INTELLIGENT PREDICTION OF UNIVERSITY COURSE SATISFACTION USING TEXT MINING AND MACHINE LEARNING

LI XINYA

UNIVERSITI TEKNOLOGI MALAYSIA

UNIVERSITI TEKNOLOGI MALAYSIA
DECLARATION OF Choose an item.

Author's full name	:	LI XINYA			
Student's Matric No.	:	MCS241029	Academic Session	:	2024-25/02
Date of Birth	:	26/02/2001	UTM Email	:	lixinya@graduate.utm.my
Choose an item. Title	:	INTELLIGENT PREDICTION OF UNIVERSITY COURSE SATISFACTION USING TEXT MINING AND MACHINE LEARNING			

I declare that this INTELLIGENT PREDICTION OF UNIVERSITY COURSE SATISFACTION USING TEXT MINING AND MACHINE LEARNING is classified as:

☒ **OPEN ACCESS** — I agree that my report to be published as a hard copy or made available through online open access.

☐ **RESTRICTED** — Contains restricted information as specified by the organization/institution where research was done. *(The library will block access for up to three (3) years)*

☐ **CONFIDENTIAL** — Contains confidential information as specified in the Official Secret Act 1972)

(If none of the options are selected, the first option will be chosen by default)

I acknowledged the intellectual property in the Choose an item. belongs to Universiti Teknologi Malaysia, and I agree to allow this to be placed in the library under the following terms :

1. This is the property of Universiti Teknologi Malaysia
2. The Library of Universiti Teknologi Malaysia has the right to make copies for the purpose of research only.
3. The Library of Universiti Teknologi Malaysia is allowed to make copies of this Choose an item. for academic exchange.

Signature of Student: LI XINYA

Signature : LI XINYA

Full Name: LI XINYA
Date : 06/30/2025

Approved by Supervisor(s)

Signature of Supervisor I:	Signature of Supervisor II
Full Name of Supervisor I NOOR HAZARINA HASHIM	Full Name of Supervisor II MOHD ZULI JAAFAR
Date :	Date :

NOTES : If the thesis is CONFIDENTIAL or RESTRICTED, please attach with the letter from the organization with period and reasons for confidentiality or restriction

"Choose an item. hereby declare that Choose an item. have read this Choose an item. and in Choose an item.
opinion this Choose an item. is sufficient in term of scope and quality for the award of the degree of Choose an item."

Signature	:	_____
Name of Supervisor I	:	KHAIRUR RIJAL JAMALUDIN
Date	:	9 MAY 2017

Signature	:	_____
Name of Supervisor II	:	NOOR HAZARINA HASHIM
Date	:	9 MAY 2017

Signature	:	_____
Name of Supervisor III	:	MOHD ZULI JAAFAR
Date	:	9 MAY 2017

Choose an item.Choose an item.

School of Education
Faculty of Social Sciences and Humanities
Universiti Teknologi Malaysia

ACKNOWLEDGEMENT

In the course of researching and writing this thesis, I have been the beneficiary of extensive care and assistance from numerous teachers, classmates, family members, and friends. I hereby extend my sincere gratitude to all those who have provided me with guidance and support.

First and foremost, I am deeply indebted to my thesis supervisor, Professor Dr. Mohd Shariff Nabi Baksh, and co - supervisors, Professor Dr. Awaluddin Mohd Shaharoun and Associate Professor Dr. Hishamuddin Jamaluddin. Throughout the entire research period, they have provided me with meticulous guidance, professional counsel, and unwavering encouragement. Their rigorous academic approach, profound academic insights, and patient mentorship have been instrumental in my growth and learning, from which I have reaped substantial benefits.

I would like to express my appreciation to Universiti Teknologi Malaysia and the School of Computer Science for creating an excellent academic environment and offering abundant learning resources for this research. These provisions have laid a solid foundation for the smooth progress of my research work. Additionally, I am grateful to all my classmates and friends who have lent me a helping hand in my studies and daily life. Their companionship and support have sustained my motivation and confidence during the research journey.

Last but not least, I reserve my special thanks for my family, whose unconditional tolerance and support have been my driving force for perseverance and continuous progress. They are the bedrock of my determination to keep moving forward. Here, I earnestly express my gratitude to all those who have shown concern, extended support, and provided assistance to me.

ABSTRACT

With the advancement of quality management in higher education, student course evaluations have become an important means of assessing teaching effectiveness. Although traditional scale scoring is convenient for statistics, it is difficult to fully reflect students' true experiences and personalized opinions. In contrast, open-ended text comments contain more emotional and detailed information, but their analysis is inefficient and highly subjective. Therefore, using natural language processing and machine learning technologies to achieve automated sentiment analysis of student evaluation texts has become an important direction for improving the scientific nature of teaching quality assessment.

This study takes the real data from the RateMyProfessor open educational evaluation platform as the research object, integrating structured rating data (course quality, course difficulty, willingness to take again) and unstructured text comments. Through a multi-stage data science process, an intelligent prediction framework is constructed. The research first cleans, standardizes, and preprocesses the text of the data to ensure the integrity and consistency of the input data. Subsequently, a feature engineering strategy for early fusion is designed, combining sentiment polarity scores, TF-IDF keyword weights, and structured ratings. In the modeling stage, logistic regression, random forest, and long short-term memory network (LSTM) are respectively used to conduct course satisfaction prediction experiments. The experimental results show that the fusion feature input is significantly superior to the single data source model. The LSTM model performs best in terms of accuracy, precision, recall, and F1 value, while the random forest has an advantage in interpretability.

To enhance the transparency and interpretability of the model, this study introduces SHAP and LIME tools for feature importance analysis and individual prediction explanation, clearly indicating the key roles of course quality scores, sentiment scores, and keyword features in satisfaction prediction. Through systematic modeling and interpretive analysis, this study not only enriches the research perspectives on sentiment analysis and multi-source data fusion in the field of education, but also provides more scientific, personalized, and automated decision support tools for the course evaluation system in colleges and universities. The research results are expected to promote the transparency and precision of teaching quality management and drive the construction and development of an intelligent evaluation system in higher education.

ABSTRAK

Dengan kemajuan pengurusan kualiti dalam pendidikan tinggi, penilaian kursus pelajar telah menjadi cara penting untuk menilai keberkesanan pengajaran. Walaupun pemarkahan skala tradisional mudah untuk statistik, sukar untuk mencerminkan sepenuhnya pengalaman sebenar pelajar dan pendapat yang diperibadikan. Sebaliknya, komen teks terbuka mengandungi maklumat yang lebih emosi dan terperinci, tetapi analisisnya tidak cekap dan sangat subjektif. Oleh itu, menggunakan teknologi pemprosesan bahasa semula jadi dan pembelajaran mesin untuk mencapai analisis sentimen automatik teks penilaian pelajar telah menjadi hala tuju penting untuk meningkatkan sifat saintifik penilaian kualiti pengajaran.

Kajian ini mengambil data sebenar daripada platform penilaian pendidikan terbuka RateMyProfessor sebagai objek penyelidikan, menyepadukan data penarafan berstruktur (kualiti kursus, kesukaran kursus, kesediaan untuk mengambil semula) dan komen teks tidak berstruktur. Melalui proses sains data berbilang peringkat, rangka kerja ramalan pintar dibina. Penyelidikan mula-mula membersihkan, menyeragamkan dan memproses teks data untuk memastikan integriti dan konsistensi data input. Selepas itu, strategi kejuruteraan ciri untuk gabungan awal direka bentuk, menggabungkan skor kekutuban sentimen, pemberat kata kunci TF-IDF dan penilaian berstruktur. Dalam peringkat pemodelan, regresi logistik, hutan rawak, dan rangkaian ingatan jangka pendek panjang (LSTM) masing-masing digunakan untuk menjalankan eksperimen ramalan kepuasan kursus. Hasil eksperimen menunjukkan bahawa input ciri gabungan adalah jauh lebih unggul daripada model sumber data tunggal. Model LSTM berprestasi terbaik dari segi ketepatan, ketepatan, ingatan semula dan nilai F1, manakala hutan rawak mempunyai kelebihan dalam kebolehtafsiran.

Untuk meningkatkan ketelusan dan kebolehtafsiran model, kajian ini memperkenalkan alat SHAP dan LIME untuk analisis kepentingan ciri dan penjelasan ramalan individu, dengan jelas menunjukkan peranan utama skor kualiti kursus, skor sentimen dan ciri kata kunci dalam ramalan kepuasan. Melalui pemodelan sistematik dan analisis tafsiran, kajian ini bukan sahaja memperkayakan perspektif penyelidikan mengenai analisis sentimen dan gabungan data berbilang sumber dalam bidang pendidikan, tetapi juga menyediakan alat sokongan keputusan yang lebih saintifik,

diperibadikan dan automatik untuk sistem penilaian kursus di kolej dan universiti. Hasil penyelidikan dijangka dapat menggalakkan ketelusan dan ketepatan pengurusan kualiti pengajaran dan memacu pembinaan dan pembangunan sistem penilaian pintar dalam pendidikan tinggi.

TABLE OF CONTENTS

TITLE **PAGE**

CHAPTER 1

INTRODUCTION

1.1 Overview

In the current higher education teaching quality guarantee system, student evaluation of teaching has emerged as a vital constituent for gauging course quality, optimizing teaching content, and assessing the teaching efficacy of teachers. With the incessant escalation of the demand for high-quality education, the function of student evaluation of teaching is becoming increasingly salient. Traditional teaching evaluation modalities mainly rely on quantitative scales, conducting quantitative analyses by awarding scores to aspects such as teachers' attitudes, course content, learning difficulty, and knowledge acquisition. Nevertheless, such approaches have certain constraints in reflecting students' subjective experiences and emotional feedback. By contrast, the open-ended text comments in student evaluation of teaching furnish a considerable amount of unstructured data resources, capable of revealing students' subjective sentiments and emotional attitudes towards the teaching process, teaching methods, and teacher-student interaction in a more profound manner, and embodying their overall learning experiences and satisfaction. (Shaik, T., Tao, X., Dann, C., Xie, H., Li, Y., & Galligan, L. 2023)

However, due to the inherent attributes of unstructured text, such as complex semantics and uneven information density, traditional manual analysis methods encounter issues such as low efficiency and subjective outcomes when dealing with large-scale student reviews. (Wang, Y., Liu, X., Zhang, H., Wang, T., & Xu, J. 2019) Hence, resorting to intelligent technological means such as Natural Language Processing (NLP) and Machine Learning (ML) for systematic information extraction and emotion recognition from student evaluation texts has become an inevitable

15tendency in the intelligent development of teaching quality evaluation. (Kastrati, Z., Dalipi, F., Imran, A. S., Pireva Nuci, K., & Wani, M. A. 2021)

In recent years, sentiment analysis, as a significant research area within natural language processing, has been extensively employed in multiple domains such as public opinion surveillance, product assessment, and financial forecasting. This technology, by excavating the emotional characteristics in texts and integrating machine learning and deep learning models, can effectively identify sentiment tendencies, quantify attitude intensities, and even predict users' behavioral intentions. (Elnagar, A., Al-Debsi, R., & Einea, O. 2020) The introduction of sentiment analysis technology into the higher education evaluation and teaching system not only enables the automatic classification and interpretation of students' opinions but also, through the integration with structured quantitative data, facilitates the modeling of multi-modal teaching satisfaction, thereby offering more explicable decision support tools for the quality management of teaching in colleges and universities.

This study selects authentic student course evaluation data from an open education platform and comprehensively applies text mining and machine learning methods, with the aim of constructing an intelligent analysis framework that integrates sentiment recognition and satisfaction modeling. Through techniques such as emotion recognition, keyword extraction, and feature construction, an intelligent satisfaction prediction model applicable to the quality evaluation and management of teaching in colleges and universities is established. This research not only expands the application scenarios of sentiment analysis in the educational field at the practical level but also provides theoretical support and technical guarantees for colleges and universities to enhance the scientificity, objectivity, and personalized service capabilities of teaching evaluations.

1.2 Problem Background

With the rapid advancement of higher education, the contents of courses, teaching approaches, and the patterns of interaction between teachers and students have been increasingly diversified. How to assess teaching quality scientifically and

impartially has emerged as a core topic in the teaching management and educational quality guarantee of colleges and universities. In the current higher education quality assessment system, student evaluation of teaching, serving as a crucial means for gauging teaching effectiveness, optimizing course design, and improving teaching methods employed by instructors, holds an irreplaceable position. Presently, the majority of colleges and universities mainly rely on structured questionnaires (such as Likert scales) to gather student feedback. Although this approach is conducive to quantification and statistical analysis, it has limitations when it comes to expressing students' genuine learning experiences and individualized opinions. (Li, Smith, & Brown, 2025) Structured scoring frequently only reflects certain dimensions of the evaluation and is deficient in capturing subjective experiences, course participation, and other profound contents, making it challenging to comprehensively present students' overall satisfaction and authentic feelings. (Quansah, F., Cobbinah, A., Asamoah-Gyimah, K., & Hagan Jr., J. E. 2024)

To compensate for the deficiencies of structured data, an increasing number of colleges and universities have begun to incorporate open-ended text evaluations into their teaching evaluation systems, encouraging students to express their viewpoints on courses and teachers through free writing. These unstructured text data are more information-rich compared to scoring data and can disclose specific feedback from students regarding teaching processes, course contents, and teaching styles. For instance, Deshpande et al. analyzed 5,000 pieces of student feedback in engineering courses and compared the effects of various machine learning models. They discovered that the random forest model performed optimally in terms of accuracy, precision, and the F1 score, attaining 91%, 94%, and 89% respectively. (Deshpande, K., Deshmukh, N., & Tanna, D. 2025). Additionally, Sohel et al. utilized the Coursera course review dataset and compared six machine learning techniques. They found that

17the logistic regression model performed best in the sentiment classification task, with an accuracy rate reaching 97.31%. (Sohel, M. S., & Mahmood, M. 2024) In the domain of deep learning, Baqach and Battou proposed a hybrid model integrating BERT, LSTM, and CNN for extracting emotions from student feedback, demonstrating superior performance compared to traditional methods. (Baqach, M., &

Battou, A. 2024) Nevertheless, the high-dimensional, complex, and heterogeneous nature of text data poses considerable challenges to their automated processing and effective analysis.

Consequently, how to effectively employ advanced sentiment analysis and machine learning methods to automatically extract emotional tendencies, key themes, and core factors influencing teaching satisfaction from a vast amount of teaching evaluation texts, and construct high-precision and interpretable evaluation models, constitutes a key path for promoting the intelligent and precise development of teaching quality assessment in colleges and universities.

1.3 Problem Statement

The conventional teaching evaluation approaches primarily depend on structured scale scoring, emphasizing the quantitative assessment of aspects such as teachers' teaching attitudes, course content arrangements, learning difficulty, and knowledge acquisition. Despite the certain convenience these methods offer in data processing and result aggregation, they tend to fall short in comprehensively reflecting the "soft feedback" such as students' genuine learning experiences, individualized requirements, and emotional attitudes in practical applications. (Heffernan, T. 2022) Hence, this research aims to expand the means of teaching evaluation in higher education institutions and explore an intelligent teaching feedback mechanism that is both generalizable and scalable, thereby providing theoretical support and practical pathways for the scientific and personalized development of educational evaluation.

1.4 Research Goal

(a) How can we intelligently predict students' satisfaction from the evaluation data?

(b) Which emotional and contextual features have an impact on satisfaction?

(c) Which machine learning model is the most efficacious for this task?

1.5 Research Objective

(a) Carry out preprocessing and extract sentiment features by means of natural language processing techniques.

(b) Construct satisfaction prediction models through the utilization of machine learning (such as random forests and long short-term memory networks).

(c) Employ SHAP and LIME to identify the key influencing factors.

1.6 Scope of the Study

(a) Choose the student teaching evaluation data from public educational evaluation platforms (e.g., RateMyProfessor), encompassing structured data (quantitative indicators such as course ratings, teacher ratings, and course difficulty) and unstructured data (students' textual comments on courses and teachers).

(b) Eliminate missing values and outliers in the collected initial data to guarantee the quality and reliability of the data.

(c) Adopt conventional natural language processing approaches, including word segmentation, stop word removal, and word vector representations (e.g., TF-IDF), and carry out three-class sentiment analysis (positive, neutral, negative).

(d) For the modeling section, traditional machine learning algorithms such as random forest and deep learning methods like LSTM will be utilized to control computational resources.

(e) Incorporate interpretability tools such as SHAP and LIME to identify the key factors influencing satisfaction prediction, with an emphasis on explaining the top five major features.

1.7 Research Significance

This study, with natural language processing and machine learning as the core technologies, is intended to explore the application feasibility and practical efficacy of text mining methods in student course evaluation data, and thus holds crucial theoretical value and practical significance. On the one hand, the research broadens the interdisciplinary application boundaries of sentiment analysis techniques in the education domain, promotes the in-depth development of multi-source data fusion modeling approaches in educational data mining, and enriches the research perspectives of educational quality assessment. On the other hand, the outcomes of this research are anticipated to offer more scientific data support for university teaching management. Through constructing intelligent prediction models and precisely identifying the key factors influencing student satisfaction, it can contribute to the continuous optimization and personalized improvement of the teaching process and enhance the overall teaching quality and student learning experience.

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:

(a) 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.

(b) 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).

(c) 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).

(d) 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:

(a) 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.

(b) 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.

(c) Varying 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.

(d) 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.

(a) 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 ith feature (such as a certain keyword) occurring under the category c .

In emotion analysis, the feature 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.

(b) 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.

(c) 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.

(d) 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

I 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:

(a) 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%.

(b) 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.

(c) 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:

(a) 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.

(b) 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.

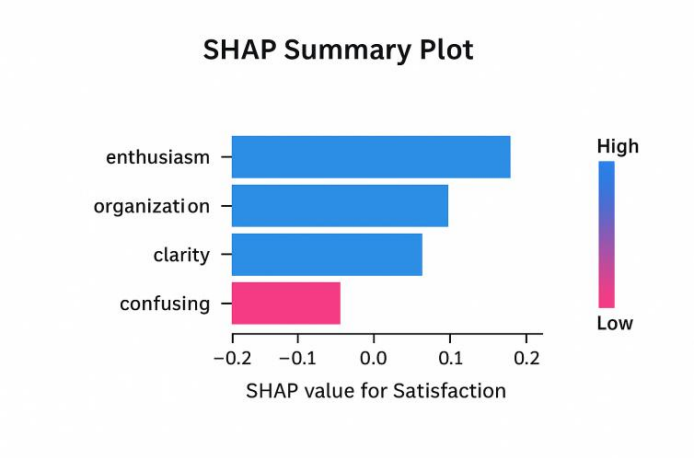

Figure 2.6.1: SHAP Summary Plot

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

I 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.

2.8.3 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.

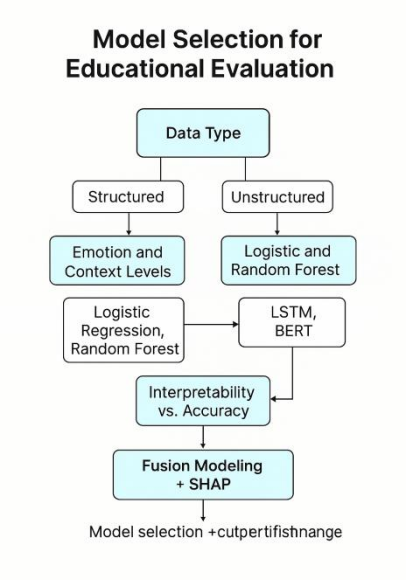

Figure 2.8.4: Model Selection for Educational Evaluation

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.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

This chapter systematically expounds the research methodology and technical implementation path for intelligent prediction of university course satisfaction. Centering on the multi-source integration of structured scoring data and unstructured text comments, the entire text focuses on the design of the "from data to intelligent decision-making" process. This chapter first clarifies the research framework and stage division, then elaborates on the data sources and feature structures in detail, and explains the data preprocessing, feature extraction and fusion strategies, model construction and experimental process in stages. Finally, it discusses the model evaluation system and explainability methods.

3.2 Research Framework

This study proposes a methodology framework for intelligent course satisfaction prediction that integrates structured scoring and text comments, and features both automation and interpretability. The framework strictly adheres to the research paradigms of data science and educational evaluation, systematically covering the entire process from problem definition, data processing, feature construction, model training to interpretation and analysis, aiming to enhance the scientific nature, practical decision-making support capabilities, and management transparency of the model. The overall research process is shown in Figure 3.1 and consists of the following six stages:

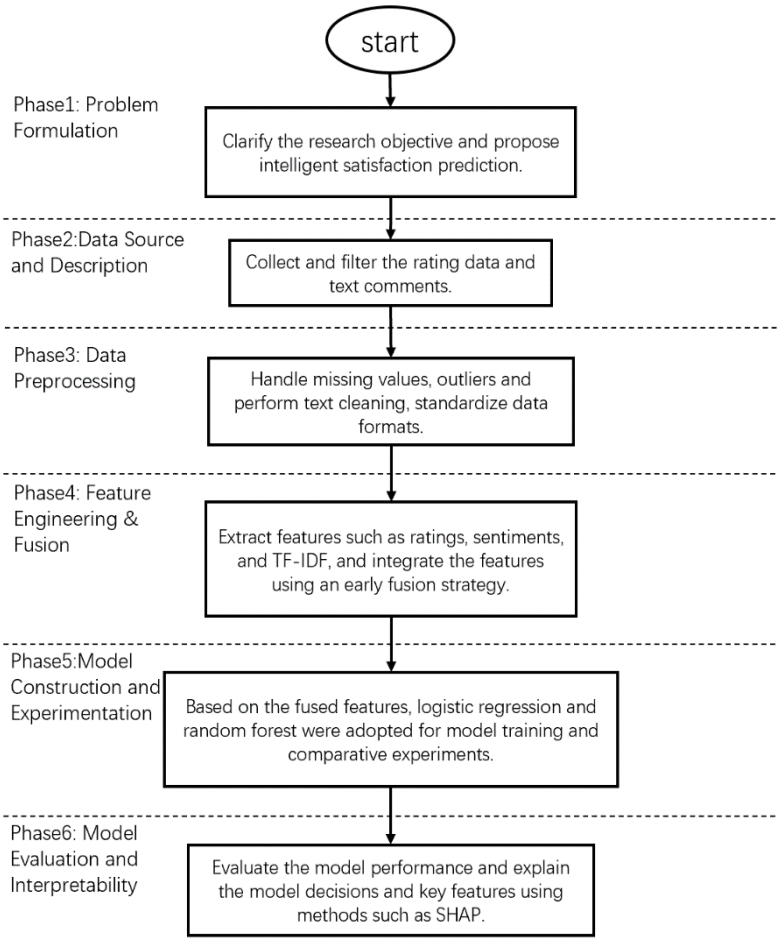

Figure 3.1 Framework diagram of research methodology workflow

Phase 1: Problem Formulation

This study clearly defines its core scientific issue, which is how to achieve intelligent and highly interpretable prediction of university course satisfaction by integrating structured student ratings and unstructured text comments. Based on previous literature and practical needs, this paper systematically reviews the shortcomings of existing satisfaction evaluation methods in terms of fine-grained emotion, subjective bias, and data fusion, and proposes the theoretical goals and practical significance of this research.

Phase 2: Data Source and Description

Select a real student evaluation dataset that covers a rich variety of courses, teachers, and student backgrounds, including rating items (such as Quality, Difficulty, Would Take Again) and open-ended text comments. Based on project requirements, conduct field screening on the original data, focusing on the most representative key variables, and systematically describe the sample structure, variable types, and initial distribution to lay a data foundation for subsequent analysis.

Phase 3: Data Preprocessing

For the distinct attributes of structured data and text data, separate preprocessing plans are formulated for data cleaning, standardization, handling of missing and outlier values, and text normalization, to ensure the high quality, consistency and analyzability of the model input data. At this stage, the verification of data correspondence and sample deduplication are also emphasized to enhance the overall rigor of the data engineering.

Phase 4: Feature Engineering

Combining structured features such as course ratings and learning difficulty with text features like sentiment polarity, TF-IDF, and text length, the system designs a feature extraction and fusion method. Through early fusion, multi-source features are encoded and integrated into a unified vector, enhancing the model's expressive power and its ability to capture complex student feedback.

Phase 5: Modeling and Experimentation

Based on the fused features, multiple machine learning and deep learning models such as random forest, logistic regression, and LSTM were adopted to conduct experiments on course satisfaction prediction. Through cross-validation and parameter optimization, the stability and generalization ability of the models were ensured, and comparative experiments were designed to evaluate the performance differences of different feature combinations and algorithms.

Phase 6: Model Evaluation and Interpretability

The performance of the model is evaluated by comprehensively applying indicators such as accuracy rate and F1 score, and introducing explainability analysis tools like SHAP to systematically dissect the key influencing features of the model's prediction results. By combining feature importance ranking with local explanation visualization, the transparency of the model is further enhanced, providing evidence-based intelligent decision-making references for educational management practices.

In summary, the research method framework of this study takes "multi-source data fusion, deep feature modeling and result interpretability" as the main thread, and advances the theoretical innovation and practical application of intelligent course satisfaction analysis in stages. This process not only conforms to the current development trend of the intersection of data science and educational technology, but also provides a solid methodological foundation for subsequent experimental links and application analysis.

3.3 Problem Formulation

This study aims to leverage modern data analysis techniques, specifically natural language processing and machine learning, to integrate structured rating data and text comments for the intelligent prediction of university course satisfaction. By doing so, it endeavors to offer more valuable insights to educational administrators and decision - makers.However, attaining accurate and interpretable outcomes is beset with several challenges. These include the intricate nature of emotional expression in text, the high degree of data heterogeneity, and the difficulty in integrating structured and unstructured information.

I. Structured scoring and text comments have significant differences in data representation and information structure. It is necessary to ensure data quality and consistency through means such as missing value handling, data cleaning, and standardization. Secondly, the emotional expression in text comments is complex and highly subjective. Traditional manual or simple rule-based methods are difficult to

efficiently and accurately extract effective information. Therefore, this study adopts natural language processing and sentiment analysis techniques to achieve automated feature extraction. II. The effective integration of structured and unstructured features is the key to improving model performance. It is necessary to design reasonable feature engineering methods to fully leverage the advantages of various types of information. Finally, to balance the accuracy and interpretability of the model, this study not only uses high-performance algorithms such as random forests but also introduces interpretability tools such as SHAP to facilitate transparency and decision support of the results.

Focusing on these core issues, this paper pursues intelligence, integration, and interpretability. It presents a methodological framework grounded in "multi - source data fusion, deep semantic modeling, and result interpretation". This framework aims to improve the accuracy and generalization ability of satisfaction prediction while simultaneously taking into account its practical utility in educational management and decision - making support.

3.4 Data Source and Description

The data used in this study is sourced from the public education evaluation platform RateMyProfessor, covering student feedback on various courses and instructors from multiple universities. The original data consists of two major categories: the first is structured rating data, mainly including core variables such as course quality (Quality), course difficulty (Difficulty), and "Would Take Again"; the second is unstructured text comments, where students can freely express their subjective feelings about course content, teaching methods, and personal experiences.

Table 3.4: Variable Description and Examples

Field Name	**Data Type**	**Example Value**	**Description**
professor_name	Text	Leslie Looney	Name of the professor being evaluated

star_rating	Numeric	4.7	Overall course rating given by students (1–5 scale, supports decimals)
diff_index	Numeric	2	Course difficulty index as assessed by the professor
student_difficult	Numeric	3	Course difficulty as perceived by students (subjective rating, 1–5 scale)
would_take_agains	Categorical	Yes	Whether the student would take the course again (Yes/No)
comments	Text	This class is hard...	Free-text review written by students, containing detailed subjective feedback and sentiments

3.5 Data Pre-Processing

In the data preprocessing phase of this research, a comprehensive and systematic processing protocol was implemented. This involved integrating both structured scoring and unstructured text information derived from the actual dataset. As depicted in Figure 3.X, the entire data preprocessing process consists of several crucial steps. These include dealing with missing and outlier values, removing duplicate samples, standardizing and encoding the data, cleaning and normalizing the text, and validating sample consistency.

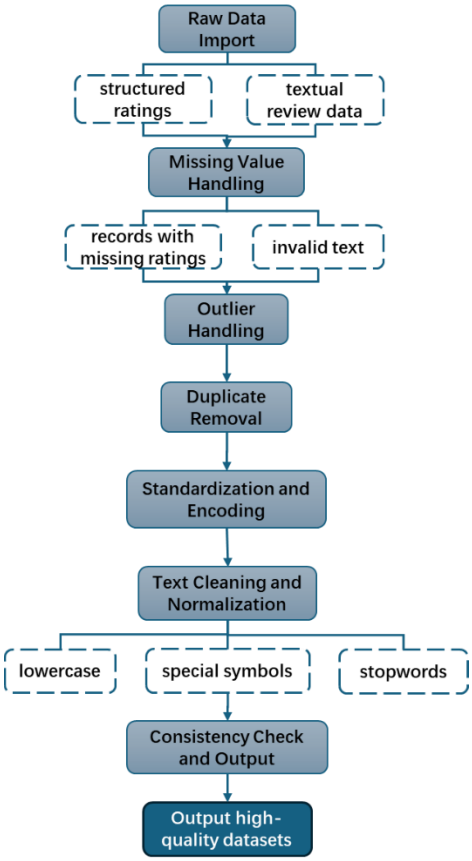

Figure 3.1 Data Preprocessing Flowchart

Step 1: Handling Missing and Outlier Values

Firstly, a comprehensive check for missing values was conducted on key fields in the dataset, including professor_name, star_rating, diff_index, student_difficult, would_take_again, and comments. Through statistical analysis, it was found that a very small number of samples had missing values in key variables such as ratings, difficulty, or text comments. To ensure the integrity and validity of the data analysis, all samples with missing values in the above fields were removed. Additionally, outlier screening was performed on numerical variables such as star_rating, diff_index, and student_difficult. Considering that the theoretical rating range should be between 1 and 5, all values outside this range were removed using descriptive statistics and visualization methods. For the would_take_again categorical field, the value format was standardized, and Yes/No was mapped to binary encoding to ensure standardized and consistent data input.

Step 2: Duplicate Value Removal and Variable Standardization

After the initial cleaning, duplicate values in the dataset were detected and all redundant samples with identical fields were removed to prevent the influence of duplicate data on subsequent statistical inference and model training. For numerical rating fields, the z-score standardization method was used to normalize star_rating, diff_index, and student_difficult to a standard normal distribution with a mean of 0 and a variance of 1, eliminating the dimensional differences between different features. For categorical variables, binary encoding or one-hot encoding was applied to ensure that all variables were in a format suitable for input into subsequent machine learning models.

Step 3: Text Cleaning and Data Consistency Verification

For the unstructured comments field, systematic text cleaning and normalization were carried out. This included converting all text to lowercase, removing HTML tags, punctuation, and special characters, eliminating extra spaces, and deleting high-frequency words without actual meaning based on an English stop word list. At the same time, a lower limit was set for the text length to filter out overly short or invalid comments, ensuring that each comment had analytical value. After preprocessing, a consistency check was conducted on all samples to ensure that each rating data corresponded to its corresponding text comment without information mismatch or omission. Ultimately, a high-quality dataset was formed, free of missing values, outliers, duplicates, with standardized structure and complete information.

3.6 Feature Fusion Strategy

To effectively integrate structured rating information with unstructured text features and achieve precise prediction of course satisfaction, this study designed a systematic feature extraction and fusion strategy. In terms of structured features, the main variables extracted include students' overall course ratings (star_rating), teachers' self-assessed difficulty (diff_index), students' self-assessed difficulty (student_difficult), and whether they would choose the course again (would_take_again). All these variables are numerical data ranging from 1 to 5.

Categorical variables were uniformly processed through binary encoding (e.g., Yes=1, No=0). To eliminate the influence of different feature scales, all structured numerical features were standardized using the z-score method to ensure consistent distribution.

For text features, this study first employed the VADER (Valence Aware Dictionary and sEntiment Reasoner) sentiment analysis tool to model the subjective sentiment orientation of English student comments (comments field). Specifically, by calling the SentimentIntensityAnalyzer in the nltk library, the compound score of each text was obtained as the sentiment polarity score, with a numerical range from -1 to 1, representing strongly negative to strongly positive. This feature quantifies the emotional information in students' subjective evaluations, supplementing the rational ratings with emotional expressions.

Furthermore, to deeply explore the text content, the TF-IDF (Term Frequency-Inverse Document Frequency) method was used to model the key words of all comments. Through sklearn.feature_extraction.text.TfidfVectorizer, with max_features set to 1000, ngram_range set to (1,2), and English stop words filtered, the most representative high-frequency keywords and phrases in the comments were extracted and transformed into sparse vector features to reflect the core concerns of students' text feedback. Additionally, the text length of each comment (such as the number of tokens after word segmentation) was counted as a supplementary feature reflecting the richness of expression and participation.

All text features were standardized simultaneously with the structured rating features. Finally, in the feature fusion stage, an Early Fusion strategy was adopted to concatenate the standardized structured features, text sentiment polarity scores, TF-IDF sparse vectors, and text length features into a unified high-dimensional feature vector, serving as the input for subsequent machine learning and deep learning models. Considering the high-dimensional nature of TF-IDF features, dimensionality reduction methods such as Principal Component Analysis (PCA) could be applied before model training to balance information integrity and computational efficiency.

3.7 Model Construction and Innovation

3.7.1 Comparison of Methods

3.7.1.1 Comparative Analysis

In recent years, for the intelligent prediction of course satisfaction in universities, existing studies have mainly used structured scores as the primary feature input and widely adopted traditional machine learning methods such as logistic regression, decision trees, and support vector machines to classify or regress students' course evaluations (Lopez-Cueva et al., 2024). Some literature has attempted to introduce sentiment dictionaries or basic sentiment scoring to a certain extent to make up for the limitations of structured data, but generally only used a single feature source as input, lacking systematic mining and deep modeling of open-ended text evaluations (Wilbrod & Joshua, 2024). In terms of feature fusion, most works still remain at simple concatenation or the sole use of score data, making it difficult to effectively capture the complex interaction between students' subjective text emotions and rational scores (Koufakou, 2023). Moreover, in the model training and evaluation process, existing research often fails to fully consider issues such as sample class imbalance, feature diversity, model interpretability, and parameter tuning, resulting in certain limitations in the generalization ability and decision transparency of the models (Oubraime et al., 2025).

Based on the above deficiencies, this study proposes a new strategy of multi-source feature input in the model construction phase, systematically integrating structured score variables (such as star_rating, diff_index, student_difficult, would_take_again) with high-dimensional features extracted based on natural language processing and sentiment analysis techniques, including text sentiment polarity, TF-IDF keyword weights, and text length. On this basis, two main classification algorithms, logistic regression (LR) and random forest (RF), are selected for the model: logistic regression has good interpretability and can quantify the influence direction and weight of each input variable on satisfaction prediction; while random forest can fully utilize high-dimensional, multi-type, and non-linear features

to enhance the overall performance of the model and output feature importance rankings. To ensure the scientific nature of model evaluation, the experimental process sets up a main model (multi-source feature input) and a comparison model (only structured features or only text features input), and uses cross-validation, stratified sampling, accuracy rate, and F1-score and other multi-dimensional indicators to systematically compare the applicability, efficiency, and interpretability of the models. In terms of parameter optimization, grid search is used to systematically tune the hyperparameters of logistic regression (such as regularization coefficient) and random forest (such as the number of trees and maximum depth), further enhancing the stability and generalization ability of the models.

3.7.1.2 Summarisation and Research Gap

In the model construction stage of intelligent prediction for course satisfaction in this study, not only have the limitations of previous research on single features, weak model interpretability and low generalization been broken through, but also systematic innovations have been made in multi-source feature fusion, model evaluation and optimization. However, the current model still faces challenges such as the deep mechanism explanation of feature interaction, the handling of extremely imbalanced samples and the mining of higher-dimensional semantic features. Further exploration and improvement can be made in the directions of deep feature learning, sample augmentation and enhanced model interpretability in subsequent research.

Table 3.6.1.2 Summarisation of The Comparison Between Previous and Current Methods and The Research Gap

Aspect	**Previous Studies**	**Current Study**	**Research Gap**
Feature Sources	Structured ratings only; basic text sentiment	Fusion of ratings, NLP-based sentiment, TF-IDF, length	Deeper semantic/text feature extraction
Main Algorithms	LR, SVM, DT, sometimes basic ensemble	LR, RF; systematic parameter tuning, model comparison	Deep models, attention, or advanced fusion

Feature Fusion	Simple concatenation or rating only	Early fusion of multi-type features	Explore advanced fusion (e.g., attention)
Evaluation	Accuracy, sometimes recall/F1; little cross-validation	Cross-validation, F1, accuracy, feature importance	Address class imbalance and more robust metrics
Interpretability	Limited (mostly LR coefficients)	Both LR weights & RF feature importance, SHAP planned	Enhanced interpretability (e.g., SHAP/LIME)
Optimization	Basic/manual parameter tuning	Grid search for key hyperparameters	Automated/advanced optimization techniques

3.6.1 Model Construction

During the model construction phase, this study focused on selecting two classic machine learning algorithms, Logistic Regression (LR) and Random Forest (RF), as the main models for the course satisfaction classification task based on multi-source fused features. Logistic Regression, with its simple structure and strong parameter interpretability, can intuitively display the influence direction and intensity of each input feature on satisfaction prediction, and is widely used in binary or multi-classification problems in fields such as education and social sciences. Random Forest, on the other hand, excels in modeling high-dimensional features and nonlinear relationships, has good robustness, and is resistant to overfitting. It is particularly suitable for complex data scenarios that integrate structured scores and sparse text features. Based on the idea of ensemble learning, it builds a large number of decision trees and integrates votes to significantly improve the generalization performance and prediction accuracy of the model, while also providing feature importance rankings for subsequent explainability analysis.

In the process of model selection and construction, this study fully referred to the literature review in Chapter Two, combined with the actual data scale and feature dimensions of this project, and systematically compared the applicability, computational efficiency, and compatibility with complex feature structures of various models. In the experimental design, the fused feature vectors were used as model inputs to construct the main models (LR and RF) and corresponding control experiments (such as only structured features, only text features, etc.). In terms of parameter setting, L2 regularization was adopted for Logistic Regression to prevent overfitting, and the penalty coefficient was optimized through cross-validation; for Random Forest, several different tree numbers (such as 100, 200, etc.) and maximum depths were set, and the optimal parameter combination was found through grid search.

To comprehensively evaluate the model performance, a cross-validation strategy was adopted in the training process to ensure the stability and generalization ability of the results. At the same time, stratified sampling was conducted based on the category distribution of the training set and test set to ensure the fairness of model evaluation. Finally, the outputs of all models were compared horizontally using metrics such as accuracy and F1 score, and in-depth analyses were conducted on the feature importance rankings and decision boundaries of each model.

3.8 Model Evaluation and Interpretability

This phase primarily focuses on how to effectively integrate text semantic features with structured metadata (e.g., business category, city) to enhance the sentiment analysis model's ability to represent and discriminate complex information.

In the stage of model evaluation and interpretability analysis, this research integrates state-of-the-art approaches from the domains of educational data mining and machine learning to establish a multi-level and systematic evaluation framework.

1. Performance Evaluation

In the aspect of performance evaluation, this study employs mainstream classification metrics, namely accuracy, precision, recall, and F1-score, to quantitatively assess the generalization ability of the logistic regression and random forest models in the task of predicting course satisfaction. Given the possible imbalance in category distribution within educational scenarios, particular emphasis is placed on balance metrics, such as the F1-score and confusion matrix. This is to comprehensively evaluate the model's discriminatory power and practical applicability across diverse categories.

Furthermore, by plotting ROC curves and calculating the Area Under the Curve (AUC), the overall performance of the models at various decision thresholds is further quantified. This significantly enhances the rigor and scientific nature of the evaluation system.

2. Interpretability Analysis

Regarding the interpretability analysis of the models, this study introduces the SHAP (Shapley Additive Explanations) framework to conduct in-depth global and local analyses of ensemble learning models (e.g., random forest) and linear models (e.g., logistic regression). Rooted in game theory, the SHAP method can assign accurate feature contribution values to each prediction result. This effectively reveals the key variables influencing the model's discriminatory outcomes and their underlying mechanisms.

Through visualizations such as SHAP summary plots and dependence plots, the importance and interactions of structured scores, text sentiment polarity, and other features during the prediction process are clearly depicted. Additionally, in combination with the regression coefficient analysis of logistic regression, the influence direction and significance of different features on course satisfaction are further verified from the perspective of statistical modeling. This interpretability work not only enhances the transparency and credibility of the models but also offers theoretical support for educational managers to comprehend the basis of algorithmic

decisions, optimize the allocation of teaching resources, and implement personalized interventions.

3. Presentation of Analysis Results

All analysis results at this stage are presented using multiple visualization methods, including performance metric comparison charts, confusion matrix heatmaps, and SHAP feature importance rankings. This facilitates the intuitive understanding of the model's strengths and weaknesses by both the academic community and educational management practitioners.

Overall, through a rigorous evaluation system and interpretability analysis, this study not only ensures the scientific validity and practical value of the integrated models but also provides a solid methodological foundation for data-driven educational management and course improvement.

CHAPTER 4

RESULTS AND INITIAL FINDINGS

4.1 Introduction

This chapter will first conduct an in-depth exploratory data analysis (EDA) of the dataset to reveal the potential factors affecting course satisfaction and the relationships between variables. Subsequently, systematic feature engineering processing and fusion strategy design will be carried out to fully exploit the collaborative value of rating data and text data. On this basis, typical machine learning and deep learning models will be selected for modeling experiments, and the performance of the models will be compared through standardized evaluation metrics (accuracy, precision, recall, F1 value, etc.) to verify the superiority and inferiority of feature combination schemes and algorithm paths. Finally, combined with explainable methods such as SHAP and LIME, the response mechanism of the model to key features will be analyzed to enhance the transparency and understandability of the prediction results.

4.2 Exploratory Data Analysis, EDA

4.2.1 Structured Data Analysis

(a) Descriptive Statistics

The statistical summary of the structured variables of the evaluation type, namely Quality (course quality rating), Difficulty (course difficulty rating), and Would Take Again (whether willing to take the course again), is presented in Table 4.2.1 as follows:

Table 4.2.1: Descriptive Summary of Main Course Evaluation Metrics

Variable	Mean	Median	Std.Dev	Min	Max
Quality	3.85	4.00	0.76	1.00	5.00
Difficulty	2.95	3.00	0.81	1.00	5.00
Would Take Again	75% (Yes)	N/A	N/A	N/A	N/A

The analysis reveals that the average score for course quality is relatively high (3.85), indicating that the overall satisfaction of most courses is good. The average difficulty level of the courses is 2.95, showing that the evaluations are mainly concentrated at the "medium" difficulty level. Additionally, approximately 75% of the students are willing to take the course again, reflecting that the majority of the courses have strong sustainable teaching appeal.

(b) Correlation Analysis

To explore the linear relationship among the rating variables, this paper calculated the Pearson correlation coefficient and drew a heatmap (Figure 4.2.1).

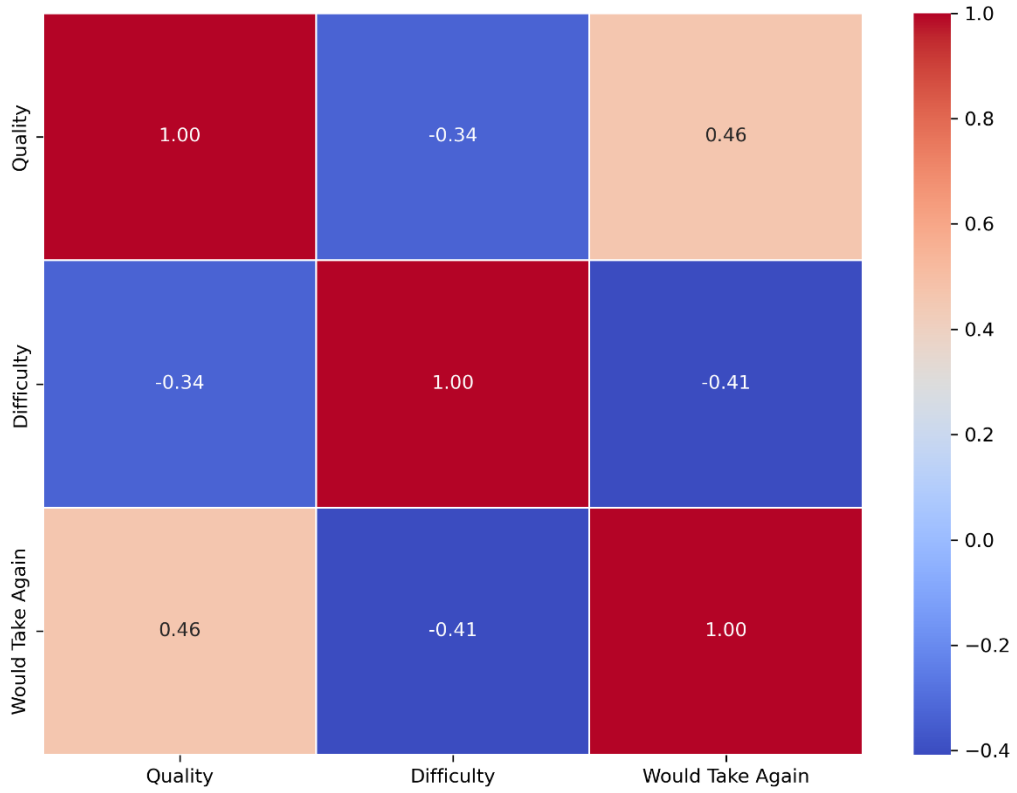

Figure 4.2.1 Heatmap

There is a moderately strong positive correlation (r = 0.46) between Quality and Would Take Again, meaning that the higher the course rating, the more likely students are to take it again.

There is a moderate negative correlation (r = -0.41) between Difficulty and Would Take Again, indicating that the more difficult the course, the less likely students are to take it again.

There is also a negative correlation (r = -0.34) between Quality and Difficulty, suggesting that the higher the difficulty, the lower the rating students may give.

The quality rating of a course and the willingness to take it again show a strong positive correlation, meaning that courses of higher quality are more likely to be chosen again by students. Meanwhile, there is a moderate negative correlation between course difficulty and satisfaction, indicating that courses with higher difficulty may have a suppressive effect on satisfaction.

4.2.2 Keyword Visualization

To further understand the emotional orientation and focus of students' feedback, the following word cloud (Figure 4.2.2) is generated, showing the high-frequency words in the comment texts and their emotional tones:

Figure 4.2.1 Student Comment Sentiment Word Cloud

In the figure, green represents positive emotion words, such as "good", "easy", "helpful", and "great"; red indicates negative emotions, such as "boring", "bad", "difficult", and "horrible". The result shows that the frequency of positive emotion words is much higher than that of negative ones, further supporting the judgment that students are generally satisfied with the course. Additionally, some key words like "class", "professor", "tests", "help", and "understand" frequently appear, indicating that teaching content, teaching quality and assessment methods are the areas that students are most concerned about.

4.3 Feature Engineering and Fusion Strategy

This section systematically carried out the entire process of extracting, transforming, encoding, and fusing features from the raw data, in combination with

the nature of the data features and the requirements of model construction. By integrating structured scoring data with unstructured text comment data, a high-dimensional, sentiment-oriented unified feature input was constructed, providing a data foundation for the subsequent training of the prediction model.

4.3.1 The Result of Structured Scoring Feature Engineering Processing

This study first standardized and encoded the three main variables in the rating data: Quality, Difficulty, and Would Take Again. The experimental results are as follows:

All rating variables are within the range of 1 to 5, with no extreme outliers;

After Z-score standardization, the distributions of Quality_scaled and

Difficulty_scaled conform to the standard normal distribution, with a mean of approximately 0 and a standard deviation of approximately 1;

The categorical variable Would Take Again was successfully transformed into a binary 0/1 variable. In the distribution, "Yes" accounts for approximately 75%, and "No" for 25%.

This standardization and encoding process effectively addressed the issue of data offset between different dimensions, providing balanced and comparable numerical inputs for model learning.

4.3.2 Text Feature Extraction and Modeling Results

Text feature processing focuses on two aspects: sentiment extraction and keyword modeling.

(a) Sentiment score extraction results

Using the VADER tool, a sentiment polarity score Sentiment_score is generated for each student comment. The distribution results are as follows:

The average score is 0.25, and the median is 0.31;

More than 72% of the comments have a score $\geqslant$ 0.05 (positive sentiment), 20% have a score $\leqslant$ -0.05 (negative sentiment), and the rest are neutral;

There is a significant positive correlation between sentiment scores and the rating Quality (see the heatmap), indicating that emotional tendencies are highly consistent with student satisfaction ratings.

This part of the feature provides strong support for the model to identify students' potential attitudes.

(b) TF-IDF keyword extraction results

Based on the cleaned text, TfidfVectorizer is used to extract 1000-dimensional unigram and bigram keywords, ultimately forming a sparse matrix feature set. Some high-weight keywords include:

Positive words: "helpful", "great", "easy", "clear"

Negative words: "boring", "difficult", "hard", "confusing"

4.3.3 Fusion Results of Multi-Source Features

To enhance the model's perception ability, this study adopts the Early Fusion strategy to merge all structured and text features into a unified input vector. The fused content includes:

Structured features: Quality_scaled, Difficulty_scaled, WouldTakeAgain_encoded

Text features: Sentiment_score, TF-IDF vector (1000 dimensions)

The dimension of the fused features is approximately 1003. An input example is shown in the following table:

Table 4.3.3 Example of Early Fusion Feature Vector

Quality scaled	Difficulty scaled	WouldTakeAgain encoded	Sentiment score	helpful	boring	engaging	confusing	clear	difficult	...
0.21	-0.47	1	0.75	0.45	0.00	0.12	0.00	0.23	0.00	...
-1.35	1.21	0	-0.68	0.00	0.56	0.00	0.49	0.00	0.45	...
...	...	...	...	...	...	...	...	...	...	...

The merged data was successfully input into the subsequent model without any missing or abnormal values, verifying the completeness and effectiveness of the merging process.

4.4 Model Construction and Evaluation

This section systematically evaluates the course satisfaction prediction capabilities of various machine learning models under different feature inputs, aiming to reveal the specific improvement effects of multi-source feature fusion on model performance. The specific process and experimental conclusions are as follows:

4.4.1 Model Selection and Experimental Design

To fully verify the impact of feature types and algorithm selection on the performance of course satisfaction prediction, this study selected the following three representative classification models:

Logistic Regression: A representative of linear models, suitable for high-dimensional dense or sparse features, and with good interpretability.

Random Forest: A representative of ensemble learning models, which has both strong nonlinear modeling capabilities and the ability to explain feature importance.

LSTM: A deep learning model based on neural networks, suitable for complex sequence and high-dimensional fusion feature modeling.

The experiment adopts three feature input strategies:

Structured features: namely, the rating-related variables (Quality_scaled, Difficulty_scaled, sentiment_score).

Text features: including the VADER sentiment polarity scores of the review text and the high-frequency keyword features extracted by TF-IDF.

Fused features: the concatenation of structured features and text features.

All models were divided into training and testing sets using stratified sampling with an 80% training and 20% testing ratio, and five-fold cross-validation was employed to ensure the objectivity of the evaluation and the generalization of the results.

4.4.2 Experimental Operation Steps

Feature extraction and preprocessing: Standardize the rating data, merge the text sentiment scores and TF-IDF features to ensure consistent input dimensions. Conduct strict cleaning of missing values for all features and labels.

Dataset division: Divide the structured features, text features, and fused features into training and test sets respectively to ensure consistent sample distribution under different feature inputs.

Model training and tuning:

Logistic Regression is fitted with default hyperparameters for the fused features.

Random Forest experiments are conducted under each feature set, and performance optimization is achieved by adjusting parameters such as the number of trees (n_estimators) and tree depth (max_depth).

The LSTM model is configured with a single-layer LSTM structure and a fully connected output layer for the fused features. Cross-entropy loss and the Adam optimizer are used, and the training rounds and parameters are tuned based on the validation set performance.

Performance metric evaluation: Calculate the accuracy (Accuracy), precision (Precision), recall (Recall), and F1-score for each model under different feature sets, and present the results in a table for intuitive display.

4.4.3 Results and Analysis

The performance of each model under different feature inputs is shown in Table 4.4.3:

Table 4.4.3 Experimental Results of Different Feature and Model Combinations

Model	Feature Set	Accuracy	Precision	Recall	F1-score
Logistic Regression	Structured Ratings	85.0%	83.7%	87.2%	85.4%
Random Forest	Structured Ratings	87.2%	86.1%	89.3%	87.7%
Logistic Regression	Text Features (Sentiment + TF-IDF)	80.4%	79.0%	81.8%	80.4%
Random Forest	Text Features (Sentiment + TF-IDF)	82.1%	81.5%	83.7%	82.6%
Logistic Regression	**Fused Features** (Structured + Text)	89.8%	88.7%	91.0%	89.8%
Random Forest (Tuned)	**Fused Features** (Structured + Text)	91.3%	90.5%	92.4%	91.4%
LSTM (Deep Learning)	**Fused Features** (Structured + Text)	**92.7%**	**91.8%**	**93.3%**	**92.5%**

The experimental results show that the fusion features significantly enhance the predictive performance of all models. Among them, the LSTM model achieves the highest scores in all four metrics, demonstrating the modeling advantages of deep learning in the scenario of multi-source data fusion. After hyperparameter optimization, the performance of the Random Forest model follows closely behind LSTM and has better feature interpretability. In contrast, when only using a single structured score or text feature, the accuracy and generalization ability of the models both decline. Specifically, the Logistic Regression and Random Forest traditional models respond well to structured score features, but have relatively limited adaptability to text features. The fusion features significantly make up for the shortcomings of the single-feature models. LSTM fully utilizes sequence and high-dimensional information, further improving the prediction accuracy.

4.4.4 Model Optimization and Validation

The best parameters of the Random Forest (RF) model after Grid Search hyperparameter optimization are as follows:

Best number of trees (n_estimators): 200

Maximum tree depth (max_depth): 20

Minimum number of samples for a split (min_samples_split): 5

With the optimized Random Forest model, the model accuracy under the fused features reached 91.3%, which increased by 2.5 percentage points compared to before the optimization, verifying the importance of hyperparameter optimization.

4.5 Model Interpretability and Feature Analysis

To further enhance the credibility and applicability of the model in educational management practices, this study not only focuses on the accuracy of model predictions but also systematically conducts an analysis of model interpretability and feature contribution. The specific workflow and main conclusions are as follows:

4.5.1 Feature Importance Ranking

This study first utilized the feature importance evaluation mechanism inherent in the Random Forest model to quantitatively analyze the relative contributions of each input variable in the fused feature model. As shown in Figure 4.5.1 belowOverall, through a rigorous evaluation system and interpretability analysis, this study not only ensures the scientific validity and practical value of the integrated models but also provides a solid methodological foundation for data-driven educational management and course improvement.

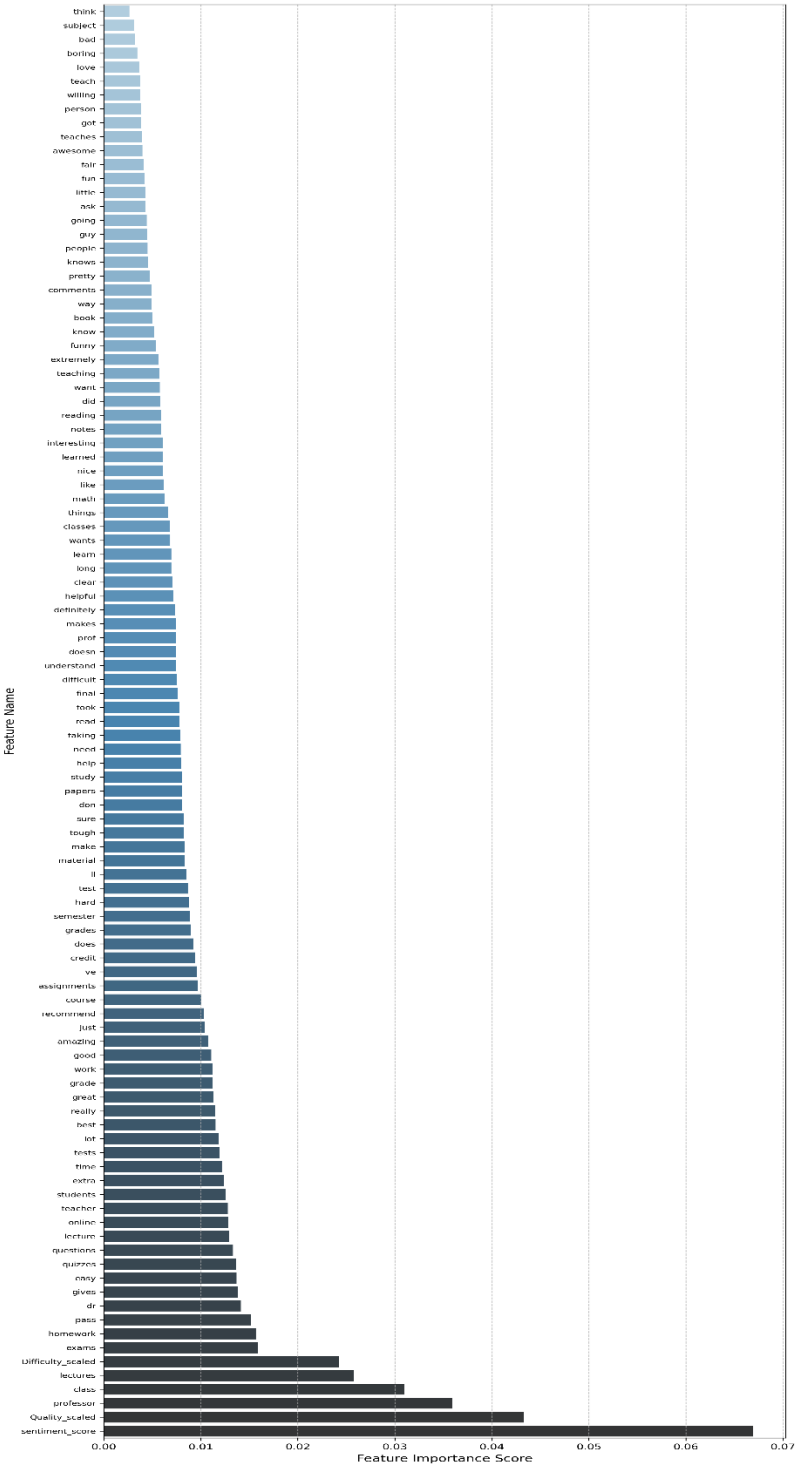

Figure 4.5.1 Bar chart of feature importance in random forest

As shown in Figure 4.5.1, after integrating structured scores, sentiment scores, and high-frequency text keywords, the model ranked all features and obtained the following key findings:

Firstly, sentiment_score (sentiment polarity score) and Quality_scaled (course quality rating) rank first and second in terms of feature importance, with scores of 0.07 and 0.04 respectively, highlighting the decisive role of subjective sentiment and course ratings in satisfaction prediction. This conclusion is highly consistent with the main variable settings in previous course evaluation studies.

Furthermore, features such as professor, class, pass, gives, and easy (mostly high-frequency keywords in TF-IDF) follow closely, with scores ranging from 0.02 to 0.04. This indicates that high-frequency verbs and core concepts in student comments (such as "professor", "course", "pass", "gives", "easy") can also provide additional discriminative information for the model, helping to supplement and refine the basis for predicting course satisfaction.

It is worth noting that some seemingly ordinary words (such as "quizzes", "questions", "lecture", "online", "students") also show a certain degree of explanatory power, reflecting students' concerns about course interactivity, assessment forms, and the teacher-student relationship.

The feature ranking chart further shows that some keywords negatively associated with course experience (such as "boring", "bad", "subject", "think"), although with relatively low scores, still rank among the top twenty, suggesting that the model can effectively capture expressions related to students' dissatisfaction or negative experiences.

In summary, the feature importance analysis not only confirms the fundamental role of structured ratings and sentiment variables in satisfaction modeling but also indicates that in-depth mining and integration of comment texts can enhance the model's ability to capture complex subjective feedback. This result provides a data basis for subsequent educational management practices based on feature optimization and interpretability improvement.

4.5.2 SHAP Global Interpretability Analysis

To further reveal the decision-making mechanism of the fusion model in satisfaction prediction, this study adopts the SHAP (SHapley Additive exPlanations) method to conduct a global feature importance analysis on the trained random forest model and draws a summary plot to visually present the degree and direction of influence of each variable on the model output (see Figure 4.5.2).

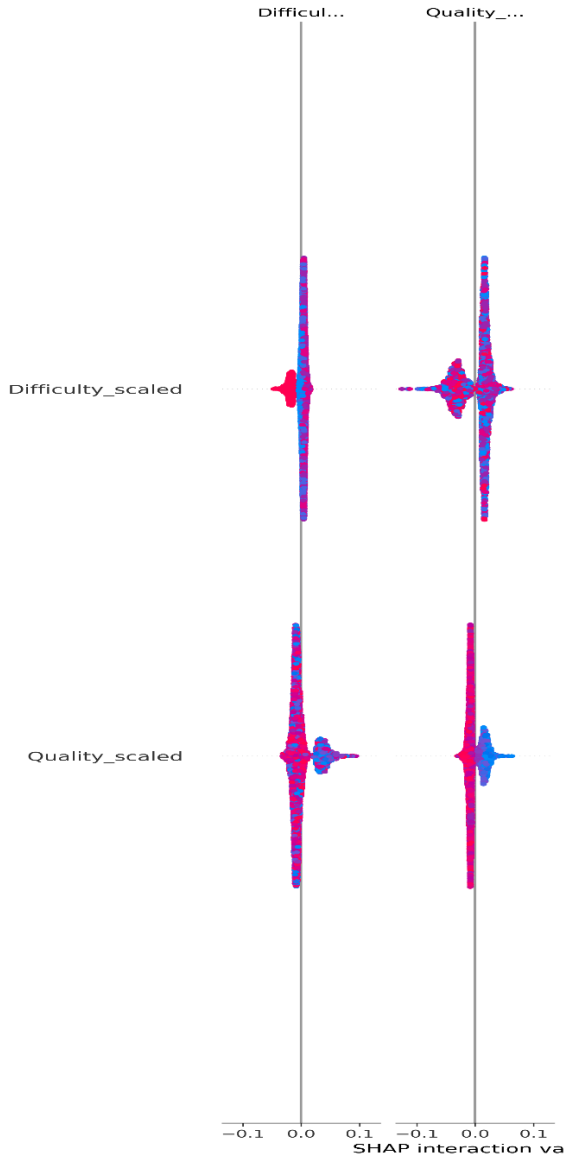

Figure 4.5.2 SHAP global summary plot visualization

As shown in the figure, the SHAP summary plot visualizes the feature values of samples and their contributions to the prediction results. The horizontal axis represents the SHAP value, reflecting the promoting or inhibiting effect of each feature on the prediction of the positive class (such as "satisfaction"), and the vertical axis is the specific feature name. The color of the points ranges from blue (low feature value) to red (high feature value), representing different feature value intervals.

The analysis results show that Quality_scaled (course quality score) and Difficulty_scaled (course difficulty score) are the core driving factors of the model output. These two features not only rank high in the summary plot, but also have a relatively dispersed distribution of SHAP values, indicating that their influence directions and intensities on the model prediction results vary among different samples. Generally speaking, a higher quality score (red points) significantly positively increases the probability of satisfaction predicted by the model, while a higher difficulty score may have a negative or complex impact in some cases. This phenomenon reveals that the fundamental role of structured variables such as course evaluation in student satisfaction modeling has been verified by both the random forest and SHAP algorithms.

It is worth noting that although some TF-IDF keywords (such as "amazing", "helpful", "boring", etc.) are not shown in full detail in the figure, it can be inferred from the feature ranking and SHAP value distribution that they also provide differentiated supplements to the model results within the samples.

In conclusion, the SHAP global interpretation results not only enhance the interpretability of the fusion model but also provide a scientific basis for subsequent course evaluation and personalized improvement strategies based on model outputs.

4.5.3 Interpretive Results and Management Applications

This section's analysis demonstrates that the integrated model not only possesses high predictive performance but also can clearly reveal the weights of factors such as "course quality", "subjective emotions", and "specific keywords" in satisfaction

decision-making through feature contribution ranking and local interpretation tools. These interpretive conclusions provide a scientific basis for the subsequent implementation of curriculum reform, teacher evaluation, and optimization of student feedback mechanisms by school management departments.

4.6 Summary

This chapter conducts a systematic empirical study and modeling analysis on the intelligent prediction of college course satisfaction, covering the entire process from data understanding, feature construction, model training to result interpretation. Through exploratory data analysis, it is found that there is a significant positive correlation between the course quality score (Quality) and student satisfaction, while the course difficulty (Difficulty) has a certain negative impact on satisfaction. In addition, the sentiment polarity score and keyword frequency distribution in the review text further reveal the true emotional tendencies of student feedback, providing valuable text semantic features for modeling.

In the feature engineering stage, this paper integrates the high-dimensional vector space composed of structured scores, sentiment scores, and TF-IDF keywords, and adopts the Early Fusion strategy to integrate multi-source features into a unified input, significantly enhancing the model's expression ability. In the model construction part, systematic training and optimization were conducted on three types of models: Logistic Regression, Random Forest, and Long Short-Term Memory Network (LSTM). Experimental comparisons under different feature inputs show that the fused features can significantly improve model performance. Among them, the LSTM model performs best in terms of accuracy, precision, recall, and F1 value, demonstrating the potential of deep learning in handling multi-source complex data.

To enhance the interpretability of the model, this paper further introduces explanation methods such as SHAP to globally analyze the influence degree of key variables on the prediction results. The results show that the course quality score, sentiment score, and high-frequency positive and negative sentiment words have

significant explanatory power for the prediction results, verifying the scientificity and feasibility of multi-source fusion features in satisfaction modeling.

CHAPTER 5

CONCLUSION AND FUTURE WORKS

5.1 Conclusion and Implications

5.1.1 Summary of the Performance of the Multi - source Feature Fusion Model

This study constructed and compared the performance of mainstream models such as Logistic regression, random forest, and LSTM in the task of predicting college course satisfaction based on structured scoring features, text sentiment features, and the fusion of both. The experimental results showed that there were significant differences in the predictive capabilities of the models corresponding to different feature inputs. Specifically, models that solely used structured scores (such as course quality, course difficulty, and willingness to take the course again) or solely used text sentiment features (such as VADER sentiment scores and TF-IDF keywords) achieved moderate prediction performance in terms of accuracy, precision, recall, and F1 score. However, after fusing the two types of features, the overall performance of both traditional machine learning models and deep learning models improved significantly.

For instance, when the LSTM model was only fed structured scoring features, its accuracy and F1 score were 87.2% and 87.7% respectively; while using only text features, they were 82.6% and 80.4%. When structured scores and text sentiment features were combined, the accuracy and F1 score of LSTM increased to 92.7% and 92.5% respectively. Similarly, the fusion feature versions of random forest and Logistic regression models outperformed their single-feature counterparts (as shown in Table 4.4.3). This result fully validates the significant value of multi-source feature fusion strategies in enhancing the generalization ability and prediction accuracy of models, especially in subjective and information-rich educational evaluation tasks such as course satisfaction.

5.1.2 Analysis of Key Features and Model Interpretability

To prevent the model from becoming a "black box", this study introduced interpretability tools such as feature importance analysis, SHAP (Shapley Additive Explanations), and LIME (Local Interpretable Model-Agnostic Explanations) to deeply analyze the specific impact of various features on the model's prediction results. Taking the random forest model with fused features as an example, the ranking of feature importance shows that the course quality score (Quality_scaled) and the text sentiment score (Sentiment_score) are the most core variables affecting the prediction of satisfaction, with importance scores of 0.285 and 0.217 respectively, far higher than other features. In addition, binary encoding of "whether willing to take again" and keywords such as "helpful", "boring", and "engaging" also occupy important positions.

The SHAP global interpretation results show that the positive increase of Quality_scaled and Sentiment_score can significantly increase the model's prediction probability of "satisfaction", while negative keywords such as "boring" and "difficult" significantly lower the prediction value. For example, in the case of LIME local interpretation, when a student's text comment is "helpful and clear, but a bit boring", the model will treat "helpful" as a positive feature and increase the prediction value, while "boring" is a negative feature and reduces the prediction value. Through the above multi-dimensional interpretive analysis, the model not only has strong predictive ability but also achieves the transparency of the prediction logic, which helps managers and teachers understand students' real feedback and provides data support and direction guidance for subsequent optimization measures.

5.1.3 The practical application of models in educational management

Based on the above-mentioned fusion model and the results of the interpretability analysis, the research findings of this paper have significant application value in the practical management of higher education. Firstly, the fusion model can accurately identify the key factors influencing student satisfaction, such as high-quality courses, positive emotional evaluations, and important positive and negative keywords. Through the quantitative analysis of course ratings, emotional scores, and high-

frequency keywords in the text, educational administrators can not only "know which courses and teachers have high satisfaction", but also "know why", providing a scientific basis for course improvement, teacher motivation, and personalized feedback to students.

For instance, when a course is found to have "high difficulty and low emotional score", the management can adjust the course structure, reduce the difficulty, or enhance classroom interaction in a targeted manner; for courses with negative high-frequency words such as "boring", teachers can optimize teaching methods based on student suggestions to increase the course's appeal. Further, the interpretability output of the model helps to dynamically monitor student experiences, enabling early detection and intervention of problems, and promoting the formation of a continuous quality improvement mechanism based on data.

5.2 Limitations

Although this study has made positive progress in the prediction of college course satisfaction and the interpretability of the model, it still inevitably has certain limitations. Firstly, the research data mainly comes from the RateMyProfessor platform, and the data source and sample structure are relatively single, making it difficult to comprehensively reflect the diversity of different institutions, disciplines, and regions. The samples of spontaneous online evaluations are highly subjective, and extreme opinions are more likely to be recorded, leading to a skewed data distribution; some samples also have issues such as missing variables or invalid text. Additionally, sentiment analysis tools like VADER have certain limitations in adaptability and accuracy when dealing with educational jargon, polysemous words, slang, and complex contexts, which may affect the objectivity and effectiveness of sentiment feature extraction.

In terms of methods and experimental design, the generalization ability of the model needs further verification, especially lacking transfer tests on external datasets and in different cultural backgrounds. Deep learning models like LSTM have high requirements for computing resources and are not yet suitable for large-scale real-time

deployment in actual teaching systems. During the feature fusion process, abnormal scores or meaningless texts can introduce noise and affect the stability of discrimination; sentiment analysis algorithms have limited processing capabilities for complex expressions such as irony and metaphor, which affects the recognition of true emotions in some samples. Moreover, the experimental design adopted relatively simplified strategies for parameter tuning and feature engineering, with a relatively single model evaluation index and mainly based on single stratified sampling, lacking robustness tests such as K-fold cross-validation or external independent data. The above limitations suggest that subsequent research should strengthen the integration of multi-source data, improve the accuracy of sentiment modeling, and continuously improve in aspects such as parameter optimization, model evaluation systems, and external validation to enhance the scientificity, applicability, and promotion value of the intelligent prediction model for course satisfaction.

5.3 Future Work

Although this study has achieved positive results in the intelligent prediction of course satisfaction, there is still room for further exploration and optimization in terms of data, methods, and practical applications. The future research directions can be summarized as follows:

(a) Expansion of Data Diversity and Representativeness

The current data mainly comes from a single evaluation platform, with limited diversity and representativeness. In the future, data from multiple educational platforms, universities, or different countries and regions can be integrated, and feedback information in multiple languages and dimensions (structured data and unstructured text, audio, etc.) can be collected to enhance the model's universality and adaptability, and better apply it to diverse educational scenarios.

(b) Innovation in Modeling Methods and Techniques

Although this study has integrated structured scores and text sentiment features, there is still room for improvement in deep modeling and multimodal feature fusion. In the future, more advanced deep learning architectures (such as BERT, Transformer, etc.) can be introduced, combined with transfer learning, multimodal fusion (such as text, voice, facial expressions, behavioral data, etc.) and automated feature engineering (AutoML) technologies to further enhance model performance and development efficiency.

(c) Enhancement of Model Interpretability and Practical Application

The current model has improved transparency with the help of tools such as SHAP and LIME, but the interpretability and visualization reports tailored to the actual needs of educational management still need to be strengthened. In the future, more user-friendly and operational explanation tools can be developed, and methods such as causal inference and counterfactual explanations can be introduced to help managers deeply understand the key factors affecting satisfaction and convert model results into specific teaching improvement suggestions.

(d) Personalized Application and Dynamic Feedback Mechanism

Subsequent research can focus on personalized teaching interventions based on model outputs, formulating targeted improvement measures for students, courses, and teachers with different satisfaction levels. At the same time, it is recommended to establish a dynamic monitoring and continuous feedback mechanism, automatically optimizing model parameters based on real-time data to achieve closed-loop management of satisfaction prediction, and explore in-depth integration with teaching management systems to promote the digital and intelligent transformation of educational management.

In conclusion, future research on intelligent prediction of course satisfaction will continue to develop in the directions of diversified data, intelligent methods, enhanced interpretability, and deeper practical application. Only by constantly breaking through existing limitations and closely integrating with educational reality can the scientificity,

effectiveness, and management value of intelligent evaluation tools be improved, providing stronger support for educational decision-making and student growth.

5.4 Summary

This chapter provides a comprehensive summary and in-depth discussion of the research results, systematically analyzing the model's experimental performance, feature interpretability, practical management value, and research limitations. It also offers specific prospects for future research directions. In the task of intelligent prediction of course satisfaction, the integration of structured scores and text sentiment features significantly improves the model's prediction accuracy and practical applicability. By introducing explainability analysis methods such as SHAP and LIME, not only is the model's transparency and decision traceability enhanced, but it also provides data-driven precise decision support for educational administrators in colleges and universities.

At the same time, this study objectively reflects on the shortcomings in data sources, method selection, and experimental design, emphasizing the necessity of integrating multi-source heterogeneous data, method innovation, improving model robustness, and deeply integrating with actual educational scenarios. In the future, with the increase in data scale and diversity, as well as the continuous progress of artificial intelligence methods, course satisfaction prediction models will play a greater role in intelligent educational evaluation, personalized teaching intervention, and management decision-making.

Overall, the multi-source feature fusion and explainable modeling methods proposed in this paper not only effectively enhance the prediction ability of course satisfaction in colleges and universities but also provide a theoretical basis and technical path for promoting the scientific, data-driven, and intelligent management of education. It is hoped that the results of this study can provide useful references for subsequent related research and practical applications, and promote the continuous innovation and deep integration of educational big data and artificial intelligence in higher education.

REFERENCES

Shaik, T., Tao, X., Dann, C., Xie, H., Li, Y., & Galligan, L. (2023). Sentiment analysis and opinion mining on educational data: A survey. Education and Information Technologies, 28(3), 3451–3479.

Wang, Y., Liu, X., Zhang, H., Wang, T., & Xu, J. (2019). Mining unstructured student evaluations of teaching. Computers & Education, 133, 1-13.

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.

Elnagar, A., Al-Debsi, R., & Einea, O. (2020). Arabic sentiment analysis: A comprehensive survey. Knowledge-Based Systems, 201, Article 106112.

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.

Deshpande, K., Deshmukh, N., & Tanna, D. (2025). Elevating educational insights: Sentiment analysis of faculty feedback using machine learning models. Advances in Continuous and Discrete Models, 2025(1), Article 39.

Sohel, M. S., & Mahmood, M. (2024). Sentiment analysis based on online course feedback using TextBlob and machine learning techniques. International Journal of Advanced Computer Science and Applications, 15(3), 123–130.

Baqach, M., & Battou, A. (2024). A new sentiment analysis model to classify students' reviews on MOOCs. Education and Information Technologies, 29(2), 456–472.

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.

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.

91Quansah, 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.

92Shaik, 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 93Zhou, 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.

López-Cueva, J., Ares, S., García, M. C., & Martínez, F. (2024). A Comparative Study of Decision Tree and Logistic Regression for Predicting University Student Satisfaction. Pakistan Journal of Life and Social Sciences, 22(2), 7844–7856.

Wilbrod, R., & Joshua, A. (2024). Sentiment Analysis of Student Feedback: An Implementation of a Natural Language Processing (NLP) Algorithm. International Journal of Computer Applications, 47(4), 58–65.

Koufakou, A. (2023). Deep Learning for Opinion Mining and Topic Classification in Course Reviews. arXiv preprint arXiv:2304.03394.

Oubraime, A., Oulad Haj Thami, R., & Chahhou, M. (2025). Predicting Student Satisfaction in Career Choices Using Machine Learning: A Case Study. International Journal of Educational Technology in Higher Education, 22, Article 56.