Macrothink Institute™

International Journal of Education
ISSN 1948-5476
2024, Vol. 16, No. 4

Method for Clustering Comments in Class Evaluation Questionnaires using Keyword Feature Scores

Maya Iwano[1,*] & Kazuhiko Tsuda[2]

[1]Organization for Education and Student Affairs, Office for Teaching and Learning Management, Yamaguchi University, Yamaguchi, 7530841, Japan

[2]Faculty of Business Sciences, University of Tsukuba, Tokyo 1120012, Japan

*Corresponding author: Organization for Education and Student Affairs, Office for Teaching and Learning Management, Yamaguchi University, 1677-1 Yoshida, Yamaguchi-shi, Yamaguchi, 7530841, Japan. Tel: 81-83-933-5261

Received: July 24, 2024   Accepted: October 3, 2024   Published: December 10, 2024

doi:10.5296/ije.v16i4.22097   URL: https://doi.org/10.5296/ije.v16i4.22097

Abstract

In Japanese universities, improving classes is a task that must be continually implemented. Broadly speaking, a university class's purpose is threefold: (1) have students understood the content of the class, (2) have they achieved their goals for the class, and (3) were they satisfied with the class? Universities must embrace these three points. For this reason, class evaluation questionnaires administered at many universities nearly always include questions on "comprehension," "achievement," and "satisfaction." However, it is not possible to collect sufficient information by simply establishing questions and rating the responses. Therefore, research analyzing free descriptions in class evaluation questionnaires is increasing. This study developed a system to classify free descriptions provided by students into "comprehension," "achievement," and "satisfaction." In addition, it proposed a method of knowledge construction for such classification. One of its benefits was that it could be implemented within a short period of time with little effort. The proposed method's accuracy was evaluated by comparing the results of the classification by means of the proposed method with the results of the classification of the class evaluation questionnaire by an analyst. The two classification results matched with a probability of 92.67% to 93.67%, confirming that the method was sufficiently practical.

Keywords: class evaluation questionnaire, keyword extraction, document clustering, text mining

1. Introduction

In Japanese universities, improving teaching is a permanent task to be performed (Ministry of Education, Culture, Sports, Science and Technology, 2015; 2023). The plan, do, check and action (PDCA) cycle is a useful tool for applications in the improvement process. In class improvement:

- Plan (*P*) pertains to constructing improvement methods for current issues;
- Do (*D*) relates to applying the proposed improvement methods to actual classes;
- Check (*C*) verifies whether the improvement methods were effective in the classes in which they had been introduced and to understand their effects and reactions; and
- Action (*A*) involves scrutinizing the verification results and considering future improvement measures.

If the verification and analysis in *C* are not performed properly during the PDCA cycle for class improvement, erroneous countermeasures might be implemented. Therefore, the *C* component is crucial, and many universities administer class evaluation questionnaires, using these questionnaires to obtain evaluations from students. Many class evaluation questionnaires use a five-point grading scale. However, it is common to find surveys in which more than 90% of the items are rated 4 or 5, raising doubts as to whether the grading system captures students' true evaluations. Consequently, comments in the free-description section of class evaluation questionnaires have been recently analyzed using text mining techniques. Recent text-mining techniques have made it possible to obtain positive and negative emotion information (Anil et al., 2017; Johnson-Laird & Oatley, 1989; Maya & Kazuhiko, 2023; Nozomi et al., 2005; Ryuichiro et al., 2014; Saif, 2016; Yla & James, 2010). Therefore, it is possible to extract a large number of items that students are satisfied or dissatisfied with after taking a class. However, the large number of extracted items pose a problem because of the difficulty in effectively organizing them. From the perspective of class improvement, it is crucial that students comprehend the class content, achieve class goals, and are satisfied with attending the class. Against this background, this study proposes an algorithm that analyzes comments in the free description of class evaluation questionnaires and classifies them into "comprehension," "achievement," and "satisfaction." The algorithm was applied to develop the classification system, making it possible to improve classes based on students' objective responses.

2. Issues and Solutions in Higher Education

2.1 Initiatives in Higher Education

In recent years, it has become common in the educational field to quantitatively measure students' learning outcomes. This idea stems from enrolment management and institutional research that originated in the USA. This concept is similar to the PDCA cycle of business improvement used in private companies, which has been adopted in the education field across the globe (Bennett, 2001; Douglass et al., 2012; George et al., 2014). Quantitative measurements of learning outcomes are performed by collecting information using class evaluation questionnaires (Marsh, 1983). Since class evaluation questionnaires were introduced, many studies have been conducted on compiling and analyzing class evaluation questionnaires (Davis, 2009; Maya & Kazuhiko, 2022; Ruriko, 2012; Yukimasa & Yahachiro, 2004). These studies have three common points:

1. Did the students understand the class's content?

2. Did they achieve their goals regarding the class?

3. Were they satisfied with the class?

It follows that it is essential to obtain students' evaluations of comprehension, achievement, and satisfaction through class evaluation questionnaires.

2.2 Analysis of Comment Descriptions

Problems experienced with class evaluation questionnaires include a low response rate and high burden on students. Davis (2009) noted that, because students need to answer surveys for all classes, the questions should be simple. Therefore, class evaluation questionnaires that provide set questions and must be answered on a five-point scale have become quite common. However, commonly, more than 90% of respondents award 4 or 5 points (out of 5) regarding the items measured. To address this situation, attention has been focused on the inclusion of free-description sections in class evaluation questionnaires. By including free-description sections, it is possible to obtain perspectives not previously anticipated by questions with options provided as well as a variety of student opinions (Keigo & Hironori, 2015). Recently, text mining technology has been used to obtain emotional information from documents. In addition, this technique has been applied to the analysis of comments provided by students in class evaluation questionnaires (Hideya & Takahiro, 2011; Jyunichi et al., 2013; Koji et al., 2015; Rumiko, 2017). Most of these studies have aimed to extract individual information about students' positive or negative experiences but did not extract information on whether students understood the class, achieved their personal goals, or were satisfied with the class in question.

2.3 Issues in Comment Description Analysis

A single free description in a class evaluation questionnaire has the advantage of requiring fewer items for students to complete. However, it is difficult for those who design and analyze questionnaires to classify such descriptions into the categories of comprehension, achievement, and satisfaction with such limited information. Requiring students to provide separate free descriptions for comprehension, achievement, and satisfaction would be advantageous for analysis as the required clustering would already have been done. This presents a challenge in practice, as students' perceptions of what constitutes comprehension, attainment, and satisfaction might differ. For example, University A employed a class evaluation questionnaire in which students wrote separate, free descriptions of comprehension, achievement, and satisfaction. When the content of these responses was examined, it was found that the responses contained many statements that did not match the analysts' perceptions of comprehension, achievement, and satisfaction. Therefore, questionnaire analysts still had to manually classify those items into categories of comprehension, achievement, and satisfaction. In addition, many students declined to describe the items in full, as requested; they responded only to one item, and answered the remaining two items with "same as above." For these reasons, providing separate free-description sections for comprehension, achievement, and satisfaction in the class evaluation questionnaire was not advisable.

2.4 Use of Class Evaluation Questionnaire

To determine whether the free descriptions in the class evaluation questionnaire pertained to comprehension, achievement, or satisfaction, it would be beneficial to develop a computerized classification system that could automatically classify the contents of free descriptions into those categories. Recently, major strides have been made concerning technology (e.g., artificial intelligence document classification using prompt engineers). Document classification instruments that learn from large amounts of text data (e.g., large language models [LLMs]), have also been developed. However, in the analysis of the free descriptions in the class evaluation questionnaire handled in this study, there were not enough data to use LLM in the first place. In addition, the free-description section for comprehension contained large amounts of data on achievement and satisfaction. In other words, the data contained many errors. Prompts learned from data with many errors could not be used because they led to incorrect answers. Therefore, in this study, we adopted a document classification algorithm that used keyword matching to produce easy-to-understand classification results. This study classified the collected information into three categories: comprehension, achievement, and satisfaction. It was difficult to achieve sufficient accuracy by using a simple bag-of-words (BoW) algorithm. Consequently, feature scores were assigned to each keyword for each category of comprehension, achievement, and satisfaction, to construct classification knowledge comprising a set of keywords. To calculate these feature scores, we attempted to learn by incorporating the concepts of term frequency and inverse document frequency (TF/iDF). Section 3 presents the proposed method for constructing classification knowledge. Section 4 discusses an evaluation of classification accuracy using the proposed classification knowledge.

3. Constructing Classification Knowledge

3.1 Outline of Procedure

In this section, we demonstrate how to construct classification knowledge that classified the content of free descriptions in a class evaluation questionnaire into comprehension, achievement, and satisfaction. As discussed in the previous section, classification knowledge consisted of keywords that were assigned feature scores for comprehension, achievement, and satisfaction. This method is used by survey analysts to construct this type of classification knowledge. However, if the keywords used by the survey analyst and those written by the students were not written in the same way, it could be possible that the BoW algorithm would not match the keywords. Therefore, we attempted to construct classification knowledge from actual class evaluation questionnaires despite the problem that the free descriptions written by students for comprehension, achievement, and satisfaction, containing descriptions of categories that might differ from the survey analyst's perception. To use these answers as learning data, it was necessary to eliminate divergent descriptions of these different categories. Figure 1 shows the algorithm used to learn the classification knowledge.

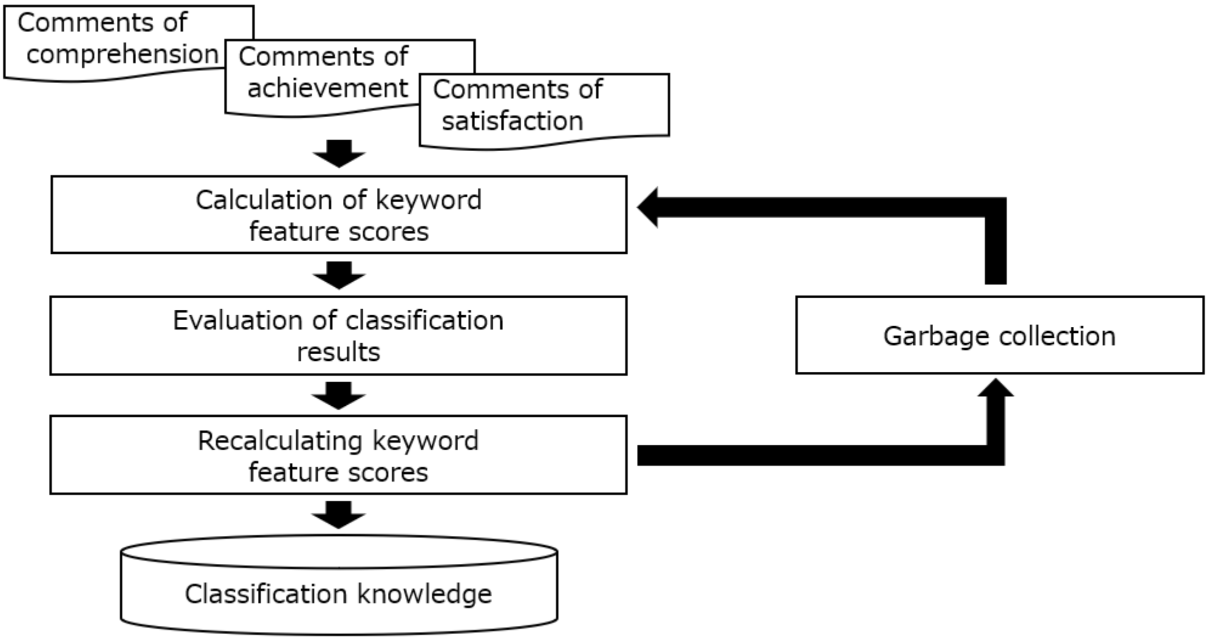
Comments of comprehension
Comments of achievement
Comments of satisfaction
Calculation of keyword feature scores
Evaluation of classification results
Recalculating keyword feature scores
Classification knowledge
Garbage collection

Figure 1. Overview of Classification Knowledge Learning Flow

3.2 Training Data

To build the classification knowledge, a class evaluation questionnaire was administered at University A during the second semester of 2022 after students completed the class. This class evaluation questionnaire had four free-description writing sections: one each for comprehension, achievement, satisfaction, and "anything else." Table 1 shows the statistical information that resulted from the survey conducted in the second semester of 2022. "TermExtract" was used to extract keywords. TermExtract is an "automatic technical terminology extraction system" jointly developed by Professor Hiroshi Nakagawa of the Department of Digitalization of the Information Infrastructure Library at the University of Tokyo and Assistant Professor Tatsunori Mori of the Graduate School of Environment and Information Sciences at Yokohama National University.

Table 1. Statistics on Comments in the Class Evaluation Questionnaire for the Second Semester of 2022

	Comments of comprehension	Comments of achievement	Comments of satisfaction
Number of questionnaires	8 800	8 800	8 800
Number of comments	3 162	2 805	2 999
Comment description rate	35.93%	31.88%	34.08%
Average number of characters	24.05	22.83	25.55
Average number of keywords	3.00	2.84	3.05

3.3 Keyword Feature Score Calculation Method

The keyword feature score was based on the TF/iDF concept but the calculation was modified to consider the classification into three categories. The procedure for calculating the feature scores for the keywords is shown in Figure 2.

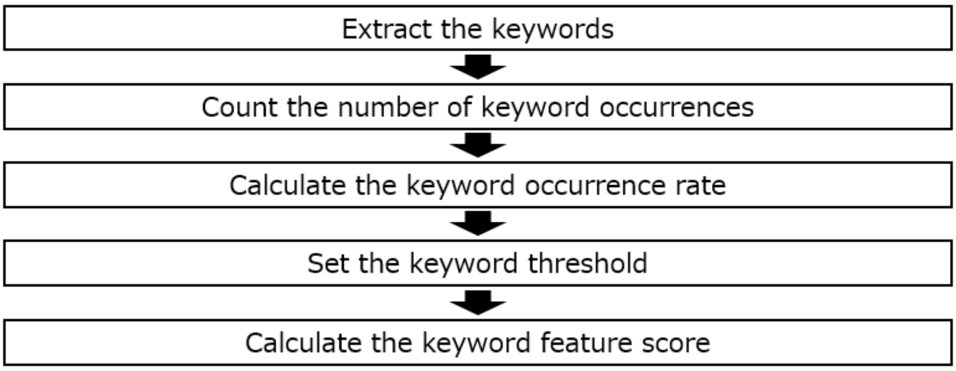
Extract the keywords
Count the number of keyword occurrences
Calculate the keyword occurrence rate
Set the keyword threshold
Calculate the keyword feature score

Figure 2. Procedure for Calculating Keyword Feature Scores

In Step 1 (extracting keywords), TermExtract was used to extract keywords from the free descriptions provided in the three categories. In Step 2 (counting the number of keyword occurrences), the extracted keywords were used as labels to count the number of keywords that appeared in the three categories. Table 2 shows the frequency of some keywords that appeared in the comprehension, achievement, and satisfaction categories.

In Step 3 (calculating the keyword occurrence rate), the number of instances of each keyword was converted into a ratio. This converted value was used as the base number for the feature score. This measure was adopted because, in the sentence evaluation that was performed later, the feature score was added by the number of occurrences, and the frequency information was only used to evaluate the similarity of the keyword in question in the three categories. To emphasize the difference in the frequency of occurrence of the keywords in each category, a process was performed in Step 4 (setting the keyword threshold) to round the base number of the feature score obtained in Step 3 if there was a specific difference from the top numbers.

The threshold for this rounding process was set at a frequency > 2/3 of the top frequencies.

Table 2. Frequency of Keywords in the Comprehension, Achievement, and Satisfaction Categories

Keyword	Comprehension	Achievement	Satisfaction	Total
"knowledge"	71	85	126	282
"tasks"	61	56	41	158
"understand"	180	21	79	280
"interesting"	2	2	211	215
"class content"	51	11	22	84
"syllabus"	0	132	0	132
"hard"	48	27	5	80
"engage"	53	71	17	141
"cooperate"	8	28	14	50
"useful"	0	1	35	36

If the frequency of the second-most frequent category for the keyword was 2/3 or less than that of the most frequent category, the feature scores of the second- and third-most frequent categories were set to zero. If the frequency of the third-most frequent category was 2/3 or less than that of the second-most frequent category, the feature score of the third category was set to zero. The threshold was set at 2/3 because when the second-most frequent category exceeded 2/3 of the most frequent category and the third-most frequent category exceeded 2/3 of the second-most frequent category, with the three categories remaining in the feature scores, the most frequent category would be about half the total frequency of the most to third-most frequent categories.

Table 3. Feature Score of Keywords

Keyword	Comprehension	Achievement	Satisfaction
"knowledge"	0.083924	0.100473	0.148936
"tasks"	0.128692	0.118143	0.086498
"understand"	0.642857	0	0
"interesting"	0	0	0.981395
"class content"	0.607143	0	0
"syllabus"	0	1	0
"hard"	0.6	0	0
"engage"	0.187943	0.251773	0
"cooperate"	0	0.56	0
"useful"	0	0	0.972222

In Step 5 (calculating the keyword feature score), if only one category feature score remained from Step 4, that feature score was used. If two category feature scores remained, each feature score was divided by two. If three category feature scores remained; each feature score was divided by three. This method equalized the total value of the feature scores given when one keyword was matched in the sentence evaluation and emphasized the difference in feature scores between categories. Table 3 shows some of the feature scores for the keywords.

Through the above process, classification knowledge consisting of 3 492 keywords was obtained, 2 749 of which had feature scores in only one category, 572 had feature scores in two categories, and 171 in three categories.

3.4 Sentence Evaluation of Comments

Using the word feature scores constructed in the previous section, we classified the free descriptions in the class evaluation questionnaire as either comprehension, achievement, and satisfaction. Then, we verified whether the classification was correct.

Sentence evaluation was the same as for the BoW algorithm, using the procedure in Figure 3.

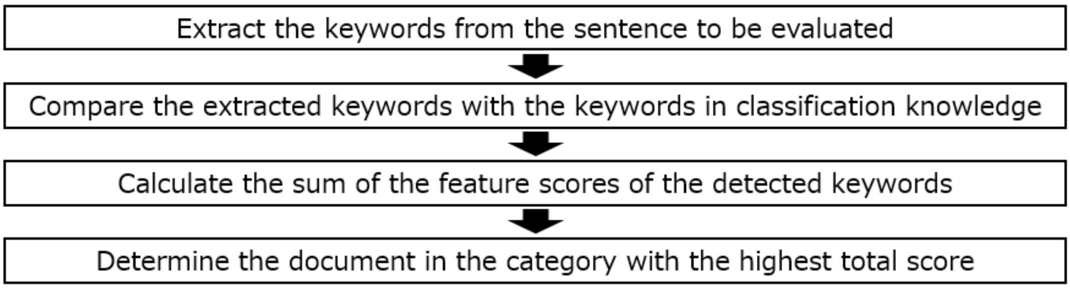
Extract the keywords from the sentence to be evaluated
Compare the extracted keywords with the keywords in classification knowledge
Calculate the sum of the feature scores of the detected keywords
Determine the document in the category with the highest total score

Figure 3. Procedure for Calculating Sentence Ratings

The previous section noted that there was a problem regarding students' responses to the class evaluation questionnaire with diverse perceptions of comprehension, achievement, and satisfaction. This would have been an issue in the survey conducted during the second semester of 2022, which was used as learning data to build classification knowledge. Using this classification knowledge, we attempted to classify the descriptions in the free-description section of the class evaluation questionnaire. In other words, it amounted to a self-evaluation. Table 4 summarizes the results.

Table 4. Sentence Evaluation Results Using Learning Data

Judgment result	Comments of comprehension		Comments of achievement		Comments of satisfaction	
Comprehension (%)	2 227	72.92%	741	28.17%	2 227	18.52%
Achievement (%)	364	11.92%	1 583	60.19%	364	9.52%
Satisfaction (%)	463	15.16%	306	11.64%	463	72.13%
Total	3 054		2 630		3 054	

Over 70% of the free descriptions about comprehension and satisfaction were judged to fall into these categories. Among the free-description responses about comprehension that were judged to be about satisfaction and the free-description responses about satisfaction that were judged to be comprehension, many of them expressed the notion that “I was satisfied because I understood.” In other words, many statements confirmed that the participating students were satisfied as a result of their understanding. In the free-description responses on achievement, only 60% of the statements were judged to fall into this category, and just under 30% were judged to be about comprehension. Among the statements that were judged to be about comprehension, many expressed the idea that “I felt a sense of accomplishment because I understood.” It was confirmed that there was a large amount of writing on comprehension, which was the reason for this result. In other words, the statements that were judge as not belonging to a given category were not scattered but were clustered together around similar sentiments. One possible reason was that some student groups had the same understanding of the meanings of the categories as did the class evaluation questionnaire analyst, whereas other student groups attached different connotations to it.

3.5 Garbage Collection and Reconstruction of Classification Knowledge

As reported in the previous section, the descriptions in the categories contained a mixture of descriptions that matched and did not match the meaning of the category assumed by the analyst of the class evaluation questionnaire. Consequently, classification knowledge was reconstructed using only the matched descriptions, as it would enable the construction of classification knowledge that mapped the meaning of the category assumed by the questionnaire analyst. The learning data used for construction consisted only of descriptions judged as “understood” in the free descriptions for comprehension, those judged as “achieved” in the free descriptions for achievement, and those judged as “satisfied” in the free descriptions for satisfaction. Table 5 presents the statistical data.

Table 5. Statistical Information of Comments for Classification Knowledge Reconstruction

	Comments of comprehension	Comments of achievement	Comments of satisfaction
Number of comments	2 227	1 583	2 068
Comment description rate	35.93%	31.88%	34.08%
Average number of characters	25.38	25.38	28.20
Average number of keywords	3.24	3.29	3.50

The classification knowledge construction algorithm was the same as that described in Section 3.3. The reconstructed classification knowledge comprised 3 436 keywords, 2 993 of which had feature scores in only one category, 375 in two categories, and 68 in three categories. Compared with the previously constructed classification knowledge, the reconstructed classification knowledge had the same number of keywords. However, the number of keywords with feature scores in only one category increased, whereas the number of keywords with feature scores in two or three categories decreased, resulting in feature scores with clearer meanings for keyword classification. Subsequently, we checked whether the constructed classification knowledge could be correctly evaluated and classified into comprehension, achievement, or satisfaction. The method of confirmation was to evaluate whether the free descriptions in the class evaluation questionnaire used in learning the classification knowledge could be correctly classified. Table 6 lists the evaluation results.

Table 6. Results of Sentence Evaluation on Training Data

Judgment result	Comments of comprehension		Comments of achievement		Comments of satisfaction	
Comprehension (%)	2 179	97.80%	151	9.53%	76	3.68%
Achievement (%)	14	0.63%	1 402	88.51%	21	1.02%
Satisfaction (%)	35	1.57%	31	1.96%	1 971	95.31%
Total	2 228		1 584		2 068	

Of the 2 228 free descriptions of comprehension, 2 170 sentences were judged to be about "comprehension," reflecting an accuracy rate of 97.80%. These figures confirmed that the free descriptions of comprehension could be accurately classified as "comprehension." Of the 1 584 free descriptions of achievement, 1 402 were judged to be "achievement." This represented a correct answer rate of 88.51%. Most incorrect judgments were related to "comprehension," and an analysis of the reasons for this result was conducted. Figure 4 shows examples of free descriptions of achievement that were judged to be about "comprehension." In the example sentences shown in the Figure 4, a single underline indicated a description of the reason and a double underline indicated a description of the conclusion. Because this was a class evaluation questionnaire, the ability to understand the content of the class was a valid reason for achievement. When this description of the reason was clearly written, it was often judged to be "comprehension." A total of 1 971 of 2 068 free description of satisfaction were judged to be about "satisfaction." This represented a correct response rate of 95.31%, which confirmed that free descriptions of satisfaction could be accurately categorized as "satisfaction." The number of data points used for this evaluation was 5 880, and the total number of data points that could be correctly classified was 5 552, resulting in an overall correct answer rate of 94.42%. Based on these figures, it is safe to assume that the classification knowledge was the result of correct learning.

Because I understood the lesson content and actively participated in class activities and group work.
Because I understood the class and was able to use it.
Because I learned how economics works and became interested in economics.

Figure 4. Example of a Description of "Achievement" That Was Judged to be "Comprehension"

4. Estimation

4.1 Overview of Evaluation Procedure

In this section, the classification knowledge constructed in the previous section was used to evaluate whether the free responses in the class evaluation questionnaire during the first half of 2023 could be classified into comprehension, achievement, and satisfaction. As mentioned in the previous section, the accuracy rate was close to 95% when classification knowledge was used to classify data. However, it was necessary to evaluate whether data other than the training data could be correctly classified.

4.2 Evaluation Data

The data used in this evaluation were obtained from a class evaluation questionnaire administered after the end of classes in the first semester of 2023 (April to August). This questionnaire had four free-description sections: (1) writing about comprehension, (2) achievement, (3) satisfaction, and (4) anything else. Table 7 presents statistical information on free-description writing in the class evaluation questionnaire conducted for classes offered during the first semester of 2023.

Table 7. Statistics on Comments from the Class Evaluation Questionnaire for Evaluation

	Comments of comprehension	Comments of achievement	Comments of satisfaction
Number of comments	12 791	12 791	12 791
Number of sentences	7 107	6 479	6 773
Comment description rate	55.54%	50.65%	52.95%
Average number of characters	20.37	20.31	20.24

We evaluated whether the statements written in the free-description section of the class evaluation questionnaire could be correctly classified into comprehension, achievement, and satisfaction. Figure 3 in Section 3.4 shows the algorithm used to evaluate the statements written in the free descriptions of the class evaluation questionnaire (the subject of evaluation) for the first semester of 2023. Table 8 lists the evaluation results.

Table 8. Classification Results of Comments from Class Evaluation Questionnaire for the First Semester Of 2023

Judgment result	Comments of comprehension		Comments of achievement		Comments of satisfaction	
Comprehension (%)	4 452	64.40%	2 571	41.83%	1 672	27.30%
Achievement (%)	928	13.42%	2 252	36.64%	732	11,91%
Satisfaction (%)	1 533	22.18%	1 324	21.54%	3 743	60.89%
Total	6 913		6 147		6 147	

As seen in the Table 8, the majority (60%) judged the free descriptions of comprehension as "comprehension" and the free descriptions of satisfaction as "satisfaction"; however, more of the free descriptions of achievement were judged as "comprehension" than as "achievement." A similar situation was confirmed in constructing the classification knowledge, as explained in the previous section. The reason for these evaluation results might be that there was a difference between students' perceptions of comprehension, achievement, and satisfaction, as stated in the free descriptions in the class evaluation questionnaire, and the perceptions of the person analyzing the questionnaire.

4.3 Evaluation Results

Whether the classification system using the classification knowledge proposed in this study could classify the descriptions written in the free-description section of the class evaluation questionnaire into comprehension, achievement, and satisfaction should be evaluated. Therefore, the person analyzing the class evaluation questionnaire visually checked the sampled data.

The sampling procedure was as follows:

- One sentence in the free description comment section. Some free description comments included multiple statements about comprehension, achievement, and satisfaction. The purpose of this evaluation was to determine whether the comments could be classified into comprehension, achievement, and satisfaction categories. Therefore, statements longer than two sentences were likely to include content from multiple categories and were eliminated.
- A sentence with two or more keywords. The proposed method employed the BoW algorithm. In this algorithm, keywords in the description and those registered in the classification knowledge might match. Hence, one or two keywords in the description were eliminated because the impact of accidental matches was significant.

Table 9. Judgments by the Automatic Classification System and Experts

Automatic classification system		Expert judgment				Matching rate
Judgment category	Number of judgments	Compre-hension	Achievement	Satisfaction	Others	
Comprehension	112	106–108	2	2–4	0	94.64%–96.43%
Achievement	58	3–4	51	3	0–1	87.00%
Satisfaction	130	3	4	121–122	1–2	93.08%–93.85%
Total	300	112–115	57	126–129	1–3	92.67%–93.67%

Next, 100 statements were selected from each of the free descriptions of comprehension, achievement, and satisfaction, for a total of 300 statements. Then, multiple-class evaluation questionnaires were used to judge whether the categories output by the classification system were correct. Table 9 summarizes the results. In some documents, judgments differed among the class evaluation questionnaire analysts. The agreement rate between the judgments of the automatic classification system and those of the class evaluation questionnaire analyst was 92% to 93%. The agreement rate for comprehension was 95% and that for satisfaction was 93%, both exceeding 90%; however, achievement had a low agreement rate of 87%.

Automatic classification system: "Comprehension"/Expert judgment: "Satisfaction"	
	I gradually became accustomed to the process from experiment to report writing and also became accustomed to writing reports.
	I was able to hear first-hand from people in various positions, including companies and faculty members in my department, about how they use data.
Automatic classification system: "Satisfaction"/Expert judgment: "Achievement"	
	Because my knowledge of exercise and health science is increasing and I am able to apply that knowledge to my daily life and live a healthier life.
	I became very interested in the content covered in the lectures and was able to gain a multifaceted perspective as well as knowledge about the fields covered in each lecture.
Automatic classification system: "Achievement"/Expert judgment: "Comprehension"	
	Until now, when I read English texts, I would just read them vaguely, without paying much attention to the context, but through this class I was able to understand how to read carefully.
	I was able to understand the various data in today's highly information-driven world and how to use them, as well as the advantages and disadvantages of each type.

Figure 5. Cases Where Judgments Did Not Match

Figure 5 displays the specific contents of the cases in which the judgments of the automatic classification system and the analyst differed. The sentences that the system judged as "comprehension" and the analyst judged as "satisfaction" both contained the notation, "because I was able to do it." The analyst judged this to be satisfaction based on the verb phrase "because I was able to do it." However, the verb phrase "because I was able to do it" was not registered in the classification knowledge, which appeared to be the reason for the differences in judgment. Sentences that the system judged as “satisfaction” and the analyst judged as “accomplished” were written based on the knowledge gained in the class. However, there was no description within that statement that expressed the feelings of “I was able to do it,” “it was good,” or “I was happy.” The analyst judged this point to reflect “accomplished,” which was deemed to be the cause of differences in judgment. The sentences that the system judged as “accomplished,” and the analyst judged as “understood” were sentences that concluded with “I was able to understand it.” The analyst attached importance to this point, which was the conclusion, whereas the system placed importance on a long description of the reasons for understanding. This was thought to be the cause of differences in judgment.

5. Discussion

For universities, improving classes is essential. To make such improvements, many universities use class evaluation questionnaires, which are administered to students immediately after they have attended a class. Various studies have been conducted on class evaluation questionnaires and their use is expected. It is particularly important to obtain an evaluation of the three perspectives of student “comprehension,” “achievement,” and “satisfaction.” These questions are often asked about using a five-point scale, with many students choosing 4 or 5. Another challenge is that the responses are typically simply tabulated and analyzed. Therefore, there is room for additional analysis of free descriptions.

However, with regard to the analysis of the free descriptions, many research perspectives have been based word frequency, the relationships between words, and what the students had positive or negative feelings about. Because of this, the analysis of the three perspectives (“comprehension,” “achievement,” and “satisfaction”) have been dependent on the analyst’s interpretation and point of view. Although it is possible to solicit free descriptions for each of the three perspectives, as in the case of University A discussed here, this places an increased burden on students, and their descriptions can be shaky. Therefore, in this study, we proposed to use a system that automatically categorizes the free-response statements so that even if there is only one questionnaire item, the responses can be classified into “comprehension,” “achievement,” or “satisfaction.”

If what the students comprehended, whether they felt a sense of achievement, and what they were satisfied with can be clarified, it should be possible to discover improvement items with a common understanding of what to improve and how to improve it and link the results to specific improvements in classes.

6. Conclusion

In university education, it is important that students who take classes understand the content of the classes, achieve their goals, and are satisfied with their enrollment. Therefore, we conducted a class evaluation questionnaire to understand the situation of students who had taken certain classes. In this study, we built a system that automatically classified the content written in the free-description section of the class evaluation questionnaire into categories of comprehension, achievement, and satisfaction. The proposed automatic classification system was confirmed to be able to classify documents with an accuracy of > 92%. For this evaluation, the system was designed to categorize comprehension, achievement, and satisfaction responses. Therefore, if there were no differences in the characteristic scores for comprehension, achievement, and satisfaction, the category with the highest score was the output. Most descriptions for which the judgment of the classification system and that of the analyst differed in this type of situation. Therefore, if the system were designed to allow a small amount of output of "other" to be included, more accurate judgments would be possible. In the future, we would like to use this automatic classification system in an educational setting. Furthermore, we would like to investigate the causes of the classification errors that we noticed when using the system and work on improving them.

References

Anil, B., Nirmalie, W., Stewart, M., & Deepak, P. (2017). Lexicon Generation for Emotion Detection from Text. *IEEE Intelligent Systems, 32*(1), 102-108. https://doi.org/10.1109/MIS.2017.22

Bennett, D. C. (2001). Assessing Quality in Higher Education. *Liberal Education, 87*(2), 40-45.

Davis, B. G. (2009). *Tools for teaching.* (2nd ed.). San Francisco, CA: Jossey-Bass.

Douglass, J. A., Thomson, G., & Zhao, C-M. (2012). The Learning Outcomes Race: The Value of Self-Reported Gains in Large Research Universities. *Higher Education, 64*, 317-335. https://eric.ed.gov/?id=EJ972412

George, D. K., Natasha J., Stanley, O. I., & Jillian, K. (2014). *Knowing what students know and can do: The current state of student learning outcomes assessment in U.S. colleges and universities*. National Institute for Learning Outcomes Assessment. Retrieved from https://www.learningoutcomesassessment.org/wp-content/uploads/2019/02/2013AbridgedSurveyReport.pdf

Hideya, M., & Takahiro, S. (2011). Development and Evaluation of a Class Evaluation Questionnaire Feedback System using Data and Text Mining. *Journal of the Japan Society of Educational Technology, 35*(3), 217-226.

Johnson-laird, P. N., & Oatley, K. (1989). The Language of Emotions: An Analysis of a Semantic Field. *Cognition and Emotion, 23*(2), 81-123. https://doi.org/10.1080/02699938908408075

Jyunichi, M., Chiyuki, N., & Koji, E. (2013). Criteria for Evaluating Classes of Junior College Students Enrolled in Childcare Training Departments: Study using Text Mining Methods. *Miyagi University of Education Information Processing Center Research, 20*, 15-18.

Keigo, T., & Hironori, W. (2015). Analysis of Course Evaluation for Improving On-demand Academic Writing Class: Text Mining Approach from the Perspective of Customer Satisfaction Analysis. *Kyoto University Higher Education Research*, 21, 1-14. 2015. http://hdl.handle.net/2433/210139

Koji, E., Toshiko, T., Hidetoshi, K., Akinobu, A., Yoshi, T., Kenichi, T., Masaaki, O., & Kimiharu, I. (2015). Analysis of Class Evaluation Questionnaire by Text Mining: An Attempt to Visualize Free Texts by Co-Occurrence Network. *Miyagi University of Education Information Processing Center: COMMUE, 15*, 67-74.

Marsh, H. W. (1983). Multidimension Ratings of Teaching Effectiveness by Students from Different Academic Settings and their Relation to Student/Course/Instructor Characteristics. *Journal of Educational Psychology, 75*(1), 150-166. https://doi.org/10.1037/0022-0663.75.1.150

Maya, I., & Kazuhiko, T. (2022, December). *Evaluation of questionnaires for measuring the learning outcomes of educational activities*. 13th International Congress on Advanced Applied Informatics (IIAI-AAI-Winter), Phuket, Thailand, 124-129. https://doi.org/10.1109/IIAI-AAI-Winter58034. 2022.00034

Maya, I., & Kazuhiko, T. (2023). Knowledge Discovery to Improve Students' Class Achievement using Questionnaires. *Procedia Computer Science, 225*, 1862-1871. https://doi.org/10.1016/j.procs.2023.10.176

Ministry of Education, Culture, Sports, Science and Technology. (2023). *Status of reform of educational content etc. at universities FY2021* (in Japanese). MEXT. Retrieved from https://www.mext.go.jp/content/20230908-mxt_daigakuc01-000031526_1.pdf

Ministry of Education, Culture, Sports, Science, and Technology. (2015). *Status of reform of educational content others at universities 2013* (in Japanese). MEXT. Retrieved from https://www.mext.go.jp/a_menu/koutou/daigaku/04052801/_icsFiles/afieldfile/2016/05/12/1361916_1.pdf

Nozomi, K., Ryu, I., Kentaro, I., & Yuji, M. (2005, October). *Opinion extraction using a learning-based anaphora resolution technique*. International Joint Conference on Natural Language Processing, Jeju Island, Korea. Retrieved from https://aclanthology.org/I05-2030.pdf

Rumiko, A. (2017). *Analysis of relationship between learner's characteristics and level of understanding using text-mining*. Japan Society for Information and Systems in Education: 42 National Conference Keynote Collection, 73-74.

Ruriko, T. (2012). Basic Analysis of Course Evaluation Questionnaire Data. *Research Report of the Japan Society for Education Technology, 12*(3), 1-6.

Ryuichiro, H., Nozomi, K., Toru, H., Chiaki, M., Toyomi, M., Toshiro, M., & Yoshihiro, M. (2014). Syntactic Filtering and Content-Based Retrieval of Twitter Sentences for the Generation of System Utterances in Dialogue Systems. In A. Rudnicky, A. Raux, I. Lane,

& T. Misu (Eds.), *Situated dialog in speech-based human-computer interaction*. Signals And Communication Technology Book Series. Cham: Springer. https://doi.org/10.1007/978-3-319-21834-2_2

Saif, M. M. (2016). 9 - Sentiment analysis: Detecting Valence, Emotions, and Other Affectual States from Text. *Emotion Measurement*, 201-237. https://doi.org/10.1016/B978-0-08-100508-8.00009-6

Yla, R. T., & James, W. P. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology, 29*(1). https://doi.org/10.1177/0261927X09351676

Yukimasa, M., & Yahachiro, T. (2004). Analysis of Lecture Evaluations and Point Quantification for Teaching Improvements Based on the Concept of Customer Satisfaction Analysis. *Kyoto University Researches in Higher Education, 10*, 21-32. Retrieved from http://hdl.handle.net/2433/53924

Acknowledgments

Not applicable

Authors contributions

Not applicable

Funding

Not applicable

Competing interests

Not applicable

Informed consent

Obtained.

Ethics approval

The Publication Ethics Committee of the Macrothink Institute.

The journal’s policies adhere to the Core Practices established by the Committee on Publication Ethics (COPE).

Provenance and peer review

Not commissioned; externally double-blind peer reviewed.

Data availability statement

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

Data sharing statement

No additional data are available.

Open access

Copyrights