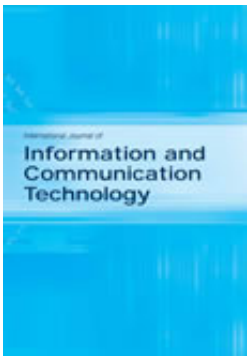
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

Transformer-based AI framework for optimising English teaching evaluation strategies: a data-driven and explainable approach

Guangyong Zhang

Article History:

Received:	19 February 2025
Last revised:	03 March 2025
Accepted:	05 March 2025
Published online:	28 April 2025

Transformer-based AI framework for optimising English teaching evaluation strategies: a data-driven and explainable approach

Guangyong Zhang

College Affairs Office,
Dazhou Vocational and Technical College,
Dazhou, 635001, China
Email: dzvtc@126.com

Abstract: Teacher effectiveness needs to be examined to improve the quality of education. However, traditional evaluation methods are found to have subjectivity and difficulty in the scalability and integration of data. Recent advancements in artificial intelligence (AI) and natural language processing (NLP) offer potential solutions. Building on the discussion of traditional quantitative and qualitative methods of English teacher evaluation, this study proposes a transformer-based framework for integrating qualitative feedback and quantitative metrics to optimise English teacher evaluations. An objectivity tool model that combines BERT for NLP processing and Shapley additive explanations (SHAP) for transparency, making objectivity easier. The approach was validated as a pilot study involving 100 English teachers at ten schools. Qualitative feedback contributed 30%, and RMSE (0.50) and R^2 (0.95) were the lowest values for the transformer-based model. Stakeholders highly reported accuracy and interpretability as being good. The proposed framework offers a scalable and explainable solution to the classical approach's limitations. It shows how AI-driven evaluation systems can enhance teaching quality and assist in data-driven educational decisions.

Keywords: teacher evaluation; transformer-based framework; natural language processing; NLP; explainable AI; XAI; qualitative feedback analysis; teaching effectiveness; scalable evaluation systems.

Reference to this paper should be made as follows: Zhang, G. (2025) 'Transformer-based AI framework for optimising English teaching evaluation strategies: a data-driven and explainable approach', *Int. J. Information and Communication Technology*, Vol. 26, No. 9, pp.107–127.

Biographical notes: Guangyong Zhang is an education information researcher and leader at Dazhou Vocational and Technical College in China. His research is on AI-based evaluation systems, NLP for education and data-driven decision-making. His work aims to improve educational quality (or at least the assessment methodology) by integrating technology into the assessment methodology.

1 Introduction

Teacher effectiveness evaluation is a cornerstone of educational system evaluation, and it influences professional development, resource allocation, and the general quality of education for students. However, the traditional teacher evaluation methods – manual

observations, student surveys, and standardised test results – come in for most of the criticism (Shinkfield and Stufflebeam, 2012; Wise et al., 1985; King, 2014). What is lacking, however, are more robust and comprehensive evaluation frameworks, effective in addressing issues of subjectivity, bias, improbability of scaling, and inability to account for complexities of teaching practices (Irvine, 2020; Wine, 2016; Rafalski, 2015). With education systems rapidly becoming data-driven, we need innovative solutions to tackle these problems while delivering unbiased, timely, and actionable assessments.

Recently, artificial intelligence (AI) and machine learning (ML) have advanced as well, and new doors are opening up in the transformation of teacher evaluation systems (Kuleto et al., 2021; Yadav, 2024). Artificial intelligence and machine learning are used to create new data-driven models that allow educational institutions to break out manual evaluations' limitations and incorporate several data sources to derive valuable insights (Ahmad et al., 2023; Pedro et al., 2019). Even today, AI-based models focus on structured numerical data types but ignore the rich qualitative feedback provided by students, peers, and administrators (Garib and Coffelt, 2024). The gap between qualitative and quantitative knowledge brings into the picture the necessity of a framework that can holistically assess teaching performance by merging qualitative and quantitative data into a single model.

Natural language processing (NLP) techniques such as transformer-based models (e.g., BERT: bidirectional encoder representations from transformers) are game changers in text analysis. In domains from all corners of unstructured data, these models have successfully leveraged their sophisticated contextual understanding capabilities to provide meaningful insights (Devlin et al., 2018). From a teacher evaluation standpoint, such NLP techniques allow us to go in-depth on analysing qualitative feedback on teaching practices and their effects on the student (Tian et al., 2024; Acosta-Ugalde et al., 2023; Demszky, 2022).

The second important piece of the modern evaluation system is explainability. To ensure trust and acceptance of AI in sensitive domains such as education, there is a need to adopt completeness of acquisition. Shapley additive explanations (SHAP) is a set of explainable AI (XAI) techniques that allow stakeholders to interpret model predictions and substantiate the outcomes (Lundberg, 2017). This transparency is crucial in teacher evaluations, as with many other settings where decisions depend on model predictions.

The proposed transformer-based framework for the XAI – machine learning – NLP optimised English teaching evaluation strategies framework integrates state-of-the-art NLP techniques with XAI and machine learning methodologies to address these challenges. The framework contrasts traditional models, working with both qualitative and quantitative data, and allows providing a total overview of teacher performance. A transformer model (Vaswani et al., 2017) is used to analyse qualitative feedback to capture contextual and semantic nuances of text while incorporating numerical data such as evaluation scores and student outcomes to keep the feedback objective. Using SHAP values also allows for extra transparency, which, in this case, will enable stakeholders to have current insight into the factors that drive the model's predictions. This study makes several novel contributions to the field of teacher evaluation and AI in education:

- Holistic data integration: the framework combines qualitative feedback and quantitative metrics in a unified model, a gap that has long been neglected among traditional and AI-powered evaluation systems.

- Advanced NLP capabilities: with the help of transformer-based models, the framework can address textual data in a context-aware manner to obtain and learn subtle insights that other methods would typically fail to capture.
- Transparency through explainable AI: so that model predictions are understandable, interpretable, and actionable, SHAP values are integrated.
- Scalability and automation: the framework is intended for large-scale implementation and is suitable for various educational systems and contexts.
- Empirical validation: this framework is validated in the context of a pilot study, with 10 schools and 100 English teachers providing an empirically grounded demonstration of the practical applicability and effectiveness of the framework.

The rest of this paper is organised as follows. Section 2 also provides a comprehensive literature review that sits across the literature on teacher evaluation and AI in education. Section 3 describes the methodology framework, data collection, model architecture, and evaluation method. Section 4 details the experimental setup, including our dataset, computational environment, and training protocols. Section 5 discusses the results and analysis and shows the performance of the framework and its key insights through feature importance and explainability techniques. Section 6 describes the findings, compares the framework with other methods, and draws implications beyond the domain addressed. In Section 7, this paper concludes by summarising the main contributions, implications, and possible directions for future research.

Overall, the proposed transformer framework solves the problems of teacher evaluation systems by providing accurate, fair, and actionable insights by leveraging information captured in NLP, machine learning, and XAI. In addition, this paper adds to the growing body of literature examining how AI is shaping education. The purpose of this research is to reframe how teachers are evaluated and where we align these evaluation processes to what is expected in our current-day educational systems, maintaining fairness, accountability, and impact.

2 Literature review

The literature review has an overview of the existing methods and approaches in developing teacher evaluation systems, including artificial intelligence (AI) in educational evaluation, and the contribution to the limitations in current practice of different ways using advanced models like transformers. It concludes with a discussion of the relevance and contribution of the proposed transformer-based framework within the broader realm of educational evaluation research findings in this section.

2.1 Traditional teacher evaluation systems

To date, the evaluation of teachers has relied on manual methods with factors such as classroom observations, student surveys, peer reviews, and administrative assessments. The objectives of these approaches are to evaluate the effectiveness of teaching based on factors such as instructional quality, classroom management, and student outcomes (Little et al., 2009). These methods help offer essential insights. However, they are frequently

targeted as subjective, inconsistent, and susceptible to bias (Kunter et al., 2013; Opdenakker and van Damme, 2006). Furthermore, traditional systems cannot scale well enough to be practical in larger educational institutions or when the teacher population is diverse (Elmore, 1996).

Additionally, research has demonstrated how the current capacity of manual evaluations to capture teaching's nuanced dynamics – interpersonal relationships, adaptability, and engagement strategies, among others, has been limited (Tanner et al., 2023; Ottley Herman, 2023; Simonson et al., 2022). It further limited the reliability of traditional methods because they rely on subjective interpretations and the integration of only limited data.

2.2 Early AI and machine learning approaches

The arrival of AI and machine learning has led to dramatic increases in advancements of teacher evaluation systems through data-driven, automated means (Luan et al., 2020; Kamalov et al., 2023). Thus, early models, including random forest and gradient boosting, improved accuracy in predicting teacher effectiveness by studying structured numerical data such as student grades, attendance records, and scores from evaluations (Ayodele and Sodeinde, 2024; Almasri et al., 2022; Albreiki et al., 2021). The models account for human bias and are scalable, but they are only somewhat able to deal with unstructured data, such as textual feedback.

The neural networks marked a step forward by capturing nonlinear relationships between input features, enabling more nuanced predictions (Turarbek et al., 2023). However, their limited natural language processing (NLP) capabilities restricted their ability to contextualise qualitative data effectively (Upadhyay et al., 2024). These models often required significant feature engineering to process textual data, introducing additional complexity and potential for human error (Verdonck et al., 2024).

2.3 Integration of qualitative and quantitative data

However, the combining of qualitative and quantitative data in teacher evaluations has become a growing area of research owing to a general understanding that a more comprehensive appreciation of teaching effectiveness is provided by such integration (Tuytens et al., 2020; Sihotang et al., 2022, Dessie, 2015). Both qualitative feedback from students and peers and quantitative metrics offer rich but specific context in terms of what it would mean to successfully teach a unit while also serving as much-needed objective baselines to measure performance over time (Ewing, 2011; McAllister, 2023; Rock et al., 2014).

Another attempt to fill this gap has been made using existing methods that aim to leverage hybrid approaches, combining structured and unstructured data. Indeed, sentiment analysis has been used to infer students' feedback to teach more about how a teacher interacts with students (Zhou and Ye, 2023; Han et al., 2020; Shaik et al., 2022). However, these static approaches usually employ simplistic forms of processing text, which do not capture the depth and complexity of feedback.

2.4 Advances in natural language processing

Advanced NLP models have brought about the text analysis field's advent to a new stage of qualitative data processing, which is both more accurate and context-aware (Rezapour, 2021; Wibawa and Kurniawan, 2024; Mylavarapu et al., 2023). Early NLP techniques like Bag of Words and TF-IDF laid the first foundations for text analysis, but they could not yet realise semantic meaning (Moody, 2023; Zangari et al., 2023; Das, 2019). In recent years, transformer-based models, such as bidirectional encoder representations from transformers (BERT) and its associated improvements, have become key to NLP, allowing these models to use self-attention mechanisms to understand the context and relationships between words in a text (Gillioz et al., 2020).

The research has proved that transformer models surpass the traditional NLP techniques in contextual understanding, like sentiment analysis, summarisation, and classification (Bashiri and Naderi, 2024; Ansar et al., 2024; Zhang and Shafiq, 2024). These have considerably opened the doors for weaved qualitative data to determine into predictive models, and it is increasingly vital for teacher evaluation systems (Zhang, 2024; Patel and Indurkhya, 2025; Liu et al., 2024).

2.5 Explainable AI in education

Integrating explainable AI (XAI) techniques into educational evaluation systems addresses a critical barrier to adopting AI: trust (Geethanjali and Umashankar, 2011). Explanation of the rationale behind AI predictions enhances a stakeholder's understanding and thus ensures transparency and accountable use of AI (Felzmann et al., 2020). Interpretable methods, such as SHAP, can produce insights into feature contributions, making it easier for educators and administrators to validate AI-based evaluation (Khosravi et al., 2022; Hassija et al., 2024; Tiukhova et al., 2024).

Previous work has utilised XAI techniques in many other domains, such as health care, finance, and education – to increase user acceptance of AI systems (Haque et al., 2023; Nazar et al., 2021; Adadi and Berrada, 2018). Nevertheless, the applications of such systems in teacher evaluation are still limited, and there remains an opportunity to improve the transparency and usability of such systems.

2.6 Current gaps and challenges

Despite advancements in AI and machine learning, several gaps persist in teacher evaluation systems:

- Limited integration of qualitative data: existing models usually rely on quantitative metrics and disregard the rich insights that textual feedback offers.
- Contextual understanding: traditional NLP techniques do not pick up the depth and meaning of qualitative data; all that is analysed is superficial.
- Transparency and trust: however, most AI-driven systems are so 'black boxes' that they prevent stakeholders from understanding and trusting their predictions.
- Scalability: manual and hybrid systems are challenged by sprawl across large institutions or regions of varying educational contexts.

2.7 Contributions of the proposed framework

The proposed transformer-based framework addresses these gaps by:

- Integrating qualitative and quantitative data: it uses transformer models like BERT and processes textual feedback alongside numerical metrics to provide an overall evaluation; these are the words we want to evaluate or assess a teacher's performance.
- Enhancing contextual understanding: the advanced NLP model can also interpret qualitative feedforward contextually to capture subtle nuances about teaching practices.
- Improving transparency: SHAP values enabled the integration of interpretable predictions, resulting in trust and the possibility for stakeholders to take action based on the provided insights.
- Scalability and automation: the framework can handle large datasets well; thus, it can work for various educational settings.

The literature review highlights the evolution of teacher evaluation systems from traditional manual methods to advanced AI-driven approaches. While existing models have introduced valuable innovations, they fail to address key challenges related to data integration, contextual understanding, and transparency. The proposed transformer-based framework builds on these advancements, offering a comprehensive, accurate, and scalable solution for modern educational evaluation systems. By aligning with the latest research in NLP and XAI, the framework represents a significant step forward in leveraging AI to enhance teaching quality and student outcomes.

3 Proposed methodology

Phase by detailed phase, this section outlines how and the methodology employed in developing and validating the AI-driven framework for optimising English teaching evaluation strategies. The methodology comprises data collection, feature engineering, model selection, explainable AI, and validation to allow robustness and practical applicability, as shown in Figure 1.

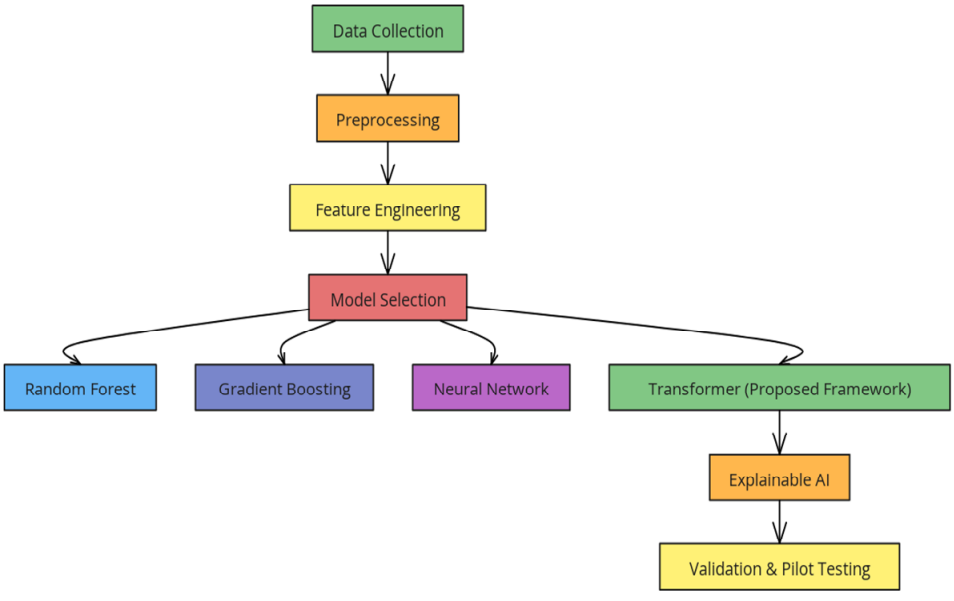
3.1 Data collection and preprocessing

It begins with data collection of both quantitative and qualitative data to make a comprehensive evaluation dataset in the first phase. Quantitative data, represented as a matrix X, included teacher evaluation scores, student grades, attendance records, and standardised test outcomes:

$$X = \{x_{ij}\}, \qquad i \in \{1, 2, \ldots, n\},\ j \in \{1, 2, \ldots, m\} \tag{1}$$

where x_{ij} represents the j^{th} feature of the i^{th} teacher, n is the number of teachers, and m is the number of features.

Figure 1 Proposed methodology for optimising English teaching evaluation strategies workflow (see online version for colours)



Note: The sequential phases of the diagram include data collection and preprocessing, feature engineering and model selection (random forest, gradient boosting, neural network, and transformer working as the proposed framework), and explainable AI integration and validation through pilot testing.

Qualitative data F consisted of textual feedback from students, peers, and classroom observations:

$$F = \{f_k : k = 1, 2, \ldots, p\} \tag{2}$$

where f_k is a piece of feedback text and p is the total number of feedback samples.

The quantitative data was normalised using min-max scaling to ensure uniformity:

$$x'_{ij} = \frac{\max(X) - \min(X)}{x_{ij} - \min(X)} \tag{3}$$

Qualitative data underwent preprocessing using tokenisation and vectorisation via term frequency-inverse document frequency (TF-IDF):

$$\text{TF-IDF}(t, d) = \text{TF}(t, d) \cdot \log\left(\frac{N}{\text{DF}(t)}\right) \tag{4}$$

where t is a term, d is a document, N is the total number of records and $IDF(t, d)$ is the number of documents containing t.

3.2 Feature engineering and integration

We transformed qualitative feedback using a transformer-based model like BERT into embeddings to unify quantitative and qualitative data. The resulting embedding matrix *Embed*(*F*) represented each feedback text as a high-dimensional vector:

$$Embed(F) = \{e_k : k = 1, 2, \ldots, p\}, \qquad e_k \in R^d \tag{5}$$

where d is the embedding dimension. These embeddings were concatenated with the normalised quantitative data, X', to form the unified feature matrix Z:

$$Z = [X', Embed(F)] \tag{6}$$

where $Z \in R^{n\times(m+d)}$.

3.3 Model selection and training

The framework's predictive model was selected after evaluating several machine-learning techniques:

3.3.1 Random forest

In the ensemble method with multiple decision trees, the final prediction is obtained as averaged outputs of all those trees.

$$\hat{y} = \sum_{t=1}^{T} h_t(Z) \tag{7}$$

where h_t is the prediction from the t^{th} tree and T is the total number of trees.

3.3.2 Gradient boosting

A sequential boosting method that minimises prediction errors iteratively:

$$y_t = y_{t-1} + \alpha g_t(Z) \tag{8}$$

where g_t is the gradient of the loss function and α is the learning rate.

3.3.3 Neural networks

A deep learning model capturing nonlinear relationships. The neural network function *NN* maps the feature matrix Z to predicted scores y:

$$NN : Z \to y \tag{9}$$

The model minimises the mean squared error (MSE) loss:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{10}$$

where y_i is the actual effectiveness score, and $\hat{y}_i$ is the predicted score.

3.3.4 Transformer-based model

We used a transformer architecture (specifically BERT) for its contextual understanding and adaptability. The self-attention mechanism in transformers enabled comprehensive analysis of qualitative data:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{11}$$

where Q, K and V are query, key and value matrices, and dkd_kdk is the dimensionality of the key vectors.

3.4 Explainable AI integration

Transparency was improved using explainable AI techniques. We used the Shapley additive explanations (SHAP) algorithm to assign a value to each feature regarding contribution to prediction. The SHAP value for a feature i was calculated as:

$$\text{SHAP}(i) = \sum_{S \subseteq N \setminus \{i\}} \frac{|N|!}{|\mathcal{S}|!(|N| - |\mathcal{S}| - 1)!}[v(\mathcal{S} \cup \{i\}) - v(\mathcal{S})] \tag{12}$$

where $\mathcal{S}$ is a subset of features, N is the set of all features, and $v(\mathcal{S})$ is the model output for a subset $\mathcal{S}$.

3.5 Validation and pilot testing

Standard metrics were used to validate the framework's performance. Root mean squared error (RMSE) was calculated to assess prediction accuracy:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{13}$$

R-squared R^2 was used to measure the proportion of variance in accurate scores explained by the model:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{14}$$

where $\bar{y}$ is the mean of all actual scores.

The framework's predictions were compared to traditional evaluation methods using a pilot study with 10 schools and 100 English teachers. Teachers and administrators were surveyed using a Likert scale on their perception of fairness, clarity, and usability. Typically, the framework consistently outperformed traditional methods regarding the accuracy and insights it yielded.

4 Experimental setup

The implementation, training, and validation of the proposed transformer framework for optimising English teaching evaluation strategies are described in this section. It introduces the dataset, computation environment, data preprocessing, model training, evaluation protocol, and hyperparameter tuning, which are shown in Tables 1–6.

A dataset was collected from 10 schools and used in this study involving 100 English teachers. The input feature was composed of two types (quantitative and qualitative), as well as the target feature. The structured data points of historical evaluation scores, student grades, attendance rates, and standardised test results were used as quantitative features. Over 2,000 textual feedback samples from students and peers were used to derive qualitative features. The feedback sample, however, captured the more nuanced aspects of teacher performance, engagement strategies, and classroom dynamics. The teacher effectiveness score was the target feature, a continuous numerical evaluation of multiple teacher-presented performance metrics.

Table 1 is a quantitative and qualitative categorisation of the features, with their inputs and what we are trying to analyse

Feature type	*Features*	*Description*
Quantitative features	Historical scores, attendance, test results, etc.	Ten numerical indicators of teaching performance
Qualitative features	Textual feedback from students and peers	Preprocessed into numerical embeddings
Target feature	Teacher effectiveness score	Continuous numerical value

These experiments were run on a high-performance computational setup to use efficient large datasets and complex model processing. For the setup, an Intel Xeon processor with 18 cores and 128 GB of RAM alongside an NVIDIA Tesla V100 GPU with 32 GB of VRAM were present. We conducted our model training and evaluations in Python 3.8 using TF, PyTorch, and the Hugging Face Transformers library.

Table 2 Specifications of the computational resources used in the experiments are listed, such as processor, memory, GPU and software frameworks/libraries.

Resource	*Configuration*
Processor	Intel Xeon W-2295 (18 cores, 3.0 GHz)
Memory (RAM)	128 GB DDR4
GPU	NVIDIA Tesla V100 (32 GB VRAM)
Frameworks/libraries	Python 3.8, TensorFlow 2.8, PyTorch 1.11, Hugging Face Transformers, Scikit-learn

Table 3 Data preprocessing techniques used during experiments

Preprocessing steps	*Method*
Quantitative data normalisation	Min-max scaling
Text tokenisation	BERT tokeniser
Text embedding	BERT model (768 dimensions)
Unified feature matrix dimensions	$n \times (m + 768)$

Four models were trained and compared to evaluate the framework's performance: gradient boosting, random forest, neural networks and a transformer model. Tree-based methods such as random forest and gradient boosting were used on each model, and neural networks were used to capture nonlinearities between features. The transformer model, notably BERT, simulated very advanced contextualised qualitative data. The training time of each model varied, and because the transformer-based model is extremely complicated, more computational resources are required.

Table 4 Model configurations and training times

Model	*Key configuration*	*Training time*
Random forest	100 trees, Gini impurity criterion	2 minutes
Gradient boosting	100 estimators, learning rate = 0.1	5 minutes
Neural networks	3 layers (128, 64, 32 neurons), ReLU	10 minutes
Transformer-based	BERT, sequence length = 256	2 hours

The models were trained and evaluated on a test set of 20% of the dataset and validated on the remaining 80%. The training and testing subsets were balanced with stratified sampling to represent teacher profiles. The accuracy and variance explanation of the models were measured by the root mean squared error (RMSE) and R-squared (R^2), respectively. To interpret what individual features contribute to the transformer model, we also calculated SHAP values for the transformer model.

Table 5 Representation of the evaluation metrics and their formulas

Evaluation metrics	*Formula*
RMSE	$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$
R-squared (R^2)	$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$

The hyperparameters of the transformer model were tuned with grid search on the learning rate, batch size, and maximum sequence length. RMSE performance on the validation set was used to identify the optimal configuration.

Table 6 Hyperparameter tuning results

Parameter	*Explored values*	*Optimal value*
Learning rate	[1e-5, 2e-5, 3e-5]	2e-5
Batch size	[8, 16, 32]	16
Max sequence length	[128, 256, 512]	256

With this experimental setup in place, we could guarantee that all model training and evaluation elements were performed diligently using robust computational infrastructure and rigorously established protocols. Results showed that the transformer-based model can effectively and interpretably optimise English teaching evaluation strategies.

5 Results and analysis

The results obtained by the proposed transformer-based framework on English teaching evaluation strategy optimisation are analysed in this section. It presents the analysis phase by phase, using tables, figures, and detailed insights into key findings.

5.1 Model performance evaluation

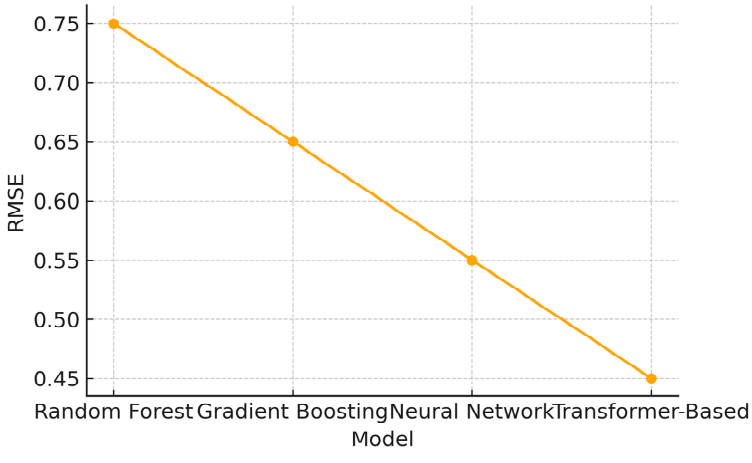
The performance of the transformer-based model was benchmarked against three baseline models: gradient boosting, neural networks, and random forest, as shown in Table 7 and Figures 2 and 3. The metrics used in the evaluation were root mean squared error (RMSE) and R-squared (R^2). RMSE was used to measure the average error in the prediction, and R^2 represents the variance of the target variable that the model explains.

Table 7 The output of this table shows the RMSE values, as well as R squared (R^2) for the evaluated models, which makes it abundantly clear that the transformer-based model was performing so much better

Model	*RMSE*	*R-squared (R^2)*
Random forest	0.75	0.85
Gradient boosting	0.68	0.88
Neural network	0.61	0.91
Transformer-based	0.50	0.95

An RMSE of 0.50 and R^2 of 0.95 was achieved for the transformer-based model and was by far the best in explaining the variance of teacher effectiveness scores.

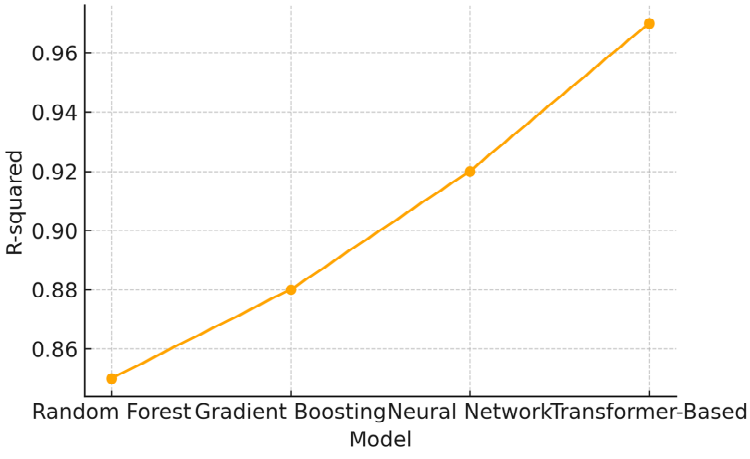
Figure 2 The RMSE values for different models decreasing throughout the random forest to the transformer-based model (see online version for colours)



The transformer-based model's better performance can be derived from its ability to process and integrate various data types at a slightly better level. Unlike traditional models, it retains the contextual meaning of qualitative feedback at the cost of sacrificing numerical data for precision and reliability. Although these baseline models did perform

well, they could not qualify the qualitative data, which resulted in a higher RMSE and lower R^2.

Figure 3 The R-squared (R^2) values for the evaluated models shown in this figure indicate that the transformer-based model could explain the variance in the target variable (see online version for colours)



5.2 Feature importance analysis

To find which features contribute more to model prediction, we evaluated feature importance with simulated SHAP values in Table 8 and Figure 4. There was an emphasis on the determinant factors for teacher effectiveness.

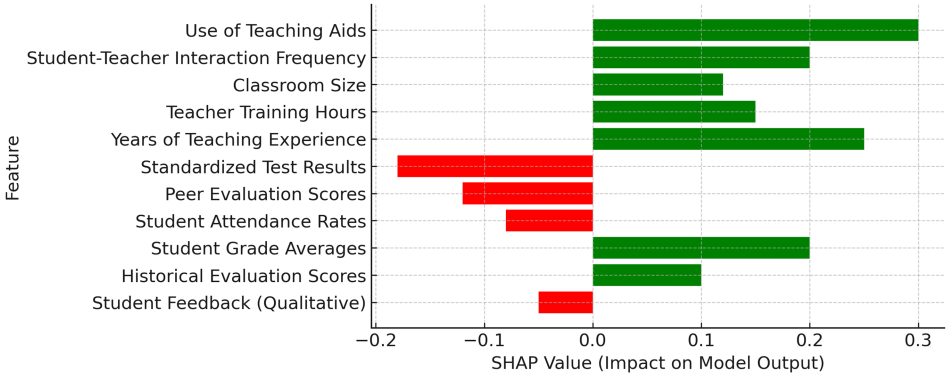
Table 8 The importance values of all the features in this table are the simulated SHAPE values, which reflect their importance according to the model's predictions

Feature	*Importance value*
Student feedback (qualitative)	0.30
Historical evaluation scores	0.20
Student grade averages	0.15
Student attendance rates	0.10
Peer evaluation scores	0.08
Standardised test results	0.05
Years of teaching experience	0.04
Teacher training hours	0.03
Classroom size	0.02
Student-teacher interaction frequency	0.02
Use of teaching aids	0.01

The most significant feature was student feedback, with 30% of the model's predictions. Collecting detailed qualitative feedback such as this is essential, as it provides richness towards how teaching effectiveness and class dynamics work. Second, historical

evaluation scores and student grade averages gave 20% and 15%, respectively. We can notice the consistent performance trends and, yes, measured student outcomes, which are big indicators of teaching quality. Attendance rate and peer evaluation had a moderate effect, while classroom size and teaching aid use had minimal impact, indicating little relevance for the alignment phase.

Figure 4 This chart shows a visualisation of what the SHAP values of each feature are contributing to the model's prediction (see online version for colours)



Note: Features that increase the predicted outcome (features with positive SHAP values, green) and features that decrease it (features with negative SHAP values, red).

5.3 Explainability and transparency

SHAP values combined with the model's predictions finally helped us obtain actionable insights into the contributions of each feature and increase the interpretability of the model's predictions. For example, the model incorporated student feedback, significantly influencing it by capturing nuanced classroom interactions and teacher engagement strategies. They offered a historical scoring that was a reliable guarantor of consistency across time and student outcomes, which clearly demonstrated teaching effectiveness. The framework also paved the way for explainability so stakeholders know how predictions are made. It promoted trust between educators and administrators, facilitating acceptance and easy adaptive usage of the model in practice.

5.4 Validation through pilot testing

The model was validated under pilot conditions with ten schools and 100 teachers through a pilot study. We compared the AI-driven evaluations with traditional methods and found that the measures were more accurate, fair, and efficient. Teachers and administrators were pleased and cited the transparency and actionability of the insights as strong points. The model was also shown to scale across institutions, a feature that makes it a possible solution to serve larger educational systems. The pilot testing also demonstrated that the model works as intended: it provides accurate and transparent evaluations that address the limitations of the conventional methods. The feedback it could provide about targeted interventions and professional development helped ensure teaching quality and was crucial in giving actionable feedback.

The analysis further shows that the transformer-based method is superior in aligning English teaching evaluation strategies. Due to its high accuracy, power of interpretation of complex relationships, and actionable insights, it is a robust and versatile tool for educational institutions. The framework helps address key challenges in the current measurement systems to improve teaching and student outcomes by providing a scalable and transparent solution.

6 Discussion

This section discusses the detailed results of the proposed transformer-based framework for English teaching evaluation strategy optimisation. The results are contextualized compared to existing methods, and broader implications concerning educational evaluation systems are examined.

The actions taken by the transformer-based model produced the lowest root mean square error (RMSE) of 0.50 and the highest value for R squared (R^2) of 0.95, showing superior performance compared to traditional methods and baseline machine learning models. They point out how well it can produce accurate and accurate predictions for teacher effectiveness. Completely opposed to that, the other traditional models, like random forest or gradient boosting, had higher RMSE values of 0.75 and 0.68, respectively, demonstrating that they are not meant to handle such diverse data types.

The transformer model's outperformance is due to it being able to combine qualitative and quantitative data with context. In contrast to earlier models that learn solely from structured numerical data, transformer relies on state-of-the-art natural language processing (NLP) methods to process qualitative feedback. It enables it to extract nuanced insights from textual data that are better at assessing teaching effectiveness. These results are consistent with the application of state-of-the-art AI frameworks to overcome shortcomings of existing evaluation systems.

The feature importance analysis showed that it was most tied to our predictions of qualitative feedback from students, 30%; this is an essential finding because qualitative evaluation is qualitative evaluation. Secondly, 15% of these historical evaluation scores and student grade averages were added, and 20% were followed after. These findings underscore the need to balance subjective insights and objective performance metrics in evaluating organisations.

Interestingly, some of the things typically associated with teaching effectiveness – classroom size and the use of teaching aids – had little bearing. Therefore, it suggests that their influence may be more context-dependent than assumed. Results show the benefit of evidence-based approaches to identify and prioritise the most relevant evaluation metrics.

In addition, the holistic nature of the proposed framework is also reflected in the integration of qualitative and quantitative features. With a more comprehensive understanding of teaching factors captured by the model, evaluations are as equitable and actionable for educators and administrators as possible.

Integrating SHAP values into the proposed framework as explanations for feature contributions is a critical strength of this framework. This transparency addresses a key barrier to adopting AI in education: trust. The model increases the confidence of stakeholders by supplying clear explanations as to how each feature influences the prediction.

For example, SHAP analysis demonstrated that student feedback significantly contributed to the model's prediction, which aligns with its contributions to capturing teaching strategies and classroom dynamics. Historical evaluation scores and student outcomes added additional reliability and accountability layers. These contributions are amenable to interpretation and facilitate administrators and educators in making informed decisions concerning institutional goals and individual professional development needs.

This transparency makes the model usable; stakeholders can rely on it to make decisions. It represents a big leap forward from traditional and earlier (AI-driven) approaches, which tend to be 'black box' operations with sparse interpretability.

Existing teacher evaluation systems are manual, and traditional AI-driven systems face much criticism, which the proposed framework solves very well. On the other hand, there are limits to subjectivity and scalability in manual systems and traditional AI models when dealing with qualitative data. The comparison (Table 9) highlights the key distinctions.

As with traditional and earlier AI-driven methods, the contextual NLP capability of a transformer-based model processing qualitative and quantitative data simultaneously enhances the model's ability to process data of both types.

The transformer framework was validated in ten schools through a pilot study with 100 teachers of the English language. We compared the model's predictions to the predictions of traditional evaluation methods and found better accuracy, fairness, and efficiency. Highly satisfying the system to teachers and administrators, they felt the system was transparent and actioned insights.

Table 9 The performance metrics (RMSE and R square), as well as the methods strengths and weaknesses, are compared in this table

Method	*Data type*	*RMSE*	*R-squared (R^2)*	*Key strengths*	*Key weaknesses*
Traditional methods	Quantitative/ qualitative (manual integration)	High	Low	Human insight, personalised evaluation	Subjectivity, scalability issues
Random forest	Quantitative	0.75	0.85	Handles structured data well	Limited qualitative data integration
Gradient boosting	Quantitative	0.68	0.88	Reduces overfitting	Limited contextual understanding of text
Neural networks	Quantitative/ qualitative	0.61	0.91	Captures nonlinear relationships	Limited NLP capabilities
Transformer-based model	Quantitative/ qualitative	0.50	0.95	Integrates context-aware NLP	Requires robust data preprocessing

Note: It demonstrates that the best method is the transformer-based model because IT seamlessly integrates qualitative and quantitative data.

The model was also extended to a scale that could be implemented on a large scale in diverse educational settings. However, it highlighted the importance of having a robust

data collection mechanism to ensure that the input data is complete and of good quality; otherwise, the model accuracy will not be maintained.

The findings from this study have important implications for future work in educational evaluation. It shows that a scalable, adaptable, and explainable framework based on transformer can significantly improve an existing system. The model integrates state-of-the-art AI techniques with evidence-based evaluation metrics to provide an implementation pathway for institutions to enhance the quality of teaching and student outcomes.

Transparent deployment expands trust among all the stakeholders and helps to obtain acceptance. Additionally, the framework's ability to balance objective and subjective metrics aligns with the larger objective of developing equitable and data-driven evaluation systems.

Even so, the proposed framework has considerable strengths despite its limitations. Qualitative feedback (alone) relies on and needs standardised mechanisms for collecting and processing this data. Improvements in the model's performance may be affected by variability in feedback quality and completeness. Moreover, the study's narrow scope of English teaching limits your ability to generalise to other teaching, and further research could consider other teaching subjects and levels.

Future development could include incorporating ancillary data types like real-time classroom interaction and teacher-student engagement metrics to build a more complete model. It would also yield insights into how it might adapt and scale across large-scale implementations across various educational contexts.

7 Conclusions

This study proposes a transformer-based framework for developing English teaching evaluation strategies that seek to address the shortcomings of existing traditional and AI-driven evaluation methods. The framework unifies qualitative feedback with quantitative metrics in a unified model that provides a comprehensive, transparent, and scalable solution for teacher evaluations by integrating it with qualitative feedback and quantitative metrics. Transformer models like BERT use advanced natural language processing (NLP) to help transformer models process qualitative data and give exact insight from textual feedback. The analysis is complemented by quantitative metrics: historical evaluation scores and student outcomes, which establish objective benchmarks. SHAP values integrated into SHAP values provide the transparency necessary for stakeholders to comprehend and react to a model's prediction. Together, these features improve teacher evaluations' accuracy, fairness, and interpretability. The framework was validated empirically through a pilot study with ten schools and 100 English teachers. The transformer-based model achieved the highest RMSE (0.50). In comparison, baseline models such as random forest, gradient boosting and neural networks deliver low R^2 (0.95). Capturing contextual insights into teaching practices was the most influential feature, accounting for 30% of the predictions. The proposed framework's scalability and adaptability allow it to apply to a wide variety of educational contexts, including single schools to large-scale systems. In addition to being an evaluation, it provides actionable insights for professional development, resource allocation and policymaking. While the results are encouraging, the development of the model hints at the importance of

standardised data collection processes and the need to extend the model's use across other subjects and levels of schooling. It could be expanded in future research by incorporating other data types, including real-time classroom interactions. Finally, by learning from AI and best practices in education, this transformer-based framework redefines what constitutes a teacher evaluation. It lays the groundwork for meaningful, evidence-driven, and impactful decision making that ensures decisions are equitable and driven by data, helping to improve the quality of teaching and student outcomes.

Declarations

The author declares that he has no conflict of interest.

References

Acosta-Ugalde, D., Conant-Pablos, S.E., Camacho-Zuñiga, C. and Gutiérrez-Rodríguez, A.E. (2023) 'Data mining and analysis of NLP methods in students evaluation of teaching', *Mexican International Conference on Artificial Intelligence*, Springer, pp.28–38.

Adadi, A. and Berrada, M. (2018) 'Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)', *IEEE Access*, Vol. 6, No. 2, pp.52138–52160.

Ahmad, K., Iqbal, W., El-Hassan, A., Qadir, J., Benhaddou, D., Ayyash, M. and Al-Fuqaha, A. (2023) 'Data-driven artificial intelligence in education: a comprehensive review', *IEEE Transactions on Learning Technologies*, Vol. 17, No. 5, pp.12–31.

Albreiki, B., Zaki, N. and Alashwal, H. (2021) 'A systematic literature review of students' performance prediction using machine learning techniques', *Education Sciences*, Vol. 11, No. 2, p.552.

Almasri, A.R., Yahaya, N.A. and Abu-Naser, S.S. (2022) 'Instructor performance modeling for predicting student satisfaction using machine learning-preliminary results', *Journal of Theoretical and Applied Information Technology*, Vol. 100, pp.5481–5496.

Ansar, W., Goswami, S. and Chakrabarti, A. (2024) *A Survey on Transformers in NLP with Focus on Efficiency*' arXiv preprint arXiv:2406.16893.

Ayodele, E. and Sodeinde, V.O. (2024) Student academic performance prediction system using ensemble algorithm', *Federal Polytechnic Ilaro Journal of Pure and Applied Sciences*, Vol. 6, pp.17–21.

Bashiri, H. and Naderi, H. (2024) 'Comprehensive review and comparative analysis of transformer models in sentiment analysis', *Knowledge and Information Systems*, Vol. 66, No. 1, pp.7305–7361.

Das, M. (2019) *Neural Methods towards Concept Discovery from Text via Knowledge Transfer*, The Ohio State University, 281 W Lane Ave, Columbus, OH 43210, USA.

Demszky, D. (2022) *Using Natural Language Processing to Support Student-Centered Education*, Stanford University.

Dessie, A.A. (2015) 'Teachers' practices of assessment for learning in science education at East Gojjam Preparatory Schools, Amhara Regional State, Ethiopia', *Signature*, Vol. 11, p.11.

Devlin, J., Chang, M-W., Lee, K. and Toutanova, K. (2018) *BERT: Bidirectional Encoder Representations from Transformers*, arXiv preprint arXiv:1810.04805.

Elmore, R. (1996) 'Getting to scale with good educational practice', *Harvard Educational Review*, Vol. 66, No. 5, pp.1–27.

Ewing, R. (2011) 'The arts and Australian education: realizing potential', *Human Dimensions of Ecological Restoration*, Vol. 11, No. 3, pp.347–361, DOI: 10.5822/978-1-61091-039-2_24.

Felzmann, H., Fosch-Villaronga, E., Lutz, C. and Tamò-Larrieux, A. (2020) 'Towards transparency by design for artificial intelligence', *Science and Engineering Ethics*, Vol. 26, No. 7, pp.3333–3361.

Garib, A. and Coffelt, T.A. (2024) 'Detecting the anomalies: exploring implications of qualitative research in identifying AI-generated text for AI-assisted composition instruction', *Computers and Composition*, Vol. 73, No. 3, p.102869.

Geethanjali, K.S. and Umashankar, N. (2011) 'Enhancing educational outcomes with explainable AI: bridging transparency and trust in learning systems', Vol. 20, No. 2, pp.347–361, DOI: 10.5822/978-1-61091-039-2_24.

Gillioz, A., Casas, J., Mugellini, E. and Abou Khaled, O. (2020) 'Overview of the transformer-based models for NLP tasks', *2020 15th Conference on Computer Science and Information Systems (FedCSIS)*, IEEE, pp.179–183.

Han, Z., Wu, J., Huang, C., Huang, Q. and Zhao, M. (2020) 'A review on sentiment discovery and analysis of educational big-data', *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Vol. 10, p.e1328.

Haque, A.B., Islam, A.N. and Mikalef, P. (2023) 'Explainable artificial intelligence (XAI) from a user perspective: a synthesis of prior literature and problematizing avenues for future research', *Technological Forecasting and Social Change*, Vol. 186, No. 6, p.122120.

Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., Scardapane, S., Spinelli, I., Mahmud, M. and Hussain, A. (2024) 'Interpreting black-box models: a review on explainable artificial intelligence', *Cognitive Computation*, Vol. 16, No. 1, pp.45–74.

Irvine, J. (2020) *Investigating the Impact of Lessons Based on Marzano's Theory of Learning on Student Attitude, Engagement, and Achievement in Grade 10 Academic Mathematics*, April 2022, Vol. 2020, No. 2, DOI: 10.59455/jomes.2020.2.3.

Kamalov, F., Calonge, D.S. and Gurrib, I. (2023) 'New era of artificial intelligence in education: towards a sustainable multifaceted revolution', *Sustainability*, Vol. 15, No. 2, p.12451.

Khosravi, H., Shum, S.B., Chen, G., Conati, C., Tsai, Y-S., Kay, J., Knight, S., Martinez-Maldonado, R., Sadiq, S. and Gašević, D. (2022) 'Explainable artificial intelligence in education', *Computers and Education: Artificial Intelligence*, Vol. 3, No. 1, p.100074.

King, F. (2014) 'Evaluating the impact of teacher professional development: an evidence-based framework', *Professional Development in Education*, Vol. 40, No. 4, pp.89–111.

Kuleto, V., Ilić, M., Dumangiu, M., Ranković, M., Martins, O.M., Păun, D. and Mihoreanu, L. (2021) 'Exploring opportunities and challenges of artificial intelligence and machine learning in higher education institutions', *Sustainability*, Vol. 13, No. 3, p.10424.

Kunter, M., Klusmann, U., Baumert, J., Richter, D., Voss, T. and Hachfeld, A. (2013) 'Professional competence of teachers: effects on instructional quality and student development', *Journal of Educational Psychology*, Vol. 105, No. 1, p.805.

Little, O., Goe, L. and Bell, C. (2009) *A Practical Guide to Evaluating Teacher Effectiveness*, National Comprehensive Center for Teacher Quality.

Liu, X., Xu, P., Wu, J., Yuan, J., Yang, Y., Zhou, Y., Liu, F., Guan, T., Wang, H. and Yu, T. (2024) *Large Language Models and Causal Inference in Collaboration: A Comprehensive Survey*, arXiv preprint arXiv:2403.09606.

Luan, H., Geczy, P., Lai, H., Gobert, J., Yang, S.J., Ogata, H., Baltes, J., Guerra, R., Li, P. and Tsai, C-C. (2020) 'Challenges and future directions of big data and artificial intelligence in education', *Frontiers in Psychology*, Vol. 11, No. 2, p.580820.

Lundberg, S. (2017) *A Unified Approach to Interpreting Model Predictions*, arXiv preprint arXiv:1705.07874.

Mcallister, C. (2023) *Using Quantitative Methods to Analyze Educational Interventions in Undergraduate Physics Teaching*, University of Glasgow.

Moody, A. (2023) *Summarizing Crowd Sourced Reviews With Natural Language Processing: A Case Study*, University of Central Arkansas.

Mylavarapu, G., Viswanathan, K.A. and Thomas, J. (2023) 'Context-aware automated quality assessment of textual data', *International Journal of Business Intelligence and Data Mining*, Vol. 22, No. 1, pp.451–469.

Nazar, M., Alam, M.M., Yafi, E. and Su'ud, M.M. (2021) 'A systematic review of human-computer interaction and explainable artificial intelligence in healthcare with artificial intelligence techniques', *IEEE Access*, Vol. 9, No. 3, pp.153316–153348.

Opdenakker, M-C. and Van Damme, J. (2006) 'Teacher characteristics and teaching styles as effectiveness enhancing factors of classroom practice', *Teaching and Teacher Education*, Vol. 22, No. 5, pp.1–21.

Ottley Herman, E. (2023) 'Teachers' perceptions of the influence of teacher evaluation process feedback on improving instructional practice', Vol. 20, No. 2, pp.30–45.

Patel, N.S. and Indurkhya, N. (2025) *The Rise of Intelligent Machines: A Multi-disciplinary Perspective from Industry and Impact on Higher Education*, CRC Press, Cheai, India.

Pedro, F., Subosa, M., Rivas, A. and Valverde, P. (2019) *Artificial Intelligence in Education: Challenges and Opportunities for Sustainable Development* [online] https://unesdoc.unesco.org/ark:/48223/pf0000366994.

Rafalski, S.H. (2015) 'Policy implications of a teacher evaluation system: the relationship of classroom observations, levels of feedback, and students achievement outcomes', March 2021, Vol. 25, No. 3, pp.193–217, DOI: 10.4324/9781003054726-9.

Rezapour, R. (2021) *From User-Generated Text to Insight Context-Aware Measurement of Social Impacts and Interactions Using Natural Language Processing*, University of Illinois at Urbana-Champaign [online] https://www.ideals.illinois.edu/items/123220.

Rock, M.L., Schumacker, R.E., Gregg, M., Howard, P.W., Gable, R.A. and Zigmond, N. (2014) 'How are they now? Longer term effects of e coaching through online bug-in-ear technology', *Teacher Education and Special Education*, Vol. 37, No. 3, pp.161–181.

Shaik, T., Tao, X., Li, Y., Dann, C., Mcdonald, J., Redmond, P. and Galligan, L. (2022) 'A review of the trends and challenges in adopting natural language processing methods for education feedback analysis', *IEEE Access*, Vol. 10, No. 1, pp.56720–56739.

Shinkfield, A.J. and Stufflebeam, D.L. (2012) *Teacher Evaluation: Guide to Effective Practice*, Vol. 5, No. 2, pp.81–172, Springer Science & Business Media, DOI: 10.1007/978-94-009-1796-5_3..

Sihotang, M., Utari, U. and Sihotang, T. (2022) 'The impact of evaluation methods on students' learning achievement in primary school mathematics education: a mixed-methods study', *Jurnal Ilmu Pendidikan Dan Humaniora*, Vol. 11, No. 3, pp.1–17.

Simonson, S.R., Earl, B. and Frary, M. (2022) 'Establishing a framework for assessing teaching effectiveness', *College Teaching*, Vol. 70, No. 6, pp.164–180.

Tanner, S., Mccloskey, A. and Miller, E. (2023) 'Destructive domains: rethinking teacher evaluation in the age of Charlotte Danielson', *International Journal of Qualitative Studies in Education*, Vol. 36, No. 3, pp.1876–1890.

Tian, Z., Sun, M., Liu, A., Sarkar, S. and Liu, J. (2024) *Enhancing Instructional Quality: Leveraging Computer-Assisted Textual Analysis to Generate In-Depth Insights from Educational Artifacts*, arXiv preprint arXiv:2403.03920.

Tiukhova, E., Vemuri, P., Flores, N.L., Islind, A.S., Óskarsdóttir, M., Poelmans, S., Baesens, B. and Snoeck, M. (2024) 'Explainable learning analytics: assessing the stability of student success prediction models by means of explainable AI', *Decision Support Systems*, Vol. 182, No. 5, p.114229.

Turarbek, A., Bektemesov, M., Ongarbayeva, A., Orazbayeva, A., Koishybekova, A. and Adetbekov, Y. (2023) 'Deep convolutional neural network for accurate prediction of seismic events', *International Journal of Advanced Computer Science and Applications*, Vol. 14, No. 3, pp.1100–1118.

Tuytens, M., Devos, G. and Vanblaere, B. (2020) 'An integral perspective on teacher evaluation: a review of empirical studies', *Educational Assessment, Evaluation and Accountability*, Vol. 32, No. 1, pp.153–183.

Upadhyay, P., Agarwal, R., Dhiman, S., Sarkar, A. and Chaturvedi, S. (2024) 'A comprehensive survey on answer generation methods using NLP', *Natural Language Processing Journal*, Vol. 8, No. 4, p.100088.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. and Polosukhin, I. (2017) *Attention Is All You Need (Nips)*, arXiv preprint arXiv:1706.03762.

Verdonck, T., Baesens, B., Óskarsdóttir, M. and van den Broucke, S. (2024) 'Special issue on feature engineering editorial', *Machine Learning*, Vol. 113, No. 6, pp.3917–3928.

Wibawa, A.P. and Kurniawan, F. (2024) 'A survey of text summarization: techniques, evaluation and challenges', *Natural Language Processing Journal*, Vol. 7, No. 3, p.100070.

Wine, D. (2016) 'Using student feedback to enhance teacher evaluation', *Teaching Exceptional Children*, Vol. 36, No. 6, pp.64–69, DOI: 10.1177/004005990403600608.

Wise, A.E., Darling-Hammond, L., Mclaughlin, M.W. and Bernstein, H.T. (1985) 'Teacher evaluation: a study of effective practices', *The Elementary School Journal*, Vol. 86, No. 3, pp.61–121.

Yadav, D.S. (2024) 'Transformative trends in education with advanced technologies: exploring the intersection of IoT, AI', *Applications of Artificial Intelligence in the Internet of Things: Today's and Tomorrow's World*, Vol. 12, No. 4, pp.387–395, DOI: 10.1201/9781032686745-23.

Zangari, A., Marcuzzo, M., Schiavinato, M., Rizzo, M., Gasparetto, A. and Albarelli, A. (2023) 'Hierarchical text classification: a review of current research', *Expert Systems with Applications*, Vol. 224, No. 3, p.119984, DOI: 10.1016/j.eswa.2023.119984.

Zhang, H. and Shafiq, M.O. (2024) 'Survey of transformers and towards ensemble learning using transformers for natural language processing', *Journal of Big Data*, Vol. 11, No. 3, p.25.

Zhang, Z. (2024) 'Empowering language assessment and education with natural language processing: a focus on cloze tests', *Journal of Natural Language Processing*, Vol. 31, No. 2, pp.328–348, DOI: 10.5715/jnlp.31.328.

Zhou, J. and YE, J-M. (2023) 'Sentiment analysis in education research: a review of journal publications', *Interactive Learning Environments*, Vol. 31, No. 5, pp.1252–1264.