Minds and Machines Unite: Deciphering Social and Cognitive Dynamics in Collaborative Problem Solving with AI

Mohammad Amin Samadi
masamadi@uci.edu
University of California, Irvine
Irvine, California, USA

Spencer JaQuay
sjaquay@uci.edu
University of California, Irvine
Irvine, California, USA

Yiwen Lin
yiwenl21@uci.edu
University of California, Irvine
Irvine, California, USA

Elham Tajik
University of Rochester
Rochester, New York, USA

Seehee Park
University of California, Irvine
Irvine, California, USA

Nia Nixon
University of California, Irvine
Irvine, California, USA

ABSTRACT

We investigated the feasibility of automating the modeling of collaborative problem-solving skills encompassing both social and cognitive aspects. Leveraging a diverse array of cutting-edge techniques, including machine learning, deep learning, and large language models, we embarked on the classification of qualitatively coded interactions within groups. These groups were composed of four undergraduate students, each randomly assigned to tackle a decision-making challenge. Our dataset comprises contributions from 514 participants distributed across 129 groups. Employing a suite of prominent machine learning methods such as Random Forest, Support Vector Machines, Naive Bayes, Recurrent and Convolutional Neural Networks, BERT, and GPT-2 language models, we undertook the intricate task of classifying peer interactions. Notably, we introduced a novel task-based train-test split methodology, allowing us to assess classification performance independently of task-related context. This research carries significant implications for the learning analytics field by demonstrating the potential for automated modeling of collaborative problem-solving skills, offering new avenues for understanding and enhancing group learning dynamics.

CCS CONCEPTS

• **Applied computing** → **Collaborative learning**; • **Human-centered computing** → **Empirical studies in collaborative and social computing**.

KEYWORDS

CPS, Artificial Intelligence, Machine Learning, Learning Analytics, NLP

ACM Reference Format:
Mohammad Amin Samadi, Spencer JaQuay, Yiwen Lin, Elham Tajik, Seehee Park, and Nia Nixon. 2024. Minds and Machines Unite: Deciphering Social and Cognitive Dynamics in Collaborative Problem Solving with AI. In *The 14th Learning Analytics and Knowledge Conference (LAK '24), March 18–22, 2024, Kyoto, Japan.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3636555.3636922

LAK '24, March 18–22, 2024, Kyoto, Japan

ACM ISBN 979-8-4007-1618-8/24/03.
https://doi.org/10.1145/3636555.3636922

1 INTRODUCTION

Educational paradigms increasingly highlight the importance of teamwork, recognizing the complexity inherent in modern collaborative activities [3, 29]. Within the realms of academia and professional sectors, collaborations are often complicated by challenges such as managing culturally diverse teams [54] and remote workforces [48]. To meet these multifaceted demands, there's a pressing need for educators to provide rigorous training that transfers into the workplace [23].

Collaborative Problem-Solving (CPS) emerges as a cornerstone training method in educational settings, directly addressing the challenges of teamwork in the 21st century – be it in the workplace, academic environments, or broader sectors such as the military [7, 9, 18, 24, 29, 30, 47]. According to the Programme for International Student Assessment (PISA), collaborative problem-solving (CPS) refers to an individual's ability to participate in a process where multiple agents work together to find a solution by sharing their knowledge, skills, and efforts. The process involves effective engagement and coordination among the agents to reach a common goal [41]. Frameworks for CPS typically integrate problem-solving processes and collaborative strategies, underscoring aspects like shared understanding and team organization [50]. Given its comprehensive approach, CPS prepares future leaders to address global challenges, necessitating both hands-on training and innovative tools for evaluating collaborative communication [29].

With the shift towards a digitally connected and collaborative world, there's a heightened need for tools adept at capturing and refining dialogues, specifically in educational settings [52]. Given that communication is pivotal for collaboration, placing emphasis on language becomes a pathway to evaluating the learning process. Discourse analysis is commonly used in educational research to unpack individual learner's social, cognitive, and affective states [8, 20, 21, 26–28, 46]. In collaborative learning, analyzing communication between team members has been effective in providing deeper insights into collaborative skills exhibited by individuals [1, 2], as well as group processes such as information sharing, coordination, negotiation, monitoring progress, and so on [17, 29, 45]. Coding utterances and assigning utterances a label based on the coding framework allow us to identify nuanced collaborative processes. However, as the number of groups or the number of people

in groups increases, the amount of effort, time, and resources required in manual coding can quickly become challenging. While human coding still sets the gold standards, it is no longer viable for processing the entirety of an exponentially larger corpus of student data that is being collected.

Increasingly, computational methods are employed to process educational data at scale [11, 15–17, 19, 35]. Previous studies have attempted to classify CPS processes based on learner discourse, and researchers continue to seek approaches that improve the performance of automatic prediction. For instance, [31] leveraged Conditional Random Fields (CRF), a sequential dependent classifier, to automatically classify utterances using four CPS skills: sharing ideas, negotiating ideas, regulating, and maintaining communication. CRF was found to outperform other non-sequential methods used in their research. Another study [43] found that using a transfer learning approach for Natural Language Processing (NLP), Bidirectional Encoder Representations from Transformers (BERT) achieved improved performance compared to standard classifiers for modeling CPS language. They further suggest the necessity of considering the context of an utterance (what was said before and after) for accurately identifying particular CPS skills. This is in line with another study that also supported the importance of considering task context [51].

2 CURRENT STUDY

Our study has two overall goals. First, we aim to compare the performance and feasibility of machine learning models in automatically identifying CPS language. We evaluated several standard methods and state-of-the-art deep learning approaches. Towards this effort, we tested random forest, support vector machines (SVM), naive Bayes, recurrent neural networks (RNN), long short-term memory (LSTM), convolutional neural networks (CNN), LSTM-CNN, as well as large language models such as BERT and GPT-2. Second, as an expansion of the current literature, we explored the potential effect of context-related bias on model performance. We applied a task-based train-test split method to take into account the variation of task-related language.

Our study aims to address the following research questions:
1. How accurately can predictive modeling automate the coding of CPS interactions?
2. To what extent are predictive models sensitive to task-related context?

3 METHODS

3.1 Participants

A total of $N = 514$ undergraduate students from a large university in the southwest U.S. participated in the study. The dataset comprised interactions of these students, evenly distributed across four collaborative problem-solving (CPS) tasks—ranking apartments, professors, party venues, and job candidates, to represent diverse decision-making contexts. For consistency, we considered only four-person groups, leading to $N = 129$ teams in the study. Over half of the participants were female ($N = 347$), predominantly freshmen ($N = 342$) or juniors ($N = 128$). Racial and ethnic demographics were as follows: White (12%), Black or African American (1.7%), Asian or Asian American (40.6%), Hispanic or Latino (31.5%), and multiracial (1.7%) among those who reported (497 out of 514). Additionally, over half ($N = 277$) were first-generation students.

3.2 Procedure

After providing informed consent in line with institutional ethical guidelines for human research, participants were randomly assigned into teams to perform a decision-making task on the Education Platform for Collaborative Assessment and Learning (EPCAL). This platform facilitates collaboration, management, and research in a computer-mediated environment. A background survey was conducted to collect demographic information. In the task, students were given a problem (e.g., "choose the best apartment") with various options, each having positive and negative features. Each team member received different information relevant to the problem. During the 20-minute discussion phase, they communicated synchronously via text to share this information for optimal ranking. The CPS skills distribution within the dataset was as follows: SSI: 37%, SESU: 20%, SMC: 19%, SN: 14%, CRF: 10%, CE: 7%, CM: 6%, CP: 1%. This distribution was maintained to reflect real-world frequencies of these skills in group interactions.

3.3 Qualitative Coding

This study adopts a Collaborative Problem Solving (CPS) framework, specifically adapted from Andrews (2020), to analyze utterance data. The framework, influenced by PISA and other studies, consists of eight key skills, divided into social and cognitive aspects. We excluded the 'exploring and understanding' component due to its limited applicability in our context. Table 1 outlines the definitions and examples for each CPS skill.

Coding was performed at the utterance level. Each utterance was categorized under one primary CPS skill and one of 29 subskills. The study concentrated on the primary skills. Four undergraduate assistants, trained in the CPS framework, coded 7,711 utterance events. A 20% data sample was initially coded independently by each rater to ensure consistency, achieving a high inter-rater reliability (Kappa = .81). The remaining data was then equally divided and coded by the raters. Observed frequencies for each primary code are: SSI (2298), SESU (1009), SMC (953), SN (722), CRF (488), CE (369), CM (308), CP (28).

4 ANALYTICAL APPROACH

Using machine learning algorithms for text classification usually involves two stages: a feature extraction stage and a classification stage using a machine learning technique. More specifically, textual data containing words and characters are transformed into quantitative values that represent the text data, and then this numerical representation is used as the input values to a classification algorithm, i.e., decision tree, naive bayes, or neural network to predict the label values [10]. In the context of our study, each message sent to the group chat is treated as the unit of analysis, and the CPS skill codes assigned to each are considered the label values that are predicted through the classifier. In this study, we will be utilizing machine learning, deep learning, and large language models (LLMs) as classifiers. Each is explained in more detail in the subsequent sections.

Table 1: CPS skills description from

	CPS skill code	Definition	Examples
Social	Maintaining Communication (SMC)	Off-Topic Communication, Rapport Building Communication, Inappropriate Communication	"nice job guys" "no problem"
	Sharing Information (SSI)	Share Own Information, Share Task or Resource Information, Share Understanding	"Candidate A was listed as having good leadership skills"
	Establish Shared Understanding (SESU)	Presentation Phase, Acceptance Phase	"What skills do we need?"
	Negotiating (SN)	Express agreement or disagreement, Resolve conflicts	"My list shows that C is unwilling to further their education."
Cognitive	Representing and Formulating (CRF)	Represent the problem using words, Proposes specific conceptual thinking	"Yeah I feel that B is the best because everything is nearby"
	Planning (CP)	Set Goals, Develop Strategies	"we have to choose between A and B for best"
	Executing (CE)	Suggesting an action to a teammate, Report of own action	"Please list all your features for candidate C"
	Monitoring (CM)	Monitor progress toward the goal, Monitor whether teammates are present	"So we in agreement to make B the best?"

4.1 Machine Learning

As discussed in the previous section, the first step in text classification is feature extraction. We used the term frequencies as the criteria for this step. This was done through the sklearn python implementation [40]. 1 and 2 grams are used in this process, meaning that frequent two-word terms are also considered features along with single words. Maximum features were set as 10,000, which means the 10,000 most frequent terms were selected as the feature map list. Next, each text message is transformed into this feature map based on the existence of the selected terms in the input text. The Bag of Words (BoW) approach represents data entries based on a selected set of words, without considering their order or grammar. Once the input data is quantified through the feature selection step, the values are used as input values to a machine learning algorithm. This study uses the most popular and commonly used machine learning methods to classify the training data, namely: Gaussian Naive Bayes (GNB), Linear SVM, and Random Forest.

4.2 Deep Learning

In our deep learning approach for feature mapping, we utilized Fasttext's pre-trained word vectors, trained on the Common Crawl dataset comprising 600 billion tokens, to represent our training data. These 300-dimensional vectors, from a pool of 2 million terms, are adept at capturing word associations and semantic and syntactic similarities. The extensive vocabulary of these vectors, selected for their robustness in diverse contexts, aligns well with the varied subjects handled by our undergraduate study participants, providing reliable semantics for a wide range of contexts, as evidenced in prior research [5, 39]. For input consistency, the first 50 tokens of messages were used, based on the training data's message length distribution. Our study replaced each word in the training corpus with the corresponding 300-dimensional vector, making a sentence of length L represented by an L * 300-dimensional matrix. We employed four deep learning architectures: Recurrent Neural Networks (RNN) with three 50 sequence layers [22], Bidirectional Long Short-Term Memory (LSTM) with three 50 sequence layers [33], Convolutional Neural Networks (CNN) with four convolution and max pooling layers [37], and a Hybrid CNN + LSTM model comprising a convolution and max pooling layer followed by two bidirectional LSTM layers [49]. Each model underwent training for 30 epochs, with the best parameters on validation saved and reloaded for predictions to prevent overfitting.

4.3 Large Language Models

The NLP landscape has undergone significant transformation with the advent of transformer-based architectures [53], positioning Large Language Models (LLMs) at the forefront of technological advancements in the field. The practical application of these models often entails a delicate balance between computational resources and the advanced capabilities they offer. In our study, we utilized BERT [13] and GPT-2 [44], leveraging the pre-trained models available on the Hugging Face platform [55], using BERT[1] and GPT-2[2] implementations as a base. Subsequently, we fine-tuned a neural network classifier on top of these LLMs, maintaining the integrity of the base parameters.

Our decision to utilize GPT-2, a model representing the transformative impact of transformer technology, was influenced by both computational considerations and model accessibility [36]. While more recent models like GPT-3.5 and GPT-4 exhibit enhanced performance capabilities [6], the feasibility of their application within the context of our research resources was a critical factor. The use of GPT-2 allowed us to conduct a focused and meaningful study, providing valuable insights into the application of LLMs in NLP tasks, while acknowledging the growing potential and challenges of these models in academic research [38]. Future research endeavors will aim to build upon this foundation, exploring the capabilities of more

[1]https://huggingface.co/bert-base-uncased
[2]https://huggingface.co/gpt2

advanced models as they become more accessible and manageable within academic settings.

4.4 Evaluation Metrics

In order to evaluate the prediction quality of the classification models, we used an unseen test set of data and made predictions using the trained models on the training set. We used three main metrics to evaluate the performance of the predictions.

(1) **Accuracy**: Measures the ratio of correct predictions over the total predictions. Useful but potentially misleading in imbalanced datasets as it might overemphasize dominant labels' performance.
(2) **F1-Score**: The harmonic mean of precision and recall, calculated as

$$F1\text{-}score = 2 \times \frac{(precision \times recall)}{(precision + recall)} \quad (1)$$

It is a more balanced measure than accuracy, especially in imbalanced datasets [4].
- **F1-Macro Average**: The arithmetic mean of per-class F1-scores, treating all classes equally. It reflects the performance across all classes, regardless of their size [42].
- **F1-Weighted Average**: The average of F1-scores for each class, weighted by the number of true instances for each class. This metric accounts for class imbalance [42].

(3) **Baseline**: This metric represents a naïve model predicting every instance as the most frequent class. For our study, with SSI as the most frequent code, the baseline accuracy is approximately 37%.

4.5 Generalizability

In order to evaluate the generalizability of the classification, we implemented a task-based train-test split. We hypothesized that the classifiers might develop biases based on the training data towards the vocabulary and terms frequently present within the tasks (apartment, candidate, party venue, and professor). To minimize this bias, we implemented task-based train-test splits where we left out data from one of the tasks as the test set and used the other three tasks as the training set. We believe this could be a better criterion for the model's performance on new data since it minimizes the bias that can come from the task-specific terms and can generalize to never-before-seen data. The results are then averaged across the four classifications for each classifier. The visualization of this process is available in Figure 1.

5 RESULTS

The BERT classifier demonstrated exceptional efficacy, leading the pack with an accuracy of 73% and equally commendable f1-macro and f1-weighted scores of 61 and 73, respectively. These results are substantially superior to the baseline accuracy of 37%, which is founded upon predictions based on the most frequent class. Both the GPT-2 and LSTM models showed a commendable accuracy of 70%. However, there was a notable difference in their f1-macro scores, with LSTM registering a higher score of 58 as opposed to GPT-2's 46. Models such as LSTM-CNN, CNN, Linear SVM, RNN base model, and Random Forest presented moderate performances with accuracies ranging between 65% and 68%. Among them, CNN and Linear SVM paralleled in f1-macro scores at 54. Gaussian Naive Bayes (GNB) surpassed the baseline, albeit with modest metrics, attaining an accuracy of 48%. Although all classifiers exceeded the performance of the most frequent class baseline, the optimal classifier selection would be predicated on the specific goals of the application and any computational constraints. Classification performance evaluation results are available in Table 2.

Drawing on the findings presented, deep learning and LLM models generally outperform classical machine learning models. Yet, machine learning models can have the advantage of higher training speed and greater explainability. It's also worth noting that while the performance of both machine learning and deep learning models will improve with larger training data, deep learning models stand to benefit substantially more due to their higher complexity [34]. This suggests the performance gap might widen with a larger training dataset. Across the board, all models have smaller f1-macro values compared to the f1-weighted averages, indicating inconsistent performance across classes and better performance in more frequent classes. This expected trend might see a shift with more data on the less frequent labels, potentially enhancing performance uniformly across all classes.

To address RQ2, we evaluated the models on a task-based train-test split settings. Our assessment of multiple classifiers on the dataset revealed significant variations in their performance. The BERT classifier emerged as the most proficient, registering the highest accuracy of 70%, as well as superior f1-macro and f1-weighted scores of 57 and 70, respectively. These figures notably exceed the baseline accuracy of 37%, which represents predicting the most frequent class. GPT-2 and LSTM closely followed, both securing an accuracy of 68%, albeit with slight differences in F1 scores. Other models like the LSTM-CNN hybrid, CNN, Linear SVM, and Random Forest showed moderate performances, with accuracies oscillating between 60% and 67%. On the lower end, the Gaussian Naive Bayes (GNB) classifier, while outpacing the most frequent class baseline, registered the least favorable metrics among the evaluated models with an accuracy of 44%. While every classifier demonstrated improved utility over merely predicting the dominant class, the choice of classifier would inevitably be guided by specific application needs and computational considerations. The results are shown in Table 2.

6 DISCUSSION

The central objective of this study was to investigate the feasibility of automating the coding of collaborative problem-solving (CPS) skills, encompassing both social and cognitive aspects, through the use of machine learning, deep learning, and large language models (LLMs). Our findings revealed that LLMs, with a notable emphasis on BERT, exhibited superior performance in terms of accuracy and f1-score when compared to other methodologies, signifying their proficiency in capturing the semantic and syntactic features of text data and their ability to generalize effectively to unseen data. However, it is crucial to note that LLMs demand greater computational resources and training time, potentially limiting their applicability in certain contexts. Deep learning models, particularly LSTM, also showcased commendable performance relative to machine learning

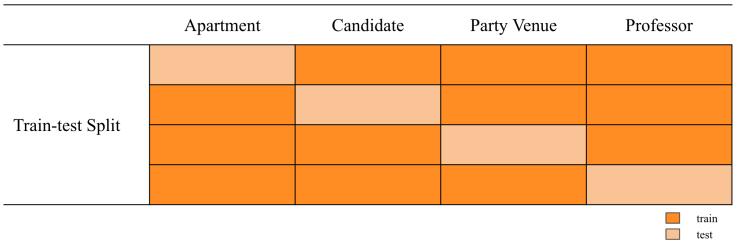

Figure 1: Visualization of the task-based train/test split methodology. Darker sections represent the training set, and lighter sections represent the test set for each task. This method ensures that each task is held out as a test set in a separate batch to evaluate the model's ability to generalize to unseen data.

Table 2: Merged Classification and Task-based training results

	Classification Results			Task-based Training Results			
Classifier	**accuracy**	**f1-macro**	**f1-weighted**	**accuracy**	**f1-macro**	**f1-weighted**	**Baseline**
BERT	0.73	0.61	0.73	0.70	0.57	0.70	0.37
GPT-2	0.70	0.46	0.69	0.68	0.55	0.68	0.37
LSTM	0.70	0.58	0.69	0.68	0.53	0.68	0.37
LSTM-CNN	0.66	0.52	0.65	0.60	0.45	0.60	0.37
CNN	0.68	0.54	0.66	0.66	0.51	0.66	0.37
RNN base	0.65	0.51	0.64	0.63	0.46	0.63	0.37
GNB	0.48	0.39	0.50	0.44	0.36	0.44	0.37
Linear SVM	0.65	0.54	0.66	0.63	0.54	0.63	0.37
Random Forest	0.67	0.51	0.65	0.67	0.51	0.67	0.37

models, underscoring the advantage of exploiting the sequential nature of text data and word embeddings to enhance classification accuracy. Nevertheless, deep learning models are hampered by their lack of interpretability and transparency, posing challenges in comprehending and elucidating classification outcomes. Machine learning models, including random forest and SVM, demonstrated moderate performance when compared to deep learning and LLMs [14]. Their primary limitation lies in their reliance on the bag of words approach, which may overlook crucial word order and contextual details [32], while also rendering them more vulnerable to data imbalances and noise [25].

The task-based train-test split technique was implemented to evaluate the generalizability of the models across different tasks. The results showed that there was a slight drop in performance for most models when using this technique, which implies that the task-specific vocabulary and terms may influence the classification results. This emphasizes the need to account for task-related context when applying automated coding of CPS skills across varying domains.

The study has several limitations that need to be addressed in future work. First, the data set used in this study was relatively small and unbalanced, which may affect the reliability and validity of the classification results. Future work should collect more data from diverse sources and tasks to increase the robustness and representativeness of the data set. Moreover, generative AI models could be further explored as possible methods for data augmentation to generate more data entries for the less represented categories [12]. Second, the study only focused on eight main CPS skills and did not consider the subskills or other factors that may influence CPS performance. In follow-up studies, we plan to explore more granular and comprehensive measures of CPS skills and examine how they relate to other variables such as team composition, task complexity, and learning outcomes.

7 CONCLUSION

Our study delved into the potential of harnessing machine learning, deep learning, and advanced language models for the automatic modeling of social and cognitive collaborative problem-solving (CPS) skills. The findings from our research, which involved the analysis of qualitatively coded utterances within teams, have several implications for the future of education and collaborative learning.

One direct application of our approach is the development of systems that can automatically generate reports for educators. Such systems could provide insights into the CPS skills exhibited by various student groups. This would be invaluable for educators overseeing multiple student groups, enabling them to gauge the proficiency of each group in CPS. By doing so, educators can pinpoint groups that might be struggling and allocate their resources more effectively to assist them. Beyond group-level insights, our approach can also be tailored to provide individualized feedback. By analyzing a student's participation and interaction patterns, the system can identify areas of strength and areas needing improvement. For instance, a student who is proactive in sharing but less involved in collaborative discussions could be nudged to be more receptive to their peers' ideas and to contribute constructively.

The future holds promise for integrating our approach with intelligent systems designed to monitor CPS in real-time. Such systems

could actively intervene during group discussions to optimize outcomes. Imagine a scenario where a group veers off-topic; the system could gently remind them to refocus. Similarly, if students remain passive, the system could encourage them to voice their thoughts. The design and implementation of these interventions, including their timing, presentation, and target (be it the entire group or an individual), require further exploration and fine-tuning.

Lastly, our research utilized prominent machine learning techniques and language models like BERT and GPT-2. Future research could delve deeper into refining these models or exploring newer models to enhance the accuracy and applicability of automatic CPS skill modeling. As we move forward, it will be crucial to design and test these systems in diverse settings to ensure their efficacy and adaptability.

ACKNOWLEDGMENTS

This research was supported by The Gates Foundation (#INV - 000752) and The Andrew W. Mellon Foundation (1806-05902). The authors would like to thank the Next Generation Undergraduate Success Measurement Project team members for help with data collection. Thanks to Educational Testing Services for the experiment information.

REFERENCES

[1] Jessica Andrews-Todd and Carol M Forsyth. 2020. Exploring social and cognitive dimensions of collaborative problem solving in an open online simulation-based task. *Computers in human behavior* 104 (2020), 105759.

[2] Jessica Andrews-Todd and Deirdre Kerr. 2019. Application of ontologies for assessing collaborative problem solving skills. *International Journal of Testing* 19, 2 (2019), 172–187.

[3] Marilyn Binkley, Ola Erstad, Joan Herman, Senta Raizen, Martin Ripley, May Miller-Ricci, and Mike Rumble. 2012. Defining twenty-first century skills. *Assessment and teaching of 21st century skills* (2012), 17–66.

[4] David C. Blair. 1979. Information Retrieval, 2nd ed. C.J. Van Rijsbergen. London: Butterworths; 1979: 208 pp. Price: $32.50. *Journal of the American Society for Information Science* 30, 6 (1979), 374–375. https://doi.org/10.1002/asi.4630300621 arXiv:https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/asi.4630300621

[5] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics* 5 (2017), 135–146. https://doi.org/10.1162/tacl_a_00051

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[7] Jeremy Burrus, Teresa Jackson, Nuo Xi, and Jonathan Steinberg. 2013. Identifying the most important 21st century workforce competencies: An analysis of the Occupational Information Network (O* NET). *ETS Research Report Series* 2013, 2 (2013), i–55.

[8] Whitney Cade, Nia Dowell, Art Graesser, Yla Tausczik, and James Pennebaker. 2014. Modeling student socioaffective responses to group interactions in a collaborative online chat environment. In *Educational Data Mining 2014*. Citeseer.

[9] Esther Care, Claire Scoular, and Patrick Griffin. 2016. Assessment of collaborative problem solving in education environments. *Applied Measurement in Education* 29, 4 (2016), 250–264.

[10] Jing Chen, James H Fife, Isaac I Bejar, and André A Rupp. 2016. Building e-rater® scoring models using machine learning methods. *ETS Research Report Series* 2016, 1 (2016), 1–12.

[11] Harshita Chopra, Yiwen Lin, Mohammad Amin, Jacqueline G Cavazos, Renzhe Yu, Spencer Jaquay, and Nia Nixon. 2023. Semantic Topic Chains for Modeling Temporality of Themes in Online Student Discussion Forums. (2023).

[12] Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Yihan Cao, Zihao Wu, Lin Zhao, Shaochen Xu, Wei Liu, Ninghao Liu, et al. 2023. AugGPT: Leveraging ChatGPT for Text Data Augmentation. *arXiv preprint arXiv:2302.13007* (2023).

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[14] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[15] Nia Dowell, Oleksandra Poquet, and Christopher Brooks. 2018. Applying group communication analysis to educational discourse interactions at scale. International Society of the Learning Sciences, Inc.[ISLS].

[16] Nia MM Dowell, Christopher Brooks, Vitomir Kovanović, Srećko Joksimović, and Dragan Gašević. 2017. The changing patterns of MOOC discourse. In *Proceedings of the Fourth (2017) ACM Conference on Learning@ Scale*. 283–286.

[17] Nia M Dowell, Arthur C Graesser, and Zhiqiang Cai. 2016. Language and discourse analysis with Coh-Metrix: Applications from educational material to learning environments at scale. *Journal of Learning Analytics* 3, 3 (2016), 72–95.

[18] Nia MM Dowell, Tristan M Nixon, and Arthur C Graesser. 2019. Group communication analysis: A computational linguistics approach for detecting sociocognitive roles in multiparty interactions. *Behavior research methods* 51, 3 (2019), 1007–1041.

[19] Nia MM Dowell and Oleksandra Poquet. 2021. SCIP: Combining group communication and interpersonal positioning to identify emergent roles in scaled digital environments. *Computers in Human Behavior* 119 (2021), 106709.

[20] Nia Marcia Maria Dowell and Arthur C Graesser. 2014. Modeling learners' cognitive, affective, and social processes through language and discourse. *Journal of Learning Analytics* 1, 3 (2014), 183–186.

[21] Sidney K D'Mello, Nia Dowell, and Art Graesser. 2013. Unimodal and multimodal human perceptionof naturalistic non-basic affective statesduring human-computer interactions. *IEEE Transactions on Affective Computing* 4, 4 (2013), 452–465.

[22] Jeffrey L Elman. 1990. Finding structure in time. *Cognitive science* 14, 2 (1990), 179–211.

[23] Michael Eraut. 2012. Transfer of knowledge between education and workplace settings. In *Knowledge, values and educational policy*. Routledge, 65–84.

[24] Stephen M Fiore, Arthur Graesser, and Samuel Greiff. 2018. Collaborative problem-solving education for the twenty-first-century workforce. *Nature human behaviour* 2, 6 (2018), 367–369.

[25] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* (2014).

[26] Arthur C Graesser, Nia Dowell, and Danielle Clewley. 2017. Assessing collaborative problem solving through conversational agents. In *Innovative assessment of collaboration*. Springer, 65–80.

[27] Arthur C Graesser, Nia Dowell, Andrew J Hampton, Anne M Lippert, Haiying Li, and David Williamson Shaffer. 2018. Building intelligent conversational tutors and mentors for team collaborative problem solving: Guidance from the 2015 Program for International Student Assessment. In *Building intelligent tutoring systems for teams*. Emerald Publishing Limited.

[28] Arthur C Graesser, Nia Dowell, and Christian Moldovan. 2011. A computer's understanding of literature. *Scientific Study of Literature* 1, 1 (2011), 24–33.

[29] Arthur C Graesser, Stephen M Fiore, Samuel Greiff, Jessica Andrews-Todd, Peter W Foltz, and Friedrich W Hesse. 2018. Advancing the science of collaborative problem solving. *Psychological Science in the Public Interest* 19, 2 (2018), 59–92.

[30] Patrick Griffin, Esther Care, and Barry McGaw. 2011. The changing role of education and schools. In *Assessment and teaching of 21st century skills*. Springer, 1–15.

[31] Jiangang Hao, Lei Liu, Alina A von Davier, Nathan Lederer, Diego Zapata-Rivera, Peter Jakl, and Michael Bakkenson. 2017. EPCAL: ETS platform for collaborative assessment and learning. *ETS Research Report Series* 2017, 1 (2017), 1–14.

[32] Zellig S Harris. 1954. Distributional structure. *Word* 10, 2-3 (1954), 146–162.

[33] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[34] Christian Janiesch, Patrick Zschech, and Kai Heinrich. 2021. Machine learning and deep learning. *Electronic Markets* 31, 3 (2021), 685–695.

[35] S Joksimović, O Poquet, V Kovanović, NM Dowell, C Mills, D Gašević, and C Brooks. 2017. How do we Model Learning at Scale? A Systematic Review of the Literature. *Review of Educational Research* (2017).

[36] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361* (2020).

[37] Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882* (2014).

[38] Jesse G Meyer, Ryan J Urbanowicz, Patrick CN Martin, Karen O'Connor, Ruowang Li, Pei-Chen Peng, Tiffani J Bright, Nicholas Tatonetti, Kyoung Jae Won, Graciela Gonzalez-Hernandez, et al. 2023. ChatGPT and large language models in academia: opportunities and challenges. *BioData Mining* 16, 1 (2023), 20.

[39] Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in Pre-Training Distributed Word Representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

[40] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[41] Ismaël Peña-López et al. 2017. PISA 2015 Results (Volume V). Collaborative Problem Solving. (2017).
[42] David MW Powers. 2020. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061* (2020).
[43] Samuel L Pugh, Shree Krishna Subburaj, Arjun Ramesh Rao, Angela EB Stewart, Jessica Andrews-Todd, and Sidney K D'Mello. 2021. Say What? Automatic Modeling of Collaborative Problem Solving Skills from Student Speech in the Wild. *International Educational Data Mining Society* (2021).
[44] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.
[45] Carolyn Rosé, Yi-Chia Wang, Yue Cui, Jaime Arguello, Karsten Stegmann, Armin Weinberger, and Frank Fischer. 2008. Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning. *International journal of computer-supported collaborative learning* 3 (2008), 237–271.
[46] Mohammad Amin Samadi, Jacqueline G Cavazos, Yiwen Lin, and Nia Nixon. 2022. Exploring Cultural Diversity and Collaborative Team Communication through a Dynamical Systems Lens. *International Educational Data Mining Society* (2022).
[47] Bertrand Schneider, Nia Dowell, and Kate Thompson. 2021. Collaboration Analytics—Current State and Potential Futures. *Journal of Learning Analytics* 8, 1 (2021), 1–12.
[48] Julian Schulze and Stefan Krumm. 2017. The "virtual team player" A review and initial model of knowledge, skills, abilities, and other characteristics for virtual collaboration. *Organizational Psychology Review* 7, 1 (2017), 66–95.
[49] Xiangyang She and Di Zhang. 2018. Text Classification Based on Hybrid CNN-LSTM Hybrid Model. *2018 11th International Symposium on Computational Intelligence and Design (ISCID)* 02 (2018), 185–189. https://api.semanticscholar.org/CorpusID:133603472
[50] Matthias Stadler, Katharina Herborn, Maida Mustafić, and Samuel Greiff. 2020. The assessment of collaborative problem solving in PISA 2015: An investigation of the validity of the PISA 2015 CPS tasks. *Computers & Education* 157 (2020), 103964.
[51] Angela EB Stewart, Zachary Keirn, and Sidney K D'Mello. 2021. Multimodal modeling of collaborative problem-solving facets in triads. *User Modeling and User-Adapted Interaction* (2021), 1–39.
[52] Angela EB Stewart, Hana Vrzakova, Chen Sun, Jade Yonehiro, Cathlyn Adele Stone, Nicholas D Duran, Valerie Shute, and Sidney K D'Mello. 2019. I say, you say, we say: Using spoken language to model socio-cognitive processes during computer-supported collaborative problem solving. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–19.
[53] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
[54] Jie Wang, Grand H-L Cheng, Tingting Chen, and Kwok Leung. 2019. Team creativity/innovation in culturally diverse teams: A meta-analysis. *Journal of Organizational Behavior* 40, 6 (2019), 693–708.
[55] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in Natural Language Processing: System Demonstrations*. 38–45.