Advances in Science, Technology and Engineering Systems Journal
Vol. 10, No. 6, 42-54 (2025)
www.astesj.com
Special Issue on Innovation in Computing, Engineering Science & Technology

ASTESJ
ISSN: 2415-6698

Identifying Comprehension Faults Through Word Embedding and Multimodal Analysis

Kento Yasuda*, Hiromitsu Shimakawa, Fumiko Harada

Ritsumeikan University, Ibaraki, 567-8570, Japan

ARTICLE INFO

Article history:
Received: 14 October, 2025
Revised: 27 November, 2025
Accepted: 29 November, 2025
Online: 05 December, 2025

Keywords:
Data science
Writing questions
Open-ended questions
Text analysis
Electrodermal Activity (EDA)
Comprehension
Individual tutoring sessions
Multimodal

ABSTRACT

This study establishes a method for determining whether learners have an understanding of data science. Data science requires knowledge in various fields, which makes many learners give up. To prevent learners from being discouraged, it is necessary to judge the comprehension of the principles in each specified skill. It is important to assess not only learner's knowledge but also the extent to which they understand the underlying principles of data analysis methods. Open-ended questions effectively assess comprehensive understanding because they require learners to construct and articulate their understanding in their own words. This study analyzes teacher–learner interaction and electrodermal activity to examine the educational significance of open-ended responses. A random forest model identifies key behaviors, which indicate that articulating thoughts in one's own words is crucial for improving learner's understanding. For each specified skill, the proposed method provides us with a way to examine whether learner's answers to descriptive questions are close to the model answer. Learner's level of understanding is determined from the document vectors of their responses using the Word2vec method. In addition, important words for each skill are extracted using a Naive Bayes model. Furthermore, the method has identified words representing ununderstood concepts and unassociated procedures with a logistic regression model. Experimental results indicate that the proposed method effectively identifies learner's comprehension levels and extracts key linguistic features for distinguishing those with insufficient understanding. Using responses from 16 learners, the method achieves an F1-score of 0.824, although the small sample size limits generalizability. The word embeddings of learners with and without understanding show markedly different distributions. It suggests that we can identify the concepts and procedures that learners do not understand from their words. It enables us to offer suggestions to assist learners who are likely to be stuck.

1. Introduction

In recent years, the demand for data scientists has been increasing not only in IT companies but also in a wide range of industries, including manufacturing, logistics, healthcare, and so on. Data scientists need knowledge in a wide range of fields and problem-solving skills to apply them to real problems. However, what they need is not only the ability to understand field-specific terminologies and concepts. They must also understand the mathematical basis of the analytical methods they use. As noted in [1], data scientists must have the ability to select the most appropriate analytical method from various options and apply it to real data. To bring up learners to skillful data scientists, it is necessary to train mathematical ability as well as knowledge specific to each problem domain.

The utilization of data science technologies requires diverse mathematical knowledge and skills in probability, statistics, optimization, and programming. Those who try to master the knowledge often get stuck along the way. Many people fail to become data scientists because of repeated dead ends. They would repeat failures because there is no way of knowing what they understand and what they do not understand.

To uncover areas of incomprehension, it would be appropriate to analyze their free-text response to descriptive questions that test the mathematical knowledge regarding data analysis. However, it is difficult to automatically judge their understanding from the text they freely write. It is also important to identify which concepts learners with poor understanding do not understand. It needs manual tasks, which require huge costs. As shown in [2], [3], and [4], most previous studies on assessing comprehension have used multiple-choice questions. Very few studies assess comprehension from open-ended questions, as shown in [5] and

*Corresponding Author Kento Yasuda, Ritsumeikan University, is0584ih@ed.ritsumei.ac.jp

https://dx.doi.org/10.25046/aj100604

[6]. The experimental results have revealed that changes in electrodermal activity are observed when teachers encourage learners to express the correct answers in their own words. The experiments have also pointed out that learners who received such instruction exhibit higher levels of understanding. It is effective to assess learner's understanding using open-ended questions that allow them to express their comprehension in their own words.

This study assumes that learner's comprehension of data analysis methods can be estimated from the lexical features that appear in their free-text responses. Many of those who get stuck are unable to organize their knowledge, explain it in order, and express it in their own words. On the contrary, many people who understand without getting stuck can organize their knowledge to explain it in an orderly manner. Those learners answer questions with statements close to the sentences in the model answers. The study focuses on the difference.

Learners would get stuck because they have a misunderstanding of certain concepts. There is a high possibility that the words used in the correct descriptions explaining these concepts are not included in the learner's descriptions. The study examines words that are not in the sentences written by misunderstanding learners but are in the correct answer. It enables us to figure out what they do not understand. Based on the observation, the study analyzes how close learner answers are to the model answers for open-ended questions. The proposed method creates a distributed representation of words using Word2Vec. As described in [7], this approach uses a corpus of explanatory sentences for the relevant concepts presented in textbooks. The similarity between the learner's response and the model answer is measured using the cosine similarity between their document vectors, which are computed based on the words appearing in the response.

Experimental results show that the study can discriminate what is understood with the cosine similarity between the document vectors of the response sentences and those of the model answers. Recent advances in educational NLP have increasingly applied transformer-based models. In [8], the authors demonstrate that Sentence-BERT is effective for evaluating the semantic validity of short-answer responses. Building on these developments, the present study compares multiple representation methods, including Doc2Vec, Sentence-BERT, and TF-IDF combined with SVM, to examine their effectiveness in estimating learner's comprehension from free-text responses. The results indicate that the approach using Doc2Vec-based document vectors and cosine similarity achieves the highest F1-score, demonstrating its usefulness for comprehension estimation based on the lexical features of open-ended responses.

Furthermore, the distributed representations place words used by understanding learners in less overlap with those used by misunderstanding learners. The distribution of the two kinds of words turns out to be extremely far apart. Learners who cannot write the words appearing in the model answers are likely to get stuck soon. Early care of such learners will prevent them from running away from data science training. In [9], the authors focus exclusively on learners. The present paper extends this analysis by adding further analyses and incorporating teachers into the discussion. It also provides new insights by employing open-ended questions.

The structure of this paper is shown below. Section 2 describes the vectorization of documents by embedding. A measure of the closeness to the correct answers is proposed in Section 3. Section 4 presents an experiment to discuss its results. Section 5 concludes the study along with statements of future work.

2. Vectorization of Documents by Embedding

2.1. Dimensional Comprehension and Dimensional Selection

Information expressed in higher dimensions is complex to process and difficult to visualize. Therefore, it is necessary to reduce the dimensionality while losing as little of the necessary information as possible. The reduction enables graphical representation. It makes it easier to understand the distribution of the sample intuitively. Dimension selection can achieve dimension reduction through dimension compression.

Dimensional selection is the process of selecting dimensions that are easy for humans to understand in order to grasp the meaning of each dimension. Correct classification is possible by selecting appropriate dimensions. However, there is the problem that information is completely lost for unselected dimensions.

Dimensionality compression refers to representing high-dimensional data using a smaller set of dimensions while minimizing information loss. In [10], the authors explain this concept and demonstrate its effectiveness in practical applications. The interpretation of the dimension after reduction facilitates the classification of the information. In artificial dimension reduction, it is desirable to compress dimensions in order to estimate similarity and classify comprehension, because the discarded dimensions may contain important information. For example, when compressing a 100-dimensional object to one dimension, the scalar values are combined by taking the inner product of the 100-dimensional vector and the weight vector. The composite scalar value becomes the dimension after compression. Naturally, the method of deriving this weight vector differs according to the application.

PCA (Principal Component Analysis) and NMF (Nonnegative Matrix Factorization) are typical models for dimensional compression. In [11], the author explains that PCA selects the axis that minimizes the mean squared error. In other words, the maximum amount of information in the high-dimensional state is retained in a lower compressed dimension. This method compresses along the axis where the variance of each point is maximized. In [12], the authors explain that PCA is an unsupervised learning model because it does not require supervised data. It is considered a typical and basic model for dimensionality compression. Compared to PCA, NMF has a non-negative value constraint, which makes the results easier to interpret. In [13], [14] and [15], the authors describe NMF as an unsupervised learning model that serves as a typical and basic model for dimensionality compression.

2.2. Embedding and Cosine Similarity

The use of corpora yields a representation of the meaning of words. A corpus is a collection of example sentences that serves as a reference for machine learning models to understand a text. Words are polysemous. The meaning of the same word often changes from field to field. This makes it difficult to obtain a distributed representation that can be applied to all fields. To accurately represent the meanings of different fields, it is necessary to prepare a corpus with a rich collection of example sentences that are commonly used in a particular field. Embedding refers to the arrangement of natural language information, such as words and sentences, in a vector space that represents the meaning of the words and sentences. Vectorization of words and sentences is commonly used to calculate the similarity between words and sentences.

Word2Vec is commonly used for vectorizing words in natural languages, which vectorizes target words based on their co-occurrence probabilities. Word2Vec vectorizes target words from the probability of word co-occurrence. The input and output layers are neural networks with as many neurons as the number of words in the corpus and only one hidden layer. Hidden layers keep the number of neurons to a small number. When a set of words appearing in example sentences in a corpus is given to the input layer as a one-hot representation, this set of words is trained to be reproduced in the output layer.

Word2Vec considers the sequence of outputs of the hidden layer of the neural network after training as a vector representing the meaning of the words. This vector is called the distributed representation. The generated vector is calculated based on the co-occurrence probability of the words. The order relations of the words are not taken into account, so the context is ignored. Nevertheless, the advantage of being able to convert words into a distributed representation vector is significant.

One of the main advantages of vectorization by embedding is that the similarity of words can be calculated in terms of cosine similarity. Cosine similarity is calculated using the inner product of two vectors. Cosine similarity is an index that expresses how much the action of one of two vectors contributes to the action of the other vector. The possible values range from -1 to 1. The higher the value, the higher the similarity, while the lower the value, the lower the similarity. Cosine similarity can be calculated using the following formula.

$$sim = \cos\theta = \frac{\vec{a}\vec{b}}{|\vec{a}||\vec{b}|} = \frac{\vec{a}^{T}\vec{b}}{|\vec{a}||\vec{b}|} \tag{1}$$

※Cosine similarity can be used to calculate vector similarity.

2.3. Vectorization of Documents

A textual description of a specialized field contains concepts and procedures from several more basic disciplines. Determining whether a given new description is correct is a matter of checking that the correct concepts and procedures are used for each basic field. In order to analyze the descriptions, a dimension is assigned to each basic field. Textual descriptions become representations in higher-dimensional spaces.

In [16], the authors explain that Doc2Vec distributed representations of sentences and documents, providing a compact vector representation that can be used for dimensionality compression. Doc2Vec vectorizes documents by embedding all words that occur in a document with Word2Vec and finding their average value. The analysis of large amounts of text data with Doc2Vec can convert sentences written in natural language into meaningful distributed representations. The transformation of the distributed representation measures the similarity between the vectors after analysis. It allows the sentence classification and the detection of similar sentences.

3. Determining Proximity to Correct Answers to Written Questions

3.1. Data Preprocessing and Signal Alignment

In the text preprocessing stage, the free-text responses are first normalized by unifying full-width and half-width characters. Noise such as symbols is then removed using regular expressions. Subsequently, morphological analysis is applied to extract nouns, verbs, adjectives, and adverbs. For vectorization, the extracted tokens are input into a Doc2Vec model to generate document embeddings. Given the limited amount of free-text data, the Doc2Vec parameters are configured to prioritize embedding stability, and the number of training epochs is increased until convergence is observed. To capture both contextual information and lexical distribution, the study employs both the PV-DM and PV-DBOW algorithms to obtain robust document representations.

To synchronize conversational data with physiological signals, the start and end times of each utterance in the conversation log are converted into seconds and mapped onto the same timeline as the EDA timestamps. This alignment allows the EDA values corresponding to each utterance to be extracted, enabling integrated analysis of linguistic behavior and physiological responses. In this study, “multimodal” refers to the integration of textual responses, conversational behaviors, and EDA signals, all aligned on a shared timeline for unified analysis.

To estimate how differences in cognitive load relate to learning performance, the ChangeFinder algorithm is applied to detect change points in the electrodermal activity. By identifying these change points, the study investigates situations in which cognitive load increases and examines how the magnitude and frequency of such changes differ between higher-performing and lower-performing learners.

In the ChangeFinder algorithm, a local autoregressive model is first constructed within a sliding window over the observed time series. The model is then used to generate a smoothed sequence, from which a second-stage AR model is learned. The change-point score is computed as the difference between the log-likelihood of the prediction model and that of the observation model, treating this difference as the loss that indicates abrupt changes in the signal.

In this study, time segments with high change-point scores are analyzed by examining teacher behaviors during the preceding one-minute interval. Furthermore, to identify more precise moments at which electrodermal activity changes occur within each one-minute window, a second pass of the ChangeFinder is applied to the EDA time series at a 0.25-second resolution. This finer-grained analysis enables the detection of specific instructional moments that induce changes in learner’s cognitive load. Based on the estimated change points, the study examines how teacher behaviors influence learning performance.

3.2. Estimation of Poorly Understood Concepts and Procedures

This paper proposes a method to evaluate whether the learner's answer is close to the correct answer in writing questions asked in order to check the learner's understanding. Furthermore, the study provides appropriate guidance for learners who do not have a good understanding. Textual analysis of the answers to the written questions determines which concepts and procedures in the specialized field are not understood. For this reason, the important words for each dimension are calculated after dimensionality reduction by dimensionality compression and dimensionality selection.

The calculation of words takes the difference between the answers of learners who understand and those who do not. The difference in answers estimates the concepts and procedures associated with the latter's lack of understanding. A schematic diagram of the proposed method is shown in Figure 1. Adaptation of this method can assist learners who are faltering if they identify concepts and procedures that they do not understand.

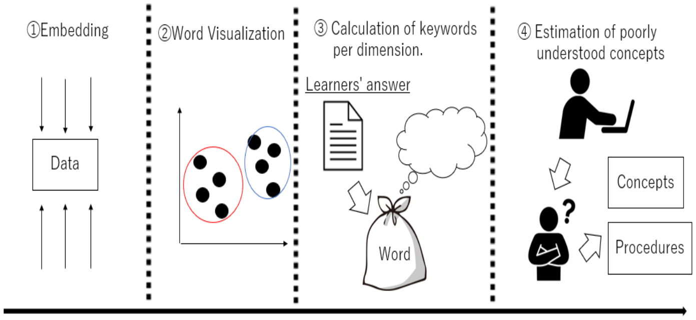
①Embedding
Data
②Word Visualization
③ Calculation of keywords per dimension.
Learners' answer
Word
④ Estimation of poorly understood concepts
Concepts
Procedures

Figure 1 : Methodological Overview Diagram.

This research requires the learner to master many concepts and procedures in a specialized field. Data analysis is taken as an example of a specialized field, but the proposed method is not specific to this field. The method gives the learner a task to analyze data and collects textual data from the learner's answers. The use of text analysis techniques determines whether an answer to a writing question is close to the correct answer.

The first step is dimensional compression and selection, followed by similarity estimation and comprehension classification. Similarity estimation calculates the cosine similarity by Doc2Vec. For comprehension classification, each word in the learner's answer and the model answer is converted into a distributed representation using Doc2Vec to visualize the similarity of the words. The visualization makes it easier to understand the classification of correct and incorrect answers. Next, after classification, this method calculates important words for each dimension. Naive Bayes lists the words that appear in the answers of correct and incorrect answers in descending order of frequency. Finally, this method extracts concepts that are poorly understood.

Logistic regression calculates the importance of important words in classifying those who answered correctly and those who answered incorrectly. Logistic regression also identifies procedures and concepts that are not understood. The evaluation of the Word2Vec and logistic regression models uses cross-validation to check the confusion matrix. The reliability and validity are ensured by calculating the F1-score from the reproducibility and fit rate, which is close to 0.8 without overfitting. Measures to reduce bias are taken to ensure that the actual level of understanding is adequately reflected in the test scores. Here, a certain number of people are ensured. The corpus is adjusted to reduce bias further.

3.3. Correct Answers and Solution Embedding

Model answers and learner answers in short answer questions are text data containing high dimensional information. The use of Doc2Vec allows the embedding of model answer examples and answer text into low-dimensional data. The compression of the dimensions allows all information to be taken into account.

First, morphological analysis using Mecab divides the learner's answer and the model example sentence into words to determine the degree of similarity. Next, Doc2vec is used to transform each sentence into a meaningful distributed representation. Finally, cosine similarity calculates the similarity between two sentence vectors obtained from Doc2Vec. The training of the Doc2Vec model requires a corpus, which is a collection of example sentences.

This research extracts multiple passages from textbooks that serve as answers to build a corpus. The way to improve the accuracy of the model is to increase the number of sentences used in the corpus. It also creates a corpus that matches the text content you want to analyze. The number of words is large when that of sentences used in a corpus is large. In addition, the amount of training for the model increases, resulting in higher accuracy. Furthermore, this method accurately distinguishes the intention of the questioner and the meaning of the question. In [17] and [18], the authors report that accuracy increases when the corpus is adapted to the content of each text so that it matches the subject matter being analyzed.

3.4. Word Visualization

When there are many dimensions, it is difficult to grasp the meaning of each dimension, so it is necessary to select dimensions that are easy for humans to understand. Doc2Vec visualizes the words that appear in the answer descriptions created by the learner, including correct and incorrect ones. The compression of high-dimensional vectors into two dimensions allows an understanding of the relationship between each text in the document vector.

The word groups are different for those who answered correctly and those who answered incorrectly. It is visualized by plotting document vectors on a two-dimensional plane. As an example, Figure 2 shows the words that appear in the learner’s answers and the model answers for a descriptive task asking why a conditional branch is necessary in a C programming exercise. This dataset contains data from 87 learners and is collected internally. The words "else" and "return" appear in the model answers, whereas the words "printf", "recognize", and "post" in the answers are words that are not necessary for a correct answer. This shows that there is a distance between words that are necessary and words that are not. Learners who do not understand answer the writing questions without using the necessary words.

They use the wrong words because they do not understand the syntax of if-else sentences.

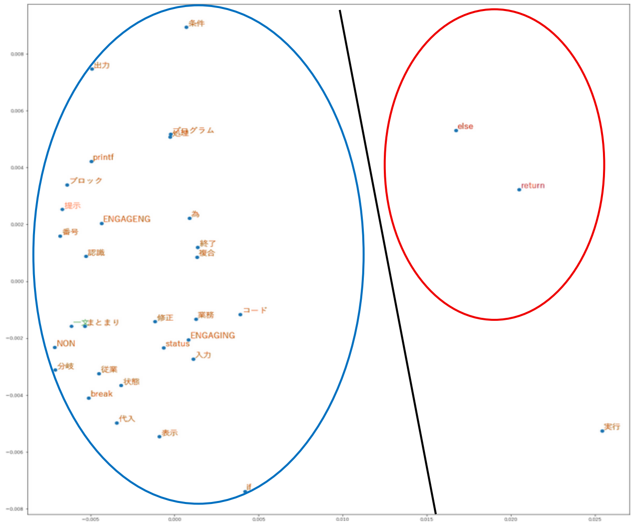

Figure 2: Similarity of Words Appearing in Example Answers and Model Answers.

3.5. Calculation of Keywords per Dimension

In [19], the authors describe Naive Bayes as a probabilistic classifier that identifies informative words based on their likelihood. Morphological analysis with MeCab is used to split sentences into words. For preprocessing, sentences are segmented into words using MeCab, which performs Japanese morphological analysis based on part-of-speech information, as explained in [20]. The result is a multinomial Naive Bayes model created from the generated word document matrix. Thus, creation involves calculating the probability of occurrence of words that appear in documents belonging to a specific field. The list of words in order of probability enables us to obtain the important words in the documents belonging to the field.

After acquiring important words, this method distinguishes the learner's writing into two classes: correct and incorrect. This method investigates what words are likely to be used in correct answers. It identifies words that are important in distinguishing between correct and incorrect answers. This important calculation uses a logistic regression model. Logistic regression is a regression model that uses qualitative variables (binary data) as objective variables.

This study uses the probability of the occurrence of words that appear in learner's descriptions as an explanatory variable. Logistic Regression calculates the weights of the explanatory variables, so it is possible to find out which words are more likely to be correct if their probability of occurrence is high.

3.6. Identification of Poorly Understood Concepts

Learners who understand are likely to use the words they need and use keywords more frequently, as described in 3.4. In addition, learners who do not understand do not use the necessary words and use important words less frequently. The logistic regression model finds the weights of words that are particularly important in separating correct from incorrect answers among the multiple important words used by the correct and incorrect respondents.

This study allows the extraction of poorly understood concepts by determining the weights of particularly important words. The calculation of word importance allows us to discover which concepts and procedures have a significant influence on the classification of correct and incorrect answers.

4. Experiments, Experimental Results and Discussion

4.1. Experimental Overview

The task is to analyze specific data with a specified algorithm and to collect textual data from the subject's answers to written questions. This dataset is collected internally. The task involves decomposing MNIST image data into its components using NMF and PCA. The tasks include fill-in-the-blank questions, discussion questions, comprehension verification questions and a confirmation test.

This research particularly focuses on comprehension verification questions and confirmation tests. The fill-in-the-blanks question involves creating multiple blanks in a Python code that uses NMF and PCA to decompose the image data of a human face into its constituent elements. The fill-in-the-blank questions check whether the learner can correctly fill in the blanks. This question estimates whether the learner correctly understands the knowledge about the code for analysis. The discussion question asks learners to describe what the numerical values output as a result of applying the above Python code to the given data mean.

The comprehension verification questions, and confirmation tests ask questions about concepts and procedures in the field of dimensional compression, to which the two specified machine learning algorithms belong. The 40-point understanding verification questions include 16 true/false questions and 4 descriptive questions that require explanations using 30 characters or more. The scores are shown in Table 1.

Therefore, setting a minimum number of characters for written questions prevents learners from answering with fewer characters. Table 2 shows the comprehension verification questions used in the experiment. In addition, the confirmation test consisted of two writing questions of 100 characters or more, as shown in Table 3. These questions test the learner's understanding of lectures that explain the specified methods for analyzing a given problem. These also estimate whether the learner can correctly interpret the results of applying the method.

In the first experiment, subjects work on the NMF (non-negative matrix factorization). The learner and the teacher solve the problem together in a one-to-one format. The learner's role solves machine learning fill-in-the-blanks, discussion, and comprehension verification problems. On the other hand, the teacher's role helps the learners with exercises so that they can solve the problems.

There should be a difference in proficiency between the learner's role and the teacher's role. Therefore, the learner's role is a learner who has been studying data science for a few months, and the teacher's role is a learner who has been studying data science for several years.

The teacher’s role is assumed to be a learner with a deep understanding of data science. There are 12 subjects in the learner’s role. Regarding the experimental procedure, learners learn about NMF in advance in class. Next, the learner’s role solves the fill-in-the-blank questions and discussion questions, and the teacher’s role provides support. Finally, learners complete comprehension verification questions to check whether they understand NMF.

The second experiment involves PCA (Principal Component Analysis). As in the first experiment, the learner’s and teacher’s roles work one-to-one. The subjects are the 12 learners who participated in the first experiment and four learners who did not participate in Experiment 1.

The learners first attend a lesson, then complete a fill-in-the-blank question, a reflective question and finally a writing test. In the second experiment, the learners do not have to answer questions to verify their understanding of the first experiment, but rather they have to answer open-ended questions of 100 words or more. This is the purpose of this research, which is to find a method to determine whether the answer to the descriptive question is close to the correct answer.

The reason for using NMF and PCA tasks is that NMF and PCA have a similar function of extracting common components. Working on both NMF and PCA may deepen the understanding of both NMF and PCA.

Table 1: Scores on the 40-points Comprehension Verification Question in Experiment 1

※Table 1 shows the results for the 12 subjects of the 40-point comprehension verification and written questions in Experiment 1.

Learner	1	2	3	4	5	6	7	8	9	10	11	12
Score	34	34	28	24	35	33	30	26	27	25	33	24

Table 2: Comprehension Verification Questions to be Used in Experiment 1

※It contains both a writing question and a correct answer question. If you consider a statement to be incorrect in a correct or incorrect question, please provide a correct explanation in the description box.

Q 1	Non-negative matrix factorization is not non-negative for all features.
Q 2	Non-negative matrix factorization decomposes the original matrix into a matrix multiplication form.
Q 3	For the non-negative matrix factorization, the product of the original matrix and the two non-negative-valued matrices after decomposition is exactly equal without error.
Q 4	Features can be extracted by decomposing the matrix in the NMF.
Q 5	Answer in at least 30 words what scaling is.
Q 6	Answer the need for scaling in at least 30 words.
Q 7	Name one scaling method.
Q 8	Non-negative matrix factorization expresses a non-negative matrix as the product of two non-negative matrices.
Q 9	A horizontal sequence is called a row vertical sequence is called a column.
Q 10	Non-negative matrix factorization features can estimate missing values and can be classified using hidden features.
Q 11	Non-negative matrix factorization is supervised learning.
Q 12	The non-negative matrix factorization is the ability to estimate and complete missing values.
Q 13	One of the features of non-negative matrix factorization is the ability to estimate and complete missing values.
Q 14	Non-negative matrix factorization has been applied in various fields such as text data, sound separation, and automatic musical notation.
Q 15	Why must the non-negative matrix factorization target non-negative values ?
Q 16	In this exercise, we extracted features from facial images, but non-negative matrix factorization can also be applied to other applications such as Amazone product recommendations. Answer why it can be applied to such a variety of examples.
Q 17	Non-negative matrix factorization is not non-negative for all features.
Q 18	Non-negative matrix factorization decomposes the original matrix into a matrix multiplication form.
Q 19	For the non-negative matrix factorization, the product of the original matrix and the two non-negative-valued matrices after decomposition is exactly equal without error.
Q 20	Dimensional compression of non-negative-valued matrix factorization is the transformation of low-dimensional data into higher dimensions.

Table 3: Questions for the Confirmation Test to be Used in Experiment 2

※Please answer at least 100 words.

Q 1	Answer why common features can be taken out by multiplying the NMF matrices.
Q 2	Answer how the NMF gradually modifies the randomly set weight and feature matrices to reduce the error.

4.2. Optimal Method for Measuring Comprehension

In this experiment, open-ended questions are employed as a means of assessing learner’s understanding. Moreover, the reason for setting open-ended questions is that learners must express correct answers in their own words.

In the first experiment, conversations between teachers and learners, which had a positive effect on learners, and electrodermal activity are recorded. In [21], the authors explain that electrodermal activity represents changes in the skin’s electrical properties caused by sweat gland activity. In addition, the teacher's behavior toward the learner is quantified based on the content of the conversation.

Specifically, features that can be expressed as quantities, such as the number of characters spoken by the teacher and the number of questions asked by the teacher, are expressed as actual numerical values. For subjective evaluation items (e.g., whether the teacher guides the learner), we manually recorded whether the behavior appeared during that period. If it exists, it is represented as 1; otherwise, it is represented as 0.

Therefore, we present the results of random forest discrimination to identify behaviors that cause cognitive load to learners by teachers. In [22], the authors explain that random forests can estimate the importance of each predictor variable for classification and often achieve better predictive performance than linear regression.

Furthermore, because cognitive load causes stress in people, changes in cognitive load affect galvanic skin response through changes in skin conductance. In [23], the authors explain that higher cognitive load leads to stronger electrodermal responses. In addition, several studies have reported a positive correlation between cognitive load and electrodermal activity, as shown in [24], [25] and [26].

Although electrodermal activity is influenced by emotional changes, research shows that it can still objectively distinguish between different levels of cognitive load. In [27], the authors demonstrate this robustness even in the presence of emotional variation. Therefore, it can be said that skin potential activity can be used to measure cognitive load significantly even in individual instruction.

The objective variable is whether the teacher caused a change in the learner's skin potential activity during the one minute, and the explanatory variable is the teacher's behavior during the one-minute exercise. The data used in this study are collected from two learner participants during an individual tutoring experiment, where the same teacher participant conducted one-on-one instructional sessions with each learner. A total of 61 data samples collected from the two learner participants are divided into training and test datasets in an 80:20 ratio. Cross-validation is performed by alternately switching the training and test datasets 12 times to evaluate prediction accuracy. The average accuracy across the 12 iterations is approximately 79%. The variable importance derived from the random forest model is presented in Figure 3.

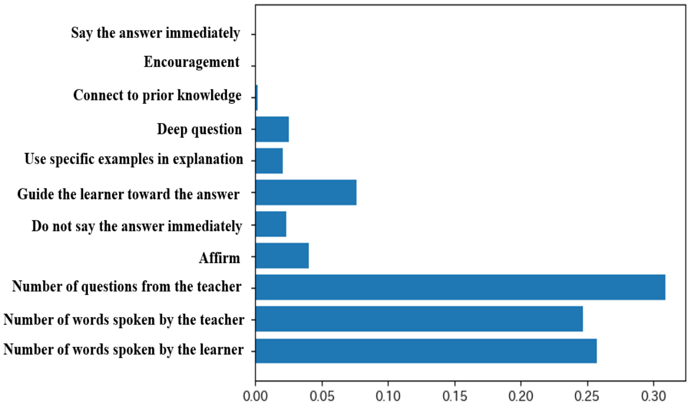

Figure 3: Variable Importance in the Random Forest Model

Among the variables with high importance, the most influential explanatory variable is the number of questions posed by the teacher to the learner during the individual tutoring sessions. When the learner is questioned by the teacher, it is assumed that cognitive load is induced as the learner attempts to formulate a response.

Subsequently, the variables that exhibited high importance included the number of words spoken by the learner, the number of words spoken by the teacher, and whether the teacher's utterances are intended to guide the learner toward the answer. It is assumed that a higher cognitive load is imposed when the learner produces a greater number of utterances. This is because the learner formulates and articulates responses while engaging in cognitive processing.

It is also observed that electrodermal activity increased when the teacher produced a larger number of utterances. This may be because the learner is cognitively processing the teacher's utterances to comprehend them.

In a similar manner to the teacher participant, additional analyses are performed for Participant B, who also acted as a teacher and provided one-on-one instruction to two learners. A total of 47 one-minute data samples obtained from the two learner participants are aggregated and split into training and test datasets in an 80:20 ratio.

A total of 47 one-minute data samples obtained from the two learner participants are aggregated and split into training and test datasets in an 80:20 ratio. Cross-validation is conducted by repeatedly exchanging the training and test datasets 12 times to evaluate the prediction accuracy. As a result, the mean prediction accuracy is approximately 40%.

The relatively low prediction accuracy may be caused by the fact that Participant B engaged in fewer instructional interactions with the learners compared to Participant A. Since the teacher's interventions are limited, when the explanatory variables are defined according to the teacher's behaviors during one-minute instructional segments, several factors other than the teacher's interventions appeared to contribute to elevated electrodermal activity. Therefore, this may have resulted in low prediction accuracy.

To identify the factors that influence performance, we first compute the correlation coefficient between EDA and the learner's scores. In this analysis, the learner's EDA during approximately 30 minutes of one-to-one tutoring is examined at 0.25-second intervals. For each time point, we compare the EDA value with that of the immediately preceding interval, and the proportion of intervals in which the EDA increases is treated as the overall rate of EDA elevation.

The learner's performance is defined as the score obtained on the forty-point test completed after the one-to-one tutoring session. The correlation between test scores and the proportion of increases in electrodermal activity is then computed. The result shows a correlation coefficient of minus zero point two four two, indicating that the frequency of increases in EDA, interpreted as increased cognitive load, does not correlate with learner's test performance. Figure 4 presents the relationship between the learner's test scores and the proportion of increases in electrodermal activity.

Even when electrodermal activity increases, this does not imply that the learner's performance improves. One possible explanation is that electrodermal activity also fluctuates in response to emotional changes. It is highly likely that the learners experienced emotional variations during their interactions with the teaching assistant, which may have contributed to changes in electrodermal activity independent of cognitive load.

Another possible explanation is that an increase in cognitive load does not necessarily lead to improved understanding. Cognitive load can have both beneficial and non-beneficial effects on learning outcomes.

According to cognitive load theory, there are three types of cognitive load, among which only germane cognitive load is considered to enhance learning. The other two types, intrinsic cognitive load and extraneous cognitive load, are regarded as unrelated to learning. During one-on-one tutoring, it is likely that not only germane load but also intrinsic load, caused by the inherent difficulty of the task, and extraneous load, arising from unclear or confusing instructions, are imposed on learners.

Furthermore, even when a teacher attempts to induce germane cognitive load, performance does not improve unless the learner's actual understanding deepens. These considerations indicate that simply increasing cognitive load does not enhance performance, and that unnecessary cognitive load should be avoided to support effective learning.

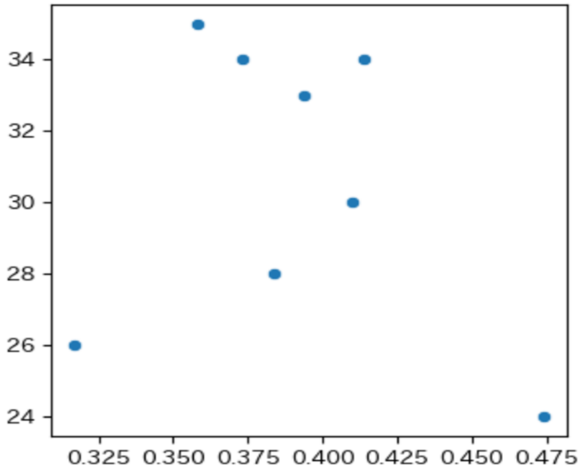

Figure 4: Scatter Plot of Scores and EDA

Next, to identify the types of cognitive load that positively influence performance, we use change point detection to extract the moments in which cognitive load increases and analyze the corresponding segments of conversation.

This study uses a total of four datasets collected from two teacher participants, Participant A and Participant B, each of whom provided instruction to two learners. Both learners taught by Participant A achieved higher test scores following the individual tutoring sessions, whereas both learners taught by Participant B showed lower test performance.

Therefore, Participant A is classified as a teacher who positively influenced learner's performance, whereas Participant B is classified as a teacher who did not exert a positive influence on learner's performance. The learner's test scores and standard scores are presented in Table 4.

Table 4: Test Scores and Standard Scores of Four Learners

Learner	Teacher	Score	Standard Score
1	Participant A	34	59.1
2	Participant A	34	59.1
3	Participant B	28	43.5
4	Participant B	24	33.2

In order to identify teacher behaviors that positively affect learner's performance, the behaviors of the two teacher participants are analyzed at the moments when electrodermal activity increased. To achieve this, change-point detection of electrodermal activity is performed using the ChangeFinder algorithm.

The teacher's behaviors in the one minute preceding each detected change point are compared to identify behavioral differences. To aid in understanding the above procedure, Figure 5 presents a graph illustrating the original electrodermal activity data and the scores generated by applying the ChangeFinder algorithm.

The left vertical axis represents the ChangeFinder change score, the right vertical axis represents electrodermal activity, and the horizontal axis represents time in 50-second intervals. Furthermore, the red line indicates the original data output, whereas the blue line indicates the change scores obtained from the ChangeFinder algorithm.

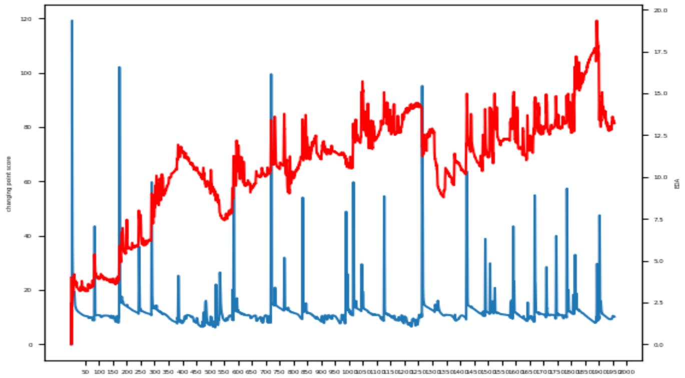

Figure 5: Electrodermal Activity and Change Scores

For the high-performing learners, the periods of increased electrodermal activity corresponded to times when the teacher asked a greater number of questions to the learners. In particular, this teacher frequently posed questions intended to guide the learners toward the correct answers.

For the low-performing learners, the periods of elevated electrodermal activity corresponded to times when the learners asked questions to the teacher or engaged in self-initiated reflection. These results are consistent with the hypothesis that teachers who positively affect learner's performance encourage learners to articulate correct answers in their own words.

These results are derived from the analysis of data obtained from four learners. To further test the hypothesis, data obtained from eight learners will be utilized. Table 5 presents the performance group classifications, test scores, and standard scores for the eight learners.

The data from the eight learners are divided into two groups, with the top four performers in one group and the bottom four performers in the other, and the behaviors observed during the one minute preceding each moment of increased electrodermal activity are analyzed.

The analyzed items include the following: (1) the proportion of learner utterances in the total teacher–learner conversation during one minute, (2) the number of words spoken by the learner per minute, (3) the number of words spoken by the teacher per minute, (4) the total number of words spoken by both the learner and the teacher per minute, (5) the number of questions asked by the teacher per minute, (6) the number of guiding questions asked by the teacher per minute, (7) the number of questions asked by the learner per minute, and (8) whether the teacher immediately provided the answer within one minute.

Table 5: Test Scores and Standard Scores of Eight Learners

Learner	Performance Group	Score	Standard Score
1	High performing	34	59.1
2	High performing	34	59.1
3	Low performing	28	43.5
4	Low performing	24	33.2
5	High performing	35	61.6
6	High performing	33	56.5
7	Low performing	30	48.7
8	Low performing	26	38.4

The analysis is conducted for each of the eight learners. The averages are calculated separately for the top four and bottom four performers. The results, rounded to four decimal places, are shown in Table 6. Compared with the low-performing group, the high-performing group showed a lower proportion of learner utterances and a higher total number of words spoken by the teacher, while the number of words spoken by the learners is nearly the same.

These findings suggest that, in the high-performing group, the teacher engaged in more frequent verbal interactions with the learners. Furthermore, compared with the low-performing group, the high-performing group exhibited a greater total number of words spoken by both the learners and the teacher, suggesting that both participants spent less time in silence.

Compared with the low-performing group, the high-performing group shows more teacher-initiated questions and fewer learner-initiated questions, indicating stronger instructional control.

The high-performing group also exhibited a greater number of guiding questions posed by the teacher. Therefore, for the high-performing group, higher levels of cognitive load are likely to occur during periods when the teacher is asking questions or engaging in verbal instruction.

In contrast, for the low-performing group, higher levels of cognitive load are more likely to occur when learners are engaged in independent thinking or when they ask questions to the teacher after a period of individual reflection.

Taken together, these findings support the hypothesis that encouraging learners to articulate correct answers in their own words is valid, and they underscore the importance of using open-ended questions to evaluate learner's understanding.

Table 6: Mean Values of the High and Low Performing Groups

Item	High Performing	Low Performing
Proportion of learner utterances	0.177	0.306
Total number of words spoken by learners	40	41.75
Total number of words spoken by the teacher	229.167	141
Total number of words spoken by both learner and Teacher	268.917	182.75
Number of questions asked by the teacher	1.125	0.15
Number of guiding questions asked by the teacher	0.75	0.05
Number of questions asked by the learner	0.083	0.25
Whether the teacher immediately gave the answer	0	0.25

4.3. Estimation of Similarity by Doc2Vec

In order to learn, the text data needs to be pre-processed so that unnecessary parts are removed from each text. This research uses processed text data from sentences as learning data and creates a learning model. The parameters and their values during training are shown in Table 7.

Table 7: Parameters during Learning

Parameter	size	window	min	workers	epochs
Setpoint	10	5	1	4	100

・size：dimensions of distributed representation

・window：number of surrounding words in context

(Decide how many words to consider at the same time.)

・min：minimum number of occurrences of words used for learning (Discard words with fewer occurrences than this value.)

・workers：number of threads in learning

・epochs：number of epochs

Doc2Vec checks the accuracy of the distributed representation of sentences. The similarity is calculated between the text written by the learner in the two confirmation test questions in Experiment 2 and the text of the correct answers. The use of Doc2Vec yields a vector representing each text. The cosine similarity checks how close the learner's text is to the correct text.

The individual learner scores are calculated by computing the cosine similarity of the distributed representations via Doc2Vec, then taking the average cosine similarity between Confirmation Tests 1 and 2. There are 12 learners in Experiment 1 and Experiment 2. Four more learners are added to the eight listed in Table 5. There are also four learners who participate in Experiment 2 only.

We analyze four learners who participated in Experiment 1 and also attended PCA classes outside the experiment. A significant difference is observed in the mean cosine similarity scores between these learners and those who participated only in Experiment 2.

Furthermore, Learners who participated in both Experiment 1 and 2 show higher cosine similarity scores than those who joined only Experiment 2. These results suggest that experiencing both NMF and PCA helps learners consolidate and deepen their understanding of NMF.

To address the need for quantitative evaluation, we conducted a performance analysis using 32 free-text responses collected from 16 learners. The proposed Doc2Vec-based cosine similarity approach is evaluated in terms of accuracy, precision, recall, and F1-score.

To provide a rigorous comparison, three baseline models are included: (1) TF-IDF combined with SVM, (2) Sentence-BERT, and (3) cosine similarity without embedding. Table 8 summarizes the results. The Doc2Vec cosine similarity with embeddings achieved the highest overall performance (F1-score = 0.824), outperforming TF-IDF + SVM (F1 = 0.800), Sentence-BERT (F1 = 0.818), and cosine similarity without embedding (F1 = 0.698). These results indicate that Doc2Vec embeddings capture lexical and semantic features that are not effectively represented by traditional bag-of-words models or simple similarity measures.

To further evaluate statistical significance, we calculated the p-value, confidence interval, and effect size for the Doc2Vec-based classifier. The method shows a statistically significant difference ($p = 0.0096$), a large effect size (Cohen’s $d = 1.286$), and a 95% CI [0.500, 0.833]. These findings demonstrate that the Doc2Vec approach discriminates reliably between learners with sufficient and insufficient understanding.

Taken together, the quantitative comparisons confirm that Doc2Vec embeddings provide the most robust representation for comprehension estimation in this dataset. Additionally, the superior performance relative to Sentence-BERT suggests that transformer-based scoring models do not necessarily generalize well in low-data settings such as ours, supporting the use of lightweight embedding models for small-scale educational assessment tasks.

Table 8: Baseline Results

Baseline	TF-IDF ＋ SVM	Sentence Bert	Doc2Vec (No Embedding)	Doc2Vec (with Embedding)
Accuracy	0.667	0.733	0.667	0.700
Precision	0.690	0.783	0.850	0.700
Recall	0.952	0.857	0.650	1.000
F1-score	0.800	0.818	0.698	0.824

4.4. Extracting Important Words Using Naive Bayes

Experiment 2 consisted of two descriptive questions as a confirmation test. Doc2Vec converts the similarity of words that appear between correct and incorrect answers into a distributed representation and visualizes it. The visualization confirms whether it is possible to classify those who answer correctly and those who answer incorrectly. Figure 6 shows the classification results for Confirmation Test 1.

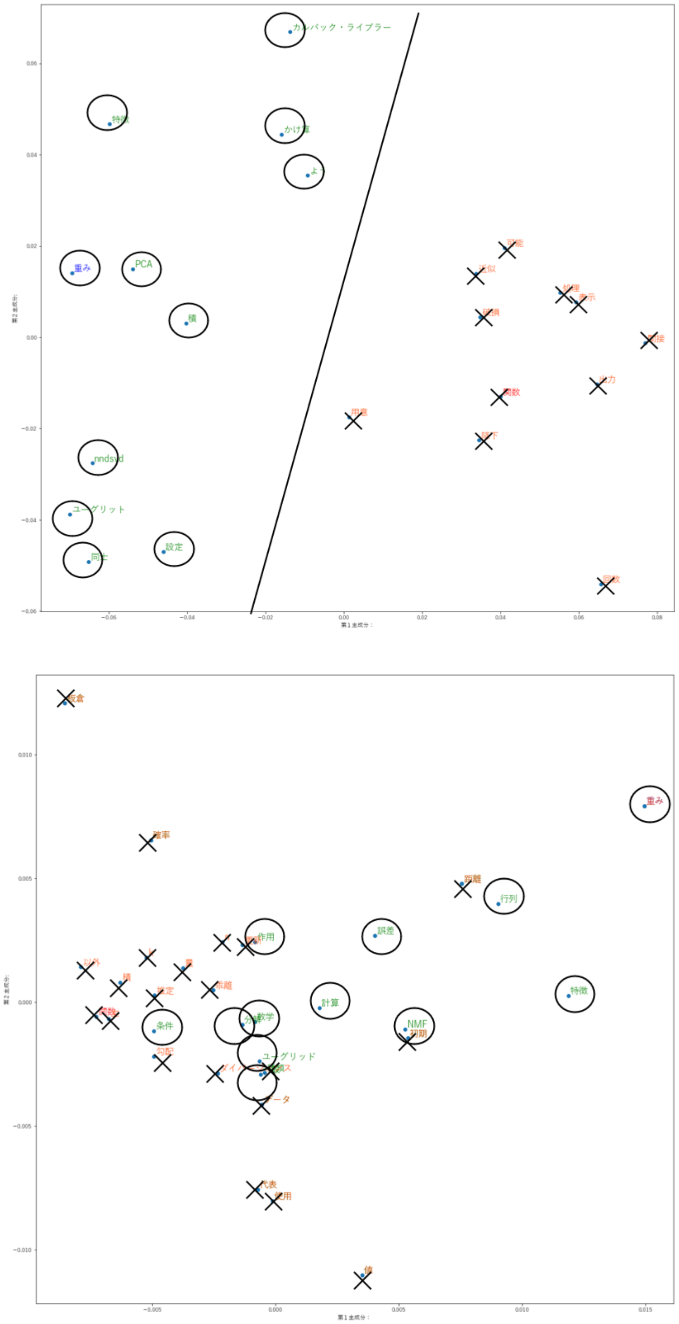

Figure 6: Distributed Representation of the Words in Each of the Subjects' Answers and Correct Answers in Confirmation Test 1(with and without Corpus)

In Confirmation Test 1, words frequently used by correct answers are marked with black circles, and words frequently used by incorrect answers are marked with black crosses. The left-hand diagram shows the case where no corpus is provided for each question, and the right-hand diagram shows the case where a corpus is provided.

Figure 6 shows that with a question-by-question corpus, it is possible to use a straight line to classify the words most frequently used by both correct and incorrect answers. This shows the necessity of having a corpus for each question. Consistent classification patterns are also observed in Confirmation Test 2.

4.5. Extracting Important Words Using Naive Bayes

Words whose appearance probability exceeds 0.02 are considered important words. Learners with high cosine similarity to the correct answer have high comprehension, while learners with low cosine similarity have poor comprehension. In both Confirmation Test 1 and Confirmation Test 2, the important words used by learners with high cosine similarity are positioned on the left side. For Confirmation Test 1, these words include “non-negative”, “data”, “factorization”, “weight”, and “multiplication”. For Confirmation Test 2, the important words are “matrix”, “feature”, “use”, “weight”, “data”, “be”, “modification”, and “calculation”, “multiplication.” In contrast, the important words used by learners with low cosine similarity are placed on the right side. For Confirmation Test 1, these words are “base” and “vector”, and for Confirmation Test 2, the words are “function” and “minimum”.

Although some words overlap between learners with high and low cosine similarity, the non-overlapping words represent the truly important terms. Accordingly, learners with high cosine similarity to the correct answers use a larger number of important words, whereas learners with low cosine similarity use far fewer of these important terms.

4.6. Calculation of importance by Logistic Regression

We examine the words with the largest absolute weights in the logistic regression model used to classify correct and incorrect answers. These are important words for classifying correct and incorrect answers. The words 'non-negative' and 'factorization' are more important in Confirmation Test 1, and the words 'error', 'random', and 'use' are more important in Confirmation Test 2. The important words in Confirmation Test 1 and Confirmation Test 2 relate to the features and mechanisms of the NMF. The difficult concepts for learners are the core properties of NMF, such as low dimensionality, non-negativity, and the mechanisms for exploring feature and weight matrices. These concepts are found by identifying several key terms. Upper-division university learners studying data science are selected for the analysis.

4.7. Significance of Results

The results of all the analyses allow us to identify the words that are important for classifying correct and incorrect answers. This identification compares the words used by learners with a high cosine similarity to the correct answer with those used by learners with a low cosine similarity. This can identify areas and concepts that learners with low cosine similarity do not understand.

The method uses the results of responses to writing questions from previous learners who have experienced the same lesson and task. Examining new learner’s writing can find those who seem to be faltering. Learners with a poor understanding do not use key vocabulary. They do not know what the key concepts are that they need to understand in the area.

In [28], the authors introduce transformer-based NLP models such as BERT for the automatic scoring of short-answer responses. However, we do not adopt BERT-based approaches because lighter methods—especially Doc2Vec—show higher performance while requiring fewer computational resources.

In [29], the authors explain that ChatGPT, which has recently received significant attention, can generate fluent and high-quality responses to user queries. However, in [30], the authors report that ChatGPT shows limitations in paraphrasing and in handling tasks that require processing semantically similar expressions. Because of these limitations, obtaining an accurate response often requires that the questioner clearly identify what is not understood and explicitly specify the relevant concept when asking a question. Mastering ChatGPT therefore requires understanding one’s own points of confusion and recognizing the key concepts within each content area. In this respect, the responses of past learners to writing questions provide valuable insight into which keywords are essential for understanding.

In [31], [32] and [33], the authors indicate that many prior studies on comprehension measurement rely on mark-test formats. This is to minimize the degree of freedom in the notation of the analyzed subject. This method analyzes answers to descriptive questions with a high degree of freedom. Open-ended questions are advantageous because they reduce bias and allow the discovery of unanticipated learner ideas. For learners, the open-ended nature of the questions means that they have to express themselves in their own words and explain logically, which is thought to improve their understanding.

For data science educators, the results can be used to create practice questions to check learner’s understanding, for example, by using the questions in the examinations as questions on points that many learners did not understand.

For data science learning, especially when it comes to the motivation and engagement of learners who have difficulty with the material, the open-ended questions in this study reveal key points in those who do and do not understand the material. Therefore, it can be assumed that teaching these points will have a more positive impact on motivation and commitment.

On the other hand, subjects could be asked to write Python code that solves an appropriate data analysis problem. With Python code, there are limited ways to express it, and the characteristics become clearer when it comes to understanding and not understanding. The analysis of the code allows for a more accurate analysis than analyzing descriptions in natural language. Therefore, this study also found it necessary to analyze the codes.

Although this study is able to identify the concepts needed for understanding, it is possible that it may not be applicable to a diverse range of learners at different levels. Therefore, as a future research project, although this experiment is conducted with about 15 learners, it may be necessary to conduct the experiment with a separate group of learners, e.g., 100 learners.

Given the work required to analyze the open-ended responses, it is clear that the method can be applied to around 15 upper-year university learners. However, it is not known whether this would make the possible analysis more accurate in larger classes or in an online environment. This method does not require training, so real-time evaluation is possible. However, large corpora are needed beforehand.

In this study, all participants are fully informed about the purpose of the research, the use of EDA, conversational data, and free-text response data, as well as the measures taken to protect their privacy. Written informed consent is obtained prior to participation. All collected data are anonymized and securely stored to ensure that no individual can be identified.

5. Conclusion

This study proposed a method for estimating learner's understanding of data science by analyzing their open-ended responses using document embeddings.

By transforming free-text answers into distributed representations and evaluating their cosine similarity to model answers, the method successfully identifies learners who demonstrate correct conceptual understanding. The quantitative evaluation shows that the Doc2Vec-based approach achieves the highest performance among multiple baselines, with an F1-score of 0.824, statistically significant improvement ($p = 0.0096$), and a large effect size (Cohen's $d = 1.286$; 95% CI [0.500, 0.833]). These results indicate that the embedding-based similarity measure provides a reliable metric for discriminating between adequate and insufficient comprehension.

The analysis further reveals that learners with low similarity scores tend to omit key conceptual vocabulary used in correct answers, allowing the method to automatically identify specific concepts or terminology that learners misunderstand or fail to recall. This demonstrates the utility of open-ended responses for diagnosing misconceptions in ways that traditional multiple-choice assessments cannot capture.

From an educational perspective, the proposed method offers actionable benefits for instructors. By automatically extracting missing key terms and detecting conceptual gaps from learner's written explanations, instructors can more efficiently pinpoint topics that require reinforcement and design targeted feedback or supplementary exercises.

The approach also provides an immediate diagnostic tool for monitoring learner's comprehension during data science instruction, helping educators intervene before misunderstandings become entrenched. Overall, the findings support the effectiveness of embedding-based analysis for assessing conceptual understanding from free-text answers and highlight the pedagogical value of incorporating open-ended questions into data science education. Future research will focus on extending the proposed method to larger and more diverse learner populations and exploring its integration into real-time educational support systems.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

We would like to thank many people for their guidance and encouragement in the preparation of this paper. We would like to thank everyone who willingly participated in the experiments during the execution of this paper. Finally, we would like to thank the members of the Data Engineering Laboratory for their invaluable advice and cooperation.

References

[1] R. D. De Veaux et al., "Curriculum guidelines for undergraduate programs in data science," Annual Review of Statistics and Its Application, 4, 15–30, 2017, doi:10.1146/annurev-statistics-060116-053930.

[2] H. Hedges and K. Given, "Addressing confirmation bias in middle school data science education," Foundations of Data Science, **5**(2), 2023, doi: 10.3934/fods.2021035.

[3] C. E. Brassil and B. A. Couch, "Multiple-true-false questions reveal more thoroughly the complexity of student thinking than multiple-choice questions:A Bayesian item response model comparison," International Journal of STEM Education, 6, 2019, doi: 10.1186/s40594-019-0169-0.

[4] K. Inohara, C. Matsuno, M. Furuya, and I. Kutsuzawa, "Differences between yes/no and multiple-choice vocabulary tests: Examination from the perspective of familiarity with reading," Japanese. Journal of Psychology, 91(6), 367–377, 2021, (in Japanese), doi: 10.4992/jjpsy.91.19028.

[5] R. Azuma, "Analysis of relationship between learner's characteristics and level of understanding using text-mining," Journal of Japan Society for Information and Systems in Education, 2017, (in Japanese).

[6] B. R. Shapiro, A. Meng, C. O'Donnell, C. Lou, E. Zhao, B. Dankwa, and A. Hostetler, "Re-Shape: A method to teach data ethics for data science education," in Proc. ACM CHI Conference on Human Factors in Computing Systems, 1–13, 2020, doi: 10.1145/3313831.3376251.

[7] C. Xing, D. Wang, X. Zhang, and C. Liu, "Document classification with distributions of word vectors," in Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, Siem Reap, Cambodia, 1–5, 2014.

[8] A. Condor, M. Litster, and Z. Pardos, "Automatic short answer grading with SBERT on out-of-sample questions," in Proc. 14th International Conference on Educational Data Mining, 345–352, 2021.

[9] K. Yasuda, H. Shimakawa, and F. Harada, "Identifying comprehension fault from word occurrences in writing questions," in Proc. Int. Conf. Frontiers of Signal Processing, Paris, France, 133–141, 2024, doi: 10.1109/ICFSP62546.2024.10785436.

[10] A. Géron, Practical Machine Learning with Scikit-learn, Keras, and TensorFlow, 2nd ed., M. Shimoda, Supervis., T. Nagao, Transl., O'Reilly Japan, pp. 215–235, 2020.

[11] D. Hachiya, Basic Python Learning from 0, Kodansha, 2020, (in Japanese).

[12] H. Zhang, "The optimality of naive bayes," in Proc. 17th International FLAIRS Conference, 562–567, 2004.

[13] M. W. Berry, M. Browne, A. N. Langville, V. P. Pauca, and R. J. Plemmons, "Algorithms and applications for approximate nonnegative matrix factorization," Computational Statistics and Data Analysis, 52, 155–173, 2007, doi: 10.1016/j.csda.2006.11.006.

[14] M. W. Gills and F. Glineur, "Document classification using nonnegative matrix factorization and underapproximation," in Proc. IEEE International Symposium on Circuits and Systems, 2782–2785, 2009, doi: 10.1109/ISCAS.2009.5118379.

[15] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," Nature, **401**, 788–791, 1999, doi: 10.1038/44565.

[16] Q. V. Le and T. Mikolov, "Distributed representations of sentences and documents," in Proc. 31st International Conference on Machine Learning, 32, 1188–1196, 2014, doi: 10.5555/3044805.3045025.

[17] T. Sasada, S. Mori, Y. Yamagata, H. Maeda, and T. Kawahara, "Definition of recipe terms and automatic construction of a tagging corpus for automatic recognition," Journal of Natural Language Processing, **22**, 107–131, 2015, (in Japanese), doi: 10.5715/jnlp.22.107.

[18] S. K. Safa and D. R. CH, "Development of a practical system for computerized evaluation of descriptive answers of middle school level students," Interactive Learning Environments, **30**(2), 215–228, 2019, doi: 10.1080/10494820.2019.1651743.

[19] A. McCallum and K. Nigam, "A comparison of event models for naïve Bayes text classification," in Proc. AAAI Conference Artificial Intelligence, 1998.

[20] T. Kudo, K. Yamamoto, and Y. Matsumoto, "Japanese morphological analysis using conditional random fields," Journal of Natural Language Processing, 161, 89–96, 2004, (in Japanese).

[21] M. Benedek et al., "A continuous measure of phasic electrodermal activity," J. Neurosci. Methods, **190**(1), 80–91, 2010, doi: 10.1016/j.jneumeth.2010.04.028.

[22] K. Yamanishi, J. Takeuchi, G. Williams, and P. Milne, "A unifying framework for detecting outliers and change points from time series," IEEE Transactions on Knowledge and Data Engineering, **18**(4), 482–492, 2006, doi: 10.1109/TKDE.2006.1599387.

[23] P. Ayres, J. Y. Lee, F. Paas, and J. J. G. van Merriënboer, "The validity of physiological measures to identify differences in intrinsic cognitive load," Frontiers Psychology, 12, 2021, doi: 10.3389/fpsyg.2021.643265.

[24] N. Nourbakhsh et al., "Using galvanic skin response for cognitive load measurement in arithmetic and reading tasks," in Proc. 24th Australasian Computer–Human Interaction Conference, 420–423, 2012, doi: 10.1145/2414536.2414602.

[25] C. Setz et al., "Discriminating stress from cognitive load using a wearable EDA device," IEEE Transactions on Information Technology in Biomedicine, 14, 410–417, 2009, doi: 10.1109/TITB.2009.2036164.

[26] B. Mehler et al., "Sensitivity of physiological measures for detecting systematic variations in cognitive demand from a working memory task: An on-road study across three age groups," Human Factors, 54, 396–412, 2012, doi: 10.1177/0018720812442086.

[27] N. Nourbakhsh, F. Chen, Y. Wang, and R. Calvo, "Detecting users' cognitive load by galvanic skin response with affective interference," ACM Transactions on Interactive Intelligent Systems, **7**(3), 1–20, 2017, doi: 10.1145/2960413.

[28] P. Ghavidel, S. Zargari, and A. Mohammadi, "Using BERT and XLNet for the Automatic Short Answer Grading Task," in Proc. International Conference on Artificial Intelligence in Education, 125–136, 2020, doi: 10.5220/0009422400580067.

[29] G. Zuccon and B. Koopman, "Dr ChatGPT, tell me what I want to hear: How prompt knowledge impacts health answer correctness," arXiv preprint arXiv:2304.10017, 2023, doi: 10.18653/v1/2023.emnlp-main.928.

[30] Q. Zhong, L. Ding, J. Liu, B. Du, and D. Tao, "Can ChatGPT understand too? A comparative study on ChatGPT and fine-tuned BERT," 2020, doi: 10.48550/arXiv.2302.10198.

[31] S. Ashraf, S. Saleem, T. Ahmed, Z. Aslam, and M. Shuaeeb, "Iris and foot based sustainable biometrics identification approach," in Proc. International Conference on Software, Telecommunications and Computer Networks, Split, Croatia, 1–6, 2020, doi: 10.23919/SoftCOM50211.2020.9238333.

[32] S. Saleem, S. Ashraf, and M. K. Basit, "CMBA – A candid multi-purpose biometric approach," ICTACT Journal of Image and Video Processing, 2211–2216, 2020, doi: 10.21917/ijivp.2020.0317.

[33] Y. Fujita, Y. Hino, and A. Akazawa, "Multigigabit optical interconnection LSTIs," in Proc. Symposium on VLSI Circuits, Kyoto, Japan, 69–70, 1993, doi: 10.1109/VLSIC.1993.920541.