International Journal of Innovative Research and Scientific Studies, 8(2) 2025, pages: 311-320

IJIRSS
International Journal of Innovative Research & Scientific Studies

ISSN: 2617-6548

URL: www.ijirss.com

Sentiment analysis of students' CQI: A comparative study using Textom and ChatGPT models (3.5 and 4.0)

Joungmin Kim[1], Dohyun Kim[2], Yongwon Cho[3*]

[1]*Hyangseol Nanum, Liberal Art, Soonchunhyang University, Asan, Republic of Korea.*
[2,3]*Department of Computer Science, Soonchunhyang University, Asan, Republic of Korea.*

Corresponding author: Yongwon Cho (*Email: dragon1won@gmail.com*)

Abstract

This study explores text-based analysis and advanced AI-driven sentiment analysis using GPT-3.5 and GPT-4.0 models to evaluate college students' Continuous Quality Improvement (CQI) from students. The goal is to provide deeper insights into educational assessments by comparing and integrating both methods. Using Textom for keyword analysis, network visualization, and Ucinet6 NetDraw for CONCOR analysis, we processed a final dataset of 32,285 cleaned evaluations. Key terms such as "material," "test," "helpful," "liked," and "content" were identified through TF-IDF weighting, and the CONCOR analysis revealed one central opinion cluster and several sub-clusters focused on course content, teaching methods, and student participation. Additionally, sentiment analysis using GPT-3.5 and GPT-4.0 was conducted to categorize feedback into positive, negative, and neutral sentiments. The GPT-3.5 model demonstrated higher accuracy in understanding contextual nuances and detecting emotional intensity than traditional methods, highlighting areas of satisfaction like course materials and instructor engagement and identifying areas of dissatisfaction linked to evaluations and assignments. Integrating traditional Textom analysis and GPT-based sentiment analysis provides a comprehensive and actionable framework for understanding student feedback. This integration enables institutions to design targeted interventions, such as refining teaching practices, improving course content, and tailoring assessments to enhance student satisfaction and learning outcomes. The findings are particularly valuable in addressing challenges in remote and hybrid learning contexts, offering scalable solutions for adapting to evolving educational needs. By bridging traditional methods with AI-powered insights, this study underscores the transformative potential of AI in advancing academic quality.

Keywords: CQI, Educational assessment, GPT model, Sentiment analysis.

DOI: 10.53894/ijirss.v8i2.5158
Funding: This study received no specific financial support.
History: Received: 16 January 2025 / **Revised:** 17 February 2025 / **Accepted:** 24 February 2025 / **Published:** 7 March 2025

Competing Interests: The authors declare that they have no competing interests.
Authors' Contributions: J.K. and Y.C.; Methodology, J.K. and Y.C.; Data collection, J.K.; Data analysis, D.K. and Y.C.; Writing original draft preparation, J.K..; Writing review and editing, J.K. and Y.C.; supervision, J.K. All authors have read and agreed to the published version of the manuscript.
Transparency: The authors confirm that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained. This study followed all ethical practices during writing.
Acknowledgments: This research was supported by Soonchunhyang University. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. RS-2023-00218176).
Publisher: Innovative Research Publishing

1. Introduction

As the Ministry of Education removes restrictions on distance learning and allows online graduate courses, the need to improve the quality of both face-to-face and online lectures must continue. Lecture evaluations have been a critical tool for monitoring and enhancing the quality of lectures for over 30 years. Above all, since the ultimate purpose of lecture evaluation is to improve the quality of the class, it is crucial to deliver meaningful feedback to faculty for practical improvements rather than using the evaluations solely for ranking purposes [1-3].

2. Literature Review

A review of previous research on lecture evaluations reveals three main issues. First, lecture evaluation items are uniformly applied across all courses without considering the characteristics of each class, raising concerns about the appropriateness of the evaluation tools [4-7]. Second, a system has been implemented to increase participation rates by withholding grade access if students do not complete the evaluation. However, this mandatory system compromises the reliability of the results [8, 9]. Third, controversy exists over linking lecture evaluation results with teaching performance scores, particularly regarding the evaluation's role in summative or formative assessments [10, 11].

While quantitative evaluations using Likert scales are widely used, collecting and analyzing descriptive feedback from students is essential for gaining deeper insights into their attitudes and improving lecture quality. Descriptive feedback contains nuanced information that quantitative evaluations often fail to reveal [12]. However, analyzing descriptive feedback presents significant challenges due to its unstructured nature. Text data from lecture evaluations commonly include spelling errors, informal language, and varying levels of detail, complicating consistent and accurate analysis. Moreover, feedback is often imbalanced, with positive or neutral comments dominating negative ones, which can obscure critical areas needing improvement [13]. Manual analysis of descriptive feedback from large datasets can be time-intensive and subjective, as shown in research where analysis of 10,000 student evaluations required extensive resources and led to inconsistent outcomes. These limitations underscore the need for automated methods to improve the reliability and scalability of feedback analysis [14].

However, there are few analytical studies on descriptive lecture evaluation data, both domestically and internationally. Some studies have used advanced data analysis techniques to analyze large-scale descriptive feedback. Still, most of the existing research relies on subjective responses at the end of the semester, offering limited insights into students' attitudes toward lectures. Recent developments in deep learning and machine learning have led to renewed attention to artificial intelligence (AI), which can process large datasets and extract meaningful patterns from text data. Sentiment analysis algorithms, in particular, can mechanically analyze emotions and opinions embedded in text data, and this approach has been applied in various fields, including social media analysis and educational feedback [15, 16].

AI-driven approaches, particularly those leveraging advanced deep learning techniques, address many of the limitations inherent in traditional methods by automatically identifying patterns within large and diverse datasets [17]. Unlike conventional approaches that depend on predefined lexicons, AI-based sentiment analysis adapts to contextual variations, detects subtle emotional undertones, and processes unstructured feedback with significantly higher accuracy. For instance, sentiment analysis models in education can process descriptive feedback to identify student satisfaction and dissatisfaction trends that might go unnoticed. This ability to extract actionable insights from unstructured data enhances its practical application in real-time feedback systems [18]. AI-driven sentiment analysis has achieved high accuracy in identifying key themes in lecture evaluations, outperforming manual methods in speed and reliability. Transfer learning and pre-trained language models, such as GPT, BERT, or T5, also allow for domain-specific fine-tuning, further improving accuracy and relevance in educational contexts [14].

This study bridges the gap between quantitative and qualitative approaches by integrating sentiment and text frequency analysis. While text frequency analysis highlights recurring themes or keywords, sentiment analysis provides emotional context, enabling a deeper understanding of student feedback. For instance, identifying terms like "grading clarity" with a strongly negative sentiment highlights an area needing immediate attention. This dual approach enhances the reliability of evaluation results while equipping educators with actionable data to improve teaching practices, course design, and administrative policies [19, 20].

This study applies sentiment analysis using Chat GPT 3.5 and 4.0 models to evaluate descriptive feedback collected through university CQI (Continuous Quality Improvement) processes. These models excel at capturing nuanced meanings and subtle emotions, setting them apart from traditional frequency-based keyword analyses. Unlike earlier approaches, which focused primarily on lexical patterns, GPT models offer a robust ability to contextualize sentiment based on the phrasing and structure of feedback [14]

This research identifies critical factors influencing student satisfaction and dissatisfaction through comprehensive sentiment analysis. These findings inform specific recommendations, such as refining grading rubrics, improving instructor communication, and increasing the interactivity of lectures. Furthermore, the scalability and adaptability of GPT models allow institutions to implement real-time feedback systems capable of monitoring changes in student sentiment over a semester. By leveraging AI and deep learning, educators and administrators can significantly enhance their responsiveness to student needs, ultimately fostering a more supportive and effective learning environment.

3. Research Method

3.1. Data Cleaning

This study evaluated narrative lectures in the first semester of 2022-2023 at S University. There were 85,829 lecture evaluations of the raw material. We analyzed the morphemes using R's KoNLP package for data preprocessing. We deleted words, numbers, special characters, punctuation marks, space characters, and symbols that have little or no meaning in the analysis because of their high frequency of use or their minimal contribution to the analysis. In addition, we used the KoNLP package, a representative morpheme analyzer of R, to extract the morphemes of 'normalization.' We removed 18,484 cases with nonsensical answers, punctuation, spaces, and special symbols. We divided the types of university lectures into liberal arts, significant subjects, teacher-training curricula, and introductory undergraduate courses. The number of qualitative evaluations was high in the order of significant subjects > liberal arts > undergraduate basics course > teacher training curriculum. After refining, 67,345 cases remained, as shown in [Table 1].

While nonsensical answers and noisy data were removed during the cleaning process, this step risks excluding feedback that could provide valuable insights. For example, while often dismissed as irrelevant, high-frequency words may hold contextual significance in specific feedback clusters. Similarly, informal or fragmented language, common in negative feedback, may disproportionately affect the analysis of critical comments. Although removing such noise is necessary to improve analytical clarity, it introduces the potential for unintended data loss. Future studies should explore alternative cleaning strategies to retain contextually meaningful data while minimizing noise.

Table 1.
Data cleaning.

Type of course	Number of students' descriptive lecture evaluations (%)	Noise data (%)	Final data (%)
Liberal Art	28,297 (32.95)	5,836 (6.80)	22,461 (26.17)
Major	40,219 (46.860)	8,905 (10.37)	31,315 (36.49)
Teacher Training Course	614 (0.72)	88 (0.1)	526 (.61)
Introductory Course	16,699 (19.46)	3,655 (4.26)	13,043 (15.20)
Total	85,829 (100.00)	18,484 (21.53)	67,345 (78.47)

3.2. Network Analysis

Before applying the GPT models, Textom analysis was conducted to gain an initial understanding of the CQI. The top 100 most frequently appearing words in the refined data were selected for further analysis. UCINET was then used to examine the connectivity and centrality between these words, and NetDraw was employed to visualize keywords. For network analysis, Eigenvector centrality was chosen to highlight the most influential words, while Prestige centrality was applied to indicate which nodes held the most significant influence. Prestige centrality reflects that a node's importance increases when connected to other highly central nodes.

Before the network analysis and visualization, words were organized based on frequency, but it remains to be seen which words significantly influenced the overall feedback. To address this, nodes with high prestige were highlighted. The NetDraw package in UCINET was used for visualization, and the Convergence of Iterated Correlations (CONCOR) technique was applied to cluster words based on their similarity. This clustering allowed inferences regarding the meaning of a word by examining other words within the same cluster. Since word frequency alone did not reveal the students' underlying intentions, CONCOR analysis helped interpret the contextual meaning of frequently used words.

3.3. Performance Comparison of GPT-3.5 and GPT-4.0

This study employs Chat GPT 3.5 and 4.0, provided by OpenAI, to perform sentiment analysis. The same settings and learning environments are maintained for both models to ensure that any performance differences arise solely from technological improvements between the two versions. Both models use pre-trained natural language processing-based sentiment analysis algorithms, classifying feedback data into three sentiment categories: positive, negative, and neutral. Additionally, the sentiment analysis results are used to gauge students' satisfaction or dissatisfaction with specific lectures, categorizing the results by topic to identify patterns. If Chat GPT 4.0 demonstrates superior sentiment classification performance compared to Chat GPT 3.5, reasons for recommending Chat GPT 4.0 for CQI data analysis are presented.

To further enhance the relevance of the models, each generative AI model is fine-tuned using texts related to lecture evaluation extracted from our institution and others. This process utilizes key phrases and prompts specific to lecture evaluation feedback, improving the models' ability to provide actionable insights. Despite their robust performance, both models may reflect biases inherent in their training data, such as misclassification of ambiguous or mixed sentiments. For instance, neutral feedback may be skewed toward positive or negative categories, and exaggerated expressions could result in overestimating certain sentiment types. Future studies could mitigate these limitations by incorporating domain-specific fine-tuning or applying multi-label classification techniques to better capture the nuanced nature of descriptive feedback.

4. Research Validation

4.1. Sentiment Classification Accuracy

A comparison was conducted using a manually labeled dataset to validate the performance of GPT-3.5 and GPT-4.0 in sentiment classification. A subset of student feedback was randomly selected and labeled by human annotators as positive, neutral, or negative. The AI-generated sentiment classifications were then compared against human-assigned labels, and precision, recall, and F1-score were computed for each sentiment category to evaluate model performance.

To ensure a fair comparison, GPT-3.5 and GPT-4.0 were tested under identical settings and learning environments, preventing external factors from influencing performance differences. Each model employed pre-trained natural language processing (NLP) based sentiment analysis algorithms to classify feedback data into three sentiment categories: positive, negative, and neutral.

Additionally, sentiment analysis results were leveraged to assess students' satisfaction or dissatisfaction with specific lectures, categorizing feedback by topic to identify sentiment patterns across different course components. The models were also fine-tuned using domain-specific texts related to lecture evaluations extracted from institutional and external sources. This fine-tuning process incorporated key phrases and prompts specific to lecture evaluation feedback, enhancing the models' ability to extract actionable insights.

Despite these optimizations, both models exhibited limitations, such as potential biases in sentiment classification. For example, neutral feedback was sometimes misclassified as either positive or negative, and exaggerated expressions led to an overestimation of sentiment intensity. Addressing these challenges through domain-specific fine-tuning or implementing multi-label classification techniques could improve sentiment analysis accuracy in future studies.

GPT-3.5 achieved an overall accuracy of 60.57%, while GPT-4.0 significantly underperformed, achieving only 6.87% accuracy. This suggests GPT-3.5 was more effective at capturing contextual nuances in student feedback.

4.2. Cosine Similarity Analysis

Cosine similarity was employed to further assess how closely AI-generated sentiment classifications aligned with human-labeled data. This metric calculates the cosine of the angle between two vectors, providing a measure of textual similarity.

The results indicated that:

- GPT-3.5 exhibited a cosine similarity score of 28.90%, indicating a relatively strong alignment with human sentiment classifications.
- GPT-4.0 achieved only 7.60%, demonstrating significantly lower alignment with ground truth data.

These findings reinforce the conclusion that GPT-3.5 provides superior sentiment classification for CQI analysis, while GPT-4.0 struggles with accurate sentiment identification.

4.3. Error Analysis and Misclassification Patterns

An error analysis was conducted on misclassified instances to gain deeper insights into the limitations of AI-based sentiment classification. A sample of incorrectly classified feedback was reviewed to identify recurring misclassification patterns.

Key observations include:

- Neutral Sentiment Misclassification: GPT-4.0 frequently misclassifies neutral feedback as positive or negative, suggesting an inability to recognize balanced or non-opinionated language.
- Overestimation of Sentiment Intensity: Both models occasionally misclassified slightly positive comments as highly positive, leading to an inflated representation of student satisfaction.
- Handling of Ambiguous Statements: Feedback containing mixed sentiments (e.g., “The professor is great, but the grading is harsh”) is often misclassified due to the coexistence of both positive and negative elements.

These insights highlight areas where further model fine-tuning could enhance classification performance. Future research could improve sentiment analysis by incorporating contextual sentiment weighting, which would help models handle nuanced feedback with mixed opinions better.

4.4. Comparative Performance with Traditional Text Analysis

A comparative analysis using Textom keyword frequency analysis was conducted to evaluate whether AI-driven sentiment analysis offers improvements over traditional methods.

Textom, a widely used TF-IDF-based keyword extraction tool, identifies commonly occurring words but does not interpret sentiment. A comparison of GPT models and Textom revealed the following:

- Keyword-Based Analysis (Textom): Effectively identified frequently mentioned terms (e.g., “exam,” “assignment,” “fair”).
- Lacked sentiment interpretation and contextual analysis.
- GPT Sentiment Classification: Provides an additional layer of sentiment classification, allowing for nuanced feedback interpretation beyond keyword frequency.
- Combined Insights: Integrating keyword frequency analysis with AI-driven sentiment analysis enabled a more comprehensive understanding of student feedback.

This comparison underscores the advantages of using AI-driven methods in lecture evaluations, particularly for processing qualitative data to detect sentiment trends. By combining Textom's keyword extraction with GPT's sentiment classification, universities can gain deeper insights into student concerns and areas for improvement.

5. Results

This section presents the findings of the analysis, focusing on keyword extraction, sentiment distribution, and the CQI network analysis. The results are organized to address the research questions regarding the distinctive characteristics of positive, negative, and neutral feedback, the keyword patterns in the responses, and the sentiment variations across different review lengths.

5.1. Number of Letters

The descriptive lecture evaluations reveal that most feedback consists of short comments, with 67.11% being ten characters or less [Table 2]. Such brevity often lacks actionable detail, limiting its usefulness for meaningful class improvement. Given this, instructors are encouraged to guide students in providing more constructive feedback by introducing prompts or examples that emphasize specific areas of improvement, such as course content, assessment methods, or teaching style. This approach could enhance the quality of feedback while maintaining student engagement.

Table 2.
Number of letters in descriptive lecture evaluation.

Less than six letters (%)	10-Jun	20-Nov	21-50	51-100	101-150	Total
	letters (%)	letters (%)	letters (%)	letters (%)	letters (%)	(%)
35,062	10,136	10,874	8,031	2,193	49	67,347
-52.06	-15.05	-16.15	-11.92	-3.25	-0.07	-100

5.2. Keyword Extraction

The keywords extracted using TF-IDF values (Table 3) show significant focus on terms like "material," "test," "helpful," and "liked," indicating heightened attention to course content and evaluation methods, particularly in the context of online learning. Complaints regarding assignments and system issues were also prevalent. These findings suggest actionable strategies such as:

- Conducting workshops for instructors to refine course materials and ensure assignment clarity.
- Improving the user experience of digital platforms to reduce system-related complaints.

Table 3.
Top 60 words corresponding word by TF-IDF weight value.

Keyword	TF-IDF	Keyword	TF-IDF	Keyword	TF-IDF
Material	0.965	difficult	0.712	thanks	0.476
Test	0.964	easy	0.688	write	0.465
Helpful	0.961	better	0.675	satisfy	0.444
Liked	0.922	speed	0.632	hope	0.381
Content	0.882	team	0.629	hard	0.37
Sorry	0.868	challenge	0.623	fun	0.361
Practice	0.854	video	0.619	work	0.353
Assignment	0.841	appropriate	0.614	helpful	0.32
Improvement	0.828	great	0.606	listening	0.296
Understand	0.826	easily	0.576	professor	0.286
Hardship	0.826	teaching	0.548	problem	0.269
Feedback	0.803	time	0.535	course	0.246
Online	0.776	think	0.517	fast	0.231
Covid-19	0.772	offline	0.513	good	0.223
Various	0.757	PPT	0.505	many	0.22
Face-to-face	0.755	evaluation	0.505	interested	0.166
Explain	0.739	feel	0.503	funny	0.13
Knowledge	0.738	exam	0.487	take	0.127
Wish	0.736	course	0.485	difficulty	0.093
Disappointed	0.727	learn	0.482	data	0.081

5.3. Sentiment Analysis

Sentiment analysis categorized 67,347 responses into positive, neutral, and negative sentiments (Table 4). The study found that shorter comments (e.g., 5-10 characters) were predominantly positive (81.1%), while longer comments (101-150 characters) were overwhelmingly negative (86.1%), suggesting that students express dissatisfaction more thoroughly when providing detailed feedback.

To address this, institutions can implement mid-semester feedback sessions that encourage longer, more thoughtful

responses. Additionally, sentiment-based summaries can be shared with instructors, highlighting specific pain points like assignments or evaluation criteria to target improvement efforts.

Table 4.
Differences by length of descriptive lecture evaluation.

Group	Negative	Neutral	Positive	Total
1	0.331	0.533	0.773	0.753
2	0.278	0.531	0.811	0.664
3	0.221	0.529	0.799	0.537
4	0.149	0.525	0.826	0.41
5	0.075	0.517	0.846	0.242
6	0.026	0.501	0.861	0.172
Total	0.193	0.53	0.785	0.645

The visualization of sentiment keywords using GPT-3.5 and GPT-4.0 models (see Figure 1 and Figure 2) showed that positive feedback frequently included words like "thank," "helpful," and "good." In contrast, negative feedback often featured terms related to evaluations and assignments, such as "test" and "difficult." Neutral feedback included more general impressions, reflecting the overall course experience.

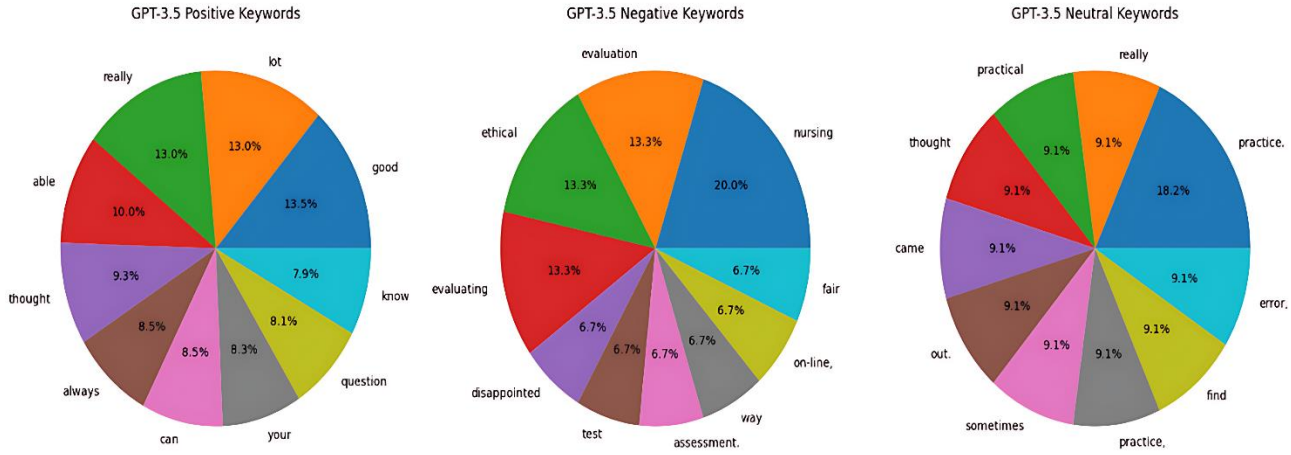

Figure 1.
Sentiment keywords GPT-3.5.

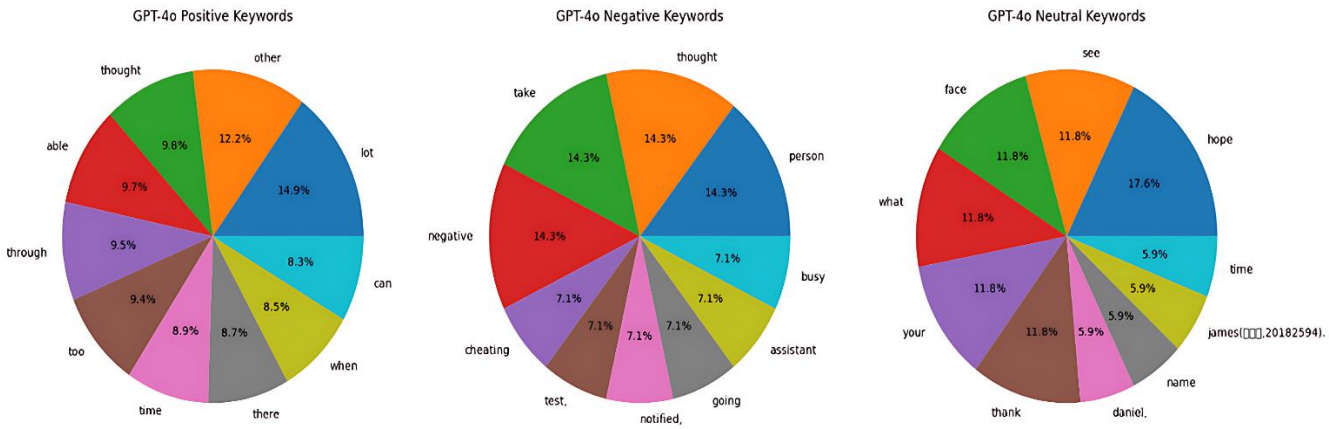

Figure 2.
Sentiment keywords GPT-4.0.

5.4. Network Analysis

Network analysis using UCINET and CONCOR was conducted to examine the relationships between key terms and to cluster similar words. Network analysis revealed key clusters of terms that highlight recurring themes in student feedback, as shown in Table 5. The Primary Cluster included terms such as “lecture,” “passion,” and “thank,” which reflect a general appreciation for instructor efforts. Meanwhile, Sub-cluster 4 focused on evaluation-related dissatisfaction, with keywords like “feedback,” “criteria,” and “improvement” pointing to concerns about inconsistencies in evaluation methods. These findings suggest actionable strategies, including introducing standardized rubrics to ensure fair and consistent evaluation criteria and the provision of detailed feedback on assignments to enhance student satisfaction. Furthermore, the N-gram

visualization [Figure 3] demonstrated the strong interconnectivity of terms like “professor,” “lecture,” and “feedback,” underscoring the critical role of instructors in shaping students' learning experiences. This highlights the need for targeted faculty training programs to improve teaching methods and communication skills, especially in remote and hybrid learning environments.

Table 5.
The result of UCINET NetDraw CONCOR.

Cluster	Keywords
Primary cluster 1 (Impression of lecture)	Passion, teach, situate, enthusiasm, chance, effort, thank, lecture, hardship, face, explanation, knowledge, experience.
Sub-cluster 2 (Prefer career-related content)	Develop, depart, relate, career, college, guide, job, graduate.
Sub-cluster 3 (Improving the quality of lectures)	Attendance, semester, professor, method, online, trouble, quality, course.
Sub-cluster 4 (Improvement of class feedback and evaluation methods)	Selected text in Word: Lesson, problem, manage, evaluate, feedback, assign, review, communicate, criteria, student, class, improvement, difficulty, announce, assignment.
Sub-cluster 5 (Differentiation of textbooks according to class types)	Textbook, process, experiment, theory, issue, video, voice, PPT, fast, image, clip, different, material, example.
Sub-cluster 6 (Student participatory class method preference)	Opportunity, thought, participate, concentrate, opinion, perspective, discussion, activity, addition, interaction.

The N-gram network visualization in Figure 3 further highlighted the interconnections between crucial terms. For example, “professor” was often associated with “lecture,” “course,” “content,” and “feedback.” This analysis reinforces the idea that students emphasize the role of instructors and teaching methods, particularly in remote learning contexts.

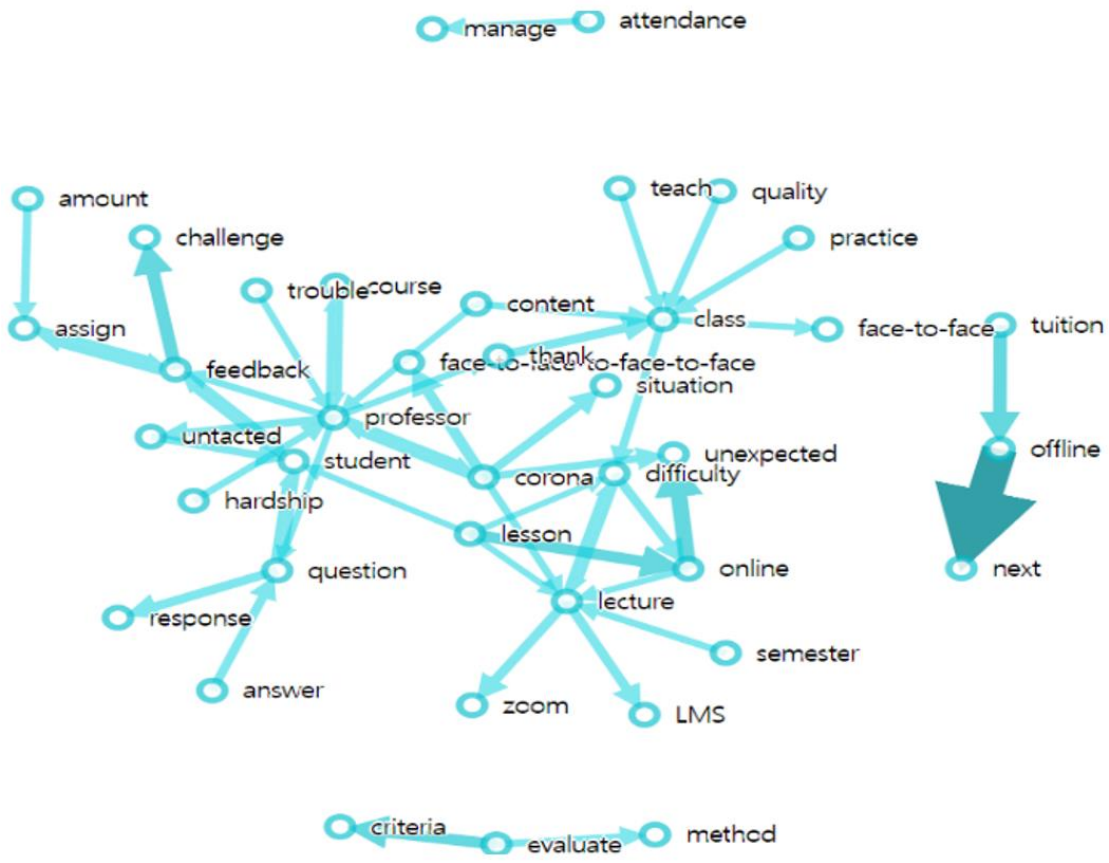

Figure 3.
N-gram network visualization.

5.5. GPT-3.5 vs GPT-4.0 Performance

Cosine Similarity was employed to compare the model-generated responses with the actual feedback to assess the accuracy of the GPT models in analyzing feedback. Cosine Similarity measures the similarity between two vectors using the cosine of the angle between them, quantifying the semantic similarity between two pieces of text. The formula for Cosine

Similarity is as follows (1):

$$\frac{A \cdot B}{\|A\| \times \|B\|} \quad (1)$$

Each generative AI (GPT-3.5 and GPT-4.0) model is fine-tuned using texts related to lecture evaluation extracted from our institution and others. The training utilizes key phrases presented in the lecture evaluation and various prompts. In addition, each result can be compared and analyzed through fine-tuning using open source-based natural processing (LLAVA). The new fine-tuning model provides lecture improvements based on learners' evaluations.

Additionally, Result Segmentation was used to measure how accurately each model classified feedback into positive, negative, and neutral sentiments. This metric, commonly referred to as Accuracy, is defined by the following formula (2):

$$\frac{\text{True Positives} + \text{True Negatives}}{\text{Total Sample}} \quad (2)$$

AAA represents the vector of the model-generated response, and BBB represents the vector of the actual feedback. The Cosine Similarity score ranges from 0 to 1, where a value closer to 1 indicates a higher similarity between the two texts. The performance comparison of GPT-3.5 and GPT-4.0 revealed that GPT-3.5 significantly outperformed GPT-4.0, achieving a Cosine Similarity score of 28.90% and an accuracy of 60.57%, compared to GPT-4.0's 7.60% and 6.87%, respectively, as shown in Table 6. This indicates that GPT-3.5 was more effective in capturing the contextual nuances of student feedback.

Despite GPT-4.0's underperformance in this study, its advanced contextual capabilities may still provide valuable insights if fine-tuned for domain-specific data. Therefore, the following strategies are proposed based on the analysis:

- GPT-3.5 is recommended for analyzing descriptive feedback for immediate application due to its higher accuracy and contextual understanding in this dataset.
- Future studies should focus on fine-tuning GPT-4.0 with lecture-specific datasets to leverage its potential to capture nuanced feedback.

Table 6.
The performances of GPT-3.5 vs GPT-4.0.

	Cosine Similarity	Result Segmentation (Accuracy)
GPT-3.5	28.90%	60.57%
GPT-4.0	7.60%	6.87%

6. Discussion

This study stands out by moving beyond traditional Textom-based network and sentiment analysis to utilize GPT-3.5 and GPT-4.0 models for a more in-depth analysis of descriptive lecture evaluations. The focus was on comparing and evaluating the differences and resulting distinctions between conventional text analysis methods and advanced AI-based natural language processing (NLP) techniques.

Regarding the existing Textom-based analysis, network analysis revealed students' most frequently used words and the associations between these words. Standard terms such as "thanks," "good," "face-to-face," and "hardship" were prominent, with students generally commenting on the overall class impression rather than specific aspects such as teaching methods or difficulty. Additionally, TF-IDF analysis highlighted frequently occurring keywords such as "materials," "exam," and "helpful," suggesting that students placed considerable emphasis on course materials and evaluation methods.

While the Textom-based analysis effectively identified recurring keywords in the feedback, it needed to be improved to capture the contextual meaning or nuances of emotions expressed in the input. Especially for negative feedback, analyzing the intensity or specific details of the feelings involved was challenging. The analysis using GPT-3.5 and GPT-4.0 models demonstrated distinct differences and advantages over the traditional Textom-based analysis, as outlined below:

- Deeper Contextual Understanding and Sentiment Analysis: GPT models are not limited to counting word frequencies; they can also comprehend the context and capture subtle nuances of emotions within the text. For instance, while the Textom analysis identified "thanks" as a frequently occurring word, the GPT models could differentiate whether "thanks" was used as a formal expression or a genuine indication of satisfaction with the course. GPT-3.5, in particular, exhibited superior performance in interpreting contextual nuances and emotional intensity, making it a more reliable tool for understanding the true sentiments expressed in feedback. In addition to lecture evaluation, GPT-3.5 has demonstrated superior performance in analyzing customer feedback on e-commerce platforms, accurately identifying strongly negative and positive sentiments. These capabilities underscore its versatility and highlight the importance of tailoring GPT models to specific domains for optimal performance.
- In-depth Analysis of Negative Feedback: Textom indicated that negative feedback was infrequent, but it struggled to assess the seriousness or specifics of dissatisfaction. GPT models, especially GPT-3.5, provided a more nuanced understanding of negative feedback by identifying specific emotions and dissatisfaction drivers. For instance, GPT-3.5 detected concerns about exam fairness and highlighted students' discomfort with the exam process and evaluation methods. This capability suggests that GPT-3.5 can assist educators in pinpointing actionable areas for improvement in course delivery and evaluation practices.
- Practical Implications of GPT-3.5's Superior Performance: The performance comparison revealed GPT-3.5's notable accuracy in contextual understanding, as evidenced by a Cosine Similarity score of 28.9% and a segmentation accuracy of 60.57%. Conversely, GPT-4.0 scored significantly lower, with a similarity score of 7.6% and an accuracy of 6.87%. These results emphasize GPT-3.5's practicality in real-world applications, as its higher accuracy allows for more reliable insights into student feedback. For immediate implementation, GPT-3.5 is recommended for descriptive feedback analysis, offering institutions a valuable tool for improving course quality. While GPT-3.5 demonstrated

superior performance in this study, GPT-4.0 holds significant potential for future applications. Training GPT-4.0 with domain-specific datasets, such as descriptive lecture evaluations, could significantly enhance its contextual understanding and sentiment analysis capabilities. This approach may allow GPT-4.0 to surpass GPT-3.5 in effectively analyzing nuanced feedback and generating actionable insights.

- Training GPT-4.0 with Domain-Specific Data: GPT-3.5 showed better results than GPT-4.0 by using two different models for lecture evaluation and sentiment analysis. Additionally, GPT-3.5 more accurately analyzed positive, negative, and neutral reviews on an e-commerce site and identified feedback, including strong negative emotions, in customer feedback analysis. These examples illustrate that generative AI models can produce varying results depending on the task. While continuous updates to generative AI models can improve their overall capabilities, it remains critical to develop customized models tailored to the specific data and objectives of the task to ensure more objective and precise analyses.
- Automation in Topic Extraction and Categorization: Unlike Textom, which depends on predefined word lists or manual input, GPT models automatically extract topics from feedback data, enabling greater flexibility and depth. For instance, GPT models can categorize feedback into specific issues, such as course content, assignments, or grading criteria, and analyze the emotional tone associated with each topic. This automation reduces the need for manual preprocessing and enhances the scalability of feedback analysis, making it a promising tool for large-scale datasets.
- Broader Implications for Educational Practice: The practical applications of GPT models extend beyond lecture evaluations, particularly in hybrid and remote learning contexts. For example, GPT-3.5's ability to detect dissatisfaction trends, such as inconsistent grading or inaccessible teaching methods, highlights the need for standardized rubrics and better communication between instructors and students. Furthermore, its nuanced understanding of positive feedback, such as appreciation for interactive teaching methods, can guide institutional efforts to promote best practices across all courses.

Future research should expand the use of GPT models by incorporating topic modeling to automatically categorize descriptive feedback and systematically analyze emotional patterns within each category. This approach could provide educators and administrators with targeted insights into specific aspects of courses that influence student satisfaction or dissatisfaction. Moreover, fine-tuning GPT-4.0 with domain-specific data may unlock its potential, address its current limitations, and enable more advanced contextual analysis.

Through these enhancements, GPT-based feedback analysis can support actionable educational practice improvements, ensuring institutions respond more effectively to students' needs and expectations.

7. Conclusion

This study highlights the potential of utilizing GPT-3.5 and GPT-4.0 models to analyze descriptive lecture evaluations, offering empirical evidence of how AI-based NLP techniques can outperform traditional Textom analysis in improving lecture assessments. Notably, GPT-3.5 demonstrated superior performance in understanding the context and emotions of feedback, making it a powerful tool for analyzing descriptive feedback data. These findings align with the study's goals and suggest actionable strategies for enhancing student-teacher feedback systems.

The findings emphasize several practical applications to improve feedback systems. First, GPT-3.5's ability to accurately identify dissatisfaction trends, such as inconsistent grading and exam-related discomfort, can guide the development of standardized rubrics and more transparent evaluation criteria. Additionally, its nuanced understanding of positive feedback, such as appreciation for interactive teaching methods, provides an opportunity to reinforce and expand these practices across different courses. Institutions can leverage GPT-3.5's insights to implement faculty workshops and targeted interventions to improve teaching practices and communication.

Additionally, integrating AI-driven tools into feedback systems could automate the analysis of large-scale datasets, reducing reliance on manual processes. Automated topic categorization and sentiment analysis would enable institutions to identify trends across diverse topics such as course materials, assignments, and instructor effectiveness, offering actionable insights that drive meaningful improvements. This approach enhances efficiency and ensures instructors receive real-time, detailed feedback that can directly inform their teaching strategies.

Further research is essential to expand and refine these findings. For instance, integrating advanced topic modeling techniques could systematically analyze emotional patterns within specific feedback categories. Such advancements would allow educators to better understand the nuances of student responses, address dissatisfaction in real-time, and promote positive changes in the learning environment. Moreover, refining the analysis of negative feedback intensity and context is crucial, as it could provide targeted solutions for addressing critical issues that impact student satisfaction and learning outcomes.

Developing an automated, real-time feedback analysis system could revolutionize student-teacher interactions by providing immediate insights for lecture improvement. Such systems would allow instructors to adapt their teaching strategies dynamically, improving student satisfaction and learning outcomes. Additionally, expanding fine-tuned AI models, such as GPT-4.0 with domain-specific datasets, could enhance the contextual understanding and overall accuracy of feedback analysis, making these tools even more effective in diverse educational settings.

In conclusion, this study demonstrates how AI-driven insights can be directly linked to actionable strategies for enhancing student-teacher feedback systems. Providing scalable and efficient solutions paves the way for significant improvements in educational practices, ensuring that institutions remain responsive to the evolving needs of students and educators.

References

[1] Ministry of Education, *Guidelines for university academic operations under COVID-19*. South Korea: Ministry of Education, 2020.

[2] J. Lee and H. Kim, "The role of lecture evaluations in enhancing teaching quality: A historical perspective," *Journal of Higher Education Research*, vol. 34, no. 2, pp. 134-148, 2021.

[3] J. M. Kim, "Research trends in liberal art education applying word cloud and KJ method," *KALCI*, vol. 21, pp. 1209-1231, 2021. https://doi.org/10.22251/jlcci.2021.21.4.1209

[4] Y. Park and S. Lee, "A study on the reliability of lecture evaluation systems," *Educational Measurement and Evaluation*, vol. 45, no. 3, pp. 223-240, 2019.

[5] J. Kim, J. Cheong, and H. Jeong, "University narrative lecture evaluation status and network analysis: A case study of S university," *The Journal of Learner-Centered Curriculum and Instruction*, vol. 21, no. 15, pp. 149-164, 2021.

[6] H. Choi and K. Jung, "Standardization issues in lecture evaluation criteria," *International Journal of Educational Development*, vol. 50, pp. 92-103, 2018.

[7] D. L. Jackson, C. R. Teal, S. J. Raines, T. R. Nansel, R. C. Force, and C. A. Burdsal, "The dimensions of students' perceptions of teaching effectiveness," *Educational and Psychological Measurement*, vol. 59, no. 4, pp. 580-596, 1999. https://doi.org/10.1177/00131649921970035

[8] J. Brown and S. Lee, "Mandatory participation policies and their impact on student feedback," *Higher Education Insights*, vol. 22, no. 4, pp. 301-317, 2021. https://doi.org/10.1007/s11162-016-9429-8

[9] J. Seo, "The effectiveness of incentivized lecture evaluation systems," *Journal of Institutional Research*, vol. 33, no. 2, pp. 78-93, 2019.

[10] S. Choi, "Linking evaluation scores to teaching performance: A double-edged sword?," *Teaching and Learning in Higher Education*, vol. 18, no. 2, pp. 105-120, 2021. https://doi.org/10.1007/s10734-013-9650-8

[11] S. Park and J. Lee, "Understanding the benefits of descriptive feedback in lecture evaluations," *Educational Review*, vol. 15, no. 5, pp. 93-110, 2020.

[12] A. Smith and L. Davis, "Analyzing feedback variability in lecture evaluations," *Journal of Applied Educational Analysis*, vol. 10, no. 3, pp. 45-62, 2017.

[13] S. Gupta, R. Ranjan, and S. N. Singh, "Comprehensive study on sentiment analysis: From rule-based to modern llm based system," *arXiv preprint arXiv:2409.09989*, 2024. https://arxiv.org/abs/2409.09989

[14] N. Rane, P. Desai, J. Rane, and S. Mallick, "Using artificial intelligence, machine learning, and deep learning for sentiment analysis in customer relationship management to improve customer experience, loyalty, and satisfaction," *Trustworthy Artificial Intelligence in Industry and Society*, pp. 233-261, 2024. https://doi.org/10.70593/978-81-981367-4-9_7

[15] J. Park and H. Choi, "AI-based sentiment analysis in higher education: Applications for student feedback," *Journal of Artificial Intelligence in Education*, vol. 26, no. 3, pp. 134-150, 2021. https://doi.org/10.1016/j.compedu.2021.104023

[16] A. Humphreys and R. J.-H. Wang, "Automated text analysis for consumer research," *Journal of Consumer Research*, vol. 44, no. 6, pp. 1274-1306, 2018. https://doi.org/10.1093/jcr/ucx104

[17] Y. Kim, "Balancing summative and formative assessment goals in lecture evaluations," *Review of Educational Research*, vol. 49, no. 3, pp. 285-303, 2018.

[18] J. Zhou and J.-m. Ye, "Sentiment analysis in education research: A review of journal publications," *Interactive Learning Environments*, vol. 31, no. 3, pp. 1252-1264, 2023. https://doi.org/10.1080/10494820.2020.1826985

[19] A. Ortigosa, J. M. Martín, and R. M. Carro, "Sentiment analysis in Facebook and its application to e-learning," *Computers in Human Behavior*, vol. 31, pp. 527-541, 2014. https://doi.org/10.1016/j.chb.2013.05.024

[20] D. Cao, R. Ji, D. Lin, and S. Li, "Visual sentiment topic model based microblog image sentiment analysis," *Multimedia Tools and Applications*, vol. 75, pp. 8955-8968, 2016. https://doi.org/10.1007/s11042-014-2337-z