Leveraging Machine Learning to Predict Roles and Knowledge Construction Process

Abstract

Collaborative learning in asynchronous environments offers opportunities for deeper reflection and engagement but presents challenges due to lack of immediate communication and feedback. This study leverages Machine Learning (ML) models to predict participants' roles and knowledge construction processes within asynchronous discussions. Textual data from structured and unstructured group discussions was analyzed with roles such as Summarizer, Skeptic, and Theoretician being predicted with high accuracy. Knowledge construction processes, including Externalization, Elicitation, and Conflict Consensus, were also moderately predicted. Interestingly, contributions from unstructured groups (No Role) exhibited meaningful involvement across all knowledge construction processes, challenging assumptions about the necessity of predefined roles. These findings underscore the potential of ML to analyze and support collaborative learning, providing insights into how structured and unstructured participation influences group interactions. While the small dataset limits generalizability, the study still serves to validate the approach laying the groundwork for future research with larger datasets and more diverse contexts.

Introduction

Collaborative learning has long been recognized as a powerful approach for fostering critical thinking, problem-solving skills, and deep engagement with content (Dillenbourg, 2002; Scardamalia & Bereiter, 1994). Asynchronous discussions provide a unique platform for collaboration, enabling participants from diverse backgrounds and geographical locations to engage with one another at their own pace. This flexibility supports the deliberate crafting of responses and promotes reflective thinking, which is essential for constructing meaningful knowledge. However, the asynchronous nature of these discussions also introduces challenges, such as delays in communication and fragmented interactions, which can disrupt group cohesion and impede effective collaboration (Anderson et al., 2001; Garrison & Cleveland-Innes, 2005).

Despite these challenges, asynchronous collaborative learning environments hold tremendous potential when carefully designed. Research in computer-supported collaborative learning (CSCL) emphasizes the importance of structuring discussions and assigning roles to participants, such as moderator, summarizer, or skeptic, to foster balanced participation and sustain goal-oriented interaction (Strijbos & Fischer, 2007). Such strategies are particularly critical in asynchronous settings, where the absence of real-time feedback requires deliberate scaffolding to sustain engagement and ensure productive collaboration. By facilitating the co-construction of knowledge, these environments can transform into rich, reflective spaces that support both individual and collective learning.

In recent years, the integration of advanced technologies, particularly machine learning (ML), has opened new avenues for enhancing collaborative learning. ML offers the potential to analyze patterns of interaction, predict participant roles, and identify knowledge construction processes with remarkable precision. For example, supervised learning algorithms have been widely used to predict behaviors and outcomes in educational settings, enabling data-driven interventions that address diverse learner needs

(Zawacki-Richter et al., 2019; Romero & Ventura, 2013). Natural Language Processing (NLP), a subfield of ML, has proven particularly effective in analyzing textual data from asynchronous discussions, revealing themes, sentiments, and conversational dynamics that correlate with learning success (Roll & Wylie, 2016). These innovations provide educators with actionable insights, facilitating the design of more inclusive and effective collaborative learning environments.

This study explores the potential of ML to predict participant roles and knowledge construction processes in asynchronous discussions, aiming to address a gap in the literature. While prior research has demonstrated the benefits of role assignments and structured interactions in collaborative learning, no study has examined how these dynamics can be analyzed and supported using ML techniques. By leveraging ML to analyze textual data from structured and unstructured group discussions, this study seeks to uncover patterns of interaction that can inform instructional design and enhance the effectiveness of asynchronous learning environments. Through this investigation, the study contributes to the growing body of literature on CSCL and ML applications with empirical evidence on the predictive capabilities of ML models and offers practical insights into the design of collaborative learning environments. Ultimately, this research aims to bridge the gap between technological innovation and pedagogical practice, ensuring that ML-driven solutions are grounded in robust theoretical frameworks and contribute meaningfully to educational effectiveness.

Asynchronous collaborative discussion

Collaborative learning supported by technology emphasizes the importance of social interaction in constructing knowledge within meaningful contexts. Early research in computer-supported collaborative learning (CSCL) demonstrated how engaging students in problem-solving and reflective tasks through digital tools fosters deeper learning outcomes (Scardamalia & Bereiter, 1994). Over time, advancements in learning analytics, artificial intelligence, and adaptive technologies have refined these environments, offering increasingly personalized and effective support for collaborative learning.

While initial research in CSCL often centered on synchronous settings, where immediate feedback and dynamic exchanges are possible, the increasing demand for flexible learning options has shifted attention to asynchronous environments. These settings allow learners from diverse temporal and geographical contexts to engage at their own pace, fostering deliberate contributions and reflection. However, the asynchronous nature of these interactions can present challenges, including delayed feedback and fragmented participation, which may hinder group cohesion and sustained engagement (Garrison & Cleveland-Innes, 2005).

Asynchronous environments, despite their challenges, provide unique opportunities for deeper engagement. The time afforded by asynchronous discussions allows learners to craft thoughtful, well-considered responses, enhancing the depth of discussion and promoting a better understanding of complex topics (Garrison & Cleveland-Innes, 2005; Kovanovic et al., 2015). Digital tools such as discussion boards and shared workspaces facilitate sustained engagement and support the building of online communities (Dillenbourg, 1999; Scardamalia & Bereiter, 1994). Unlike traditional e-learning approaches focused on content delivery, asynchronous collaborative learning prioritizes interaction and co-construction of knowledge (Jeong et al., 2014). When thoughtfully designed, these environments can evolve into rich, reflective spaces for meaningful collaboration.

The effectiveness of asynchronous collaborative learning is closely tied to the structure of its tasks and the roles assigned to participants. Research underscores the value of assigning roles, such as facilitator or summarizer, which can help balance participation and sustain goal-oriented discussions. Prompts for critical thinking and scaffolding tools further encourage meaningful interaction and reflection

(Jeong et al., 2014). These strategies ensure that learners engage more deeply with both the content and each other, creating a dynamic learning environment where individual and collective growth thrive.

Therefore, asynchronous collaborative learning, informed by foundational principles of CSCL, can create powerful learning experiences through careful design and implementation. By combining structured tasks, role assignments, and digital tools, these environments not only address the challenges of asynchronous participation but also capitalize on its strengths, fostering deeper engagement, balanced participation, and meaningful knowledge construction.

Structured vs. Unstructured groups

The organization of group activities has a significant impact on collaborative dynamics and learning outcomes (Kapur, 2011). Structured groups are characterized by predefined roles and scripted interactions that guide learners toward specific educational goals (Dillenbourg, 2002). These groups often assign roles, such as moderator or summarizer, to distribute responsibilities evenly and promote active participation from all members (Strijbos & Fischer, 2007). By contrast, unstructured groups operate with minimal direction, allowing participants to interact more spontaneously. This flexibility can foster creativity and adaptability but may also lead to uneven participation and less efficient progress toward learning objectives (Kreijns et al., 2003; Le et al., 2018). Without structure, certain individuals may dominate discussions, while others contribute minimally, which can hinder overall group effectiveness (Phielix, Prins, & Kirschner, 2010).

Structured groups benefit from the clarity that roles and scripts provide, enhancing focus and efficiency in achieving goals (Rummel & Spada, 2005). However, strict roles can sometimes limit spontaneous contributions, which may be more common in unstructured groups where interactions are less confined (Stahl, Koschmann, & Suthers, 2006; Plass et al., 2015). Unstructured groups offer opportunities for exploratory learning and open-ended inquiry, but the variability in participation can impact group cohesion and goal achievement (Kreijns et al., 2002). Designing effective CSCL environments requires a balance between structure and flexibility to optimize both creativity and goal-oriented collaboration (Kirschner, Sweller, & Clark, 2006). Consequently, the choice between structured and unstructured group formats in CSCL should align with the specific learning objectives, task nature, and group composition, allowing educators to maximize engagement and learning outcomes (Jeong & Hmelo-Silver, 2015).

Roles and knowledge construction processes

Knowledge construction in collaborative learning is a dynamic process that involves the co-creation and negotiation of ideas among group members. Within this context, several processes are widely recognized, including externalization, elicitation, integration consensus, and conflict consensus (Fischer et al., 2002). Externalization serves as the foundation of collaborative learning by allowing individuals to articulate their personal knowledge and ideas, providing the raw material for group discussions. Elicitation, on the other hand, involves actively drawing out information or perspectives from others through questioning and prompting, fostering deeper engagement. Integration consensus occurs when the group combines diverse viewpoints into a coherent understanding, demonstrating their ability to synthesize and build upon individual contributions. Conflict consensus, though less common, represents a resolution of disagreements that arise through critical discussions, leading to refined and mutually accepted insights. Together, these processes highlight the interplay between individual and group dynamics in the co-construction of knowledge.

Structured roles serve as a critical mechanism for facilitating these processes, particularly in collaborative learning environments. By assigning specific responsibilities to participants, structured roles help ensure that all members contribute meaningfully to the group's objectives (Kollar et al., 2007; Vogel et al., 2017). The deliberate specification of roles creates a framework that guides participants in their interactions with one another and with the learning material, fostering both individual accountability and group cohesion. For example, roles such as Moderator, Summarizer, Skeptic, Source Searcher, and Theoretician have been identified in the literature as instrumental in supporting collaborative learning (Strijbos et al., 2005; De Wever et al., 2007). Each role contributes uniquely to the group's knowledge construction efforts. Moderators facilitate the flow of discussion and maintain focus on the task at hand. Summarizers synthesize group input, consolidating ideas into a cohesive narrative. Skeptics encourage critical evaluation by challenging assumptions and ideas, promoting deeper analysis. Source Searchers provide evidence and resources that ground the discussion in factual information, while Theoreticians bridge theoretical concepts with practical applications, enriching the group's understanding.

In asynchronous learning environments, the use of structured roles is particularly valuable due to the unique challenges posed by delayed communication and feedback. Without the immediacy of face-to-face interaction, learners may encounter difficulties in maintaining engagement and focus. Structured roles provide clarity and direction, enabling participants to navigate these challenges effectively. Research suggests that assigning roles not only ensures equitable participation but also enhances the depth and quality of group interactions. For instance, roles that emphasize critical engagement, such as Skeptic, have been shown to improve analytical thinking and conflict resolution skills (Schellens et al., 2005). Similarly, roles like Summarizer and Theoretician support the integration of knowledge, facilitating a deeper understanding of the material and fostering collaboration among group members (De Wever et al., 2007; Weinberger, 2010).

Collaboration scripts further enhance the impact of structured roles by scaffolding the knowledge construction process. These scripts outline specific tasks and sequences of activities, prompting learners to engage with the content and each other in a meaningful way (Kollar et al., 2007; Vogel et al., 2017). In asynchronous settings, where learners often work independently over extended periods, collaboration scripts help emulate the immediacy and interactivity of synchronous discussions. They guide participants toward constructive engagement, ensuring that individual contributions align with the group's collective goals. The scripts are flexible in design, allowing educators to tailor them to the needs of the learners and the requirements of the task. This adaptability is particularly important in asynchronous environments, where the absence of real-time interaction necessitates a carefully structured approach to sustaining engagement and collaboration.

The examination of roles and knowledge construction processes provides valuable insights into the mechanisms that support effective collaborative learning. By understanding how roles and processes interact, educators can design interventions that foster meaningful participation and facilitate the co-construction of knowledge. Structured roles offer a practical strategy for addressing the challenges of asynchronous collaboration, ensuring that all participants are actively engaged and contributing to the group's success. Through thoughtful implementation of these roles and supporting frameworks, collaborative learning environments can be transformed into dynamic spaces of interaction, reflection, and intellectual growth.

Machine learning in educational settings

Machine Learning (ML) applications are revolutionizing educational practices, enabling more nuanced understandings of learning processes. Zawacki-Richter et al. (2019), in their systematic review, analyzed the landscape of ML and artificial intelligence applications in higher education, showcasing their potential for improving outcomes such as academic performance, engagement, and retention. Their work highlights how ML methodologies, particularly supervised learning, are being used to train models that predict student behaviors and outcomes. Techniques such as classification, regression, and clustering are widely employed. For example, classification models can predict various outcomes in educational settings, and clustering algorithms are widely used to group students based on behavioral patterns, facilitating data-driven interventions (Romero & Ventura, 2013). These approaches can offer important insights and support the creation of tailored interventions to address diverse learner needs.

Natural Language Processing (NLP), another prominent application of ML in education, is frequently utilized to analyze textual data from student interactions. NLP techniques enable educators to identify themes, sentiments, and conversational patterns that correlate with learning success (Roll & Wylie, 2016). For instance, sentiment analysis can uncover students' emotional engagement, while thematic analysis of discussion forums can highlight the depth of knowledge construction in collaborative activities. These insights allow educators to refine teaching strategies and foster meaningful interactions among learners.

While Zawacki-Richter et al. (2019) emphasize the growing prevalence of ML applications in education, their systematic review also points to an important gap: many studies focus on technological innovations without fully considering pedagogical integration. Although ML tools like adaptive learning systems and intelligent tutoring systems demonstrate potential for personalization and efficiency, the lack of connection to robust theoretical frameworks limits their impact on teaching and learning practices. This critique underscores the need for future research to prioritize the alignment of ML applications with meaningful pedagogical theories, ensuring these innovations contribute to educational effectiveness beyond technical advancements.

Purpose of the study

The primary purpose of this study is to explore the potential of ML to predict participant roles and knowledge construction processes in asynchronous discussions as ~~there was~~ no study has examined how these dynamics can be analyzed and supported using ML techniques.~~ and u~~Ultimately, this research aims to bridge the gap between technological innovation and pedagogical practice, ensuring that ML-driven solutions are grounded in robust theoretical frameworks and contribute meaningfully to educational effectiveness.

Research Questions are as follows:

- How accurately can ML models predict roles and knowledge construction processes in collaborative learning environments, and what patterns emerge from these predictions?
- How do participant roles and knowledge construction processes interact, and what insights can these patterns provide for improving collaborative learning?

Methods

Data collection

The data was collected from a series of asynchronous discussions that were structured around specific tasks designed to promote knowledge construction through assigned roles. The dataset includes textual interactions from these discussions, annotated for both role identification and knowledge construction processes.

Participants

An online introductory course in Information Systems was chosen due to its emphasis on discussion as a core component of academic achievement, requiring complex levels of thinking to analyze different information systems and their capabilities. Approximately 400 students participated over two semesters.

From this cohort, 103 undergraduate students—45 from 10 unstructured groups and 58 from 14 structured groups—gave their consent to participate in the study. Participants were randomly allocated to either structured or unstructured groups, ranging from four to six members. During the eighth week, instructors presented all groups with a problem to address over a three-week discussion period. In structured groups, students selected from four predefined roles—moderator, summarizer, theoretician, and source searcher—based on extant literature. Their task was to collaboratively solve the problem within their roles. In contrast, unstructured groups engaged in the same problem-solving task without role assignments. All groups were required to submit their final collective and individual reports by the twelfth week.

Initial coding and data preparation

Before implementing ML techniques, content analysis was employed to examine the depth and quality of discourse from asynchronous discussion. The analysis began by preparing the data: transcripts were extracted, anonymized, and segmented into "meaningful units" that captured distinct ideas or arguments, ensuring a more focused examination of interactions.

A coding framework was adapted from Fisher et al's (2002) knowledge construction categories, including Externalization, Elicitation, Conflict-Oriented Consensus, and Integration-Oriented Consensus. During the pilot study, a new category, Quick Consensus, emerged to account for rapid agreements that lacked deeper deliberation, reflecting either efficient communication or superficial alignment. For example, a participant's response such as, "I agree with what both of you said; our answers together make the report," exemplified this pattern of interaction. Introducing Quick Consensus as a distinct category allowed for a clearer differentiation between rapid alignment and more integrative or conflict-driven consensus-building processes.

Table 1. Quotes of knowledge construction process in structured and unstructured group discussions

Knowledge Construction Processes	Definition	Example Quotes
Externalization	The process of making implicit knowledge explicit by articulating thoughts, ideas, or concepts.	“I would recommend that the company perform daily checks to ensure that their database remains secure. Prevention is the best strategy." (Jen_Unst) "The firm could have installed an email client extension that team members can use to report suspected phishing attacks." (Matt_St)
Elicitation	The process of drawing out information or ideas from group members through questions or prompts.	“Should we also have a check and balance system to review one another's work/sources?" (Matt_St) "How would MFA prevent a malicious email in your mind?" (Matt_St)
Quick Consensus	The process where participants rapidly agree with minimal deliberation or discussion.	"That sounds like a reasonable thought process to use lessons 7, 9, and 10 instead of lesson 8." (Jen_Unst) "I honestly think the current rough draft sounds perfect and think that it is ready for final submission." (Jen_Unst)
Integration-oriented Consensus	The process of synthesizing diverse perspectives into a coherent understanding or solution.	"All of those points can be included in how I would reduce the risk if I was running a business." (Sophie_St) "A firewall and DMZ would have made a difference, as well as human training." (Jen_Unst)
Conflict-oriented Consensus	The process of reaching agreement through the exploration of conflicting ideas or perspectives, often leading to a more refined understanding.	"I do not think this question is asking us to address preventative actions." (Jen_Unst) "...but if they only shut down the affected servers and left the servers that were not affected by the attack, wouldn't the hackers be able to continue to hack into servers that were not affected yet?" (Sophie_St)

The data were manually coded to indicate the presence or absence of each knowledge construction process in the segmented units. This binary coding approach enabled a systematic analysis of structured and unstructured groups separately, offering comparative insights. To ensure consistency in applying the coding criteria, intracoder reliability was assessed through re-coding a subset of the data at different time points.

The frequencies of coding categories were calculated and visualized to compare structured and unstructured groups. In structured groups, the analysis also examined the distribution of categories across assigned roles, such as Moderator, Summarizer, and Theoretician, to explore how roles influenced knowledge construction. This preparatory analysis established a robust foundation for the application of ML techniques, ensuring that patterns of interaction were well-defined and aligned with the theoretical constructs underpinning the study.

Data processing for ML

The processing of data involved transforming the textual data into numerical features suitable for ML models. The study first used TF-IDF (Term Frequency-Inverse Document Frequency) to convert the text into feature vectors. This method quantifies the importance of each term within the individual documents and across the corpus. Additionally, Part of Speech (PoS) tagging was used to capture syntactic information, providing insights into the roles of different words in sentences.

The dataset consisted of 236 total data points. In each fold of 5-fold cross-validation, 189 data points (80%) were used for training, while 27 data points (20%) were reserved for testing.

The base model employed was logistic regression, selected for its simplicity and efficiency in handling both binary and multiclass classification tasks and appropriate with a categorical variable. This model was trained using the preprocessed feature vectors, with the target variable encompassing roles and knowledge construction processes annotated in the dataset. The performance of the logistic regression model was evaluated using standard classification metrics: Precision, Recall, and F1-Score. These metrics provide a comprehensive view of the model's accuracy, considering the accuracy of predictions and the balance between precision and recall.

- Precision measures the proportion of true positive predictions among all positive predictions, indicating the model's ability to avoid false positives.
- Recall measures the proportion of true positive predictions among all actual positives, indicating the model's ability to capture all relevant instances.
- F1-Score is the harmonic mean of precision and recall, providing a single metric that balances both aspects.

Findings

This section presents the results of applying machine learning (ML) models to predict participant roles and knowledge construction processes in collaborative learning environments. The analysis examined the distribution of contributions across roles and knowledge construction processes, evaluated the predictive performance of ML models for both roles and processes, and explored the interaction between them. These findings highlight the potential of ML to uncover patterns in asynchronous discussions and contribute to the design of a more effective collaborative learning environment. The results are organized into three key areas: (1) the distribution of contributions across roles and knowledge construction processes, (2) the predictive accuracy of ML models, and (3) the interaction between them.

This following table 1 represents an overview of how actual contributions are distributed across roles and knowledge construction processes.

Table 1. actual contributions across roles and knowledge construction processes

Role/Theme	Externalization	Elicitation	Quick Consensus	Integration Consensus	Conflict Consensus	Total
No Role	26	11	16	11	5	69
Moderator	12	11	3	6	2	34
Summarizer	11	5	0	9	0	25
Skeptic	9	3	2	4	2	20
Source-searcher	9	0	0	5	0	14
Theoretician	12	0	0	7	1	20
Total	79	30	21	42	10	182

Externalization is the most frequent process (79), while Conflict Consensus is the least represented (10) reflecting its rarity in the dataset. No role accounts for the highest number of contribution (69), particularly in Externalization (26) and Quick Consensus (16), representing the impact of participants without explicit roles. Moderators contribute across processes (34), though their presence in Conflict Consensus (2) and Quick Consensus (3) is limited, consistent with their focus on facilitation. Summarizer (25) is most active in Externalization (11) and Integration Consensus (9), while Skeptic (20) participates broadly but rarely in Quick Consensus (2) and Conflict Consensus (2). Source Searcher (14) and Theoretician (20) concentrate on Externalization (9 and 12, respectively) and Integration Consensus (5 and 7), aligning with their task-focused roles. Overall, the dataset reveals an imbalance with Externalization dominating and Conflict Consensus underrepresented. Roles like Summarizer and Theoretician show somewhat specialized contributions, whereas No Role and Moderator roles span more processes. This variability could impact the predictive performance of machine learning models, particularly less frequent processes.

Predicting Roles in Collaborative Learning Environments

The ML models demonstrated varying degrees of accuracy in predicting the roles assigned within collaborative learning environments (Table 2).

Table 2. Role prediction accuracy

Role	Precision	Recall	F1-Score
No Role	0.79	0.79	0.79
Moderator	0.76	0.78	0.77
Summarizer	0.88	0.88	0.86
Skeptic	0.90	0.91	0.90
Source Searcher	0.90	0.92	0.90
Theoretician	0.92	0.92	0.91

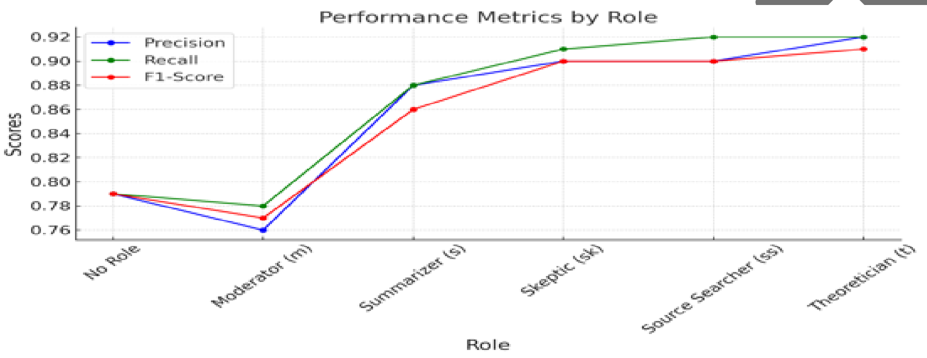

Figure 1. Performance metrics by role

Notably, the roles of Summarizer, Skeptic, Source Searcher, and Theoretician were predicted with high accuracy, as evidenced by precision, recall, and F1-scores above 0.86. Specifically, the role of Theoretician had the highest predictive accuracy with a precision and recall of 0.92, resulting in an F1-score of 0.91. However, the roles of Moderator and participants with No Role were predicted with lower accuracy. The Moderator role had an F1-score of 0.77, while the No Role category had an F1-score of 0.79. These findings suggest that these two categories may have overlapping features with other roles or are less distinct in their language use, which complicates accurate classification by the model.

Predicting Knowledge Construction Processes

The ML models also varied in their ability to predict different knowledge construction processes (Table 2).

Table 3. Knowledge Construction Prediction Accuracy

Theme	Precision	Recall	F1-Score
Externalization	0.81	0.81	0.81
Elicitation	0.83	0.85	0.84
Quick Consensus	0.79	0.80	0.79
Integration Consensus	0.73	0.72	0.71
Conflict Consensus	0.93	0.94	0.93

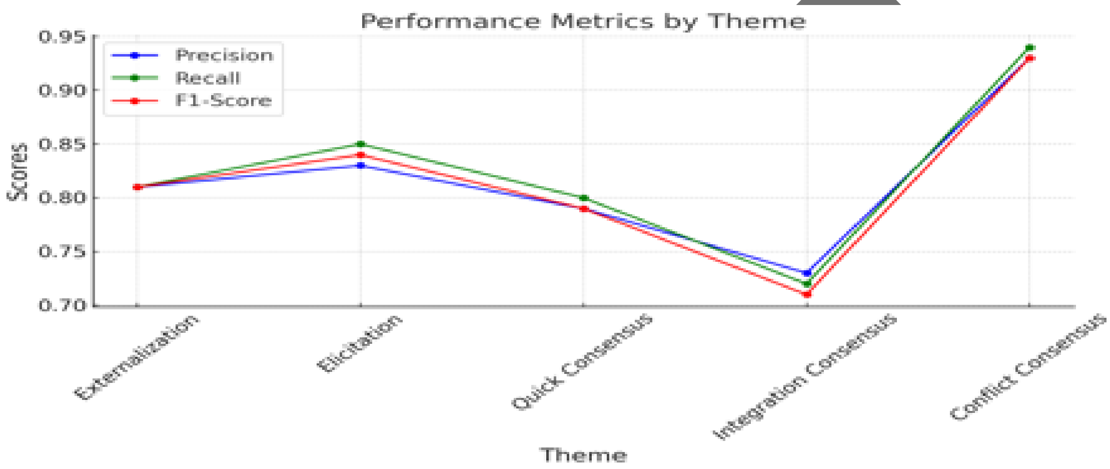

Figure2. Performance metrics by knowledge construction themes

The model predicted Conflict Consensus with the highest accuracy, achieving an F1-score of 0.93. Elicitation and Externalization were also predicted with relatively high accuracy, with F1-scores of 0.84 and 0.81, respectively. On the other hand, Quick Consensus and Integration Consensus were predicted with moderate accuracy, with F1-scores of 0.79 and 0.71, respectively. These results indicate that certain knowledge construction processes, such as Conflict Consensus, are more distinguishable and easier for the model to predict accurately compared to others.

Predicting Participant Roles in Knowledge Construction Processes

The interaction between participant roles and knowledge construction processes revealed further insights (Table 3).

Table 4: Role and Knowledge Construction Interaction

Role	Externalization	Elicitation	Quick Consensus	Integration Consensus	Conflict Consensus
No Role	0.70	0.18	0.35	0.41	0.00
Moderator	0.65	0.00	0.30	0.44	0.00
Summarizer	0.74	0.00	0.73	0.55	N/A
Skeptic	0.62	0.00	0.57	0.57	N/A
Source Searcher	0.88	N/A	0.75	0.36	N/A
Theoretician	0.84	N/A	0.67	0.53	0.33

For example, the model struggled to predict Elicitation for the No Role group, with an F1-score of 0.17. In contrast, Externalization was relatively well-predicted for the Summarizer role, with an F1-score of 0.74. Roles such as Source Searcher and Theoretician had higher predictive accuracy for Externalization, with F1-scores of 0.87 and 0.84, respectively. The role of Skeptic was moderately predicted for Quick Consensus and Integration Consensus, with F1-scores of 0.57 and 0.57, respectively, but the model could not predict Elicitation or Conflict Consensus for this role. The No Role category displayed moderate accuracy for externalization (F1-score: 0.18) and Conflict Consensus (F1-score: 0.00). These findings represent that the ML models are better at predicting certain knowledge construction processes for specific roles. The higher accuracy for roles like Source Searcher and Theoretician in predicting Externalization implies that these roles have more distinct interaction patterns that the model can recognize.

Discussion

This study explored the accuracy of ML models in predicting roles and knowledge construction processes within collaborative knowledge construction processes in collaborative learning environments, as well as the interaction between these roles and processes. The findings offer valuable insights into the predictive capabilities of ML and the dynamics of role-based interactions in asynchronous discussion addressing two research questions.

Accuracy of ML predictions and emerging patterns

The first research question explored the predictive accuracy of ML models for roles and knowledge construction processes in asynchronous discussion. The results presented that ML models were highly effective in predicting well-defined roles, such as Theoretician, Skeptics, Source Searcher, and Summarizer, achieving F1-scores above 0.86. These roles were well characterized by distinct linguistic patterns and contributions, making them easier for the model to capture and classify. For example, the high accuracy for the Theoretician role (F1 scores: 0.84) suggests a clear alignment with conceptual contributions such as articulating abstract concepts or offering deeper insights into knowledge, which generate identifiable textual patterns. Similarly, Source Searchers, with a focus on factual contributions (e.g., Externalization, F1-score: 0.88), provided consistent markers that the model could learn effectively.

In contrast, the model's relative difficulties in predicting moderator and No role categories suggest struggles related to overlapping features or the less distinct nature of these roles. For moderators, this may reflect the broad range of responsibilities they perform, such as facilitating discussions and integrating ideas, which do not always produce similar linguistic cues. Meanwhile, the relatively lower accuracy for No Role points to a deeper complexity in understanding unstructured group individuals' contributions. This relatively low accuracy for No Role could indicate that students in unstructured settings behave under implicit roles that are subtly agreed upon by the group. For example, a student may take on facilitating behaviors that resemble defined roles but are not explicitly assigned. Also, posts labeled as No Role might contain diverse contributions that lack clear linguistic or structural markers, making them harder for the model to classify accurately. Despite this, the ability of this model to predict No Role contributions with 79% accuracy remains an impressive outcome~~.~~ This finding underscores the model's capacity to infer the absence of a role as well as other explicit roles in a good amount of the posts, emphasizing its utility in analyzing group dynamics even in unstructured settings.

For knowledge construction processes, the model achieved high accuracy for conflict consensus (F1-score:0.93), despite its low representation in the dataset (10 total contributions) which reflects the distinctive nature of disagreement and resolution patterns in collaborative interactions and these are often marked by clear textual cues such as argumentative language or negotiated agreements. Similarly, processes such as Elicitation (F1-score:0.84) and Externalization (F1-score: 0.81) were well predicted, suggesting that these processes also involved clear textual markers, such as targeted questions or clear statements. However, the moderate accuracy for Integration Consensus (F1-score: 0.71) and Quick Consensus (F1-score: 0.79) presents the intricacy of these processes. Quick Consensus, for example, may reflect rapid alignment among group members, which can be efficient yet lacks the depth associated with other processes, making it harder for the model to distinguish.

The results highlight the potential of ML in identifying structured interaction patterns while also pointing to the challenges of analyzing ambiguous or implicit behaviors. Roles with well-defined tasks and contributions are relatively easier to predict, even with limited instances. Conversely, roles like Moderator and No Role, which involve diverse or unstructured behaviors, present challenges for classification due to overlapping features and implicit group dynamics. The findings emphasize the importance of considering dataset characteristics, such as the frequency and clarity of contributions, when evaluating ML performance for complex social learning processes.

Interaction between roles and knowledge construction processes

The second research question addresses the interplay between participant roles and knowledge construction processes. The interaction between participant roles and knowledge construction processes provides important insights into how specific roles align with particular processes and the performance of the ML models varied across roles and processes, reflecting differences in contribution patterns (Table 1) and the intricacy of processes (Table 4).

Roles such as Source Searcher and Theoretician demonstrated high accuracy for Externalization, with F1-scores of 0.88 and 0.84, respectively. These scores reflect the well-defined and task-specific contributions of these roles, which align closely with the process of articulating task-relevant knowledge. However, these roles show N/A values for processes like Elicitation and Quick Consensus, indicating that no contributions were recorded for these combinations in the dataset. This absence suggests a strong role specialization where these participants focus on specific processes, leaving other tasks outside their scope.

Roles like Moderator and Skeptic displayed moderate to low accuracy, reflecting their more variable contributions. For instance, Moderators contributed across processes, including Externalization (12) and Integration Consensus (6), but their facilitative role resulted in lower predictive accuracy for tasks like Quick Consensus (F1-score: 0.30) and Conflict Consensus (F1-score: 0.00). For Skeptics, predictions for Quick Consensus and Integration Consensus achieved moderate accuracy (both F1-scores: 0.57), but predictions failed for Elicitation and Conflict Consensus (both F1-scores: 0.00). These F1-scores of 0 indicate that while there were some contributions for these processes in the dataset, the model struggled to identify discernible patterns due to their low frequency or ambiguous nature.

The No Role category, with the largest number of contributions (69 total), presented a mix of outcomes. Moderate accuracy was achieved for Externalization (F1-score: 0.70) and Quick Consensus (F1-score: 0.35), reflecting the diverse and unstructured nature of No Role contributions. However, for complex processes like Elicitation (F1-score: 0.18) and Conflict Consensus (F1-score: 0.00), the model struggled, as these contributions lacked consistent patterns for effective classification.

The N/A values in the results highlight the important gaps in the dataset where certain role-process combinations were entirely absent. For example, Source Searcher and Theoretician contributed to Externalization and Integration Consensus but did not engage in Elicitation or Quick Consensus, resulting in N/A values for these combinations. Similarly, Summarizers and Skeptics did not contribute to Conflict Consensus, reflecting their limited involvement in conflict resolution. The imbalance in the dataset also influenced model performance. Externalization, as the most frequent process (79), resulted in higher predictive accuracy, particularly for specialized roles like Source Searcher and Theoretician. In contrast, rare processes like Conflict Consensus (10) led to poor or no predictions across roles due to insufficient data for the model to learn patterns effectively.

Implications

This study establishes a foundational step toward understanding collaborative learning dynamics using machine learning (ML). Although the dataset is small, it demonstrates the feasibility of leveraging ML to predict roles and knowledge construction processes in asynchronous discussions. The findings serve as proof of concept and pave the way for future research to build on these insights. Expanding datasets to include more diverse contexts, courses, and group interactions could enable the development of more robust ML models that generalize better across various educational settings.

First, regarding instructional design, the study highlights the potential for developing role-specific facilitation strategies. Roles like Source Searcher and Theoretician, which are more structured and predictable, can be supported through clearer task expectations. In contrast, less structured roles like Moderator and Skeptic may benefit from scaffolding to ensure productive engagement. Tailored facilitation strategies can help maximize the contributions of all participants and enhance group dynamics. The study also reveals that processes like Conflict Consensus and roles like Skeptic are underrepresented in the current dataset, which limits the ML model’s ability to effectively learn patterns despite achieving high accuracy when data is available. Addressing this imbalance requires deliberate strategies to ensure all roles and processes are adequately represented in future datasets. Providing clearer guidance and more concrete directions for students assigned to less structured roles, such as Skeptic, may help standardize contributions and improve their engagement.

The study underscores the importance of unstructured group participation in collaborative learning. Unlike structured settings where predefined roles like Moderator or Summarizer guide group interaction, unstructured settings without predefined interventions rely on participants taking initiative organically. The consistent involvement of No Role participants across all knowledge construction processes highlights the capacity of unstructured groups to sustain productive collaboration when groups are effectively formed (Authors, in press). Although the No Role category is moderately predictable by ML models, it demonstrates significant contributions to group discussions, especially in processes like Externalization and Quick Consensus. This finding challenges the assumption that structured interventions are always necessary, emphasizing the potential of unstructured collaboration to drive meaningful knowledge construction.

Finally, while the dataset is not sufficient to build fully automated feedback systems, the findings offer a starting point for designing incremental feedback tools. Automated tools for educators could use ML predictions to highlight group dynamics or identify areas requiring intervention (Roll & Wylie, 2016). For example, these systems could detect patterns of Externalization or Conflict Consensus and provide instructors with actionable insights to guide collaborative learning more effectively. Such tools could act as semi-automated systems, complementing human facilitation rather than replacing it, thereby enhancing the instructor’s ability to support collaborative learning environments.

Conclusion

The study provides a foundational exploration of the potential of ML in understanding collaborative learning dynamics within online discussion. By focusing on the prediction of roles and knowledge construction processes, the findings offer valuable insights that contribute to both research and practice in collaborative learning and learning analytics.

The study contributes to several aspects. First, it emphasizes the potential for developing role-specific facilitation strategies. Roles such as Source Searcher and Theoretician were easier for the ML model to predict due to their task focused contributions. This finding underscores the importance of designing clearer task expectations for structured roles while recognizing the need for scaffolding and support for less predictable roles such as Moderator and Skeptics for their productive engagement. Second, the study emphasizes the significance of unstructured participation in collaborative learning. The consistent involvement of No Role participants across all knowledge construction processes demonstrates that unstructured groups, even without predefined interventions, can sustain productive collaboration. Finally, the study establishes proof for the use of ML in analyzing collaborative learning behaviors. The ability to achieve moderate to high predictive accuracy for many roles and processes demonstrates the feasibility of ML as a tool for understanding and improving collaborative learning dynamics.

However, the study has some limitations. The primary constraint lies in the small dataset, which restricts the generalizability of the findings that can commonly be achievable in ML studies and limits the model's ability to fully capture the complexities of collaborative learning behaviors. Underrepresented roles like Skeptics and processes like Conflict Consensus, present the needs for additional data. Also, the study's focus on single course and learning environment may not fully reflect the diversity of collaborative learning practices in other contexts or disciplines.

Future research should prioritize the collection of larger and more diverse datasets. Expanding the score to include more courses, disciplines and group interactions will enable the development of more robust ML models that generalize better across different educational contexts. A larger dataset would also allow for a more balanced representation of all roles and processes, ensuring limited contributions. This approach would enhance the reliability and applicability of ML models and provide deeper insights into the CSCL.

In conclusion, this study demonstrates the feasibility and potential of using ML to analyze collaborative learning dynamics, offering insights that can inform the design of the more effective and inclusive learning environments. While the findings are exploratory, they provide a foundation for future research, structuring the way for advancements in collaborative learning. By addressing the limitation through larger datasets, future studies can build on this work to create adaptive tools that further enhance collaborative learning experiences.

References

Authors 1&2. (in press). Group formation with an algorithm. International Journal of Educational Technology in Higher Education

Anderson, L. W., & Krathwohl, D. R. (2001). A Taxonomy for Learning, Teaching and Assessing: A Revision of Bloom’s Taxonomy of Educational Objectives: Complete Edition. New York: Longman.

De Wever, B., Schellens, T., Valcke, M., & Van Keer, H. (2006). Content analysis schemes to analyze transcripts of online asynchronous discussion groups: A review. Computers & education, 46(1), 6-28.

De Wever, B., Van Keer, H., Schellens, T., & Valcke, M. (2010). Structuring asynchronous discussion groups: Comparing scripting by assigning roles with regulation by cross-age peer tutors. Learning and Instruction, 20 (5), 349-360.

Dillenbourg, P. (2002). Over-scripting CSCL: The risks of blending collaborative learning with instructional design. In P. A. Kirschner (Ed.), Three worlds of CSCL. Can we support CSCL (pp. 61-91). Heerlen: Open Universiteit Nederland.

Fischer, F., Bruhn, J., Gräsel, C., & Mandl, H. (2002). Fostering collaborative knowledge construction with visualization tools. Learning and Instruction, 12(2), 213-232.

Garrison, D. R., & Cleveland-Innes, M. (2005). Facilitating cognitive presence in online learning: Interaction is not enough. American Journal of Distance Education, 19, 133-148. http://dx.doi.org/10.1207/s15389286ajde1903_2

Jeong, H., Hmelo-Silver, C. E., & Yu, Y. (2014). An examination of CSCL methodological practices and the influence of theoretical frameworks 2005–2009. International Journal of Computer-Supported Collaborative Learning, 9(3), 305–334. https://doi.org/10.1007/s11412-014-9198-3

Jeong, H., & Hmelo-Silver, C. E. (2016). Seven Affordances of Computer-Supported Collaborative Learning: How to Support Collaborative Learning? How Can Technologies Help? Educational Psychologist, 51, 247-265. https://doi.org/10.1080/00461520.2016.1158654

Kreijns, C. J., Kirschner, P. A., & Jochems, W. M. G. (2002). The sociability of computer-supported collaborative learning environments. Journal of Educational Technology & Society, 5 (1), 8-22.

Kirschner, P.A., Sweller, J., & Clark, R.E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist discovery, problem-based, experiential, and inquiry-based teaching. Educational Psychologist 41(2), 75-86.

Kollar, I., Fischer, F., & Slotta, J. D. (2007). Internal and external scripts in computer-supported collaborative inquiry learning. Learning and Instruction, 17(6), 708-721.

Kovanović, V., Gašević, D., Joksimović, S., Hatala, M., & Adesope, O. (2015). Analytics of communities of inquiry: Effects of learning technology use on cognitive presence in asynchronous online discussions. The Internet and Higher Education, 27, 74-89.

Romero, C., & Ventura, S. (2013). Data mining in education. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 3(1), 12-27.

Roll, I., & Wylie, R. (2016). Evolution and revolution in artificial intelligence in education. International Journal of Artificial Intelligence in Education, 26(2), 582-599.

Scardamalia, M., & Bereiter, C. (1994). Computer support for knowledge-building communities. The journal of the learning sciences, 3(3), 265-283.

Schellens, T., & Valcke, M. (2005). Collaborative learning in asynchronous discussion groups: What about the impact on cognitive processing? Computers in Human Behavior, 21, 957–975.

Stahl, G. (2002, January). Contributions to a theoretical framework for CSCL. In CSCL (Vol. 2, pp. 62-71).

Strijbos, J. W., & Fischer, F. (2007). Methodological challenges for collaborative learning research. Learning and Instruction, 17(4), 389-393.

Strijbos, J. W., De Laat, M. F., Martens, R. L., & Jochems, W. M. G. (2005). Functional versus spontaneous roles during CSCL. In T. Koschmann, D. Suthers, & T. W. Chan (Eds.), Computer supported collaborative learning 2005: The next 10 years! (pp 647-656). Mahwah, NJ: Lawrence Erlbaum Associates.

Vogel, F., Wecker, C., Kollar, I., & Fischer, F. (2017). Socio-cognitive scaffolding with computer-supported collaboration scripts: A meta-analysis. Educational Psychology Review, 29, 477-511.

Weinberger, A., Stegmann, K., & Fischer, F. (2010). Learning to argue online: Scripted groups surpass individuals (unscripted groups do not). Computers in Human behavior, 26(4), 506-515.

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – where are the educators? International Journal of Educational Technology in Higher Education, 16(1), 39.