CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

The objective of this chapter is to furnish theoretical underpinnings and methodological justifications for the construction of student satisfaction prediction models.Initially, a comprehensive review is conducted on the research evolution and modeling trajectories of teaching satisfaction assessment. This includes a meticulous analysis of the distinct values and complementary relationships of structured and unstructured data within the context of teaching feedback.Subsequently, the research implications of unstructured text feedback are explored. Particular attention is paid to the challenges it poses in terms of language characteristics and processing complexity. Additionally, an overview of the current state of applications of sentiment analysis in educational settings is presented.Following this, the primary strategies and research advancements in the integrated modeling of structured scores and text emotion features are summarized. An analytical framework for model interpretability is introduced, with an emphasis on its practical significance in facilitating teaching decision-making.The content elaborated above will lay a theoretical foundation for the subsequent empirical investigations and model implementations.

2.2 Research Status of Teaching Satisfaction Assessment

Within the framework of the higher education quality assurance system, the evaluation of students' satisfaction with teaching has, for an extended period, been regarded as a crucial instrument for gauging teaching effectiveness, refining course design, and enhancing teaching methodologies (Li et al., 2025). The measurement of

student satisfaction not only bears a close relationship to the teaching improvement mechanisms within institutions of higher learning but also exerts a profound influence on aspects such as teacher evaluation, course accreditation, and the formulation of educational policies (Quansah et al., 2024). Consequently, the scientific assessment of teaching satisfaction has emerged as one of the central topics in educational research and management practice.

Traditionally, the assessment of teaching satisfaction predominantly relied on structured questionnaire instruments, particularly the rating system based on the Likert scale. Students were typically required to rate various dimensions, including the instructor's teaching proficiency, course content, teaching demeanor, difficulty level setting, and course scheduling. The merit of this approach lies in its well-defined structure, facilitating statistical analysis and quantification. It enables the collection of a substantial amount of data within a short span, rendering it suitable for large-scale educational assessments. For example, numerous universities utilize the "course evaluation form" completed by students at the conclusion of a course as a vital criterion for the annual performance evaluation of teachers.

But while the enormous power of standardized scales is standardization and comparability, their capacity for reflecting students' true experiences and individual sentiments is relatively constrained. Heffernan (2022) assumed that the rating behavior is prone to interference from extraneous variables such as students' psychological expectations, course grades, instructor gender, and language expression, thereby introducing clear-cut subjective biases. Especially when the learning material is difficult or the time to assess is limited, students are challenged in fully articulating their entire thoughts regarding the course with a single rating item. Further, methodical ratings tend to overlook intangible variables such as the feeling of course interactions, differences in learning emotions, individuality, and cultural background.

In recent years, the limitations of controlled assessment to measure teaching satisfaction have been universally acknowledged by researchers, and they started to think of open-ended text feedback as an added source of data. Open-ended text gives students the chance to say something about course content, instructor teaching strategies, classroom climate, and effectiveness of teaching without restriction in their own words. The study by Kastrati et al. (2021) found that students' text comments contain rich subjective feelings and rich proposals, offering richer feedback to teaching administrators than ratings alone.

Overall, teacher satisfaction measurement has transitioned from "structured

ratings" to "multi-source integrated feedback". The prevalent approach in current times is to use natural language processing (NLP) techniques and machine learning algorithms to extract insights from open text comments through automated sentiment analysis and then integrate the findings with rating data to develop evaluation models with enhanced explanatory and predictive capability. This integrative model not only operates to offset the blind spots of mainstream rating systems but also provides a more astute, more sensitive, and student-centered foundation for the assessment of teaching quality in universities.

2.3 The Value and Challenges of Unstructured Text Feedback

In the process of development of the current higher education quality assurance system towards perfection, the traditional method of teaching evaluation based on formally graded scales has failed to be sufficient in fully revealing deeply ingrained beliefs and specific suggestions of students. Therefore, increasingly, universities are adding open-ended comments to their standard questionnaires as an addendum form to obtain more personalized and more elaborate assessments of teaching quality (Tripathi et al., 2024). In comparison with closed-ended quantitative scales such as "overall satisfaction score" and "teaching attitude score", text comments allow students to provide their personal views on courses, teachers, pedagogy, and learning experience freely, with higher information density and personalization.

2.3.1 The Value and Functional Positioning of Text Feedback

Contrary to more traditional rating items like "teaching content satisfaction" and "teaching attitude score," unstructured text permits students to freely express and convey their own perceptions of teaching process subtleties. Such comments will typically entail multi-dimensional information across domains like course pacing, knowledge depth, teaching styles, interaction frequencies, and assessment methods. They possess more semantic richness and emotional expressiveness (Uppalapati et al., 2025). This reservoir of data can be utilized to identify curriculum issues, modify

instruction strategies, and even shed light on the operational limitations of teaching support systems. In cases where the mean ratings fall in the "4.0 - 4.5" range, qualitative text often assumes a significant discernment role for teaching administrators to separate "superficial high scores" from "substantive high-performance teaching" (Naranjo Retamal et al., 2024).

Research has revealed that written student feedback frequently over surface - level judgments and are imbued with both "emotional stances" and "suggestive inclinations." For instance, the sentence such as "The professor describes subject matter in detail, but the homework load is too heavy" not only mentions positive things but also proposes improvements (Mondal & Karri, 2025). Similarly, open-ended feedback have also been seen to be a major source of information that guides course revision, instructor evaluation, and policy refinement. A majority of universities have formally incorporated it into their pedagogical development systems as part of "curricular big data" government.

Also, the sentiment markers present in comment texts have been proven empirically to more closely align with actual course experience for students. For example, Pavankumar et al. (2024) determined that students will offer more incongruent comments regarding course design in critical texts, while positive texts are primarily aimed at instructors' personality traits and communication styles. Accordingly, within the area of educational data mining (EDM), systematic text sentiment extraction and semantic label modeling are now a universally accepted international research opinion.

Evidence from research shows that text comments made by students can capture a lot of sentiment hints and nuanced suggestions, bringing subtle aspects not readily evident from quantitative ratings to light. These may encompass such facets as classroom interactional dynamics, linguistic ability of teachers, course sequencing, and level of difficulty matching (Wang et al., 2021). Most especially in cases where course ratings reveal a high level of consensus, text comments may also serve as a "differentiating factor," providing teaching administrators with deeper and more actionable insight.Thus, numerous universities have integrated text comments into the data - support frameworks for teacher performance evaluations, curricular optimizations, and pedagogical reforms.

2.3.2 Core Challenges and Practical Limitations Encountered

Despite the fact that unstructured text feedback offers information of a higher dimensionality, its practical implementation is fraught with numerous challenges, which are principally manifested in the following aspects:

(1) Substantial Variability in Language Expression and Highly Heterogeneous Styles

Given the marked disparities among students in terms of language proficiency, expressive norms, and cultural backgrounds, the text often incorporates colloquialisms, slang, abbreviations (e.g., "prof" for "professor"), and may even contain spelling and grammatical inaccuracies (Li et al., 2024). These elements impede the efficacy of traditional dictionary - or rule - based semantic processing methodologies, thereby undermining the accuracy of sentiment analysis and keyword extraction.

(2) Complex Emotional Polarity and Ambiguous Opinion Orientations

Distinct from corpora with "explicit emotional targets," such as e - commerce reviews, emotional expressions within educational settings tend to be more rational and nuanced, frequently featuring co - existent multiple emotions or semantic shifts. For example, in the statement "The instructor lectures at a rapid pace, yet he is truly dedicated," the fast pace represents a negative sentiment, while the overall sentiment leans positive. Accurate identification thus necessitates context - based modeling (Shaik, 2022).

(3) Inconsistent Information Density and Extensive Content Scope

There are notable discrepancies in the information density and the length of meaningful information across different comments. Some comments are composed of mere words (e.g., "Great!" or "Poor."), containing minimal information, while others encompass extensive multi - paragraph texts covering aspects such as course progression, examination schedules, and methods of seeking clarification. The broad thematic scope renders the standardization of content structure a formidable task (Liet al, 2023).

(4) Inefficient Processing and Pronounced Subjective Bias

In the context of actual teaching evaluations, manually processing large volumes of open - ended text is not only time - consuming but also challenging in terms of structural analysis. Even when reviewed by seasoned educators or administrators, issues such as "interpretive arbitrariness" and "emotional projection" can arise. That is, the personal teaching experiences or preferences of the reviewers may influence the consistency of judgment. Consequently, relying solely on manual methods to handle student feedback is no longer viable in meeting the assessment demands of the contemporary teaching informatization landscape (Mondal & Karri, 2025).

2.3.3 Distinctive Attributes of Educational Evaluation Texts

The unstructured text within teaching feedback diverges significantly from other forms of user - generated content (e.g., product reviews, e - commerce ratings, and social media posts) in terms of expression style, content architecture, and pragmatic strategies. This divergence not only complicates the extraction of emotional and semantic information but also demands a higher level of adaptability in subsequent analytical approaches. Specifically, educational texts exhibit the following unique characteristics:

(1) Restrained Expressive Tone and Implicit Emotional Undercurrents

Within the realm of higher education, students, out of respect for the teacher - student relationship and teaching authority, often eschew the use of highly emotive or extreme language, opting instead for more circumspect and indirect modes of expression. While this form of emotional conveyance is subjectively authentic and valid, it may be misconstrued or overlooked during the analysis and modeling phases due to the dearth of explicit emotional markers.

(2) Frequent Incursion of Educational Terminology and Academic Jargon

Students tend to incorporate specialized vocabulary and scholarly jargon such as course titles, instruction units, concept understanding, and assignment categories in the comments. This demands a greater degree of comprehension from the analysis

system. Curiously, insofar as inter-disciplinary courses are involved, such vocabulary tends to be very context-bound. In the absence of suitable knowledge graphs or semantic recognition abilities, information fragmentation and semantic misinterpretation might ensue.

(3) arying and Intersecting Content Organization

A single open - ended comment may cover multiple aspects, including the structure of the course content, instructional methods, classroom interaction patterns, and the magnitude of the learning burden, hence creating an extremely complex semantic model. This "multi - dimensional and multi - thematic" narrative style transcends the limits of conventional single - dimensional categorization methods, raising complexity in extracting information and classifying emotions by orders of magnitude.

(4) Heterogeneous Subjective Motivations and Divergent Evaluation Tendencies

Motives behind student feedback are not monolithic. They may be driven by genuine experiential feedback or by external determinants such as performance, difficulty of course, or even learning preferences concerning teaching. This can result in evident emotional bias or selective attention within the comments. These differences in motives have the ability to influence the validity and reliability of teaching feedback outcomes. If left uncontrolled, these can lead to skewed conclusions of evaluations.

In essence, educational unstructured comments not only exhibit greater complexity in linguistic form but are also marked by greater unpredictability and variability in cognitive structure as well as emotional expression. Consequently, automatic information extraction as well as emotional cue extraction from such unstructured comments has become a key research area within the domain of Educational Data Mining (EDM) over the past few years.

2.4 Sentiment Analysis Techniques

Within the context of student response data for instruction, the open-ended student comments typically provide a rich source of subjective impressions, attitudinal inclinations, and evaluation biases. Successful elicitation of emotional traits and their transference to modeling-compatible variables is hence a key technical foundation for modeling student satisfaction. Sentiment analysis, one of the core tasks in the domain of Natural Language Processing (NLP), is primarily concerned with identifying the subjective polarity (i.e., positive, neutral, or negative) and emotional intensity expressed through text. The technology has found extensive application and continuous evolution in educational data mining, course feedback analysis, and intelligent teaching evaluation systems during the past couple of years.

2.4.1 Classification of Sentiment Analysis Approaches

The traditional methods of sentiment analysis can be categorized into three broad categories: dictionary - based approaches, machine - learning - based methods, and deep - learning - based systems.

2.4.1.1 Lexicon - Driven Approaches

Lexicon-based sentiment analysis methods estimate polarity and intensity scores using the VADER, TextBlob, and SentiWordNet tools. These tools estimate the overall sentiment through the summation of word-level scores. Experiments have proven that the performance of different lexicons varies with context: TextBlob is better suited for neutral texts, while VADER suits short or colloquial texts. Some studies, e.g., Mujahid et al. (2021), combine more than one lexicon to provide greater flexibility in emotional cases.

This approach is valued for its simplicity, effectiveness, and lack of sensitivity to labeled data, and is thus suitable for small-scale educational feedback analysis and sentiment labeling in initial stages. It is not contextual, and performs badly with negations, sarcasm, and nuanced tone changes. Therefore, recent studies

suggest combining lexicon-based with machine learning or deep learning models to bypass these limitations.

2.4.1.2 Traditional Machine Learning Approaches

These approaches typically cast the emotion classification task as a supervised learning problem. They transform annotated training data into feature vectors (e.g., TF-IDF, n-gram) and train emotion classifiers, e.g., Support Vector Machines (SVM), Naive Bayes, and Random Forests. Experiments showed that this method exhibits comparatively steady performance in the task of identifying the prevailing emotions of teaching reviews, particularly in multi - category emotion classification tasks on medium - sized corpora.

Classic machine learning methods offer good modeling, good interpretability of features, and suitability for small to medium-sized sentiment data sets. They provide good education text classification and accommodative integration with structured scores in the form of multi-class and probability outputs. Having low computational needs, they are suitable for speedy deployment in education systems.But these methods lack rich contextual knowledge and also struggle to handle negation, semantic inversion, and complex evaluative sentiment. They are strongly reliant on manual feature engineering and generalize poorly across domains. Thus, while suitable as baselines or lightweight models, traditional machine learning methods must be combined with advanced semantic modeling techniques to handle advanced sentiment hierarchies.

(1) Naive Bayes Model

The Naive Bayes (NB) model, grounded in Bayes' theorem, is a probabilistic classification model that has found extensive application in the task of text sentiment recognition. Its central tenet posits that, conditional on a given class, the input features are mutually independent.

$$\hat{y} = \arg\max_{c \in C} P(c) \prod_{i=1}^{n} P(x_i \mid c)$$

Among them:

represents the predicted emotion category.

$P(c)$ is the prior probability of the category.

$P(x_i \mid c)$ is the conditional probability of the i th feature (such as a certain keyword) occurring under the category c.

In emotion analysis, the feature x_i usually indicates whether a word or phrase appears. Therefore, it is applicable to the bag - of - words (BoW) model or the TF - IDF vector space representation. Naive Bayes can be used for binary classification (positive/negative) or multi - classification (such as emotion levels) tasks. It is especially suitable for constructing a baseline system for emotion analysis or a preliminary model for small - and medium - scale educational texts.

Notwithstanding the fact that this “naive” assumption does not invariably hold true in the realm of natural language processing, empirical studies have demonstrated that the Naive Bayes model exhibits remarkable stability in sentiment classification tasks characterized by high - dimensional text and moderately sized samples. Moreover, it demonstrates robust generalization capabilities when dealing with small - scale datasets.

(2) Support Vector Machine (SVM) Model

The Support Vector Machine (SVM) is a discriminative classification model founded on the principle of maximum margin. Its objective is to construct an optimal hyperplane within the feature space, thereby maximizing the margin separating

different classes.

In the context of sentiment analysis tasks, particularly within the high - dimensional and sparse space that emerges following text vectorization, the SVM, leveraging its formidable boundary discrimination capabilities, has established itself as one of the classical approaches for text emotion classification.

The optimization objective function under the linearly separable scenario is presented as follows:

$$\min_{\mathbf{w},\, b} \frac{1}{2} \| \mathbf{w} \|^{2} \quad \text{subject to } y_i(\mathbf{w}^T x_i + b) \geq 1, \ \forall i$$

Among them:

$\mathbf{w}$ is the normal vector of the decision hyperplane;

b is the bias term;

x_i is the input feature vector;

$y_i \in \{-1, 1\}$ is the corresponding emotion category label.

For non - linearly separable cases, SVM can map the data to a higher - dimensional feature space through the kernel function $K(x_i, x_j)$ to achieve linear separability. Common kernel functions include the Gaussian kernel (RBF), polynomial kernel, etc. In emotion analysis, SVM is usually applied to binary classification tasks and is particularly effective in judging the emotion polarity (positive/negative). It is suitable for educational review texts with clear structures and high corpus consistency.

(3) Random Forest (RF)

Random Forest is an ensemble learning model that enhances the stability and generalization ability of the model by constructing a substantial number of Decision Trees and integrating the prediction results of each tree through a voting mechanism during the classification process.

In the domain of sentiment analysis, RF proves effective in handling non - linear feature relationships. It is particularly well - suited for classification problems where both structured rating data and emotion vectors (such as emotion intensity and keyword TF - IDF) are incorporated.

The partitioning process of a single decision tree adheres to the principle of maximum information gain or minimum Gini coefficient:

$$Gini(D) = 1 - \sum_{k=1}^{K} p_k^2$$

Among them, p_k represents the proportion of samples of the k-th category in the current sample set. RF reduces the risk of overfitting through multi - tree integration and can evaluate the importance of each feature (Feature Importance). This has an interpretive advantage for sentiment analysis in the context of teaching feedback.

(4) Logistic Regression (LR)

Logistic Regression (LR) is a linear model employed to address binary classification problems. The underlying concept of LR is to utilize the log - odds function (logit function) to model the probability of a particular class occurring. The mathematical form of the model is presented as follows:

$$P(y = 1 \mid \mathbf{x}) = \frac{1}{1 + e^{-(\mathbf{w}^T \mathbf{x} + b)}}$$

Among them:

$\mathbf{x}$ denotes the feature vector of the text or rating;

$\mathbf{w}$ represents the feature weight;

b stands for the bias.

The output is the probability value of belonging to the "positive emotion" category.

In the education emotion modeling context, logistic regression is naturally adapted to the simultaneous modeling of rating items and text features. The predicted probability values of the model are very interpretable, which makes it convenient for subsequent building of interventions or recommendations for teaching feedback.

2.4.1.3 Deep Learning Approaches

Deep learning methods have been ruling sentiment analysis studies over the past few years owing to their superior context - modeling and semantic - representation abilities. Typical models like Long Short - Term Memory (LSTM), Convolutional Neural Network (CNN), and Bidirectional Encoder Representations from Transformers (BERT) can automatically identify emotional expression patterns in student evaluation texts. These models are able to capture emotional shifts and latent polarities underlying sophisticated grammatical constructs. They exhibit especially good performance in the presence of negation, irony, and lengthy texts.

In particular, large pre-trained models such as BERT, which builds semantic relations on the basis of large-scale corpora, can readily model long-distance

dependencies as well as context-dependent emotional changes. It has been empirically established that such models outperform classical counterparts overwhelmingly (Anwar et al., 2023).

Yet, these deep learning techniques have their own set of challenges as well. They are low in interpretability and high in training cost, with substantial dependence on huge datasets. In the educational context, deep learning models are particularly apt to handle large-scale corpora of teaching feedback with complex text structure. Alternatively, they can be used as feature extractors and combined with structured scoring data. This strategy reconciles the level of semantic modeling and interpretability needed in academic environments. Therefore, deep learning's importance in predicting teaching satisfaction is not in substituting conventional methods but in the manner the model learns emotional expression in language, ultimately offering great help to multimodal fusion modeling.

2.5 Research on the Fusion Modeling of Structured and Unstructured Data

Within Educational Data Mining (EDM) and Learning Analytics, fusion modeling has become an increasingly core research approach to continue improving the predictive power of student satisfaction. Early in the practice of teaching evaluation, modeling depended chiefly on tabular data, such as global course ratings, instructor performance ratings, and course difficulty ratings. This kind of data, which is defined by its standardization and receptiveness to quantification, has remained at the forefront of conventional teaching satisfaction models.

With open-ended feedback systems being implemented, students' textualized emotional stances and subjective experiences have become ever more the central part of the course feedback system. Unstructured text information can potentially reveal unobvious aspects hard to measure by numeric ratings alone, e.g., classroom atmosphere, teaching speaking style, and timeliness of teaching feedback.Thus, the integration of text-based sentiment features with structured rating data not only allows for improved predictive model performance but also their interpretability and generalizability.

2.5.1 Classification of Fusion Strategies

According to the differences between information fusion levels and process methods, current fusion modeling techniques can be mainly divided into the following three categories:

(1) Feature - level Fusion (Early Fusion)

Feature - level fusion includes concatenation of input structured data (for example, ratings, students' course willingness) and text sentiment features (for example, emotion polarity, Term Frequency - Inverse Document Frequency (TF - IDF) vectors, and VADER sentiment scores) at the preprocessing step. This leads to a merged input vector, which is then used for training a combined model. This approach is easy and effective and therefore highly appropriate for small-sample modeling problems.For instance, Deshpande et al. (2025) combined the emotion intensity scores derived from student feedback texts and rating and fed them into a Random Forest model to forecast course satisfaction. Their approach achieved an F1 score of over 91%.

(2) Decision - level Fusion (Late Fusion)

This approach involves building several sub - models, each of which is dedicated to either structured ratings or unstructured text, to create independent prediction mechanisms. The prediction outcomes of the individual models are then combined finally through ensemble methods, such as voting schemes or weighted averages. This approach is well suited to the case of complicated model structures or large variance in data dimensions, with greater flexibility.For example, Baqach & Battou (2024) employed an LSTM for text emotion analysis in the course feedback analysis of MOOC courses. They went a step further to combine the model's output with that of a rating-based model for collective decision-making, which significantly enhanced the stability and the prediction accuracy.

(3) Hybrid - level Fusion (Hybrid Fusion)

The hybrid fusion approach combines the strengths of both feature - level and

decision - level fusion. Integration of data happens at the feature input level as well as model output combination in the output layer. Although it has higher computational complexity requirements, this approach is highly appropriate to utilize when building high - accuracy prediction systems. It is able to investigate the cross-modal synergistic relations of information and thus is the best option for building high-precision satisfaction evaluation systems.

2.5.2 Construction of Key Features in Educational Data Fusion

The essence of attaining a high - performance fusion model lies in the feature complementarity and semantic connection between the two types of data. In practical implementation, it is essential to meticulously design the extraction strategies for structured and unstructured features, aiming to ensure the consistency of model input and discriminative power.

Structured data features consist of:

Overall Rating (Overall Score)

Difficulty Level (Course Difficulty)

Would Take Again (Willingness to Re - enroll), etc.

Text - based emotional features generally encompass:

Emotional polarity classification (positive, neutral, or negative)

Emotional intensity scoring (e.g., the VADER compound score)

Representation of keyword weights (vector representations such as TF - IDF

and Word2Vec)

Density of educational keywords (e.g., the frequency of emotion - related words like “organized,” “helpful,” “unclear”)

The precise extraction and encoding of these features not only influence the model's classification boundaries but also determine its capacity to integratively interpret students' subjective attitudes and objective scores. Thus, the integrity and consistency of emotional feature engineering are crucial prerequisites in the construction process of the fusion model.

2.5.3 Model Comparison and Application Achievements

Numerous studies have indicated that multimodal models integrating structured and unstructured data outperform single - modal models in the task of predicting student satisfaction. Imran and Baig (2022) discovered that when comparing structured scores with TextBlob - based emotion classification, the fusion model exhibited an accuracy 9.2% higher than that of the model relying solely on structured data. **Zyout** et al. (2024) employed a Long Short - Term Memory (LSTM) network to process students' course review texts and constructed a hybrid prediction model by incorporating rating factors. This model demonstrated superior performance on MOOC platforms, and its interpretive analysis was more readily accepted by educational administrators.

Despite the evident advantages of fusion modeling both theoretically and experimentally, several issues remain in practical applications:

Inconsistent data distribution: The number of text comments often fails to correspond one - to - one with rating records, resulting in missing training samples and misaligned features.

Ambiguous semantic label mapping: There is a lack of a strict correspondence between text emotion labels and rating scales, which affects the

standardization of model annotation.

Severe rating polarization: Structured ratings frequently concentrate in the high - score range (e.g., 4.0–4.5), which can lead to an imbalance in the training of regression models.

Poor interpretability of deep models: Although some fusion models (e.g., BERT + rating embedding) yield excellent results, their decision - making paths are difficult to trace, restricting their direct interpretation and application in educational management.

Consequently, current research has shifted from solely improving the accuracy of fusion models to enhancing their transparency, deployability, and operationality in actual teaching decision - making. Techniques such as model visualization and feature contribution analysis (e.g., SHAP) have been employed to enhance their managerial interpretability.

2.6 Model Interpretability and Its Application in Teaching Analysis

In the course of research on predicting student satisfaction by integrating structured scores and text - based emotional features, model interpretability has emerged as a critical dimension that cannot be overlooked. Although advanced models such as Random Forest, XGBoost, Long Short - Term Memory (LSTM), and Bidirectional Encoder Representations from Transformers (BERT) have demonstrated remarkable performance in emotion classification and satisfaction prediction tasks, their “black - box” nature renders the internal decision - making process of the model difficult to comprehend and trace.This lack of interpretability is particularly pronounced in the context of educational evaluation. Given that the outcomes often directly influence teacher evaluations, course adjustments, and even policy - making, which are of utmost sensitivity and significance in the educational domain (Kovalerchuk, 2024).

In the design of actual teaching feedback systems, a model is not only required to make accurate predictions but also to have the ability to clearly explain to

teachers, teaching administrators, and educational decision - makers “how it arrived at a particular judgment.” For instance, school management needs to understand “why the model deems a teacher's satisfaction to be low”; teachers themselves wish to know “which keywords or features have influenced students' judgments of the course”; and the student feedback mechanism should also provide understandable and reasonable explanations for the results.To this end, the incorporation of Explainable Artificial Intelligence (XAI) techniques not only enhances the transparency of the model but also bolsters its credibility and guiding significance in practical applications (Freiesleben& Molnar, 2024).

2.6.1 Classification of Model Interpretability Approaches

The current mainstream model interpretability methods can be broadly categorized into two types: global interpretability and local interpretability. These two types respectively concentrate on visualizing the overall patterns of the model and the decision - making paths for individual samples:

(1) Global Interpretability Methods

Global methods aim to uncover the contributions of features and decision - making tendencies of the entire model. Representative examples are as follows:

Feature Importance: This method assesses the average influence of each input variable in the model's prediction process. It is extensively applied in tree - based models such as Random Forest and XGBoost.

Partial Dependence Plot (PDP): PDP illustrates the marginal impact of a variable across different value ranges on the prediction outcomes, which is useful for exploring non - linear relationships.

(2) Local Interpretability Methods

Local interpretability methods are concerned with the rationale behind the model's judgment for a specific input sample. They help to explain, for example, "why a particular student is predicted to be dissatisfied". Notable methods include:

LIME (Local Interpretable Model - agnostic Explanations): LIME constructs a local linear model in the vicinity of the original model. By mimicking the prediction logic of the original black - box model, it analyzes the positive and negative influences of each feature within the current sample.

SHAP, which stands for SHapley Additive exPlanations, is based on the Shapley value from game theory, which assigns a "fair contribution" to each input feature for the prediction made by the model. This approach has strong mathematical interpretability and improves scalability.

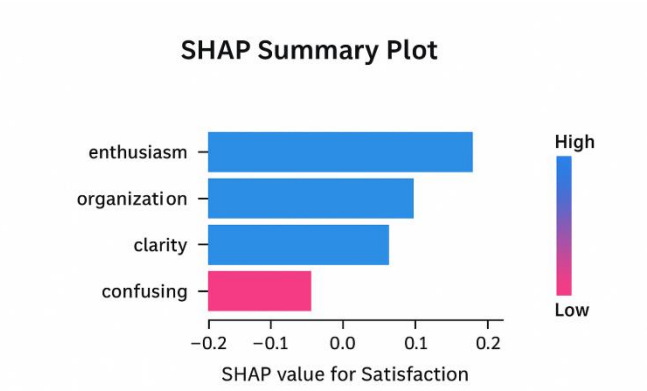


Figure 2.6.1: SHAP Summary Plot

These interpretability techniques can be applied to examine both text and structured (i.e., rating items) sentiment features. They assist in various dimensions of applications, such as model auditing, intervention guidance, and education policy refinement.

2.6.2 Application Instances of SHAP and LIME in Educational Research

In recent years, interpretability techniques like SHAP and LIME have been widely applied in educational data modeling studies to facilitate model transparency and trustworthiness to users. For instance, Teles et al.（2025） applied the SHAP technique in the study of Coursera teaching data. Based on their results, features such as "clarity", "organization", and "enthusiasm" had high positive contributions to the predictions made by the model. Conversely, adjectives "confusing", "monotonous", and "unstructured" decreased the predicted satisfaction values to a great extent. These interpretation findings not only enhance the explanatory power of the model but also provide clear teaching improvement recommendations.

Zhou and Wang（2022） employed LIME technology to carry out the local interpretation of the negative classification outcomes of specific comment texts. They identified phrases such as "poor explanation" and "no engagement" as the primary factors influencing the model's classification of negative emotions. Through the visualization of individual sample explanations, educational administrators can directly identify the core causes of student dissatisfaction and subsequently formulate personalized adjustment measures.

Furthermore, the SHAP method can be used to comparatively analyze the relative contributions of different types of features within fusion models. Research has shown that in some courses, the SHAP value of the "difficulty" rating is higher than that of the sentiment score, suggesting that structured rating items play a dominant role in determining the prediction results. In contrast, in other courses where students' text sentiment polarity is more pronounced, the weight of text features becomes relatively higher. This discovery implies that fusion models do not adopt a uniform decision - making strategy for all samples but rather dynamically balance the interaction between rating and sentiment factors.

Although SHAP and LIME have achieved initial applications in educational research, certain limitations still exist. These include insufficient stability of the interpretive results, low efficiency in interpreting deep models, and a lack of adaptability to multilingual educational data. Consequently, future research could consider integrating attention - based interpretability mechanisms or training visual models with built - in interpretable structures to further enhance the practicality and credibility of educational fusion models.

2.7 Sources and Sample Composition of Teaching Evaluation Data

In data - driven educational research, the data generation method, collection approach, and structural characteristics directly determine the analytical dimensions and modeling boundaries of the research. Especially in the process of constructing a teaching satisfaction prediction model, the dataset employed by researchers should not only possess sufficient representativeness and structural integrity but also reflect the subjectivity and emotional expressiveness of students' genuine feedback.

2.7.1 Data Generation

The data used in this study is sourced from RateMyProfessor.com, one of the largest college course evaluation platforms in North America. Since its inception in 2001, the platform has amassed over a million student evaluations of college courses and instructors, covering thousands of universities and dozens of academic disciplines. After completing a course, students can voluntarily log in to the platform to fill out structured rating items related to a specific instructor and provide open - ended text feedback. The data structure mainly consists of:

Structured Scoring Indicators: These include the overall course rating (Overall Rating), course difficulty (Difficulty), and the likelihood of taking the instructor's course again (Would Take Again).

Unstructured Text Feedback: Students' subjective descriptions regarding aspects such as the instructor's teaching style, course design, and learning experiences.

Other Auxiliary Information: Such as comment timestamps and course tags.

This data generation approach not only preserves the subjectivity of students' free expression but also integrates the platform's standardized scoring mechanism. As a result, it forms a "structured and unstructured" dual - data structure, which is highly compatible with the research objectives of multimodal fusion models.

2.7.2 Data Collection

In consideration of the scientific and ethical aspects of the research, this study adhered to the principles of open access and minimal intrusion during the data collection process. Using automated script tools, we selectively collected evaluation data published by certain higher education institutions and their instructors on public platforms from 2019 to 2024. During the collection process, only non - sensitive information fields were retrieved to ensure that no personal information of registered users or private platform resources were involved. The data collection process mainly involved the following steps:

Sample Screening Criteria: Teachers with more than 10 evaluations were selected as data subjects to ensure that the samples had a solid foundation for statistical analysis.

Field Structure Standardization: The three main rating items, namely "Overall Rating", "Difficulty", and "Would Take Again", along with the review text, were collected and standardized.

Time - Window Control: The collection time was restricted to the past five years to reflect the timeliness and dynamism of teaching feedback.

Text Anonymization: Potential sensitive information such as instructor names and course numbers in the comments was removed to safeguard data anonymity.

Sample Distribution Regulation: During the sampling process, the proportion of positive and negative sentiment comments and the distribution across different subject categories were carefully controlled to avoid label imbalance during the modeling process.

Through the above - mentioned procedures, a dataset of valid samples meeting the requirements was ultimately established, laying the groundwork for subsequent model training and performance evaluation.

Preliminary exploratory analysis reveals that the dataset exhibits a right - skewed distribution in the rating dimension, with approximately 65% of the rating records falling at 4 or above, indicating a typical "courtesy bias" characteristic. In the text part, the sentiment distribution is more dispersed, with positive comments being more numerous, but negative comments containing a higher information density. These structural characteristics provide natural support for multimodal modeling and sentiment polarity analysis.

2.8 Model Comparison and Selection Rationale

In order to accurately predict student satisfaction and provide interpretable teaching analysis, this study takes into account the adaptability and performance disparities of different modeling strategies during the model design phase. The modeling methods commonly employed in the current field of educational data mining can be categorized into three groups: ① traditional machine learning models, ② deep learning models, and ③ multimodal fusion models. Each approach has its own merits in processing structured scoring data and unstructured text sentiment data. This paper will conduct a comprehensive analysis of each model type from four dimensions: prediction accuracy, feature representation ability, interpretability, and adaptability, aiming to clarify the ultimate modeling approach to be adopted.

2.8.1 Traditional Machine Learning Models: Efficient and Stable, but Primarily Suited for Structured Data

Traditional machine learning models, such as Random Forest, Logistic Regression, and Support Vector Machines, are extensively utilized in educational scoring modeling due to their stability and remarkable generalization capabilities. Take Random Forest (RF) as an example. By constructing multiple decision trees and integrating them through a voting mechanism, it can effectively handle the multi - dimensional interactions within scoring data and demonstrates a strong resistance to overfitting. In the sentiment classification of 5,000 teacher evaluations by Deshpande et al. (2025), the RF model achieved an accuracy of 91% and a precision of 94%,

thereby validating its adaptability to structured teaching data.

However, these models encounter significant limitations when dealing with natural language text. Although techniques like TF - IDF can be employed to transform text into vectors, traditional models lack the capacity to model context semantics and syntactic structures. Consequently, they are unable to fully capture students' emotional inclinations, thereby diminishing the integrity of satisfaction modeling.

2.8.2 Deep Learning Models: High Expressive Power but with Prominent Black - Box Characteristics

In recent years, deep learning technologies have been widely applied in sentiment analysis and educational text modeling. Particularly, models such as Long Short - Term Memory (LSTM) networks and pre - trained language models like BERT have demonstrated exceptional performance in natural language understanding tasks. LSTM can effectively capture long - distance dependencies among words in a sentence, making it suitable for processing the progressive and comparative emotional expressions commonly found in educational reviews. Kastrati et al. (2021) discovered that when analyzing student feedback on Coursera, LSTM outperformed traditional methods significantly in terms of accuracy and recall.

BERT (Bidirectional Encoder Representations from Transformers), through bidirectional context modeling, further enhances semantic comprehension capabilities. The BERT - CNN hybrid sentiment model developed by Baqach & Battou (2024) achieved an F1 score of over 92% on data from multiple MOOC courses, highlighting the potential of deep models in extracting fine - grained emotional features.

Nonetheless, these models also exhibit notable drawbacks:

They are highly reliant on the quantity and quality of training samples.

The reasoning speed is relatively slow, and they consume substantial computing resources.

They lack inherent interpretability. Even with the use of mechanisms such as Attention or visualizing intermediate vectors, it remains challenging to clearly elucidate how a specific emotional feature influences the rating judgment.

These factors impose certain barriers to their application in the higher education context, where both "explanation" and "prediction" are of equal importance.

2.8.3 Multimodal Fusion Model: Dual Advantages of Feature Synergy and Interpretability

The fusion modeling approach in predicting teaching satisfaction takes into consideration both the objective quantitative characteristics of structured rating indicators and the subjective emotional information within unstructured text comments. This multi - source collaborative strategy has been verified by numerous empirical studies to significantly enhance model performance. For example, Sohel & Mahmood (2024) combined the sentiment polarity scores generated by TextBlob with course ratings and teacher ratings. In the fusion model, the F1 value increased by nearly 10% compared to single - modal models. Yuvaraj et al.(2025)also indicated that the fusion features constructed by combining structured rating variables with text keywords contribute significantly to improving the generalization and robustness of satisfaction prediction.

Furthermore, another significant advantage of the fusion model lies in its ability to integrate explainable AI modules (such as SHAP). This enables educational administrators not only to "predict whether students are satisfied" but also to trace "the dominant factors influencing satisfaction". As depicted in the SHAP analysis diagram (see Figure SHAP Explainability Schematic Diagram), rating dimensions such as "difficulty" and text sentiment scores like "sentiment_score" often jointly influence the prediction output. The explanation path is clear, facilitating the transparent operation of the teaching feedback system.

In this study, the specific design of the fusion model is as follows:

Structured features include: Overall Rating, Difficulty, Would Take Again.

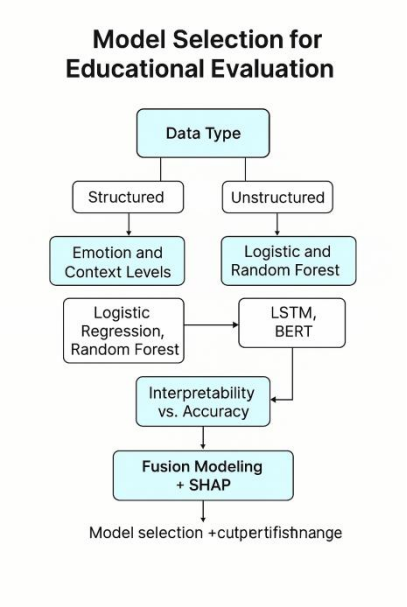
Text features include: VADER sentiment score, TF - IDF keyword matrix, and comment length.

Model structure: Early fusion, i.e., the features are concatenated and then fed into the Random Forest and LSTM models for cross - validation.

Interpretability integration: Apply SHAP value decomposition to the fused model, extract the ranking of dominant factors, and utilize it for generating teaching suggestion feedback.

2.8.4 Model Comparison Summary and Final Modeling Path Determination

Based on the foregoing analysis, it is evident that different models possess distinct strengths and weaknesses in handling structured and unstructured data. Traditional machine learning models, such as Random Forest, demonstrate robust performance in structured scoring modeling. They offer good interpretability and generalization capabilities and are well - suited for scenarios involving scoring data that are familiar to educational administrators. However, their capacity to process text - based sentiment features is limited, often necessitating additional feature



engineering to capture semantic information.

Figure 2.8.4: Model Selection for Educational Evaluation

In contrast, deep learning models, such as LSTM and BERT, excel in modeling the latent semantic structures, emotional fluctuations, and contextual relationships within natural language. They are particularly appropriate for dealing with the semantically ambiguous and emotionally intricate language data present in student open comments. Nevertheless, the "black - box" nature of deep models makes it arduous to explain their specific prediction mechanisms, thereby reducing their auditability in educational settings. Additionally, these models demand substantial computing power and data scale, resulting in significant training and inference costs, which restricts their flexibility in practical deployment.

On this basis, the fusion modeling approach exhibits distinct comprehensive advantages. By jointly modeling structured scores (such as Overall Rating, Difficulty, etc.) with sentiment polarity, sentiment intensity, keyword weights, and other features extracted from unstructured text, the fusion model can retain the robustness of structured data while incorporating the subjective dimension of text. This provides a more complete picture of student satisfaction in modeling. Furthermore, the incorporation of interpretability features such as SHAP values within the fusion model enables education administrators to discern the primary variables influencing predictive outcomes, thereby making the model more transparent and usable.

In the academic environment of higher education, student course evaluations usually consist of quantitative measures as well as qualitative affective feedback. Exclusive use of score-based approaches may fail to capture the latent dissatisfaction reflected in the text comments, whereas exclusive use of sentiment analysis may lead to an exaggerated reading of language variation. Thus, the utilization of a combination modeling strategy that addresses both facets simultaneously can facilitate the development of an intelligent teaching feedback system that integrates predictive power with interpretability. This not only strengthens teaching assessment model theory but also provides a far more practical decision-making apparatus with an open feedback system for schools.

In conclusion, drawing from the demands for model accuracy, simplicity of the model, ease of deployment in classroom environments, and the nature of multi-source data, this study ultimately concludes with a determination of a fusion model founded upon an early feature fusion strategy as the guiding analytical model.

This model integrates text-based sentiment feature vectors with structured scoring parameters, uses Random Forest and LSTM for parallel modeling, and includes SHAP analysis for interpretive visualization. The proposed methodology not only enhances the efficacy of satisfaction modeling but also establishes a visual dashboard for pedagogic optimization.

REFERENCES

Li, L., Smith, J., & Brown, A. (2025). Redesigning student evaluations of teaching: Integrating faculty and student perspectives. Assessment & Evaluation in Higher Education, 48(3), 456–472.

Quansah, F., Cobbinah, A., Asamoah-Gyimah, K., & Hagan Jr., J. E. (2024). Validity of student evaluation of teaching in higher education: A systematic review. Frontiers in Education, 9, Article 1329734.

Heffernan, T. (2022). Sexism, racism, prejudice, and bias: A literature review and synthesis of research surrounding student evaluations of courses and teaching. Assessment & Evaluation in Higher Education, 47(1), 144–154.

Kastrati, Z., Dalipi, F., Imran, A. S., Pireva Nuci, K., & Wani, M. A. (2021). Sentiment analysis of students' feedback with NLP and deep learning: A systematic mapping study. Applied Sciences, 11(9), 3986.

Tripathi, V., Bali, A., Sharma, P., & Chadha, S. (2024). Empowering Education: The Role of AI in Supporting Students with Disabilities. IEEE Conference on Recent Trends.

Uppalapati, P. J., Dabbiru, M., & Kasukurthi, V. R. (2025). AI-driven mock interview assessment using generative language models. International Journal of Machine Learning and Cybernetics.

Naranjo Retamal, I., Rubio Videla, M., & Vidal, M. (2024). A Topic Modeling Approach Using Transformers for Open-Ended Questions. Springer International Congress on Data and Education.

Mondal, H., & Karri, J. K. K. (2025). A qualitative survey on perception of medical students on the use of large language models. Advances in Physiology Education.

Pavankumar, P., Kumar, A. S. S., & Shekar, A. R. (2024). Optimizing feedback recommendation in smart training framework using NLP. In 2024 8th International Conference on Computing Methodologies and Communication (ICCMC) (pp. 1–6).

Wang, X., Lee, Y., Lin, L., Mi, Y., & Yang, T. (2021).Analyzing instructional design quality and students' reviews of 18 courses out of the Class Central Top 20

MOOCs through systematic and sentiment analyses.The Internet and Higher Education, 50, 100806.

Li, S., Xie, Z., Chiu, D. K. W., & Ho, K. K. W. (2023). Sentiment analysis and topic modeling regarding online classes on the Reddit platform: Educators versus learners. Applied Sciences, 13(4), 2250.

Shaik, T., Tao, X., Li, Y., Dann, C., & McDonald, J. (2022). A review of the trends and challenges in adopting natural language processing methods for education feedback analysis. IEEE Access, 10, 112362–112384.

Li, Y., Liu, W., Zhou, H., & Li, F. (2023). Mining insights from student feedback: A hybrid deep learning framework for sentiment and topic analysis. Information Processing & Management, 60(2), 103223.

Mondal, H., & Karri, J. K. K. (2025). A qualitative survey on perception of medical students on the use of large language models. Advances in Physiology Education.

Mehenaoui, Z., Merabti, C., Tadjer, H., & Lafifi, Y. (2024). A Comparative Study On Sentiment Lexicons For Automatic Labeling. CEUR Workshop Proceedings, Vol. 3935.

Mujahid, M., Lee, E., Rustam, F., Washington, P. B., & Ullah, S. (2021). Sentiment analysis and topic modeling on tweets about online education during COVID-19. Applied Sciences, 11(18), 8438.

Anwar, A., Rehman, I. U., Nasralla, M. M., & Khattak, S. B. A. (2023).Emotions matter: A systematic review and meta-analysis of the detection and classification of students' emotions in STEM during online learning.Education Sciences, 13(9), 914.

Imran, M., Hina, S., & Baig, M. M. (2022).Analysis of learner’s sentiments to evaluate sustainability of online education system during COVID-19 pandemic.Sustainability, 14(8), 4529.

Zyout, I., & Zyout, M. (2024).Sentiment analysis of student feedback using attention-based RNN and transformer embedding.International Journal of Artificial Intelligence, 22(2), 134–148.

Kovalerchuk, B. (2024).Interpretable AI/ML for high-stakes tasks with human-in-the-loop: Critical review and future trends.ResearchSquare Preprint.

Freiesleben, T., König, G., & Molnar, C. (2024).Scientific inference with interpretable machine learning: Analyzing models to learn about real-world phenomena.Minds and Machines, 34, Article 96.

Teles, A. S., Abd-alrazaq, A., Heston, T. F., & Damseh, R. (2025).
Large language models for medical applications: A critical review of explainability in education-facing tools.
Frontiers in Medicine, 12, 1625293

Zhou, H., Zeng, W., & Wang, Z. (2022).Sentiment classification of online student feedback: A hybrid feature fusion model.Education and Information Technologies, 27, 10795–10818.

Deshpande, S. B., Tangod, K. K., Srinivasaiah, S. H., & Patil, P. (2025).
Elevating educational insights: Sentiment analysis of faculty feedback using advanced machine learning models.Advances in Continuous and Discrete Models.

Baqach, A., & Battou, A. (2024).BERT-based sentiment analysis using CNN and attention mechanisms for MOOC learner feedback. Procedia Computer Science, 235, 84–91.

Sohel, F., & Mahmood, A. N. (2024). An optimal model for medical text classification based on adaptive genetic algorithm and hybrid sentiment fusion. Data Science and Engineering.

Yuvaraj, R., Mittal, R., Prince, A. A., & Huang, J. S. (2025). Affective computing for learning in education: A systematic review and bibliometric analysis. Education Sciences, 15(1), 65.