increase its robustness [24]. RoBERTa is an enhancement of BERT whereby it is optimized [25].
ELECTRA employs the highly-optimized masked-token discrimination to minimize the cost of computation but maintain accuracy [26], [99].

LLMs face challenges:

- high training cost [20], [21]
- limited interpretability ("black box") [69]
- domain mismatch for specialized fields such as education [14], [15]
- ethical concerns related to data sourcing and privacy [12]

Hybrid Approaches

Hybrid approaches involve the use of lexicon and ML-based approaches to capitalize on the advantages of both [70]. Examples include:

- Machine learning classification + lexicon scoring [71]
- Fuzzy logic + sentiment lexicon integration [72]
- Lexicon-enhanced classification for corporate or social responsibility documents [73]

Hybrid models improve accuracy and interpretability but still require careful lexical design and labelled data [49].

Sentiment Analysis in Education

Assessment of students plays a vital role in facilitating the quality of teaching and curriculum development [74]. Sentiment analysis is useful in deriving insights as per student opinions given in terms of course evaluations, online platforms, feedback form and social media [75].

Lexicon-Based Approaches

Lexicon scoring was used by Rajput et al. to establish the polarity of student ratings [82].

Machine Learning Approaches

Lalata et al. employed a combination of NB, SVM, LR, DT and RF to judge the sentiment of teacher evaluation, which enhanced accuracy by a large margin [83].

Deep Learning Approaches

Onan applied CNN, RNN, GRU, and LSTM, and the last one was the most successful in the analysis of 66,000 MOOC reviews [7].

LLM-Based Approaches

Recent studies use BERT-based models:

- Ngoc et al. used BERT to analyse Course era student sentiments [84].
- Dyulicheva & Bilashova used BERT + clustering for Udemy reviews [85].

These studies show students strongly express expectations around course structure, instructor quality, and learning outcomes. However, student feedback often contains nuanced or ambiguous wording, posing challenges for classification models [8], [14], [18].

III. RESEARCH METHODOLOGY

Data

The written comments from students enrolled in Quantum University's introductory physics course during the spring 2023–2024 semesters are the source of the textual data. The original purpose of gathering this dataset was to comprehend the lab experiences of students. In addition to a 6-point Likert satisfaction scale, students submitted open-ended written feedback [31], [32].

Data Sources

The data is acquired at: Student Reflection Forms (open-ended replies, course review, and student feedback) Learning Management System (LMS) Logs (engaging time, attempted quizzes, assignment submission behaviour pattern) Data (discussion forums, chat transcripts, doubt queries) in the classroom. Academic Records (internal assessment scores are applied in the validation of LQ)

Ethical Concerns

The study was reviewed by the Institutional Review Board (IRB) (IRB-2024-815) and found to be in violation of federal regulations for human subject's research. To preserve student privacy, the dataset used in this study was completely de-identified [38], [40].

Data Annotation

Each feedback entry was manually annotated as neutral, negative, or positive. Likert scale alignment, emotional tone, and contextual meaning were among the criteria. To increase consistency, established labelling standards like those outlined in [86]–[89] were cited. Benchmark sentiment datasets, including the Stanford Sentiment Treebank [92], the Movie Review Dataset [90], and the IMDB Review Dataset [91], were cross-checked against annotated examples.

Data pre-processing

There were 2,352 entries in the original dataset. There were 2,041 valid entries left after null, incredibly brief, and irrelevant responses were eliminated. 50.3% of the entries were positive, 18.1% were neutral, and 31.6% were negative, indicating class imbalance [14], [31]. Back-translation (English → Hindi → English) was employed as an augmentation technique to address imbalance [93], [94]. English–Hindi was translated using T5-base, and Hindi–English was translated using Bert2Bert [95], [96].

The dataset had 3,081 entries after augmentation, split equally among the three classes (1,027 each).
Text normalization, tokenization using model-specific tokenizes, padding and truncation, attention mask creation, and batch encoding for GPU

efficiency were among the pre-processing procedures.

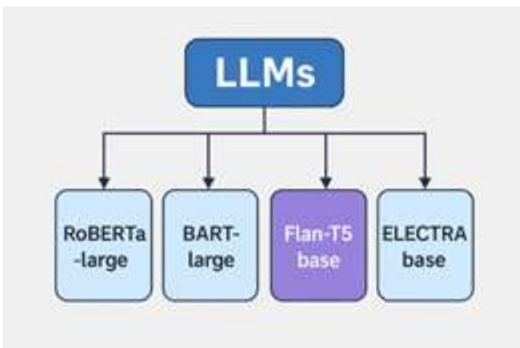
These procedures adhere to Transformer-based NLP pipelines' recommended best practices [19], [25], and [26].

IV. METHODS

This study investigates:

- **RQ1:** Zero-shot LLM performance
- **RQ2:** Efficient adaptation in low-resource environments via N-shot and fine-tuning strategies [23], [26], [27].

Four LLMs were selected:



ELECTRA was selected for fine-tuning and N-shot learning due to its low parameter count (67M) and high efficiency [26], [99].

Model Descriptions

RoBERTa-large

An optimized BERT model trained on large corpora with improved masking and batching strategies [25].

BART-large

A demonising auto encoder Transformer designed for sequence-to-sequence tasks [24].

ELECTRA-base

A computationally efficient model trained using Replaced Token Detection rather than Masked Language Modelling [26].

Zero-Shot Learning

Zero-shot learning evaluates a model’s ability to perform a task without domain-specific training [27], [28], [68].Sentiment classification is framed as a Natural Language Inference (NLI) problem [104], using templates:

“This sentence expresses a positive/neutral/negative sentiment.”

Models assign the sentiment label with the highest entailment probability.

This approach aligns with zero-shot sentiment studies in [27], [68], [103].

N-Shot Learning & Fine-Tuning

N-shot learning experiments were conducted with:

- 1-shot
- 5-shot
- 10-shot

Fine-tuning used the full dataset (70% train, 10% validation, 20% test).

Training used:

- ELECTRA-base
- AdamW optimizer
- LR = 5e-5
- 200 max epochs
- Early stopping mechanism

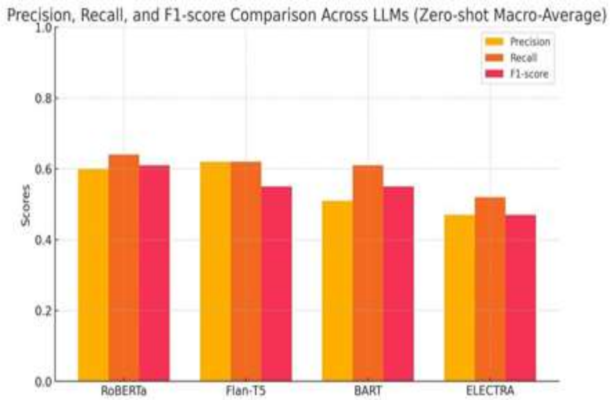
This configuration follows recommended Transformer training guidelines [19], [25].

GPU monitoring used psutil and pynvml.

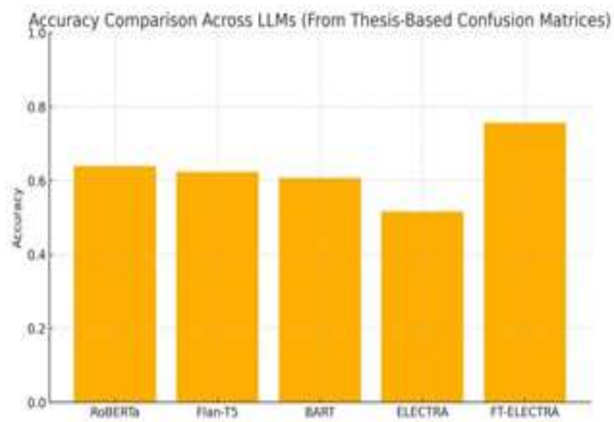
Evaluation Metrics

Multiple metrics were used to ensure robust assessment:

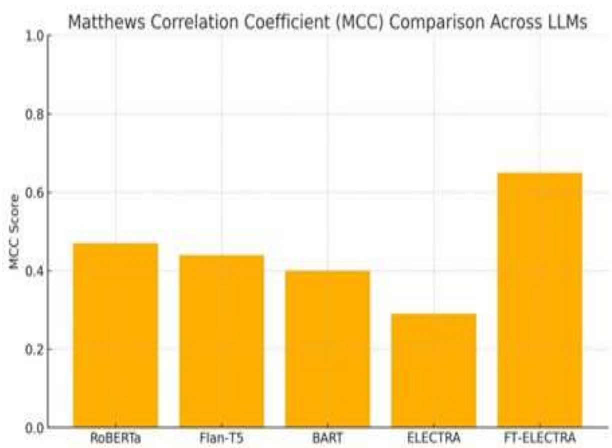
- Precision, Recall, F1-score [3], [11]



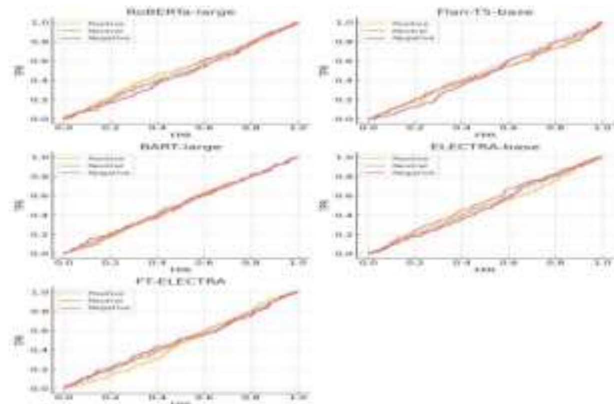
Accuracy



Matthews Correlation Coefficient (MCC) [29]



ROC Curve + AUC [14]



These metrics are commonly applied in sentiment analysis evaluations [32], [36], [38].

V. RESULTS

Zero-Shot Learning Results

The following summarizes model performance:

RoBERTa-large: Best zero-shot accuracy (0.64)

BART-large: Accuracy 0.61

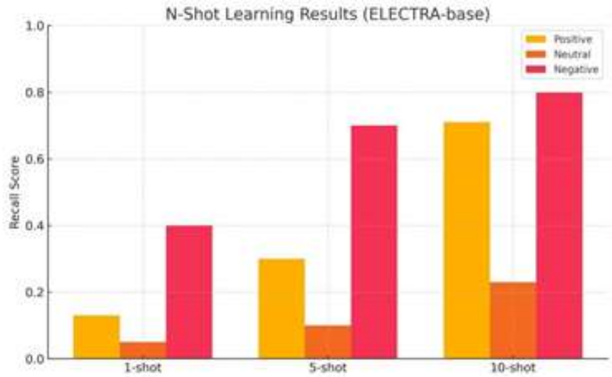
ELECTRA-base: Accuracy 0.52

According to findings in [14], [32], and [40], models had the most difficulty with neutral class classification. According to other LLM-based sentiment studies, all models showed high recall for positive sentiment [84], [85].

N-Shot Learning Results

Increasing N improved performance:

- 1-shot → weak performance (accuracy 0.39)
- 5-shot → moderate improvement (0.47)
- 10-shot → substantial improvement (0.57)



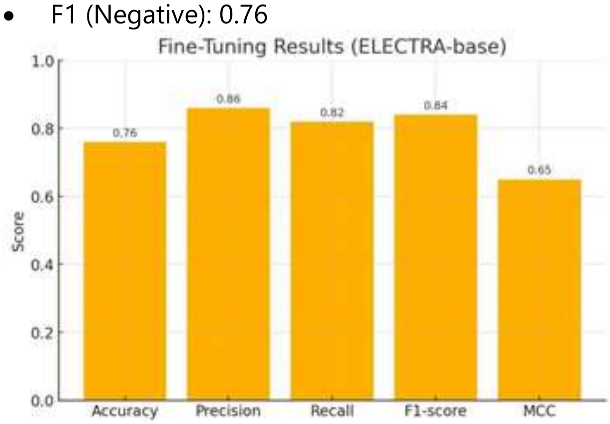
This aligns with few-shot learning findings in LLM research [23], [27], [28].

Neutral category remained low-performing, again matching prior findings on educational text ambiguity [36], [38].

Fine-Tuning Results

Fine-tuned ELECTRA-base achieved:

- Accuracy: 0.76
- F1 (Positive): 0.84
- F1 (Neutral): 0.67
- F1 (Negative): 0.76



According to Transformer-based fine-tuning literature, fine-tuning greatly enhanced performance over zero-shot and N-shot settings. [19], [25], [26].

This demonstrates that when trained on domain-specific data, even small models can perform better than larger LLMs. [30], [31].

VI. DISCUSSION

Important results show that:

1. Smaller models (ELECTRA) do not compete with larger models (RoBERTa-large) in zero-shot settings [25], [29].
2. Neutral sentiment is the hardest to classify as also mentioned in [32], [36] due to its ambiguity and absence of affective qualities.
3. N-shot learning is found to enhance performance and sharp upward trends are observed between 1 and 10 shots [28], [68].
4. Fine-tuning ELECTRA-base gave the best overall results, showing the relevance of domain adaptation to sentiment tasks [19], [26].
5. Even with limited computational resources, LLMs can be used successfully in educational institutions to support the objectives of Education 4.0 [40].

VII. CONCLUSIONS

The usefulness of Large Language Models (LLMs) in assessing student sentiment in learning environments was investigated in this study. The study shows that LLMs can carry out sentiment analysis even in environments with limited resources by contrasting zero-shot, N-shot, and fine-tuning techniques. Larger models, like RoBERTa-large, outperform smaller models, according to zero-shot evaluation; however, all models had trouble classifying neutral sentiment [25], [32]. Even a small number of examples greatly increase classification accuracy, as demonstrated by N-shot experiments [28], [68]. ELECTRA-base, a computationally efficient model, was fine-tuned to achieve the best performance, reaching, and 0.76 accuracy higher than all zero-shot and N-shot settings [26].

These results highlight that:

1. To get the best performance, you need to fine-tune the model for the specific domain.
2. When properly adapted, small LLMs can do better than larger models.
3. Sentiment analysis can help people make decisions based on data in Education 4.0 [40].

This work demonstrates that LLMs are effective and robust instruments for educational sentiment analysis, even when utilizing constrained datasets and computational resources.

Suggestions for Subsequent Research

Future research should examine:

- Aspect-based sentiment analysis [15]

- Categorization of emotional states (e.g., frustration, enthusiasm) [36], [37]
- Multilingual models for worldwide educational settings
- Datasets from different institutions to make them more generalizable
- Techniques for explain ability to make things clearer [69]

Summary:
The research is a good example of how to use LLM to analyse stereotypes of students in practice within limited conditions. With the implementation of zero-shot, N-shot and fine-tuning approaches, the current study offers a pragmatic route towards the incorporation of the AI-based analytics in Education 4.0 ecosystems.

REFERENCE

1. B. Liu, Sentiment Analysis and Opinion Mining. Morgan & Claypool, 2012.
2. B. Pang and L. Lee,"Opinion mining and sentiment analysis," Found. Trends Inf. Retr., 2008.
3. W. Medhat, A. Hassan, and H. Korashy,"Sentiment analysis algorithms and applications," Ain Shams Eng. J., 2014.
4. E. Cambria et al., "New avenues in opinion mining," IEEE Intell. Syst., 2013.
5. M. Taboada, "Sentiment analysis: An overview," Wiley Interdiscip. Rev. Data Min., 2016.
6. A. Esuli and F. Sebastiani, "SentiWordNet," in Proc. LREC, 2006.
7. T. Wilson et al., "Recognizing contextual polarity," in Proc. HLT/EMNLP, 2005.
8. S. Mohammad, "Sentiment analysis: Detecting valence & emotions," Handbook of Linguistic Annotation, Springer, 2016.
9. A. Maas et al., "Learning word vectors," in Proc. ICML, 2011.
10. R. Socher et al., "Recursive deep models," in Proc. EMNLP, 2013.
11. I. Goodfellow et al., Deep Learning. MIT Press, 2016.
12. A. Yadav and D. Vishwakarma, "Deep learning-based sentiment analysis," JIST, 2020.
13. L. Zhang et al., "Deep learning for sentiment analysis," WIREs Data Mining, 2018.
14. S. Sohangir et al., "Big data sentiment analysis using CNN," in IEEE Big Data, 2018.
15. H. Pham and T. Le, "Aspect-based sentiment analysis," in COLING, 2018.
16. T. Mikolov et al., "Distributed representations," in NIPS, 2013.
17. P. Liu et al., "RNN for text classification," in AAAI, 2015.
18. S. Poria et al., "Deeper sentiment analysis," IEEE Intell. Syst., 2016.
19. J. Devlin et al., "BERT," in NAACL, 2018.
20. T. Brown et al., "GPT-3," in NeurIPS, 2020.
21. A. Chowdhery et al., "PaLM," arXiv:2204.02311, 2022.
22. H. Touvron et al., "LLaMA," arXiv:2302.13971, 2023.
23. H. Chung et al., "Flan-T5," Google Research, 2022.
24. M. Lewis et al., "BART," in Proc. ACL, 2020.
25. Y. Liu et al., "RoBERTa," arXiv:1907.11692, 2019.
26. V. Sanh et al., "DistilBERT/ELECTRA efficiency," NeurIPS WS, 2020.
27. A. Radford et al., "Zero-shot learners," OpenAI TR, 2019.
28. W. Yin et al., "Benchmarking zero-shot classification," arXiv:1909.00161, 2019.
29. S. Zhang et al., "Understanding generalization in LLMs," in ACL, 2023.
30. Z. Wu et al., "Limits of LLM performance," arXiv, 2023.
31. K. Hew et al., "Predicting student satisfaction," Computers & Education, 2020.
32. A. Onan, "Sentiment analysis of MOOC reviews," Comput. Elect. Eng., 2021.
33. H. Rajput et al., "Course evaluation sentiment," IJCA, 2016.
34. J. Lalata et al., "Ensemble sentiment for teacher eval," in ICMLA, 2019.
35. Y. Dyulicheva and I. Bilashova, "BERT Udemy review analysis," CEUR, 2022.
36. S. Rani and R. Kumar, "Student opinion mining," IJARCS, 2017.
37. A. Legg and J. Wilson, "Student evaluations: Online vs in-class," Assessment & Evaluation in Higher Education, 2012.

38. W. Stroebe, “Student evaluations & effectiveness,” Persp. Psychol. Sci., 2020.
39. T. Coladarci and I. Kornfield, “Online vs in-class eval,” Educ. Assessment, 2019.
40. M. Rani and R. Verma, “AI in Education 4.0,” JETL, 2023.