P.A.L. Nadeesha
Bsc Honours in Computer Science, Graduate
University of Colombo School of Computing, Colombo, Sri Lanka
ORCID ID 0009-0005-6329-9091
lakshaninadeesha99@gmail.com

T.A. Weerasinghe
PhD in Computer and System Sciences, Senior Lecturer
University of Colombo School of Computing, Colombo, Sri Lanka
ORCID ID 0000-0002-7379-3916
taw@ucsc.cmb.ac.lk

W.R.N.S Abeyweera
Bsc Honours in Information Systems, Lecturer
University of Colombo School of Computing, Colombo, Sri Lanka
ORCID ID 0009-0009-9425-4450
rns@ucsc.cmb.ac.lk

AUTOMATIC SCORING OF KNOWLEDGE GAINED AND SHARED THROUGH DISCUSSION FORUMS: BASED ON THE COMMUNITY OF INQUIRY MODEL

Abstract. The Community of Inquiry (CoI) framework has been widely employed for the past two decades to assess the knowledge gained and shared through online discussion forums. The cognitive presence component of the CoI framework helps identify the evidence of thoughtful knowledge reconstructions through meaning-making during inquiry-based learning. Identifying and scoring these cognitive presences is essential for assessing the students' learning achievements through online discussion forums. Considering the difficulties associated with manual coding and identifying cognitive presences in discussion forums and the limitations in the existing techniques for auto-identifying and scoring cognitive presences, this research attempted to develop a more efficient tool to identify and score cognitive presences in online discussion forums. The research employed the constructive research approach. The methodology integrated Random Forest (RF) classification with TF-IDF feature extraction and Support Vector Machine (SVM) classification with Word2Vec embedding. A rule-based classifier, constructed upon indicator mappings, enriched the classification process. A weighted voting ensemble method was employed to combine the outputs of the individual classifiers. Our approach was trained and tested on two datasets comprising 781 messages containing 47,592 words. This ensemble method demonstrated notable efficacy, achieving a 69% accuracy rate in classification tasks. This highlights the robustness of the combined approach in enhancing classification performance. Furthermore, the study introduces a scoring model that calculates individual student scores based on post categories, enabling detailed evaluations of student engagement and participation. By assigning scores reflective of discussion contributions, this model advances comprehensive assessments of online learning interactions. Our work contributes to the ongoing conversation on leveraging machine learning for cognitive analysis in online learning environments, highlighting the importance of context-specific methodologies in advancing educational assessment practices.

Keywords: Cognitive presence; Community of Inquiry; Machine Learning; Natural Language Processing; Classification

1. INTRODUCTION

Discussion forums have proven to be an effective medium for collaborative online learning as they provide a platform for learners to share their knowledge and exchange information with peer learners. In the collaborative learning environment, learners can engage in discussions, ask questions, and provide answers to help one another learn and understand new concepts. By sharing their experiences and perspectives, learners can gain insights and deepen their understanding of a particular topic. Online discussion forums also offer flexibility and convenience, as learners can participate in discussions at any time and from anywhere, making it easier for them to learn at their own pace. A study by Peral et al. presents that using online discussion forums as an instructional tool can enhance students' learning performance by promoting cognitive and exploratory learning and fostering a sense of community and reflection [1].

The problem statement:

Despite their advantages, online discussion forums often face challenges such as managing redundant or irrelevant information and providing timely feedback to learners. The asynchronous nature of forums means discussions can remain open for extended periods, making it difficult for moderators and instructors to monitor and analyze content effectively. The analysis of discussion forum content is crucial as it offers valuable insights into student engagement, learning outcomes, and the overall effectiveness of online learning environments. By understanding the dynamics and patterns within discussion forums, instructors can gain deeper insights into student participation, interaction, and cognitive processes. However, traditional manual coding methods for analyzing forum content are labor-intensive, subjective, and infeasible for large-scale courses, creating a pressing need for more efficient solutions. Automated tools for analyzing discussion content and assessing key learning elements have the potential to address these challenges. While existing approaches, including semantic content filtering and topic extraction, have made notable progress, they often fall short in terms of scalability, precision, and adaptability to diverse contexts [2], [3], [4], [5].

Analysis of recent studies and publications:

Prior research has proposed several techniques to address the challenges of analyzing discussion forum content. Nunes et al. [2] introduced a topic extraction process to facilitate the analysis of online discussion forums. The process can help moderators identify important topics, which can be used to develop more effective moderation strategies. Distante et al. [3] leveraged information retrieval methods, such as topic modeling and concept analysis, to improve content navigation in forums. Albatayneh et al. [4] utilized Latent Semantic Analysis for semantic filtering of e-learning forum content. While these approaches provide valuable insights into analyzing discussion forum content, the content of the discussion forums can be examined using different investigative elements such as student participation and interaction, cognitive, metacognitive, and social cues, critical thinking, and group development as well [6]. These elements are crucial for understanding the depth of learner engagement and the overall quality of interactions in online discussions. The investigation of these aspects is vital for identifying key factors that contribute to fostering critical thinking, deep learning, and knowledge construction within online communities.

Students' involvement in a community of inquiry, which plays a crucial role in fostering critical thinking, deep learning, and knowledge construction within online discussions, can be analyzed using the Community of Inquiry (CoI) model. It is a widely used instrument for analyzing online discussion content, particularly in inquiry-based online learning [7] [8]. The CoI model is an educational framework introduced by Garrison, Anderson, and Archer [7] and has been further enhanced by research findings from many scholars (e.g., [8], [9]). It has been continuously refined and adapted, indicating its ongoing development and evolution [8] [10]. This framework is composed of three elements: social presence, cognitive presence, and teaching presence. While social, cognitive, and teaching presences are essential to foster meaningful interactions, cognitive presence is considered the most important element in identifying critical thinking activities in the CoI [11]. Cognitive presence refers to the extent to which learners are able to construct and confirm meaning through sustained reflection and discourse [12].

Assessing cognitive presence in online discussions is crucial for evaluating the effectiveness of instructional interventions and fostering meaningful learning experiences. However, manual coding methods have traditionally been used to assess cognitive presence in discussion forum posts within the CoI framework [13], largely due to the lack of effective automated techniques for analyzing discussions. Manual coding typically relies on qualitative analysis techniques such as content analysis and thematic coding to identify indicators of cognitive presence. These methods, while valuable, are labor-intensive, subjective, and require human annotators to classify content based on predefined categories of cognitive engagement. As a result, there is an increasing demand for automated techniques that can reliably assess cognitive presence, particularly as online learning environments scale. To facilitate the automatic analysis of cognitive presence, it is essential to develop systems that can classify student posts into the four phases of cognitive presence (triggering, exploration, integration, and resolution). The development of such automatic classifiers is critical, and there have already been several efforts to create these classifiers in languages such as English, Portuguese, and German [14] [15] [16] [17] [18].

The research goal:

While the above described classifiers have been developed in different linguistic contexts, there remains a lack of adaptability for applying them across diverse settings. Especially in terms of evaluating the cognitive presence in discussion forum posts, there is a need for approaches that assign meaningful value to each post based on cognitive presence. This study aims to address this gap by developing an automated approach to identify and classify cognitive presence within discussion forum posts, specifically within the CoI framework. The research focuses on applying machine learning techniques to improve the accuracy, precision, and scalability of cognitive presence analysis. By leveraging advanced models, the goal is to create a robust technique for evaluating online learning interactions and assessing student contributions. This work will also assess the effectiveness of existing classifiers in accurately analyzing discussion posts within our specific context. Through experiments with these existing techniques and the selection of optimal methodologies, we aim to introduce a novel approach to classify student posts and assign a rating based on their cognitive presence category.

The research questions:

The following research questions guide the study:

- Main Question:
  How can an automatic rating model be developed to assess cognitive presence in discussion forum posts within the CoI framework?
- Sub-questions:
  RQ1: What are the most appropriate existing models that can be used to assess the quality of students' contributions in online discussion forums?
  RQ2: To what extent can the most appropriate existing models identify cognitive presence elements in discussion forums?
  RQ3: What other techniques can be used to improve the performance of the selected models for identifying cognitive presence elements in discussion forums?

In the following sections, we will present an overview of the Community of Inquiry framework, with a specific focus on the cognitive presence construct. Additionally, we will review existing research on automated classifiers of cognitive presence in online discussion forums. Subsequently, we will delve into our methodological approach and our data sources. The results obtained through our automated classifier will be presented, accompanied by insights into the features utilized in the analysis.

2. THE THEORETICAL BACKGROUNDS

2.1 CoI framework and cognitive presence

The CoI framework, rooted in social constructivism, is essential for analyzing and understanding learning in asynchronous online discussion forums [11]. It provides a structured approach to examining the educational experience within online learning communities and is widely used in analyzing online discussion content. This framework emphasizes four presences that are interrelated and work together to create a rich and effective learning experience. However, in this study, we focused on fostering "critical inquiry" in text-based discussion forums, specifically emphasizing cognitive presence due to its crucial role in promoting critical thinking and knowledge construction. Cognitive presence, defined as a cycle of progressive knowledge construction, comprises four phases: 1) Triggering event (C-TE): the beginning of discussions initiated by a problem or dilemma, 2) Exploration (C-EX): students explore potential solutions, engaging in brainstorming and information exchange, 3) Integration (C-TE): students synthesize new ideas through social construction, and 4) Resolution (C-RA): students resolve the initial problem by testing and applying new knowledge, often reaching a consensus [18]. Following Figure 1 shows the subcategories of cognitive presence.

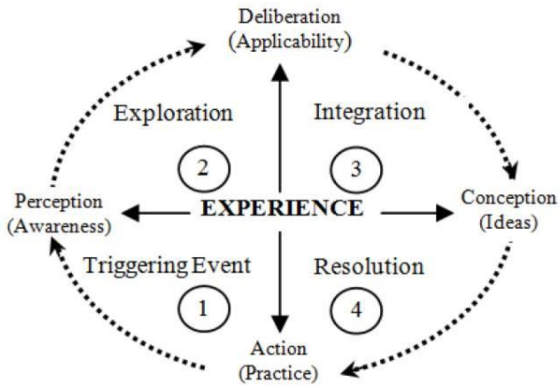

Figure 1: Categories of cognitive presence in CoI Framework

Cognitive presence focuses on the intellectual and cognitive dimensions of learning, offering tools to measure the depth of learning in online discussions. It assesses students' engagement in critical thinking, problem-solving, and knowledge construction, highlighting their contributions to knowledge development. Evaluation typically involves content analysis of online discussion messages, focusing on cognitive, metacognitive, social, and teaching presences. Entire messages are analyzed using coding schemes, but this manual process is time-consuming and limits the scalability of the CoI model for evaluating learning depth and critical thinking [19].

To address this, a recent study validated methods to categorize MOOC discussion transcripts by cognitive presence phases, suggesting improvements for using the CoI framework in MOOC contexts with larger datasets and more coders [20]. However, there remains a need for automated coding methods to broaden the use of the CoI model in assessing cognitive presence in discussion forums.

2.2 Automated classifiers of cognitive presence in online discussion forums

2.2.1 Automatic content analysis of discussion messages

Online educational platforms can effectively use machine learning due to the extensive data collected for learning purposes. Previous research has tested various machine learning models to uncover patterns in online students' learning behaviors, encouraging the use of technology to enhance educational environments [21]. The concept of learning analytics has gained prominence in higher education, with institutions leveraging substantial datasets. Researchers have applied various learning analytics techniques, including classification, clustering, and text mining [22], for purposes such as detecting student behavior and predicting performance [21], identifying students at risk [23], analyzing students' forum interactions, and providing visualization to inform instructors and other key stakeholders. However, further research is required to explore how learning analytics can enhance online instructional practices and student outcomes, emphasizing the importance of measures like cognitive presence in predicting student performance in online learning environments.

2.2.2 Supervised machine learning classifiers

Researchers have developed several automated classifiers using different ML algorithms to analyze the phases of cognitive presence in online discussion forums from small-scale courses.

McKlin [24] used a basic artificial neural network (ANN) with dictionary-based words and phrases to categorize cognitive presence phases. Human coders initially categorized messages to measure reliability, and the ANN's performance was evaluated against this measure. McKlin's [24] ANN achieved a Holsti's coefficient of reliability (CR) of 0.68 and a Cohen's k of 0.31. Corich et al. [25] employed a Bayesian network classifier, also using dictionary-based words and phrases for classification, reaching a CR of 0.71 and Cohen's k of 0.65. This indicated some improvement over McKlin's [24] results. Notably, McKlin's [24] classifier excluded the minority class (the Resolution phase), while Corich et al.'s [25] classifier analyzed cognitive presence at the sentence level instead of the message level.

Kovanović et al. [26] developed a cognitive presence classifier using the Support-Vector-Machine algorithm with n-grams and thread structures, achieving 58.4% accuracy and a Cohen's k of 0.41. Waters et al. [27] enhanced this by adding more structural features, resulting in a Conditional Random Fields classifier with 64.2% accuracy and a Cohen's k of 0.482, underscoring the significance of structural features in identifying cognitive presence phases in online discussions. However, the n-gram methods faced limitations: they generated a high-dimensional space, increasing the risk of overfitting, and rendered the classifier overly domain-specific, reducing its generalizability. Additionally, the uneven distribution of cognitive presence phases in the sample data posed challenges to the classifier's performance.

To address previous issues, Kovanović et al. [15] introduced a random forest (RF) classifier using features from computational linguistics analysis, including Coh-Metrix, LIWC, LSA, named entities, and conversational structures. They also applied over-sampling techniques to combat class imbalance, achieving 70.3% accuracy and a Cohen's k of 0.63. However, a replication study by Farrow et al. [28] showed a decline in accuracy to 61.7% and Cohen's k to 0.46, indicating that Kovanović et al.'s [15] method had produced overly optimistic results due to the over-sampling method being applied before the training-test data split.

A subsequent study by Neto et al. [18] achieved 76% accuracy and a Cohen's k of 0.55 by classifying cognitive presence phases in discussion messages from biology and technology courses. Hu et al. [20] applied their classifier to a philosophy dataset and validated it across medicine, education, and humanities, also achieving 76% accuracy and a Cohen's k of 0.54. Their method, applied in the context of MOOCs, demonstrated that a classifier designed for a specific MOOC can effectively extend to various disciplines within MOOCs.

Gorgun et al. [5] developed predictive models to automate the identification of cognitive engagement in online discussion posts using features extracted with Coh-Metrix. They achieved 68% accuracy with a decision tree, 95% with a random forest, and 85% with an SVM. The study highlighted the importance of word count, AWL (Academic Word List) count, and Flesch-Kincaid grade level in predicting cognitive engagement.

Having looked at several studies, they developed their classifiers using the same dataset, which consists of posts from the same course. For instance, several studies utilized data from a single postgraduate software engineering online course ([26], [27], [28], [29]). In contrast, Hu et al. [20] applied their classifier to a philosophy dataset and validated it across three disciplines (medicine, education, and humanities). Additionally, Neto et al. [18] applied a random forest classifier to a dataset in the field of biology and technology.

A common challenge across these studies is the imbalance in the occurrences of cognitive presence phases. To address this issue, researchers frequently employed the oversampling strategy SMOTE (Synthetic Minority Oversampling Technique), which was initially introduced by Kovanovic´ et al. [15] and elaborated upon by Farrow et al. [28]. Studies focusing on the Portuguese language also utilized SMOTE [16].

Previous research has utilized four main feature types for classifying student posts into cognitive presence phases: textual, structural, LIWC, and Coh-Metrix features. Textual features derived from raw posts include techniques like n-grams and Word2Vec, as used by Hayati et al. [30] and others, such as McKlin [24] and Kovanović et al. [26]. Structural features, employed by Kovanović et al. [26] and Neto et al. [18], capture a post's position within a discussion sequence, such as whether it is an opening message or a reply. LIWC features, generated using the Language Inquiry and Word Count (LIWC) tool, group words into categories like cognitive or affective processes. Key LIWC features for English include question marks and first-person singular pronouns, while for Portuguese, features include prepositions, conjunctions, and third-person singular pronouns [18]. Coh-Metrix features, derived from the Coh-Metrix tool, assess cohesion, language, and readability and have been integrated into classifiers by authors like Farrow et al. [28], Hu et al. [20], and Neto et al. [18]. However, combining these feature types has not consistently improved classification accuracy and limited access to LIWC and Coh-Metrix tools poses additional challenges.

The preceding discussion provided an overview of supervised machine learning classifiers across various language domains, with this study focusing specifically on the English language. The aim is to advance understanding of supervised learning methods in this domain. The study proposes a novel approach that combines machine learning model outputs with rule-based systems to refine classification outcomes. This integrated method seeks to achieve significantly better performance metrics than previous methods. The works listed in Table 1 have been considered in developing the proposed model.

Table 1

Summary of prior work reviewed

Authors	Machine learning algorithm	Main features	Accuracy (%)	Cohen' k
Corich et al. (2006)	Bayesian Network	Dictionary- based words and phrases	71	0.65
Farrow et al. (2019)	Random Forest	Same as Kovanovic´ et al. (2016)	61.7	0.38
Hu et al. (2022)	Random Forest	Same as Kovanovic´ et al. (2016)	73.6	0.54
Gorgun et al. (2022)	Decision Tree, Random Forest, and Support Vector Machine	Coh-Metrix, non- linguistic contextual features	68-DT, 95-RF, 85-SVM	0.46-DT, 0.55-RF, 0.61-SVM

3. RESEARCH METHODS

3.1. Experimental setup

3.1.1. Dataset

This research utilized data collected from online discussion forums associated with two courses offered by the University of Colombo School of Computing (UCSC) in Sri Lanka. All these discussions were conducted with no or minimal interactions from a facilitator. It is important to note that the courses are delivered in English, and English is the recommended language for online discussions. This study used two datasets: one extracted from a course in the BITVLE (https://vle.bit.lk/) that was delivered completely online, and the other obtained from an online course in the UGVLE (https://ugvle.ucsc.cmb.ac.lk/) that was used for blended learning.

The first dataset consisted of 60 forums, and the second dataset consisted of one forum. The first dataset comprises 607 messages contributed by 136 students, while the second dataset contains 210 messages from 101 students. Consent for the use of their messages in this study was obtained from both the students and the facilitator. Subsequently, the collected data was manually classified into different phases of cognitive presence. The coding categories for cognitive presence (triggering event, exploration, integration, resolution) were based on the coding scheme outlined by [6]. These forums revolve around distinct contests. Below, Table 2 illustrates the analysis of each forum sample.

Table 2

Analysis of Sample Discussions

Forum name	Factors	Total messages
Forum 01	C-TE	103
	C-IN	212
	C-EX	199
	C-RA	73
Forum 02	C-TE	03
	C-IN	91
	C-EX	100
	C-RA	16

3.1.2. Data preprocessing

In our efforts to prepare the dataset for training and classification in building a cognitive engagement classifier, we implemented several pre-processing steps. These included removing website links and instances of “see attached” notations, stripping HTML tags, eliminating new lines, white spaces, and tabs, as well as expanding contractions and removing numbering and bullet points. These steps collectively aimed to create a cleaner and more coherent textual dataset for subsequent analysis and classifier development. Additionally, in our content analysis task focusing on students’ forum posts and cognitive presence, we retained stop words to preserve cognitive nuances within messages. This decision ensures that the preprocessing steps remain straightforward, primarily involving converting uppercase text to lowercase. The preprocessed sentences are then utilized in subsequent classification stages, preserving language integrity and cognitive nuances for accurate cognitive presence analysis in student discussions.

3.2. Methods

3.2.1. Feature identification

The features examined in this study were drawn from prior research documented in the literature, specifically studies by [18], [20], and [5], focusing on theory-driven features for analysis rather than traditional text classification elements like N-grams or POS tags, to enhance interpretability. Four key feature categories were identified as follows.

1) Context features: In the context of this study, contextual features play a critical role in enriching the depth and comprehensiveness of our analysis. These features encompass the authors of both the parent and current messages, the message creation time, and the thread topic, all of which contribute to a more nuanced understanding of communication dynamics. By incorporating these contextual elements in addition to the message content, our study seeks to gain a deeper understanding of the dynamics and subtleties within the communication landscape under investigation.

2) Named entities: According to the literature by [18], it is proposed that the quantity of named entities (such as entities referring to individuals, organizations, and geographical locations) varies across different phases of cognitive presence. Consequently, Exploration messages, characterized by the exploration of new concepts and opinions, are anticipated to contain a greater number of named entities compared to Integration and Resolution messages.

3) Coherence and Cohesion: The research conducted by [18] emphasizes the significance of coherence and cohesion. Their studies reveal that the placement of a message within a discussion thread and its similarity to the subsequent message exhibit strong associations with the Triggering and Resolution phases, respectively. Additionally, Kovanović et al. [15] highlighted the importance of coherence and cohesion in the successful classification of discussion messages for cognitive presence.

4) Word count: The concept of word count involves quantifying the number of words in a given post. As per the findings of Waters et al.'s [27] study, it is anticipated that as a discussion progresses towards the integration and resolution phases, the volume of content tends to increase. This escalation is attributed to the synthesis and integration of ideas during these pivotal stages of discourse.

These features aim to deepen the understanding of communication patterns in the study.

3.2.2. Feature extraction

In this study, we employed several feature extraction methods to improve the performance of our classification model. Building on relevant precedents in the existing literature, Word2Vec was used for word embedding and achieved notable results in terms of average accuracy [30]. Building on this approach, we also harnessed the power of Word Embedding, specifically utilizing Word2Vec, to convert words and phrases into dense vector representations. In addition to this, we leveraged TF-IDF (Term Frequency-Inverse Document Frequency) as another feature extraction method. TF-IDF assigns weights to words based on their frequency within a document in relation to their importance within a larger corpus of documents. These feature extraction techniques were carefully chosen to capture the relevant linguistic characteristics and nuances within the text data, contributing to the effectiveness of our classification model.

3.2.3. Model selection

In constructing our classification model, we utilized two leading supervised algorithms: Random Forest (RF) and Support Vector Machines (SVM). While RF is commonly used in cognitive presence classification, our choice was informed by both literature and experimental evaluation, ensuring these models offered the best performance on our dataset.

Random Forest:

Random Forest (RF) was chosen because it leverages an ensemble of decision trees and provides feature importance, aiding interpretability. During the model development phase, we trained the RF classifier using the TF-IDF feature representation of the student posts. TF-IDF is a widely used technique for feature extraction in text analysis. By leveraging TF-IDF in conjunction with Random Forest, we aimed to capture the salient features of student posts that contribute to the classification of cognitive presence categories.

Support Vector Machines:

Support Vector Machines (SVM) is a powerful supervised learning algorithm used for classification tasks. SVM identifies the optimal hyperplane to separate classes and excels with high-dimensional data. We utilized the Word Embedding (Word2Vec) feature representation of the student posts for training the SVM classifier. By integrating Word2Vec with SVM, we aimed to leverage the contextual information embedded within the word embedding to enhance the classification performance of our model.

To determine the most suitable model for assessing the quality of students' contributions in online discussion forums, we conducted experiments and reviewed existing literature. Notably, RF and SVM emerged as frontrunners in achieving high accuracy. Previous studies by Kovanovic´ et al. [15], Farrow et al. [28], Hu et al. [20], and Gorgun et al. [5] showcased the effectiveness of RF models in cognitive presence classification. Additionally, commendable accuracy was observed with SVM models in the works of Gorgun et al. [5] and Kovanovic´ et al. [26]. Building upon these findings and our own experimental results, we selected RF and SVM as the most accurate models for our classification framework, addressing the first research question (RQ1) through both empirical evidence and existing literature.

3.2.4. Model tuning

Model training, hyperparameter tuning, and analyzes were conducted in Google Colaboratory using the sklearn and nltk packages. To optimize the performance of both ML classifiers, we employed grid search with 10-fold nested cross-validation. For the RF classifier, the best model performance was achieved with the following hyperparameters: max depth = 20, min sample leaf = 1, min sample split = 2, and number of estimators = 200. Similarly, the SVM classifier's optimal hyperparameters were determined as follows: regularization (C) = 0.1, degree = 2, gamma = 0.001, and kernel = 'rbf'. Upon tuning, the best-performing Random Forest model attained a 61% accuracy on the testing set, while the SVM model achieved a 45% accuracy.

3.2.5. Rule based classifier

We have developed a rule-based system inspired by previous work such as the dictionary-based methods outlined by [24] and [25]. This classification aims to enhance the manual process of categorizing student posts by automating the classification task. The rule-based classifier endeavors to categorize student posts into cognitive phrases based on predefined chunks, a technique previously employed manually as described by [6]. In this endeavor, we sought to refine the categorization process by incorporating category definitions and indicators from the cognitive presence in the CoI model, thereby enhancing the precision of chunk mapping and subsequently improving the model's performance. Furthermore, the rule-based classifier serves as a solution for scenarios where machine learning models may encounter uncertainties or misclassifications. This approach effectively addresses these issues by employing an expert-defined coding scheme in conjunction with chunk mappings, thereby ensuring more accurate classification results.

The development of our rule-based classifier serves as a direct response to the third research question (RQ3), which explores alternative techniques to enhance the performance of existing models (RF and SVM) in identifying cognitive presence elements within discussion forums. By integrating domain knowledge and expertise into the existing models, our rule-based classifier offers a novel approach aimed at improving classification accuracy.

3.2.6. Integration of classifiers

In our classification framework, we harnessed the strengths of both ML models and a rule-based classifier to develop a robust cognitive presence classifier. Combining these classifiers, we introduced a weighted voting classifier, facilitating the integration of predictions from multiple models while considering their individual performance. This ensemble learning technique enhances classification accuracy by leveraging diverse perspectives and decision-making strategies of the constituent classifiers.

3.2.7. Custom weighted voting classifier

Our implementation features a custom-weighted voting classifier, an ensemble learning method that merges predictions from multiple base classifiers, each weighted according to its perceived reliability. We incorporated predictions from three classifiers - Random Forest, Support Vector Machines, and a rule-based classifier. And then assigning weights based on their performance and relevance to the classification task. The weighted voting classifier then aggregates these predictions to determine the final class label for each instance, thereby improving overall classification accuracy.

3.2.8. Rating model

Depending on the weighted voting classifier we have developed, a rating model has been crafted to assess student contributions in online discussions by assigning ratings to various categories of posts. This model, structured into three key phases - category rating assignment, post-rating calculation, and student score aggregation, aims to provide a comprehensive evaluation framework.

In the category rating assignment phase, each category (C-TE, C-EX, C-IN, C-RA) is assigned a score reflecting its significance and contribution level to the discussion. The subsequent post-rating calculation phase determines individual post-ratings based on their assigned categories, simplifying the process with direct mappings between categories and scores. This transparent approach ensures consistency in evaluating student posts and facilitates efficient assessment across various discussion topics.

Following post-rating calculation, the student score aggregation phase consolidates ratings from all posts contributed by each student to derive a comprehensive score reflecting their overall contribution to the discussion. By summing up individual post ratings, this approach offers instructors a clear understanding of each student's engagement and performance, enabling targeted feedback and assessment. Implementing this rating model provides valuable insights into student engagement, offering instructors a quantitative measure to evaluate and enhance online discussion participation.

4. THE RESULTS AND DISCUSSION

4.1. RF with TF-IDF

The initial implementation of RF with TF-IDF yielded suboptimal results, with only 52.6% accuracy. Experiment 1, where stop words were retained, showed no improvement, maintaining a 50% accuracy rate. However, after incorporating additional features such as named entities, cohesion, coherence, and word count, accuracy improved significantly to 61%. Despite applying SMOTE to address the class imbalance, no enhancement in accuracy was observed.

After applying feature engineering techniques to enhance the performance of the RF classifier in predicting cognitive presence, the model's evaluation of feature importance revealed compelling insights. Notably, the classifier identified the student ID, representing the individual responsible for posting the message, as the most crucial predictor. This finding highlights the importance of understanding each student's contributions in shaping cognitive presence within the online learning environment. Following closely in importance were the features of coherence and word count, emphasizing the importance of linguistic coherence and the quantity of text in predicting cognitive presence. Figure 2 below illustrates the feature's importance, visually capturing the significance of these predictors in driving the classifier's decision-making process.

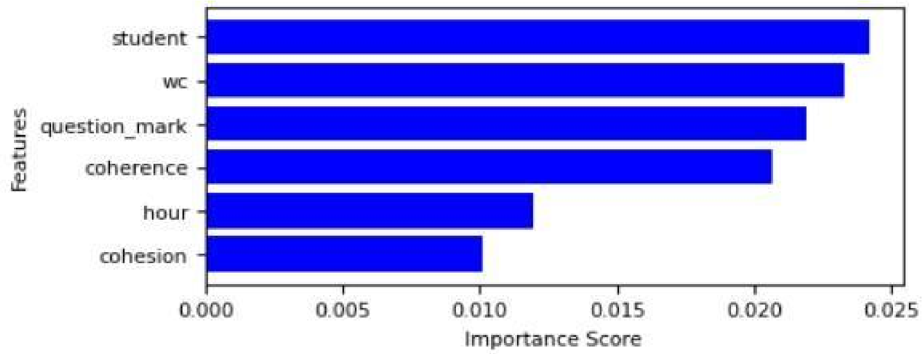

Figure 2: Feature importance

4.2. SVM with Word2Vec

SVM with Word2Vec, retaining stop words, achieved 45% accuracy. Although this accuracy is lower than that achieved by the RF with the TF-IDF model, which attained 61%, it's essential to note that the SVM model with Word2Vec captures different features compared to the RF model with TF-IDF. Despite its lower overall accuracy, the SVM model leveraging Word2Vec embedding and incorporating feature engineering directly into the SVM framework provides valuable insights into the classification task, offering a complementary perspective to the TF-IDF-based RF model. Further parameter tuning and cross-validation were applied to improve performance.

4.3. Rule-based classifier

The rule-based classifier attained 60% accuracy, showcasing potential for integration with ML models for enhanced performance. This approach leverages expert-defined rules and indicators to categorize student discussions. This rule-based classifier's accuracy of 60% highlights its potential as a valuable tool for categorizing student discussions. Further refinement and potential integration with machine learning models can enhance performance and contribute to more accurate classification outcomes.

4.4. Weighted voting classifier

Integrating RF with TF-IDF, SVM with Word2Vec, and the rule-based classifier, the weighted voting classifier achieved 69% accuracy. This ensemble approach outperformed individual models, demonstrating the efficacy of ensemble learning in improving classification accuracy. Furthermore, it leverages the strengths of each individual model while mitigating their weaknesses, resulting in improved overall performance. The diversity of the models, stemming from their distinct methodologies and feature representations, contributes to the ensemble's ability to capture a broader range of patterns and nuances in the data.

Overall, Table 3 below depicts the comprehensive accuracy of each model, alongside the precision, recall, and F1 score for every category. It is evident that the weighted voting classifier, which combines predictions from various models, has surpassed the performance of the state-of-the-art model.

Table 3

Accuracy of Each Classifier

	Category	**Precision**	**Recall**	**FI Score**	**Accuracy (%)**
RF with TF-IDF	C-TE	0.62	0.57	0.59	61.06
	C-IN	0.60	0.76	0.67	
	C-EX	0.60	0.65	0.62	
	C-RA	1.00	0.12	0.21	
SVM with Word2Vec	C-TE	0.14	0.88	0.24	45.73
	C-IN	0.44	0.94	0.60	
	C-EX	0.64	0.12	0.20	
	C-RA	0.12	0.76	0.20	
Rule Based	C-TE	0.75	0.53	0.62	60.36
	C-IN	0.54	0.82	0.65	
	C-EX	0.71	0.52	0.60	
	C-RA	1.00	0.06	0.11	
Voting Classifier	C-TE	0.75	0.88	0.81	69.51
	C-IN	0.66	0.88	0.75	
	C-EX	0.71	0.62	0.62	
	C-RA	1.00	0.24	0.38	

4.5. Rating model

The rating model quantifies individual student scores based on their engagement in discussions, offering a comprehensive assessment of participation. It provides valuable insights into both the quality and quantity of students' contributions. By assigning scores to each student's posts according to predefined categories, the model enables a thorough evaluation of their engagement and participation levels, as previously outlined.

Figure 3 below illustrates the results obtained from applying the rating model to the test dataset. Each student is assigned a score calculated by aggregating the ratings of all their posts within the discussion. This score serves as a quantitative measure of the student's overall contribution, reflecting the quality and quantity of their engagement.

```
Students with Percentage Scores More Than 20:
Student       Score
S1            80.0
S10           40.0
S100          30.0
S101          30.0
S11           70.0
              ...
student 85    45.0
student 92    30.0
student 96    40.0
student 97    40.0
student 98    40.0
Length: 95, dtype: float64
```

Figure 3: Results of the rating model

4.6. Discussion

The goal of this study was to develop a rating model that measures students' contributions and participation levels based on cognitive presence within the CoI framework. To achieve this goal, we developed classification models capable of identifying cognitive engagement in online discussion forums. Leveraging additional features extracted from student posts, we trained three classifiers: RF with TF-IDF, SVM with Word2Vec, and a rule-based classifier. The use of transparent classifiers enabled a more straightforward evaluation of feature importance for predicting cognitive engagement. By integrating these models into a weighted voting classifier, we achieved 69% accuracy on the testing dataset, outperforming individual models and highlighting the effectiveness of ensemble learning. Using insights from these classifiers, we constructed a rating model to quantify student engagement scores in discussions. The findings and their implications for education and research are discussed further.

Key findings and conclusions have emerged in response to the main research question, “How can an automatic rating model be developed to assess cognitive presence in discussion forum posts within the CoI framework?” along with its corresponding sub-questions. Throughout this study, we systematically addressed three crucial research questions (RQ1, RQ2, and RQ3) to discover the efficacy of existing models in assessing the quality of students’ contributions in online discussion forums and identifying cognitive presence elements.

In addressing Research Question 1 (RQ1), the study investigated existing literature and experimental results to determine the efficacy of both the RF and SVM models in assessing the quality of student contributions. Previous studies by [15], [28], [20], [30], [5] and [26] highlighted the effectiveness of RF in cognitive presence classification and demonstrated the effectiveness of Word2Vec for word embedding, as well as the commendable accuracy achieved with SVM models. Drawing from these precedents and experimental results, the study selected both RF and SVM with Word2Vec as suitable candidates for the classification framework, aligning with the literature and experimental findings and indicating their appropriateness for evaluating student contributions in online discussion forums.

In relation to Research Question 2 (RQ2), we assessed the RF and SVM models' ability to classify cognitive presence elements in discussion forums. The RF model with TF-IDF achieved 61% accuracy, while the SVM model reached 45%, highlighting challenges in accurately identifying cognitive presence elements. Although SVM had lower accuracy, it offered complementary insights, underscoring the complexity of the task.

To address Research Question 3 (RQ3), we developed a rule-based classifier leveraging domain knowledge to improve categorization accuracy. This classifier extracted indicators using predefined rules and achieved 60% accuracy, complementing the existing models. These findings align with [6] emphasizing the value of integrating domain-specific knowledge into classification frameworks.

Next, we devised ensemble techniques such as the weighted voting classifier to combine predictions from multiple models, strengthening overall classification accuracy. This weighted voting classifier demonstrated promising capabilities in identifying cognitive presence elements in discussion forums. The results of the weighted voting classifier demonstrate its efficacy in achieving promising outcomes. By combining predictions from the RF with TF-IDF, SVM with Word2Vec, and rule-based classifiers, the ensemble achieved an accuracy of 69% on the testing dataset. This accuracy represents a significant improvement compared to the individual models' performances and underscores the value of ensemble learning in strengthening classification outcomes.

Later, we introduced a novel rating model that assigns scores to student posts based on their cognitive presence category, offering a nuanced perspective on contribution quality. This rating model quantifies individual student scores based on engagement in discussions, providing a comprehensive assessment of student participation and offering insights into the quality and quantity of their contributions. As previously discussed, the model facilitated the calculation of individual student scores by assigning scores to each student's posts according to their respective categories. In conclusion, this study has successfully developed an automatic rating model that effectively assesses cognitive presence in discussion forum posts within the CoI framework. By systematically addressing each sub-question, we have demonstrated the viability of the selected approach and its potential to enhance online learning assessment practices.

5. CONCLUSIONS AND PROSPECTS FOR FURTHER RESEARCH

5.1. Limitations

While our study has yielded valuable insights, its approach inherits a few limitations. Firstly, the model we have developed can capture only the cognitive presence of students in online discussion forums. The accuracy and applicability of the model is conditional upon the choices made during its construction. Therefore, our model's efficacy may be constrained in accommodating presences beyond cognitive presence within the CoI framework. Moreover, its performance may falter when applied to datasets from different contexts, such as varying languages or subject areas. To mitigate this challenge, we included a rule-based classifier; however, generalizability concerns may persist. Furthermore, our model's dependency on the cognitive presence category for assigning ratings to student posts may be overlooked in student learning within discussion forums. While we have made great efforts to address these limitations to the best of our ability, they may still require consideration in interpreting the findings and future research directions.

5.2. Conclusion

Through a systematic exploration of three key research questions, we identified the strengths and limitations of existing models, shedding light on the complexities of cognitive presence identification in discussion forums. The study revealed that RF with TF-IDF and SVM with Word2Vec are more appropriate for identifying cognitive presence in online discussion forums. By integrating these classification models, with a rule-based classifier, we achieved a significant milestone in automated assessment techniques for online discussion forums. These individual classifiers were combined into a weighted voting classifier, resulting in impressive accuracy on the testing dataset and surpassing the performance of the standalone models. This ensemble approach highlighted the efficacy of ensemble learning in enhancing classification accuracy and laid the groundwork for more robust assessment practices in online learning environments.

Our work contributes to the ongoing conversation on leveraging machine learning for cognitive analysis in online learning, with a particular focus on the integration of rule-based classifiers alongside traditional machine learning models. This study's novelty lies in the integration of machine learning techniques with a rule-based approach, providing a more reliable and scalable method for assessing cognitive presence in online discussions. This integration not only improves the reliability and scalability of cognitive presence assessment but also paves the way for more personalized and context-sensitive educational practices in online learning environments. The development of a rating model also offered a nuanced perspective on students' engagement and participation, further enhancing our understanding of online learning dynamics. Ultimately, this research contributes to the growing field of automated assessment tools for online learning, fostering more comprehensive and scalable methods for evaluating student engagement and learning outcomes in digital classrooms.

5.3. Future directions

For future research, addressing the limitations identified in our study presents a promising direction for advancement. Specifically, developing similar models that can assess other presences within the CoI framework, such as social and teaching presences, holds the potential to enhance our understanding of online learning dynamics. Additionally, developing an automated rating model that evaluates students' online learning contributions and quality of e-learning while considering factors beyond cognitive presence categories could be instrumental. By analyzing variables such as language types and discussion points, researchers can gain deeper insights into student engagement and learning outcomes in online forums. These avenues of exploration broaden the scope of research in online learning assessment and provide valuable insights for improving educational practices and methodologies.

ACKNOWLEDGEMENT

We would like to express our sincere gratitude to Mrs. M. Krishnathasan for her contribution as the external coder during the evaluation process. We also appreciate all individuals who provided feedback to enhance our model.

REFERENCES (TRANSLATED AND TRANSLITERATED)

[1] J. Peral, A. Ferrández, H. Mora, D. Gil and E. Kauffmann, "A Review of the Analytics Techniques for an Efficient Management of Online Forums: An Architecture Proposal," *IEEE,* vol. 7, pp. 12220-12240, 2019. (in English)

[2] B. P. Nunes, A. Mera, R. Kawase, B. Fetahu, M. A. Casanova and G. H. B. d. Campos, "A topic extraction process for online forums," in *2014 IEEE 14th International Conference on Advanced Learning Technologies*, 2014. (in English)

[3] D. Distante, A. Fernandez, L. Cerulo and A. Visaggio, "Enhancing Online Discussion Forums with Topic-Driven Content Search and Assisted Posting," in *International Joint Conference on Knowledge Discovery, Knowledge Engineering, and Knowledge Management*, Rome, Italy, 2015. (in English)

[4] N. A. Albatayneh, K. I. Ghauth and F.-F. Chua, "Utilizing Learners' Negative Ratings in Semantic Content-based Recommender System for e-Learning Forum," *Journal of Educational Technology & Society,* vol. 21, no. 1, p. 112–125, 2018. (in English)

[5] G. Gorgun, S. N. Yildirim-Erbasli, and C. D. Epp, "Predicting Cognitive Engagement in Online Course Discussion Forums," в *International Educational Data Mining Society*, Durham, United Kingdom, 2022. (in English)

[6] T. A. Weerasinghe, K. P. Hewagamage and R. Ramberg, "Re-evaluation of community of inquiry model," *The International Journal on Advances in ICT for Emerging Regions,* vol. 5, no. 4, pp. 1-13, 2012. (in English)

[7] D. R. Garrison, T. Anderson and W. Archer, "Critical Inquiry in a Text-Based Environment: Computer Conferencing in Higher Education," *The Internet and Higher Education,* vol. 2, no. 2-3, p. 87–105, 1999. (in English)

[8] D. H.-C. Cheung, A. K.-L. Ng, K.-M. Kiang and H. H.-Y. Chan, "Creating a community of inquiry in the science classroom: an effective pedagogy for teaching diverse students?," *Journal of Further and Higher Education,* vol. 44, no. 1, pp. 1-13, 2020. (in English)

[9] D. c. K. Kozan and J. C. Richardson, "Interrelationships between and among social, teaching, and cognitive presence," *The Internet and Higher Education,* vol. 21, pp. 68-73, 2014. (in English)

[10] P. Shea, J. Richardson and K. Swan, "Building bridges to advance the Community of Inquiry framework for online learning," *Educational Psychologist,* vol. 57, no. 3, pp. 148-161, 2022. (in English)

[11] D. R. Garrison, T. Anderson and W. Archer, "Critical thinking, cognitive presence, and computer conferencing in distance education," *American Journal of Distance Education,* vol. 15, no. 1, pp. 7-23, 2001. (in English)

[12] Y.-L. Lee, "Co-Creating Educational YouTube Videos as Site as a Community of Inquiry," *Scholarly and Research Communication,* vol. 13, no. 2, pp. 15-pp, 2022. (in English)

[13] Y. Taniguchi, S. Konomi and Y. Goda, "Examining Language-Agnostic Methods of Automatic Coding in the Community of Inquiry Framework," in *International Association for Development of the Information Society*, Cagliari, Italy, 2019. (in English)

[14] K. F. Hew, X. Hu, C. Qiao and Y. Tang, "What predicts student satisfaction with MOOCs: A gradient boosting trees supervised machine learning and sentiment analysis approach," *Computers & Education,* vol. 145, p. 103724, 2020. (in English)

[15] V. Kovanović, S. Joksimović, Z. Waters, D. Gašević, K. Kitto, M. Hatala and G. Siemens, "Towards automated content analysis of discussion transcripts: a cognitive presence case," *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge,* pp. 15-24, 2016. (in English)

[16] V. Neto, V. Rolim, R. Ferreira, V. Kovanović, D. Gašević, R. D. Lins and R. Lins, "Automated Analysis of Cognitive Presence in Online Discussions Written in Portuguese," in *European Conference on Technology Enhanced Learning*, 2018. (in English)

[17] V. Dornauer, M. Netzer, É. Kaczkó, L.-M. Norz and E. Ammenwerth, "Automatic Classification of Online Discussions and Other Learning Traces to Detect Cognitive Presence," *International Journal of Artificial Intelligence in Education,* vol. 34, p. 395–415, 2023. (in English)

[18] V. Neto, V. Rolim, A. Pinheiro, R. D. Lins, D. Gaševic and R. F. Mello, "Automatic Content Analysis of Online Discussions for Cognitive Presence: A Study of the Generalizability Across Educational Contexts," *IEEE Transactions on Learning Technologies,* vol. 14, no. 3, pp. 299-312, 2021. (in English)

[19] D. Gašević, O. Adesope, S. Joksimović and V. Kovanović, "Externally-facilitated regulation scaffolding and role assignment to develop cognitive presence in asynchronous online discussions," *The Internet and Higher Education,* vol. 24, pp. 53-65, 2015. (in English)

[20] Y. Hu, C. Donald and N. Giacaman, "Cross Validating a Rubric for Automatic Classification of Cognitive Presence in MOOC Discussions," *International Review of Research in Open and Distributed Learning,* vol. 23, no. 2, pp. 242–260, 2022. (in English)

[21] R. Al-Shabandar, A. J. Hussain, P. Liatsis and R. Keight, "Detecting At-Risk Students With Early Interventions Using Machine Learning Techniques," *IEEE Access ,* vol. 7, pp. 149464 - 149478, 2019. (in English)

[22] P. Leitner, M. Khalil та M. Ebner, «Learning Analytics in Higher Education—A Literature Review,» *Learning Analytics: Fundaments, Applications, and Trends,* p. 1–23, 2017. (in English)

[23] B. Chen, Y.-H. Chang, F. Ouyang and W. Zhou, "Fostering student engagement in online discussion through social learning analytics," *The Internet and Higher Education,* vol. 37, pp. 21-30, 2018. (in English)

[24] T. E. McKlin, Analyzing cognitive presence in online courses using an artificial neural network, Georgia State University, 2004. (in English)

[25] S. Corich, K. Hunt and L. Hunt, "Computerised Content Analysis for Measuring Critical Thinking within Discussion Forums," *Journal of e-Learning and Knowledge Society,* vol. 2, no. 1, 2006.

[26] V. Kovanovic, S. Joksimovic, D. Gasevic and M. Hatala, "Automated cognitive presence detection in online discussion transcripts," *CEUR Workshop Proceedings,* vol. 1137, 2014. (in English)

[27] Z. Water, V. Kovanović, K. Kitto and D. Gašević, "Structure Matters: Adoption of Structured Classification Approach in the Context of Cognitive Presence Classification," in *Information Retrieval Technology: 11th Asia Information Retrieval Societies Conference, AIRS 2015*, Brisbane, QLD, Australia, 2016. (in English)

[28] E. Farrow, J. Moore and D. Gašević, "Analysing discussion forum data: a replication study avoiding data contamination," in *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, 2019.

[29] G. Barbosa, R. Camelo, A. P. Cavalcanti, P. Miranda, R. F. Mello, V. Kovanović and D. Gašević, "Towards automatic cross-language classification of cognitive presence in online discussions," *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge,* pp. 605-614, 2020. (in English)

[30] H. Hayati, A. Chanaa, M. K. Idrissi and S. Bennani, "Doc2Vec & Naïve Bayes: Learners' Cognitive Presence Assessment through Asynchronous Online Discussion TQ Transcripts," *International Journal of Emerging Technologies in Learning,* vol. 14, no. 8, 2019. (in English)

Text od the article was accepted by Editorial Team 27.11.2024.

АВТОМАТИЧНЕ ОЦІНЮВАННЯ ЗНАНЬ, НАБУТИХ ТА ПОШИРЕНИХ ЗАСОБАМИ ДИСКУСІЙНИХ ФОРУМІВ НА ОСНОВІ МОДЕЛІ ЗАПИТІВ СПІЛЬНОТИ

П.А.Л. Надіша
бакалавр комп'ютерних наук з відзнакою, випускник
Школа комп'ютерних наук Університету Коломбо, м. Коломбо, Шрі-Ланка
ORCID ID 0009-0005-6329-9091
lakshaninadeesha99@gmail.com

Т.А. Вірасінгхе
кандидат наук з комп'ютерних та системних наук, старший викладач
Школа комп'ютерних наук Університету Коломбо, м. Коломбо, Шрі-Ланка
ORCID ID 0000-0002-7379-3916
taw@ucsc.cmb.ac.lk

В.Р.Н.С. Абервіра
бакалавр з інформаційних систем з відзнакою, викладач
Школа комп'ютерних наук Університету Коломбо, м. Коломбо, Шрі-Ланка
ORCID ID 0009-0009-9425-4450
rns@ucsc.cmb.ac.lk

Анотація. Протягом останніх двох десятиліть концепція «Спільнота запитів» (СЗ) широко використовується для оцінювання знань, отриманих і поширених через дискусійні онлайн-форуми. Компонент когнітивної присутності в системі СЗ допомагає виявити докази вдумливої реконструкції знань через осмислення під час навчання, заснованого на дослідженні. Виявлення та оцінювання цієї когнітивної присутності має важливе значення для оцінювання навчальних досягнень учнів за допомогою онлайн-дискусійних форумів. Враховуючи труднощі, пов'язані з ручним кодуванням та ідентифікацією когнітивних присутностей на дискусійних форумах, а також обмеження в існуючих методах автоматичної ідентифікації та оцінювання когнітивних присутностей, у цьому дослідженні була зроблена спроба розробити більш ефективний інструмент для ідентифікації та оцінювання когнітивних присутностей на онлайн-дискусійних форумах. У дослідженні використовувався конструктивний підхід. Методологія поєднувала класифікацію методом випадкового лісу (Random Forest-RF) з вилученням ознак TF-IDF та класифікацію методом опорних векторів (Support Vector Machine-SVM) із вбудовуванням Word2Vec. Класифікатор на основі правил, побудований на основі відображень індикаторів, збагатив процес класифікації. Для об'єднання результатів окремих класифікаторів було використано ансамблевий метод зважених голосів. Наш підхід було протестовано на двох наборах даних, що складалися з 781 повідомлення, які містили 47 592 слова. Ансамблевий метод продемонстрував помітну ефективність, досягнувши 69% точності в задачах класифікації. Це свідчить про надійність комбінованого підходу в підвищенні ефективності класифікації. Крім того, у дослідженні представлено модель підрахунку балів, яка обчислює індивідуальні бали студентів на основі категорій дописів, що дає змогу детально оцінити залученість та участь студентів.

Присвоюючи бали, що відображають внесок в обговорення, ця модель сприяє всебічному оцінюванню взаємодії в онлайн-навчанні. Наша робота є внеском у постійну дискусію про використання машинного навчання для когнітивного аналізу в середовищах онлайн-навчання, яка свідчить про важливість контекстно-специфічних методологій для розвитку практики освітнього оцінювання.

Ключові слова: когнітивна присутність; спільнота запитів; машинне навчання; обробка природної мови; класифікація