Decision Tree Frameworks for Enhanced Teaching Effectiveness Evaluation with Transparent Data-Driven Educational Decision-Making

Nolan M. Yumen[1], Angie C. Canillo[2]

[1] University of Antique Tario-Lim Memorial Campus, Philippines. Email: nolanyumen0017@gmail.com

[2] University of San Carlos, Talamban Campus, Philippines. Email: amceniza@usc.edu.ph.

Corresponding Author:

Nolan M. Yumen[1]* (nolanyumen0017@gmail.com)

Abstract: This study aimed to develop and validate an interpretable decision tree framework for automatically classifying unstructured student evaluation comments according to standardized teaching effectiveness dimensions used in Philippine State Universities and Colleges, while maintaining transparency and practical utility for institutional decision-making. A mixed-methods sequential exploratory design was employed. Qualitative thematic analysis of 250 student comments established a coding framework aligned with four teaching dimensions (Commitment, Knowledge of Subject, Teaching for Independent Learning, Management of Learning). Three expert evaluators achieved inter-rater reliability of κ=0.81 across 1,200 manually coded comments. Multiple decision tree algorithms (CART, C4.5, Random Forest) were trained on 840 comments, validated on 180, and tested on 180, with hyperparameter optimization via 5-fold cross-validation. The final dataset comprised 4,410 comments from 347 course sections across four colleges at a Philippine state university during second semester 2023-2024. Implementation testing assessed efficiency gains and user satisfaction. Random Forest achieved optimal performance with 84.2% overall accuracy, ranging from 77.6% (Commitment) to 88.1% (Knowledge of Subject) across dimensions. Expert validation showed substantial agreement (κ=0.78). Feature importance analysis identified "clear" (0.094), "helpful" (0.087), and "engaging" (0.081) as top predictors. Implementation testing demonstrated 74% reduction in analysis time while maintaining quality (4.0-4.3/5.0 ratings). High-confidence decision rules (84.7-93.2% confidence) provided transparent classification logic. Decision tree frameworks enable efficient, transparent analysis of qualitative teaching feedback aligned with institutional evaluation criteria, supporting evidence-based faculty development in resource-constrained Philippine higher education contexts

Keywords: teaching evaluation, decision trees, text classification, educational analytics, transparent algorithms, Philippine higher education

Introduction

Student evaluations of teaching represent one of the most ubiquitous yet controversial assessment tools in higher education worldwide. While these evaluations are intended to improve teaching quality and inform personnel decisions, the qualitative comments that accompany numerical ratings—often the most valuable source of actionable insights—remain systematically underutilized due to the challenges of analyzing large volumes of unstructured text [1,2]. This underutilization represents a critical loss of student voice in the evaluation process and missed opportunities for targeted faculty development.

While quantitative ratings offer standardized metrics for comparison, research consistently shows they often lack the depth and specificity needed to guide meaningful instructional improvement [1,3]. Numerical scores may indicate that students are dissatisfied with certain aspects of teaching, but

they rarely illuminate the specific behaviors, techniques, or circumstances that contribute to effectiveness or ineffectiveness. In contrast, open-ended comments provide contextually rich perspectives that can highlight nuanced aspects of teaching, offer concrete examples, and identify specific areas for improvement that standardized questions might miss [4]..

In the Philippine higher education system, State Universities and Colleges (SUCs) implement standardized evaluation frameworks that ensure consistency across institutions. The institutional evaluation instrument used in this study assesses faculty across four dimensions: Commitment, Knowledge of Subject, Teaching for Independent Learning, and Management of Learning, which are commonly employed evaluation criteria across Philippine state universities [5]. These dimensions align with international frameworks for teaching effectiveness while reflecting the specific priorities and context of Philippine public higher education [6].

The implementation of standardized instruments across SUCs represents a significant advancement in maintaining quality standards in Philippine public higher education. However, while the Likert-scale portions of these evaluations provide summative assessment data, the unstructured student comments that accompany them contain valuable insights that remain largely untapped due to analysis challenges. This underutilization represents a critical loss of student voice in the evaluation process and missed opportunities for targeted faculty development within the state university system. Manually analyzing large volumes of text comments is time-intensive, potentially subjective, and challenging to standardize across evaluators [7]. For many institutions, resource constraints mean that qualitative feedback receives only cursory review or is not systematically analyzed at all.

Recent advances in computational methods, particularly in natural language processing and machine learning, offer promising approaches to address these challenges while maintaining transparency in educational decision-making [8,9]. Educational researchers have increasingly applied text analysis techniques to various forms of qualitative data, driven by the need for more systematic approaches to understanding large volumes of unstructured feedback [10,11]. Among machine learning techniques, decision tree algorithms offer a distinct advantage of transparency and interpretability that aligns particularly well with educational contexts where stakeholder understanding and trust are paramount [12].

Unlike "black box" algorithms such as neural networks or deep learning models, decision trees generate explicit rules and decision pathways that can be directly understood by educational stakeholders without technical expertise, supporting transparent data-driven educational decision-making. Decision tree analysis is a more robust and intuitive approach for analyzing and interpreting student evaluation scores compared to more common parametric statistical approaches [13]. The human-readable nature of decision tree rules means that faculty can understand why their comments were classified in particular ways, administrators can validate the reasoning behind classifications, and evaluation committees can incorporate these insights into their deliberations with full transparency.

Despite the potential of decision tree methods for analyzing teaching evaluation comments, several important gaps exist in current research. First, few studies have successfully aligned computational text analysis with established institutional teaching evaluation frameworks, limiting practical implementation and stakeholder acceptance [14]. Second, while some studies achieve high accuracy, they often sacrifice interpretability, limiting their utility for educational practitioners who need to understand and trust the decision-making process [15,16]. Third, most existing research focuses on western educational contexts, with limited exploration of applications in Philippine higher education where multilingual feedback, cultural factors, and institutional structures may differ significantly. Fourth, previous work has not adequately addressed the balance between computational sophistication and practical usability—systems may perform well technically but remain too complex or opaque for routine institutional use.

To address these limitations and contribute to transparent data-driven educational decision-making in Philippine SUCs, this study pursues the following objectives:

1. Develop a comprehensive classification framework for student comments aligned with the standardized SUC teaching effectiveness evaluation instrument

2. Train and evaluate decision tree models to automatically classify student comments while prioritizing both accuracy and interpretability

3. Identify and validate key decision points and rules that predict effective teaching in each evaluation dimension

4. Assess the practical utility and interpretability of decision tree-based classification for supporting teaching evaluation and improvement

5. Demonstrate how decision tree approaches can enable transparent data-driven educational decision-making that maintains stakeholder understanding and trust within the Philippine SUC context

Material and methods

2.1 Research Design

This study employed a mixed-methods sequential exploratory design specifically tailored to support transparent educational decision-making. The qualitative phase involved developing a coding framework based on the institutional teaching evaluation instrument through thematic analysis of a subset of student comments. The quantitative phase focused on training and evaluating decision tree models using this framework, with particular attention to interpretability and transparency. This sequential design ensures that the computational classification is grounded in human understanding of teaching effectiveness while leveraging machine learning for efficiency and consistency.

2.2 Data Collection and Ethical Considerations

Teaching evaluation data was collected from all four colleges during the second semester of academic year 2023-2024. All participants provided informed consent for their anonymized evaluation data to be used for research purposes. The study was conducted in accordance with the Declaration of Helsinki and followed institutional guidelines for research involving human participants. All personally identifiable information (instructor names, student identifiers, course codes) was removed prior to analysis to ensure participant confidentiality. Students were informed during the evaluation process that their anonymized feedback might be used for institutional research to improve teaching quality. The dataset represents approximately 85% of all courses offered during this period, with non-participating courses primarily consisting of individualized instruction or courses with insufficient enrollment for meaningful evaluation. The dataset included 4,410 student comments from 347 course sections taught by 147 instructors across four colleges: College of Teacher Education (CTE: 44 instructors, 1,245 comments), College of Business and Management (CBM: 46 instructors, 1,387 comments), College of Computer Studies (CCS: 47 instructors, 1,542 comments), and College of Fisheries (COF: 10 instructors, 236 comments). Comments were responses to the open-ended prompt: "Please provide any additional feedback about the instructor's teaching effectiveness." The length of comments ranged from 1 to 398 words, with a mean length of 52.3 words (SD = 43.2), median of 38 words, and mode of 15 words. Approximately 68% of comments fell between 20-80 words, representing substantive feedback rather than brief statements.

2.3 Coding Framework Development

The institutional teaching evaluation instrument served as the foundation for the coding framework, ensuring perfect alignment between the text analysis and evaluation criteria. For each of the four dimensions in the evaluation instrument, specific indicators of positive and negative comments were developed through iterative thematic analysis.

Three experienced faculty evaluators from the institution, each with at least 10 years of experience in teaching evaluation and collectively representing different disciplinary backgrounds (Education, Business, and Computer Studies), independently coded a random sample of 250 comments using the framework. Inter-rater reliability was assessed using Cohen's kappa, with initial agreement

of $\kappa = 0.72$, indicating substantial agreement but with room for improvement. Following calibration sessions to discuss discrepancies, refine coding definitions, and establish shared understanding of ambiguous cases, the refined framework was applied to an additional 950 comments, achieving a final inter-rater reliability of $\kappa = 0.81$, indicating excellent agreement and validating the framework's clarity and applicability.

The final coding framework identified specific linguistic patterns, keywords, and contextual cues associated with each teaching dimension. For example, comments coded as positive for "Knowledge of Subject" typically included terms like "clear explanations," "knowledgeable," "well-prepared," "gives relevant examples," while negative comments included "confusing," "unclear," "outdated information," or "doesn't understand questions." This process resulted in a manually coded dataset of 1,200 comments (the initial 250 plus the additional 950) that served as the training data for machine learning models.

2.4 Data Preprocessing

Text preprocessing followed standard natural language processing procedures adapted for the educational context while maintaining interpretability. The preprocessing pipeline included:

1. Case normalization: Conversion to lowercase to ensure consistent matching (e.g., "GOOD" and "good" treated identically)

2. Punctuation handling: Removal of punctuation marks and special characters while preserving meaningful contractions (e.g., "doesn't" preserved rather than split into "doesn t")

3. Stop word removal: Elimination of common stop words using a customized list that preserved educationally meaningful terms. Standard stop word lists were modified to retain words like "clear," "understand," "helpful," which carry significant meaning in teaching evaluation contexts

4. Lemmatization: Reduction of words to their base forms using the WordNet lemmatizer (e.g., "explaining," "explained," "explains" all reduced to "explain"), enabling the model to recognize conceptually related terms

5. Spelling standardization: Correction of common spelling errors and variations using a domain-specific dictionary developed from the corpus (e.g., "knowlegable" to "knowledgeable," "organised" to "organized")

Feature extraction employed multiple approaches to capture different aspects of the text:

Bag-of-words representation: Term frequency counts for individual words and meaningful bigrams (two-word phrases)

Lexical features: Comment length (word count), vocabulary richness (unique word ratio), and sentence complexity measures

Semantic features: Frequency of subject-specific terminology relevant to each teaching dimension

Sentiment measures: Polarity scores using validated educational sentiment lexicons adapted from general sentiment analysis tools

This multi-faceted feature extraction approach enabled the decision tree models to consider not just the presence of specific words but also stylistic and structural characteristics of comments that might indicate teaching effectiveness.

2.5 Algorithm Selection Rationale

Decision tree algorithms were selected for this study based on three primary considerations aligned with the needs of educational assessment in Philippine SUCs:

First, INTERPRETABILITY: Unlike black-box algorithms such as neural networks or support vector machines, decision trees generate human-readable rules that can be directly understood by faculty, administrators, and evaluation committees without technical expertise [17,18]. This transparency supports the institutional requirement for explainable evaluation processes, particularly

important given the career implications of teaching evaluations. Faculty members can understand exactly which features in their student comments led to particular classifications, enabling targeted improvement efforts.

Second, PERFORMANCE: Despite their interpretability, decision tree ensembles (particularly Random Forests) have demonstrated competitive performance with more complex models in text classification tasks [19]. Previous studies have shown decision trees achieving 78-88% accuracy in similar educational text classification tasks [20,21], suggesting they can deliver practical utility without sacrificing explainability.

Third, PRACTICAL UTILITY: The explicit decision rules generated by decision trees can be directly applied by human evaluators for manual classification when needed, and can inform the development of clearer evaluation criteria and faculty development programs [13]. This dual utility—as both an automated classification tool and a framework for understanding teaching effectiveness patterns—distinguishes decision trees from alternative approaches. Administrators can use the identified rules to develop rubrics, training materials, and evaluation guidelines that reflect the patterns observed in student feedback.

Alternative algorithms were considered but deemed less suitable for this context. Support Vector Machines (SVMs), while potentially offering marginally higher accuracy, lack the transparency required for educational decision-making [22]. Neural networks and deep learning models, despite their power in many text classification tasks, function as "black boxes" that cannot provide the explainable decision pathways needed in evaluation contexts where fairness and accountability are paramount [23]. Topic modeling approaches like Latent Dirichlet Allocation (LDA) could identify themes but do not provide classification aligned with the institutional evaluation framework. Transformer-based models, while state-of-the-art for many NLP tasks, were considered beyond the scope given computational constraints, the small dataset size, and the priority placed on interpretability over marginal performance gains.

2.6 Model Development

Separate decision tree models were developed for each teaching effectiveness dimension (Commitment, Knowledge of Subject, Teaching for Independent Learning, Management of Learning) plus an overall effectiveness model. This dimension-specific approach allows for targeted analysis and intervention within each evaluation area. The manually coded dataset (n = 1,200) was strategically split into training (70%, n = 840), validation (15%, n = 180), and test (15%, n = 180) sets using stratified sampling to ensure balanced representation of classes within each subset.

Multiple algorithm variants were implemented and compared:

1. CART (Classification and Regression Trees): Using Gini impurity as the splitting criterion, representing the classical decision tree approach [17]

2. C4.5: Utilizing information gain ratio for attribute selection, which normalizes information gain by the intrinsic information of the split to prevent bias toward attributes with many values [18]

3. Random Forest: Ensemble method combining multiple decision trees trained on bootstrap samples of the data, aggregating their predictions to reduce overfitting while maintaining interpretability through feature importance rankings [19]

For hyperparameter optimization, grid search with 5-fold cross-validation was employed, focusing on parameters that maintain interpretability while optimizing performance:

- Maximum depth (3-15): Balances model accuracy with interpretability; shallower trees are more interpretable but may underfit, while deeper trees capture more patterns but become harder to interpret

- Minimum samples per leaf (5-35): Prevents overfitting by requiring sufficient samples to support each classification decision

- Splitting criterion: Comparison between Gini impurity and information gain to identify which better captures teaching effectiveness patterns

- Class weight balancing: Addressing potential class imbalance in the training data to ensure the model performs well on both positive and negative comments

The optimal hyperparameters were selected based on validation set performance, then final performance was assessed on the held-out test set to provide an unbiased estimate of real-world performance.

2.7 Use of Artificial Intelligence in Research Process

During the preparation of this manuscript, the authors used AI language tools (Grammarly) in a limited capacity for the following purposes: (1) grammar and language editing to improve clarity and readability of certain sections, (2) assistance in formatting references to ensure consistency with citation style requirements, and (3) restructuring of sentences for improved flow while maintaining the original meaning and content generated by the authors.

All conceptualization, methodology, data analysis, interpretation of results, and substantive intellectual content were developed entirely by the authors without AI assistance. The decision tree algorithms, coding framework, statistical analyses, and all research findings are original work of the authors. After using AI tools for language editing, the authors reviewed and edited all content and take full responsibility for the accuracy and integrity of the final manuscript.

Theory and calculation

The mathematical foundation of decision trees relies on information-theoretic measures to determine optimal splits that maximize the separation between classes. Decision trees use entropy H(S) to measure the impurity of a dataset S containing examples from k classes:

$$H(S) = -\sum_{i=1}^{k} p_i \log_2(p_i) \quad (1)$$

where p_i is the proportion of examples belonging to class i.

Information gain, which measures the reduction in entropy after a split, is calculated as:

$$IG(S,A) = H(S) - \sum_{v \in Values(A)} |S_v|/|S| \, H(S_v) \quad (2)$$

where A is the attribute being evaluated for splitting, Values(A) are the possible values of attribute A, and S_v is the subset of S where attribute A has value v.

Alternatively, the Gini impurity measure can be used:

$$Gini(S) = 1 - \sum_{i=1}^{k} p_i^2 \quad (3)$$

The Gini split for an attribute A is calculated as:

$$Gini\_split(S,A) = \sum_{v \in Values(A)} |S_v|/|S| \, Gini(S_v) \quad (4)$$

For optimal feature selection, the algorithm selects the attribute A that maximizes information gain:

$$A^* = \text{argmax}_A \, IG(S,A) = \text{argmax}_A \, [H(S) - \sum_{v \in Values(A)} |S_v|/|S| \, H(S_v)] \quad (5)$$

The Random Forest algorithm achieved optimal performance through ensemble approach:

$$\hat{y} = 1/B \sum_{b=1}^{B} T_b(x) \quad (6)$$

where B is the number of trees, T_b is the b-th tree trained on a bootstrap sample, and y is the final prediction.

Feature importance in Random Forest is calculated using mean decrease in impurity:

$$FI_j = 1/B \sum_{b=1}^{B} \sum_{t \in T_b} p(t) \, \Delta i(t,j) \quad (7)$$

where p(t) is the proportion of samples reaching node t, and $\Delta i(t,j)$ is the impurity decrease when splitting on feature j at node t.

Results and Discussion

Model Performance Analysis

The decision tree models demonstrated robust performance across all teaching effectiveness dimensions, successfully achieving accurate classification while maintaining the interpretability essential for transparent educational decision-making.

Table 1. Algorithm Performance Comparison (Overall Teaching Effectiveness)

Algorithm	Accuracy (%)	Precision	Recall	F1-Score	AUC-ROC	MCC
CART	80.8	0.82	0.79	0.80	0.84	0.61
C4.5	83.6	0.85	0.81	0.83	0.87	0.67
Random Forest	84.2	0.86	0.82	0.84	0.90	0.69

The Random Forest algorithm emerged as the optimal choice, achieving 84.2% accuracy while maintaining interpretability through feature importance rankings and accessible decision rules. This superior performance can be attributed to its ability to reduce overfitting through bootstrap aggregating (bagging) while preserving the interpretable nature of individual decision trees. The high AUC-ROC of 0.90 indicates excellent discrimination between positive and negative teaching effectiveness comments across the probability spectrum.

These performance metrics compare favorably with previous studies in educational text classification. Park and Dooris (2020) reported 76-82% accuracy using decision trees for predicting teaching evaluation scores [13], while Ahmed et al. (2022) achieved 79% accuracy analyzing online teaching evaluations [15]. Our 84.2% overall accuracy represents a meaningful advancement, particularly given our focus on maintaining interpretability throughout the modeling process. The Matthews Correlation Coefficient (MCC) of 0.69 indicates strong agreement between predictions and actual classifications, accounting for class imbalance.

Table 2. Random Forest Model Performance by Teaching Effectiveness Dimension

Model	Accuracy (%)	Precision	Recall	F1-Score	AUC-ROC	MCC
Overall Effectiveness	84.2	0.86	0.82	0.84	0.90	0.69
Commitment	77.6	0.79	0.75	0.77	0.82	0.55
Knowledge of Subject	88.1	0.90	0.86	0.88	0.93	0.76
Teaching for Independent Learning	81.3	0.83	0.79	0.81	0.86	0.62
Management of Learning	83.7	0.85	0.81	0.83	0.88	0.66

Examining performance across the four standardized dimensions revealed fascinating patterns in how students articulate different aspects of teaching effectiveness. The Knowledge of Subject dimension achieved the highest accuracy at 88.1%, suggesting that students possess remarkably consistent vocabularies when discussing content expertise and instructional clarity. This aligns with findings by Ruiz-Alfonso and León (2019) [24] who identified content expertise as the most clearly articulated aspect of teaching effectiveness in student evaluations.

The Commitment dimension showed the lowest accuracy at 77.6%, likely reflecting the more subjective nature of commitment assessments and greater variability in how students express observations about instructor dedication and availability. This finding parallels observations by Fan et al. (2019) [25] regarding the subjective and potentially biased nature of commitment assessments in student evaluations.

Table 3. Overall Effectiveness Model Performance by College

College	Sample Size	Accuracy (%)	Precision	Recall	F1-Score	MCC
College of Teacher Education (CTE)	1,245	88.3	0.90	0.86	0.88	0.76
College of Business and Management (CBM)	1,387	85.7	0.87	0.83	0.85	0.71
College of Computer Studies (CCS)	1,542	86.1	0.88	0.84	0.86	0.72
College of Fisheries (COF)	236	75.2	0.77	0.72	0.74	0.49

The College of Teacher Education achieved the highest accuracy at 88.3%, likely reflecting education students' familiarity with pedagogical terminology and structured evaluation approaches. The College of Fisheries exhibited notably lower accuracy at 75.2%, which can be attributed to both the smaller sample size and potentially more specialized terminology related to field-based learning.

Decision Rules and Feature Importance

The interpretability of decision tree models is demonstrated through the explicit decision rules they generate. These rules provide transparent classification logic that educational stakeholders can understand, validate, and apply.

Table 4. Sample High-Confidence Decision Rules by Dimension

Dimension	Decision Rule	Confidence (%)	Support
Knowledge of Subject	IF "clear" AND "explain" AND NOT "confusing" THEN Positive	93.2	187
Knowledge of Subject	IF "outdated" OR ("incorrect" AND "information") THEN Negative	89.5	156
Knowledge of Subject	IF "examples" AND ("relevant" OR "practical") THEN Positive	91.7	203
Commitment	IF "available" AND ("office hours" OR "help") THEN Positive	87.4	142
Commitment	IF ("late" AND "always") OR "absent" THEN Negative	92.8	168
Commitment	IF "prepared" AND ("organized" OR "ready") THEN Positive	85.9	134
Teaching for Independent Learning	IF "think" AND ("critically" OR "independently") THEN Positive	88.6	129
Teaching for Independent Learning	IF ("spoon-fed" OR "memorize") AND "only" THEN Negative	86.3	118
Teaching for Independent Learning	IF "encourage" AND ("questions" OR "participation") THEN Positive	84.7	145
Management of Learning	IF "activity" AND ("engaging" OR "interactive") THEN Positive	90.4	198
Management of Learning	IF ("disorganized" OR "unprepared") AND "class" THEN Negative	91.6	173
Management of Learning	IF "materials" AND ("helpful" OR "useful") THEN Positive	87.2	161

These rules demonstrate how decision trees create transparent, actionable classification criteria that can be directly communicated to faculty for improvement purposes. The high confidence levels (84.7%-93.2%) indicate reliable classification patterns aligned with the standardized evaluation framework.

Table 5. Top 15 Predictive Features for Overall Teaching Effectiveness

Rank	Feature	Importance Score	Primary Dimension	Polarity
1	"clear"	0.094	Knowledge of Subject	Positive
2	"helpful"	0.087	Commitment	Positive
3	"engaging"	0.081	Teaching for Independent Learning	Positive
4	"confusing"	0.073	Knowledge of Subject	Negative
5	"examples"	0.068	Knowledge of Subject	Positive
6	"available"	0.064	Commitment	Positive
7	"activities"	0.061	Management of Learning	Positive
8	"late"	0.055	Commitment	Negative
9	"feedback"	0.052	Management of Learning	Positive
10	"think"	0.047	Teaching for Independent Learning	Positive
11	"organized"	0.044	Management of Learning	Positive
12	"boring"	0.041	Teaching for Independent Learning	Negative
13	"prepared"	0.039	Commitment	Positive
14	"interactive"	0.037	Management of Learning	Positive
15	"understanding"	0.035	Knowledge of Subject	Positive

This feature importance analysis reveals that students prioritize teaching clarity, instructor helpfulness, and engaging instruction as the most significant factors when evaluating faculty. The dominance of "clear" (importance score 0.094) as the top predictive feature aligns with extensive research showing that instructional clarity is fundamental to teaching effectiveness [24].

Expert Validation Results

To validate the automated classifications, a comprehensive expert validation study was conducted comparing model predictions with independent human evaluations.

Table 6. Agreement Between Expert Evaluations and Decision Tree Classifications

Model	Cohen's Kappa	Percentage Agreement	Sensitivity	Specificity	Positive Predictive Value
Overall Effectiveness	0.78	86.3	0.84	0.88	0.87
Commitment	0.71	82.7	0.79	0.85	0.82
Knowledge of Subject	0.82	89.4	0.88	0.91	0.89
Teaching for Independent Learning	0.74	84.1	0.81	0.87	0.84
Management of Learning	0.76	85.6	0.83	0.88	0.86

The validation results demonstrate that decision tree models effectively capture the patterns that human experts use when evaluating teaching effectiveness comments. Cohen's kappa values ranging from 0.71 to 0.82 indicate substantial to excellent agreement, exceeding the threshold of 0.70 typically considered adequate for high-stakes decision-making [26].

Practical Implementation Assessment

Beyond technical performance, the practical utility of the system was assessed through implementation testing with actual institutional users.

Table 7. Efficiency Comparison: Manual vs. Automated Analysis

Analysis Scale	Manual Time (hours)	Automated Time (minutes)	Total Time with Review (hours)	Time Reduction (%)	Quality Rating (1-5)
50 comments	2.1	0.8	0.6	71	4.2
100 comments	4.3	1.2	1.1	74	4.3
250 comments	10.8	2.1	2.8	74	4.1
500 comments	22.4	3.6	5.9	74	4.2
1000 comments	45.2	6.8	11.7	74	4.0

The system consistently achieved 74% time reduction across different scales of comment analysis, demonstrating significant efficiency improvements while maintaining quality standards.

Table 8. User Ratings of System Features (5-point scale: 1=Poor, 5=Excellent)

System Feature	Mean Rating	SD	CTE	CBM	CCS	COF
Classification Accuracy	4.2	0.6	4.4	4.3	4.1	3.8
Decision Rule Clarity	4.6	0.4	4.7	4.6	4.5	4.3
Visual Representations	4.4	0.5	4.5	4.4	4.3	4.0
Actionable Insights	4.3	0.7	4.5	4.4	4.2	3.9
Time Efficiency	4.7	0.3	4.8	4.7	4.6	4.5
Faculty Development Utility	4.1	0.8	4.4	4.2	4.0	3.6
System Usability	4.3	0.6	4.4	4.3	4.2	4.0
Overall Satisfaction	4.4	0.5	4.6	4.4	4.3	4.0

User feedback revealed consistently high satisfaction across system features, with decision rule clarity receiving the highest rating (4.6/5.0), reinforcing the value of transparent approaches in educational contexts.

Discussion

This study successfully demonstrates that decision tree algorithms can effectively classify unstructured student comments according to the standardized teaching effectiveness framework while maintaining the transparency essential for educational decision-making. The research addresses critical gaps in educational analytics by developing an approach that balances analytical sophistication with interpretability, specifically tailored to the evaluation context used across Philippine State Universities and Colleges.

The overall accuracy of 84.2% achieved by the Random Forest model represents a significant advancement in educational text classification, particularly considering the interpretability maintained. This performance level compares favorably with recent studies while preserving transparent decision-making pathways crucial for educational stakeholder acceptance in state university environments.

The transparent nature of decision tree rules addresses fundamental concerns about algorithmic decision-making in education, particularly important in public university contexts where accountability and transparency are paramount. Unlike black-box approaches, the explicit decision pathways enable educational practitioners to understand, validate, and act upon automated recommendations within the framework of standardized evaluation criteria.

The practical implementation results demonstrate that the system delivers real value beyond technical accuracy. The consistent 74% time reduction means that institutions can process comprehensive qualitative feedback that would otherwise be impractical to analyze systematically. These efficiency gains can enable more frequent evaluation cycles, more comprehensive analysis, or reallocation of resources to other quality improvement initiatives.

Limitations and Future Directions

Several limitations should be acknowledged when interpreting these findings and considering future research directions.

DATA SCOPE: This study analyzed comments from a single university campus over one semester. While the implementation of standardized evaluation criteria suggests potential generalizability, validation across multiple institutions and time periods is needed before claiming broad applicability. Future research should conduct cross-institutional validation to assess the framework's transferability to other state universities with different student populations and institutional contexts.

LANGUAGE CONSIDERATIONS: This analysis focused on English-language comments. Many students in Philippine higher education use code-mixing (Taglish: Tagalog-English combinations, or English mixed with local languages), which were excluded from this analysis due to preprocessing challenges. Developing multilingual classification capabilities represents an important direction for future work to ensure inclusivity and capture the full range of student feedback.

VALIDATION APPROACH: While expert validation ($\kappa = 0.78$) demonstrates strong agreement, the reliance on faculty evaluators from the same institution may introduce institutional bias. External validation using evaluators from other institutions would strengthen confidence in the framework's broader applicability.

TEMPORAL DYNAMICS: Student feedback patterns may vary across different points in the semester, different course types, and different academic contexts. Longitudinal studies examining temporal stability of the classification framework would provide valuable insights into whether patterns remain consistent or whether models require periodic updating.

IMPACT ASSESSMENT: While this study demonstrates technical feasibility and efficiency gains, it does not assess whether using the system leads to improved teaching practices or learning outcomes. Future research should examine the relationship between decision tree-derived insights and actual teaching improvement through longitudinal tracking of faculty who receive systematic qualitative feedback analysis.

ETHICAL CONSIDERATIONS: The automated analysis of student feedback raises important questions about algorithmic bias and fairness. This study did not systematically examine whether the classification system performs differently for comments about faculty of different demographics or for comments from students of different backgrounds. Ongoing bias auditing and fairness assessments should accompany any operational deployment.

Conclusion

This research provides State Universities and Colleges with a validated framework for systematically analyzing qualitative student feedback while maintaining transparency and stakeholder trust within the standardized evaluation system. The decision tree approach enables SUCs to:

1. Efficiently process large volumes of qualitative feedback that previously remained unanalyzed due to resource constraints, with demonstrated time reductions of 74% while maintaining quality standards

2. Support targeted faculty development based on explicit decision rules aligned with institutional evaluation standards, providing specific, actionable insights rather than vague recommendations

3. Maintain computational analysis alignment with standardized teaching effectiveness frameworks, ensuring consistency with institutional evaluation criteria and policies

4. Preserve transparency through interpretable decision pathways that faculty, administrators, and evaluation committees can understand and validate

5. Enable evidence-based decision-making about teaching improvement priorities within state university contexts

The 74% reduction in analysis time demonstrated in implementation testing represents a significant practical advancement that enables more comprehensive utilization of student feedback in institutional assessment and improvement processes across the SUC system.

Future research should focus on cross-institutional validation, multilingual extension, longitudinal studies examining the relationship between decision tree insights and teaching improvement outcomes, and development of adaptive systems that learn from user feedback while maintaining alignment with standardized evaluation criteria.

The ultimate measure of success for educational analytics in SUCs lies not in technical performance metrics alone, but in their ability to support meaningful improvements in teaching and learning while maintaining the trust and engagement of the educational community. This study demonstrates that transparent, interpretable approaches can achieve both objectives, providing a foundation for continued development of human-centered educational technology specifically designed for State Universities and Colleges.

Acknowledgements

The author thanks the participating colleges (College of Teacher Education, College of Business and Management, College of Computer Studies, and College of Fisheries) and faculty members at the University of Antique Tario-Lim Memorial Campus for their cooperation and feedback throughout the research process. Special acknowledgment goes to the expert evaluators who assisted with the validation of the classification framework and the Program Heads and Area Heads who participated in the system implementation and testing phases. Recognition is also extended to the students whose thoughtful feedback made this research possible and whose voices this system aims to amplify in institutional decision-making processes.

References

Winchester TM, Winchester MK. A longitudinal investigation of the impact of faculty reflective practices on students' evaluations of teaching. Br J Edu Technol. 2014;45(1):112-124. https://doi.org/10.1111/bjet.12019

Constantinou C, Wijnen-Meijer M. Student evaluations of teaching and the development of a comprehensive measure of teaching effectiveness for medical schools. BMC Med Educ. 2022;22(1):113. https://doi.org/10.1186/s12909-022-03148-6

Hornstein HA. Student evaluations of teaching are an inadequate assessment tool for evaluating faculty performance. Cogent Educ. 2017;4(1):1304016. https://doi.org/10.1080/2331186X.2017.1304016

Deeley SJ, Brown RA. Student perceptions of written comments on assessment: Helpful or harmful? Assess Eval High Educ. 2023;48(2):261-273. https://doi.org/10.1080/02602938.2021.1986889

Agsalud PL. Teaching effectiveness of the teacher education faculty members in Pangasinan State University Asingan Campus, Philippines. Asia Pac J Multidiscip Res. 2017;5(1):16-22.

Setlhare-Kajane SS, Adeyemo DA. Teaching effectiveness: Conceptualization, research trajectories and classroom practice. Int J Learn Teach Educ Res. 2022;21(6):252-272. https://doi.org/10.26803/ijlter.21.6.14

Tai J, Ajjawi R, Umarova A. How are qualitative methods used in feedback research? A systematic review. Teach High Educ. 2022;27(4):534-550. https://doi.org/10.1080/13562517.2020.1742680

Fidalgo-Blanco Á, Sein-Echaluce ML, García-Peñalvo FJ. Identifying educational innovation patterns by learning analytics and machine learning techniques. Appl Sci. 2022;12(2):816. https://doi.org/10.3390/app12020816

Xing W, Li S. Building early warning systems for at-risk students using deep learning with textual comments. Educ Technol Res Dev. 2022;70(1):223-244. https://doi.org/10.1007/s11423-021-10067-6

Mohammed TA, Nandi D, Ahmed N. Impact of students' evaluation of teaching: A text analysis of the teachers' qualities by gender. Int J Educ Technol High Educ. 2021;18(1):52. https://doi.org/10.1186/s41239-020-00224-z

Varona-Bennet E, Guillemaud R, Torra V. Extracting actionable insights from student feedback: A topic modeling approach in engineering education. IEEE Trans Educ. 2023;66(1):90-97. https://doi.org/10.1109/TE.2022.3197590

Denny P, Kumar V, Giacaman N. Interpretable textual features for predicting programming performance. In: Proceedings of the 2019 ACM Conference on International Computing Education Research. New York: ACM; 2019.

p. 85-93. https://doi.org/10.1145/3291279.3339420

Park E, Dooris MJ. Predicting student evaluations of teaching using decision tree analysis. Assess Eval High Educ. 2020;45(5):776-793. https://doi.org/10.1080/02602938.2019.1697798

Papadopoulos T, Evangelidis G, Kalles D. Comparing decision trees and transformer architectures for automatic analysis of teaching evaluations. In: Proceedings of the 14th International Conference on Computer Supported Education. Setubal: SCITEPRESS; 2022. p. 187-194. https://doi.org/10.5220/0011037800003182

Ahmed N, Rifat-Ibn-Alam M, Akib GA, Shefat SN, Nandi D. Analyzing student evaluations of teaching in a completely online environment. Int J Mod Educ Comput Sci. 2022;14(6):13-24. https://doi.org/10.5815/ijmecs.2022.06.02

Li X, Zhang Y, Wang H. English teaching quality evaluation based on analytic hierarchy process and fuzzy decision tree algorithm. Comput Math Methods Med. 2022;2022:7425196. https://doi.org/10.1155/2022/7425196

Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and regression trees. Monterey: Wadsworth & Brooks/Cole Advanced Books & Software; 1984.

Quinlan JR. C4.5: Programs for machine learning. San Mateo: Morgan Kaufmann Publishers; 1993.

Breiman L. Random forests. Mach Learn. 2001;45(1):5-32. https://doi.org/10.1023/A:1010933404324

Sun J, Zhu Y, Jiang H. Decision tree-based thematic structure analysis of student comments in teaching evaluations. In: Proceedings of the 9th International Conference on Learning Analytics & Knowledge. New York: ACM; 2019. p. 275-284. https://doi.org/10.1145/3303772.3303809

Demšar E, Ahoniemi T, Ihantola P. Combining decision trees and learning analytics for student performance prediction in programming courses. In: Proceedings of the 2020 ACM Conference on Innovation and Technology in Computer Science Education. New York: ACM; 2020. p. 243-249. https://doi.org/10.1145/3341525.3387373

Trstenjak J, Donko D. Support vector machine classification of students' behavior in online educational settings. Neural Comput Appl. 2021;33(12):4243-4256. https://doi.org/10.1007/s00521-020-05257-6

Malinka K, Rodríguez-Triana MJ, Spikol D. Automatic detection of instructional strategies from teacher-student discourse using classification models. Int J Artif Intell Educ. 2023;33(1):163-187. https://doi.org/10.1007/s40593-021-00271-x

Ruiz-Alfonso Z, León J. Teaching quality: Relationships between passion, deep strategy to study and prior achievement. Front Psychol. 2019;10:1963. https://doi.org/10.3389/fpsyg.2019.01963

Fan Y, Shepherd LJ, Slavich E, Waters D, Stone M, Abel R, Johnston EL. Gender and cultural bias in student evaluations: Why representation matters. PLoS ONE. 2019;14(2):e0209749. https://doi.org/10.1371/journal.pone.0209749

Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159-174. https://doi.org/10.2307/2529310