Detecting Disruptive Talk in Student Chat-Based Discussion within Collaborative Game-Based Learning Environments

Kyungjin Park
Department of Computer Science,
North Carolina State University

Hyunwoo Sohn
Department of Computer Science,
North Carolina State University

Bradford W. Mott
Department of Computer Science,
North Carolina State University

Wookhee Min
Department of Computer Science,
North Carolina State University

Asmalina Saleh
Center for Research on Learning and
Teaching, Indiana University
Bloomington

Krista D. Glazewski
Center for Research on Learning and
Teaching, Indiana University
Bloomington

Cindy E. Hmelo-Silver
Center for Research on Learning and
Teaching, Indiana University
Bloomington

James C. Lester
Department of Computer Science,
North Carolina State University

ABSTRACT

Collaborative game-based learning environments offer significant promise for creating engaging group learning experiences. Online chat plays a pivotal role in these environments by providing students with a means to freely communicate during problem solving. These chat-based discussions and negotiations support the coordination of students' in-game learning activities. However, this freedom of expression comes with the possibility that some students might engage in undesirable communicative behavior. A key challenge posed by collaborative game-based learning environments is how to reliably detect disruptive talk that purposefully disrupt team dynamics and problem-solving interactions. Detecting disruptive talk during collaborative game-based learning is particularly important because if it is allowed to persist, it can generate frustration and significantly impede the learning process for students. This paper analyzes disruptive talk in a collaborative game-based learning environment for middle school science education to investigate how such behaviors influence students' learning outcomes and varies across gender and students' prior knowledge. We present a disruptive talk detection framework that automatically detects disruptive talk in chat-based group conversations. We further investigate both classic machine learning and deep learning models for the framework utilizing a range of dialogue representations as well as supplementary information such as student gender. Findings show that long short-term memory network (LSTM)-based disruptive talk detection models outperform competitive baseline models, indicating that the LSTM-based disruptive talk detection framework offers significant potential for supporting effective collaborative game-based learning through the identification of disruptive talk.

CCS CONCEPTS

• Applied computing; **• Education**; **• Collaborative learning; Interactive environments**; **• Computers in other domains**; **• Personal computers and PC applications**; **• Computer games**; **• Computing Methodologies**; **• Artificial Intelligence**; **• Natural Language Processing**; **• Discourse, dialogue and pragmatics**;

KEYWORDS

Collaborative Game-Based Learning, Disruptive Talk Detection, Text Analytics

ACM Reference Format:
Kyungjin Park, Hyunwoo Sohn, Bradford W. Mott, Wookhee Min, Asmalina Saleh, Krista D. Glazewski, Cindy E. Hmelo-Silver, and James C. Lester. 2021. Detecting Disruptive Talk in Student Chat-Based Discussion within Collaborative Game-Based Learning Environments. In *LAK21: 11th International Learning Analytics and Knowledge Conference (LAK21), April 12–16, 2021, Irvine, CA, USA*. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3448139.3448178

1 INTRODUCTION

Computer-supported collaborative learning (CSCL) fosters the social aspects of learning using a variety of technological and constructive pedagogical strategies, such as problem-based learning and inquiry learning [9, 18, 22]. In collaborative learning, students engage in problem solving, artifact design, and inquiry in small groups. Small groups have been proven to be effective for collaborative learning and developing deep disciplinary engagement [10, 18]. Collaborative game-based learning environments bring game-based learning to small groups, providing immersive virtual learning experiences with progressively advanced challenges focused on the desired learning objectives [48]. Collaborative game-based learning environments often provide students with in-game chat facilities

LAK21, April 12–16, 2021, Irvine, CA, USA

ACM ISBN 978-1-4503-8935-8/21/04…$15.00
https://doi.org/10.1145/3448139.3448178

that support open discussions and negotiations among team members to support the coordination of their in-game learning activities. However, students can abuse the chat system to engage in undesirable communicative talk, which can negatively affect the group learning experience.

Thus, it is critical to examine collaborative discussions to understand how certain types of talk can promote learning outcomes [30]. In high quality collaborative talk, students build on each other's ideas and move toward deep disciplinary engagement [41]. However, negative socio-emotional engagement can manifest as disruptive talk and can be a barrier to high-quality collaborative talk. Disruptive talk can include insults, bullying, and negative expressions. These utterances can generate frustration, harm communication, and produce a negative group atmosphere [20]. Disruptive individuals or groups who engage in unproductive social processes may distract others from learning and can interfere with deeper learning by continually interrupting learning activities [4]. For example, students who engage in disputational talk focus on finding the flaws in others' opinions, which can impede the overall collaborative learning process [13, 31]. The ability to detect disruptive talk is therefore critical for achieving high-quality collaborative learning. However, determining whether talk is acceptable or not is context-dependent. For example, conflicts that are either process-related (e.g., disagreement on the collaboration process) or relationship-related (e.g., interpersonal clashes) negatively impact group learning. On the other hand, task-related conflicts, which are induced during group problem solving, can positively impact students' learning outcomes [21, 25, 38]. Thus, we need refined models that surface the types of interactions that potentially interfere with learning, rather than identifying disruptive talk in general.

Researchers have investigated various approaches to mitigate disruptive talk in collaborative learning environments by devising both non-computational and computational techniques. Incorporating peer assessments during the CSCL process is an attractive non-computational option for reducing students' disruptive talk as it can encourage positive attitudes toward collaborative problem solving [39]. Previous work has investigated computational approaches for automatically detecting bullying and off-task behavior in collaborative learning environments with language models ranging from classic approaches (e.g., *n*-grams) to more recent state-of-the-art contextualized word embeddings (e.g., BERT), which were used in conjunction with machine learning classification techniques (e.g., logistic regression, random forest, long-short term memory networks) [5, 35].

In this work, we present a disruptive talk detection framework for collaborative game-based learning that automatically detects disruptive talk during in-game chat-based group conversations. We evaluate the effectiveness of the model with a dialogue dataset collected from students' interactions with a CSCL-based educational game for middle-grade science. We investigate two classification models (i.e., logistic regression, and long short-term memory networks) utilizing a range of dialogue-based linguistic representations and student attributes. We also consider how disruptive talk intersects with gender, prior knowledge, individual outcomes, and group learning outcomes. The research questions (RQs) we address in this work are as follows:

- **RQ1.** To what extent does disruptive talk influence learning outcomes?
- **RQ2.** To what extent do the classification models accurately detect disruptive talk?

2 RELATED WORK

There is a growing literature on using natural language processing techniques for learning analytics and automatic analysis of talk in collaborative learning environments. We discuss each of these in turn.

2.1 Natural Language Processing in Learning Analytics

A wide range of natural language processing (NLP) approaches have been used in learning analytics [27]. Automated essay scoring and short answer grading have been widely investigated. Prior work has presented algorithms that use latent semantic analysis (LSA) for assessing the similarity of an essay to benchmark essays [24], and explored text cohesion for assessing essay quality [28, 29]. More recently, advanced NLP techniques, such as neural network-based distributed language representation learning approaches (e.g., word2vec) and transfer learning approaches (e.g., BERT), have been applied to short answer grading [34, 44, 45]. In massive open online courses (MOOCs), NLP techniques along with classification algorithms (e.g., logistic regression, random forest) have examined data from discussion forums for a wide range of tasks such as predicting students' learning outcomes, sentiment analysis [27], confusion detection [14], and cognitive presence [3, 12]. NLP techniques, such as LSA and other machine learning techniques have also been applied to group discourse in collaborative learning environments, particularly in CSCL systems [5, 7, 27, 35, 47]. Trausan-Matu et al. presented a tool that incorporates textual and gestural interactions within collaborative groups, where LSA was used to identify topics, semantic similarities, and links between utterances [47]. Later, Dascalu et al. presented a cohesion network analysis method to evaluate students' participation in CSCL conversations, utilizing basic text mining techniques such as tokenization, lemmatization, part-of-speech tagging, and LSA and latent Dirichlet allocation (LDA)-based semantic similarity scores as cohesive links used to build a cohesion network analysis graph [7].

2.2 Natural Language Processing in Automatic Analysis of Talk in Collaborative Learning Environments

Effective facilitation is important in CSCL environments to guide discussions such that positive learning experiences are provided for all students. To support the facilitation process, efforts have been made to explore the automatic detection of specific types of talk in collaborative learning. Ai et al. presented the analysis of transactive conversation within group classroom discussions in middle school math classrooms. They focused on exploring how cognitive conflict can benefit the problem-solving process and investigated the use of automatic analysis of transactive utterances using naïve Bayes, support vector machines, and decision trees [1]. Gweon et al. investigated the automatic analysis of transactive contributions from

Figure 1: EcoJourneys collaborative game-based learning environment.

speech data of undergraduate-level dyad discussions for controversial topics and created a generalizable approach for measuring the prevalence of transactive contributions utilizing an unsupervised dynamic Bayesian network modeling approach [17]. At times, students' talk can cause disruptions that impede collaborative learning processes. Recent work investigating bullying and off-task behavior in collaborative learning environments used word representation approaches along with classic machine learning and deep learning techniques [5, 35]. Nikiforos et al. investigated the automatic detection of aggressive actions within student dialogues in CSCL environments. They adopted unigrams as the word-representation approach and investigated machine learning techniques including naïve Bayes, decision trees, and feedforward neural networks [35]. However, they did not investigate recent word embedding techniques, and they focused on individual messages without considering the dialogue context. Carpenter et al. investigated the use of dialogue analysis to determine if students' messages were either on-task or off-task during collaborative game-based learning. They introduced an off-task behavior detection system, which utilizes three different distributed word representation approaches (i.e., word2vec, ELMo, BERT) and various contextual information extracted from chat messages [5]. However, the authors did not compare their results with classical approaches (e.g., bag-of-words), which could be more effective at modelling a domain-specific language dataset limited in size. Furthermore, none of the previous work investigated disruptive talk. In this work, we introduce a disruptive talk detection framework that can automatically detect disruptive talk that can negatively impact the learning process, by exploring recent dynamic (i.e., contextualized) and static word embedding approaches as well as bag-of-words approach. We then investigate the predictive performance of the models for disruptive talk detection.

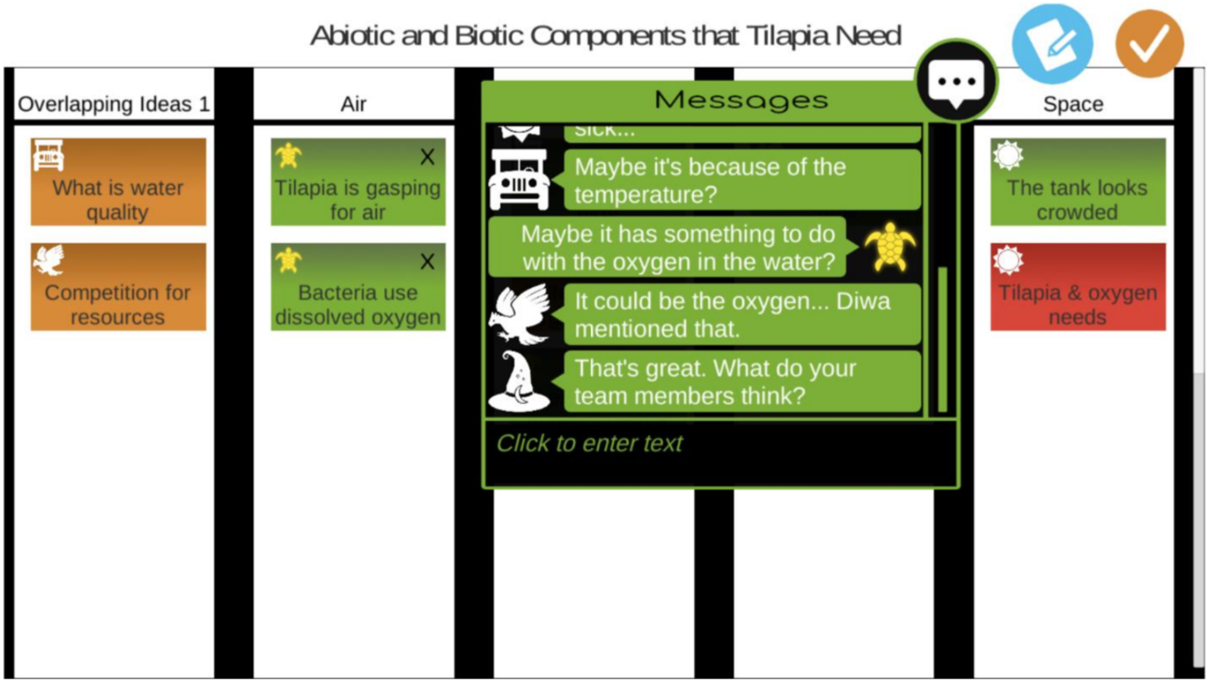
3 METHOD

To investigate disruptive talk in collaborative game-based learning, we collected a corpus of middle school student dialogue as students interacted in a collaborative game-based learning environment for learning ecosystem concepts. Below we describe the collaborative game-based learning environment and the collaborative dialogue corpus, we introduce a disruptive talk annotation protocol, and we describe analyses of student disruptive talk and the disruptive talk detection framework.

3.1 EcoJourneys: Collaborative Game-Based Learning Environment

EcoJourneys is a collaborative game-based learning environment that has a curricular focus on middle school ecosystem science (Figure 1). In this learning environment, students visit a virtual remote island and are asked to investigate what is causing a mysterious illness with fish throughout the island. To solve the mystery, students collaborate in groups of four within the game, where each student works on a separate laptop and meets peers in the virtual game world.

During gameplay, individual students investigate the fish illness by gathering information and interacting with virtual characters. The virtual characters act as local experts who provide details regarding ecosystem concepts and the unfolding narrative (e.g., "Dissolved oxygen is an abiotic factor, or non-living components that animals and plants need to live."). After investigating and collecting


Abiotic and Biotic Components that Tilapia Need
Overlapping Ideas 1
What is water quality
Competition for resources
Air
Tilapia is gasping for air
X
Bacteria use dissolved oxygen
X
Messages
Space
The tank looks crowded
Tilapia & oxygen needs

sick...

Maybe it's because of the temperature?

Maybe it has something to do with the oxygen in the water?

It could be the oxygen... Diwa mentioned that.

That's great. What do your team members think?

Click to enter text

Figure 2: In-game chat interface and virtual whiteboard.

information, students gather at a virtual whiteboard within the game to share and categorize the information they have collected into different components and discuss the most-likely cause of the fish illness. Throughout the problem-solving activity within the game, students are encouraged to share ideas, ask questions, and negotiate with their team members via the in-game chat interface (Figure 2). This built-in chat system is available throughout gameplay. For each group, a facilitator, who is either a researcher or a teacher, asks questions and encourages students to share their thoughts with other group members via the in-game chat interface. Facilitators can monitor students' activities and their conversation using a separate in-game view and intervene in order to guide students' learning. Facilitators can select messages from a set of pre-authored messages or provide free-form messages using the in-game chat interface.

3.2 Study Procedure

The EcoJourneys collaborative game-based learning environment was used in a classroom-based study. Students were either in the sixth or seventh grade (11-12 years old) and played EcoJourneys during 6 classroom periods. Among 14 groups with 56 students (4 students per group) who participated in the study, we used the group discourse data from 11 groups with 44 students (21 female and 23 male) who consented to the study and completed all the activities in the game-based learning environment. There was a total of 3,749 chat messages available in the study dataset, with 831 messages from a facilitator and 2,918 of them from students. We only consider the students' messages in the disruptive talk detection modeling. On average, students in each group sent 265.3 messages (min = 83, max = 722, SD = 166.1). Before and after the game, students took the same pre- and post-test on ecosystems to determine whether engaging in the collaborative game-based learning environment improved their science learning. A paired t-test comparing pre-test (M = 13.36, SD = 3.92) and post-test (M = 15.64, SD = 3.54) of students who completed both the tests showed that student learning gains were statistically significant (i.e., t (39) = 2.70, $p < 0.001$).

3.3 Disruptive Talk Annotation of Group Discourse

We adopted a binary annotation scheme, *disruptive talk* and *non-disruptive talk*, adapted from previous work on disruptive talk analysis [4]. We labeled student utterances as *disruptive talk* if they could distract the rest of the group members from learning and could interfere with deeper learning by continually interrupting the learning activity, and otherwise we labeled the utterance as *non-disruptive talk*. Table 1 shows examples of disruptive and non-disruptive messages along with the definition of each class.

Two human annotators labeled the students' chat-based dialogue collected during the study. Both annotators labeled approximately 20% of the entire corpus, achieving an inter-rater agreement of 0.63 using Cohen's kappa, which indicates substantial agreement among the annotators [6]. All utterances labeled differently between the two annotators were discussed, and a label was eventually chosen for each utterance for which there was not agreement before labeling the remainder of the corpus. Then, each annotator labeled approximately 40% of the remaining utterances. The distribution of disruptive and non-disruptive utterances among the dataset is 370 (12.7%) and 2,548 (87.3%), respectively.

Table 1: Example utterances of disruptive and non-disruptive talk.

Class	Definition	Example
Disruptive Talk	Talk that generates frustration, annoyance, harming communication, or contributes to an increasingly bad mood among the group members	“I want to be right I’m gonna correct you I am right” “Um yea. yep, you can’t work”
Non-Disruptive Talk	Normal talk that does not create bad moods among the group members.	“if [the fish] don’t come to the top to breath then they are not going to breath at all” “It could also be in the water quality section”

3.4 Disruptive Talk Analysis

Before we apply machine learning techniques to automatically detect disruptive talk within the group discourse, we attempt to analyze the disruptive talk in our dataset to better understand the dataset and determine which features can be helpful for model training. We analyze how each groups’ disruptive talk ratio (i.e., total number of disruptive messages / total number of messages) affects individual and group learning outcomes, as well as how the facilitators (i.e., researchers or teachers) intervention ratio (i.e., total number of facilitator messages / total number of messages) affects the amount of disruptive talk in student groups. Lastly, we analyze the distribution of words that appeared in both disruptive and non-disruptive talk in our dataset.

3.5 Automatic Disruptive Talk Detection

3.5.1 Feature Extraction. Our disruptive talk detection framework utilizes both the natural language features from student utterances as well as student attributes to determine how those features collectively contribute to prediction performance. First, we preprocess students’ utterances with NLTK for tokenizing utterances, removing stop words, lemmatization, and removing white space and punctuation [2]. In addition, we adopt a sentiment-aware tokenizing method to handle Twitter-like informal messages (e.g., emoticons, lengthening words) so we can capture the sentiment of the messages effectively [40]. The decision to use sentiment-aware tokenizing was motivated by the fact that a large portion of middle school students’ chat messages, especially the disruptive messages, are informal [40].

Next, we transform the tokenized chat message into a vector representation. We investigated three commonly used representations: (1) bag-of-words, which only considers the occurrence of words in the message, (2) static word embedding, which maps each word to a fixed, distributed vector representation, and (3) contextualized word embedding, which maps each word to a distributed vector representation that dynamically changes depending on the context. First, the bag-of-words approach creates a word dictionary containing words that are used in the training corpus only. Then each tokenized message is converted to a vector with the dimension of the number of words in the dictionary plus one additional dimension, representing the “unknown” word, which covers all words that are not present in the training dataset. This addresses the out-of-vocabulary issue that can emerge during the testing phase. The vector for the bag-of-words approach only contains zero or a positive number indicating how many times the specific word appears in the message. The number of distinct words across folds in our cross-validation evaluation ranges from 1,546 to 1,557. Second, we adopt word2vec as our static word embedding methods, which uses neural network models to represent words in a continuous vector space where semantically similar words are mapped to nearby points [32]. We employed a word embedding set that was trained with the Google news dataset, which consist of over 100 billion words and each word is represented in 300-dimensional latent features [32]. For the sentence embedding of each chat message in our work, we averaged each words’ embedding. Third, we adopt a state-of-the art contextualized word-embedding approach, Bidirectional Encoder Representations from Transformers (BERT) [8], which utilizes masked language models with self-attention mechanism to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context of input sentences through multiple layers. Pre-trained BERT or fine-tuned BERT models have been used in a wide range of NLP tasks including text classification and question answering, where they achieved state-of-the-art performance over other contextualized embeddings, such as ELMo and GPT [8, 11]. We adopted a pre-trained BERT model (i.e., BERT-base) that was trained with the Wikipedia dataset, which consists of 12 layers in the encoder with 110 million parameters and outputs 768 dimensional vectors for each word. For the sentence embedding of each chat message in our work, we used an output vector of the very first token (i.e., [CLS]), which is a special token inserted in front of the input sentence in the BERT architecture, rather than taking the average of the output embeddings of all the sentence words, since it effectively represents the essence of the input sentence and thus has been often used for BERT-based classification tasks [8]. It should be noted that we used the pre-trained BERT model without a fine-tuning process to ensure a fair comparison to the competitive methods based on bag-of-words and word2vec with respect to the representational capacity and impact on the performance of disruptive talk prediction. For the inputs for BERT models, we tokenized words into word pieces using standard BERT tokenizer.

Furthermore, we use additional linguistic features for the chat messages inspired by prior work in off-task behavior prediction [5]. First, we include the number of characters in the original chat message. We observed from the dataset that the disruptive messages tend to be shorter than the non-disruptive messages since students often try to annoy other team members by sending multiple short messages. Second, we include the Jaccard similarity [36] between the chat message and the game’s text content (e.g., virtual

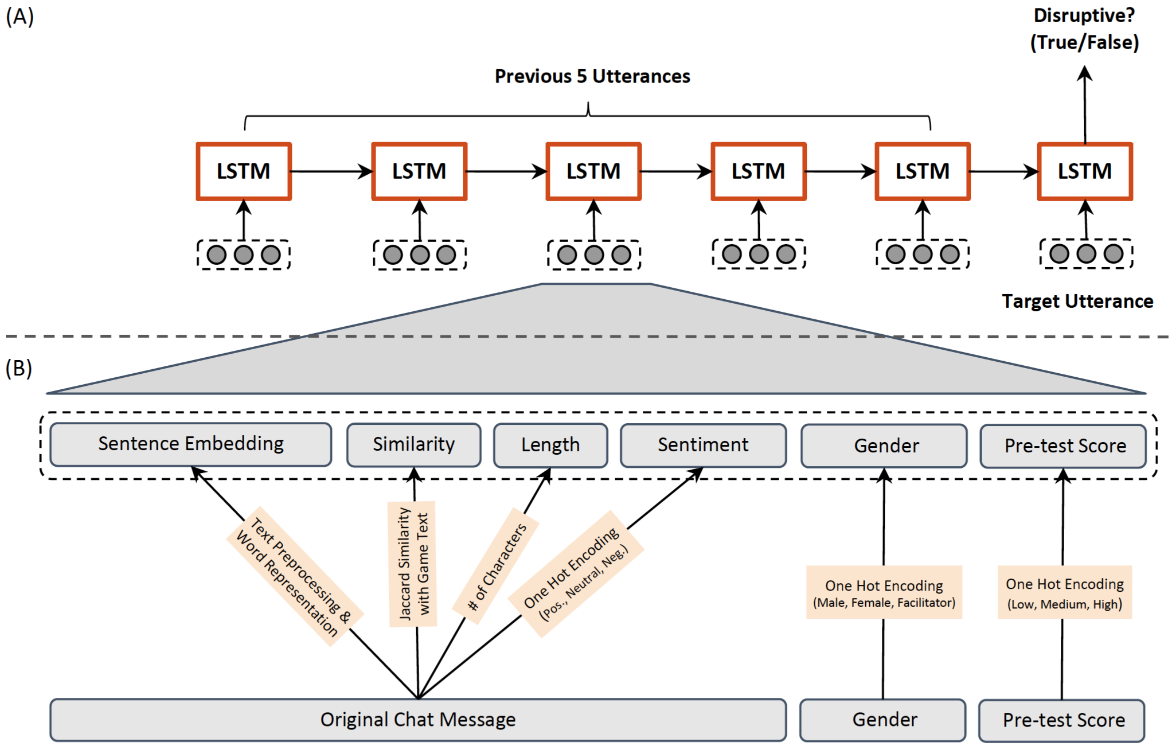


Figure 3: Illustrated LSTM-based disruptive talk detection model with a context length of 5. (A) In order to classify whether the target utterance is disruptive or not, previous 5 utterances are added sequentially to each LSTM cell. (B) Input vector of each LSTM cell consists of six parts; 1) Sentence embedding: each pre-processed chat message is converted to a vector representation based on the word representation choice (i.e., bag-of-words (BoW), word2vec, BERT), 2) Jaccard similarity between the original chat message and the game text content, 3) Original chat message length, 4) One-hot encoded sentiment of the chat message (Positive, Neutral, Negative), 5) One-hot encoded gender with an additional feature for the facilitator, and 6) One-hot encoded pre-test score based on the tertile split.

character dialogue, narration) to capture content similarity, since the messages that contain the same words as the game content are more likely to be non-disruptive. Third, we include the sentiment of the chat message. The sentiment of messages can serve as a predictor for the disruptive talk detection model since negative sentiment messages are more likely to be disruptive than positive messages. For measuring sentiment, we used a valence aware dictionary of sentiment reasoning (VADER) model [15] included in NLTK. We conjectured that these additional linguistic features might provide additional evidence for the disruptive talk detection models. In addition to the features used in off-task behavior prediction, we examined two additional features, students' gender and their ecosystems pre-test score. These features were chosen since previous studies suggested that gender and prior knowledge level can influence students' negative behavior patterns in CSCL environments [16, 26, 46]. Lastly, because disruptive talk is judged by the annotators considering a series of chat messages from the group, rather than a single message, we pre-process the data so that each data point retains the contextual information containing the history of previous messages from the group communication.

3.5.2 Modeling Disruptive Talk Detection. In order to identify the best performing model for detecting disruptive talk within collaborative game-based learning, we compare the models by varying 1) linguistic representations of chat messages: bag-of-words, word2vec, and BERT, 2) machine learning algorithms for classification: logistic regression, long short-term memory recurrent neural networks (LSTMs), and 3) the number of the previous messages utilized as context for classifying the current message: 5 or 20. The linguistic features introduced in the previous section (i.e., the number of characters, the content similarity, and the sentiment), as well as gender and pretest features, are used consistently for all models. Figure 3 shows our framework using a context length of 5.

LSTMs are a variant of recurrent neural networks (RNNs) designed to capture long-term sequential patterns, by overcoming conventional RNNs' vanishing and exploding gradient problems [19]. LSTM-based approaches have been introduced in a variety of computational sequence-labeling tasks, including speech recognition and machine translation, where LSTM-based approaches have shown state-of-the-art performance [42, 49]. As we described above, disruptive talk is highly dependent on the context of messages that dynamically changes across time. Thus, we hypothesize

Table 2: Descriptive statistics of average number of disruptive talk per gender and pre-test score.

		Number of Students	Disruptive Talk Distribution	Avg. # of Disruptive Talk
Gender	Male	23	70.3%	11.3
	Female	21	29.7%	5.2
	Low	11	44.6%	13.57
Pre-test Score	Medium	19	42.4%	7.85
	High	14	13.0%	3

that an LSTM-based approach will achieve higher performance than classical approaches, by capturing the latent sequential patterns embedded in the series of chat messages. We set the number of hidden units to 50, and the number of epochs to 20 while using early stopping with a patience of 5 to avoid overfitting. The models were trained and evaluated using 10-fold cross-validation.

Furthermore, to alleviate the data imbalance issue in our dataset (12.7% of disruptive, and 87.3% of non-disruptive messages), we adopt a random oversampling approach for the minority class (i.e., disruptive talk) to have a 50-50 distribution between the two labels for the training set in each fold. This oversampling approach has been demonstrated to be effective for achieving a higher recall rate [33], which is a relatively more important measure in disruptive talk detection (i.e., we desire to detect disruptive talk as accurately as possible). Please note that we did not oversample during either validation or testing to avoid data leakage. As a baseline model, we adopted logistic regression (LR), which has been successfully used in prior text classification work [5, 14, 23].

3.5.3 Evaluation. We evaluate the performance of the disruptive talk detection models on five evaluation metrics, accuracy, precision, recall, f1-score, and area under the ROC curve (AUC), which are typically used to assess the model's generalization performance. We apply data level 10-fold cross-validation after processing each data point to be a set of messages containing selected context length (i.e., 5 or 20) of previous utterances, and report the average of 10 folds' results for the final result. These settings were consistent across different modeling techniques and sentence representation approaches. It should be noted that we expect prediction models trained with the oversampling technique to achieve higher recall over precision since we expect the model to be trained to minimize the false negatives (i.e., missing disruptive talks that exist in the dataset), rather than false positives.

4 RESULT AND DISCUSSION

4.1 Disruptive Talk Analysis

4.1.1 Disruptive Talk distributions. Table 2 shows the overall distribution and the average number of individual student's disruptive talk across gender and pre-test scores based on tertile split (i.e., Low, Medium, High). Tertile split was adopted to create a balanced distribution among all classes while avoiding potential data sparsity issues. The distribution shows that male students are more likely to engage in disruptive talk (70.3%) than female students (29.7%). Furthermore, it also shows that students who achieved lower performance on their pre-test assessment tend to engage in more disruptive talk. These patterns suggest that students' gender and prior knowledge can serve as strong predictive features to detect whether a message is disruptive or not.

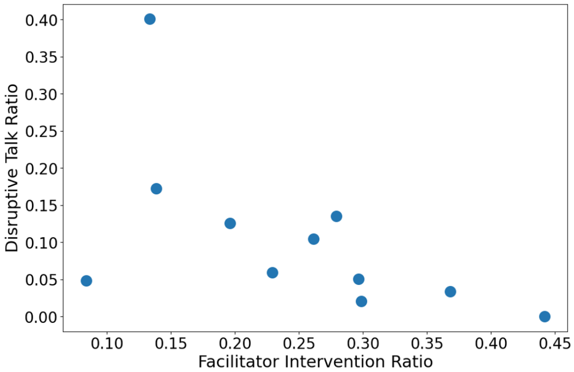


Figure 4: Facilitator intervention ratio (# of facilitator's chat messages / total # of messages), and the disruptive talk ratio (# of disruptive chat messages / total # of student messages).

4.1.2 Disruptive Talk and Facilitator Intervention Frequency. Figure 4 shows the distribution of the ratio of the facilitator's messages over the disruptive talk ratio of each group, which shows a moderate negative correlation (r = -0.56, p < .01). This indicates that an appropriate level of facilitator intervention is needed to reduce students' disruptive talk that could interfere with the group learning process. This also suggests that without proper facilitation, we would observe the groups exceeding their tolerance. The automatic disruptive talk detection framework can 1) help teachers by informing them when to support students by identifying disruptive moments, 2) assist teachers to preemptively guiding students to avoid engaging in disruptive talk, and 3) play the role of the facilitator by providing students with adaptive feedback when disruptive talk is detected.

4.1.3 Disruptive Talk and Individual/Group Learning Outcomes. Figure 5 and Figure 6 shows the distribution of disruptive utterances of each group over the individual and group learning gains calculated by averaging individual learning gains. Figure 5 (Individual learning gain) and Figure 6 (Group learning gain) show that there is no strong relationship between the disruptive talk ratio and the individual and group learning gain (r = -0.06, p < .001 and r = -0.18, p < .001, respectively). However, it should be noted that the groups had human facilitators who monitored the chat and intervened

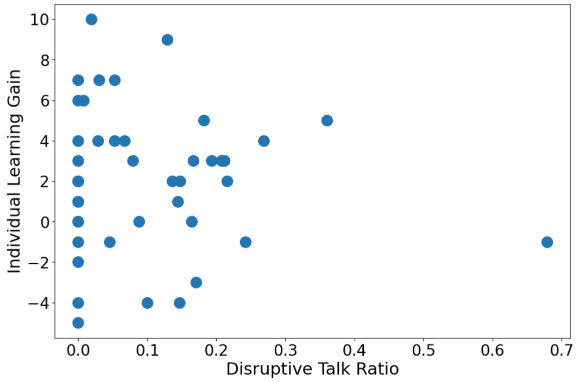


Figure 5: Individual learning gain over disruptive talk ratio of each group (# of disruptive chat messages / total # of student messages).

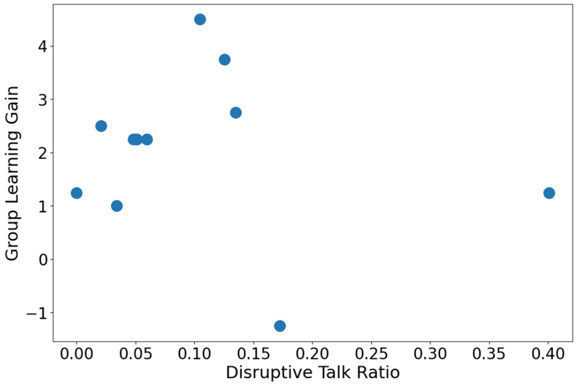


Figure 6: Group learning gain over disruptive talk ratio of each group (# of disruptive chat messages / total # of student messages).

continuously throughout the learning process. Thus, as discussed in previous section (i.e., 4.1.2), the presence of the facilitation might have helped students from being distracted by disruptive talk. On the other hand, this weak correlation might also suggest that in some groups, students independently decide on what is an acceptable level of disruption and then allowed those disruptions to occur, or they agreed to ignore it and focus on the assigned task, meaning that students themselves can effectively manage disruptions.

4.1.4 Linguistic Analysis of Disruptive Talk. We analyze the distribution of words that appeared in both disruptive and non-disruptive talk. Table 3 shows the top 15 discriminative unigrams of disruptive and non-disruptive messages, obtained based on the word frequencies. The set of top 15 words from the non-disruptive talk mostly contains game-related words (e.g., “note”, “column”, “food”). This suggests that a large part of non-disruptive messages is task related. On the other hand, the set of top 15 words in disruptive messages contains several negative words (e.g., “boo”, “nopy”), while none of these words were used in the content of the game. Although “chat” and “enter” are highly ranked in the disruptive category they are seemingly non-disruptive, we observed from the dataset that students disturbed other group members by typing “[entered chat]” constantly, which is also the message that the system generates whenever a student enters a message. The nicknames ranked in positions 11 and 12 of disruptive talk were observed from the dataset where one student kept annoying another student by calling them with unwanted nicknames. Although we rarely observed bullying-related words in our dataset, student conversation can be detrimental depending on the students in a group. In such situations, our disruptive talk detection model can be especially beneficial for enhancing both the student behavior and the learning experience of group members.

Table 3: Top words from disruptive and non-disruptive talk.

Rank	Disruptive Talk	Non-Disruptive Talk
1	“chat”	“note”
2	“enter”	“think”
3	“name”	“water”
4	“like”	“yes”
5	“boo”	“need”
6	“brother”	“column”
7	“nopy”	“food”
8	“yes”	“okay”
9	“nope”	“space”
10	“do”	“air”
11	Nickname1	“one”
12	Nickname2	“let”
13	“exit”	“fish”
14	“beep”	“say”
15	“exit”	“know”

4.2 Automatic Disruptive Talk Detection

Table 4 shows the results of 12 different models for detecting disruptive talk. Each value is the average of the results from the 10-fold cross-validation. Overall, models using LSTMs performed better than the ones with the logistic regression models suggesting that LSTMs effectively model the long-term dependency in the sequence of chat messages. The model using bag-of-words and logistic regression with the context length with 5, and word2vec+LR with the context length of 20 achieved comparative accuracy with ones with LSTMs but achieved poor performance on the other metrics. This might be because of the imbalanced dataset, where the model still can achieve high predictive accuracy if it predicts most messages as non-disruptive (i.e., the majority class; 87.3%). With respect to the number of previous chat messages (i.e., context), models with a shorter message sequence length (i.e., 5 previous chat messages) achieved worse predictive results than the ones with the longer message sequence length (i.e., 20 previous chat messages), which indicates that modeling more conversational history is beneficial for the model to determine whether the current message is disruptive or not. Furthermore, performance of models using BERT contextualized word embeddings generally outperform the models using word2vec embeddings. These results suggest that considering

Table 4: Prediction results of automatic disruptive talk detection models across language representations (bag-of-words (BoW), word2vec, and BERT), context window length (5, 20), and classifiers (LSTM, logistic regression (LR). The best performance of each evaluation metric is marked in bold.

Model	Input	Context Length	Accuracy	Precision	Recall	F1-Score	AUC
LSTM	BERT	5	0.887	0.564	0.605	0.578	0.877
LSTM	BoW	5	0.910	0.640	0.716	0.670	0.884
LSTM	word2vec	5	0.825	0.381	0.562	0.447	0.815
LSTM	BERT	20	0.912	0.637	0.722	0.674	0.922
LSTM	BoW	20	0.922	0.682	0.746	0.710	0.933
LSTM	word2vec	20	0.887	0.544	0.722	0.617	0.906
LR	BERT	5	0.752	0.271	0.568	0.366	0.749
LR	BoW	5	0.840	0.401	0.522	0.449	0.802
LR	word2vec	5	0.679	0.239	0.697	0.355	0.737
LR	BERT	20	0.768	0.259	0.443	0.325	0.712
LR	BoW	20	0.686	0.228	0.619	0.333	0.697
LR	word2vec	20	0.844	0.390	0.425	0.404	0.772

the meaning of words, that can vary in different contexts, is helpful to predict disruptive talk.

The best performance was achieved by the model using LSTM and bag-of-words across all evaluation metrics which might indicate that simply knowing what words appeared in the message is more useful in detecting disruptive talk, than considering the syntactic and semantic relationship between words. A contributing factor could also be that the student chat data examined in this work contains domain-specific language as well as various colloquial or misspelled words, which are not likely to be effectively captured by the word2vec pre-trained word embeddings or the BERT-models that were pre-trained with either news articles or Wikipedia, where colloquial or misspelled words rarely appear. This finding is echoed in [33, 43], which found that pre-trained word embedding models (e.g., BERT, GloVe [37]) exerted a detrimental impact on models' predictive performance compared to equivalent models utilizing bag-of-words or *n*-gram-based language features that directly capture domain-specific characteristics in language data. Overall, the prediction results suggest that LSTMs are an effective modeling technique for disruptive talk detection, while the bag-of-words representation can sufficiently capture salient features characterized in disruptive talks that appear in student dialogue. LSTMs were capable of modeling the temporal context of the conversation data, which can significantly improve the predictive accuracy for disruptive talk prediction.

5 CONCLUSION

Collaborative game-based learning environments can offer students engaging group-based learning experiences in immersive virtual environments. However, collaborative game-based learning, which elicits communication between students as they work toward achieving shared goals, often creates situations in which students are engaged in behaviors that are not conducive to effective learning. In this work, we analyzed disruptive talk within a collaborative game-based learning environment for middle school ecology science learning. We presented an automatic disruptive talk detection framework that utilizes LSTMs adopting one of three language representation techniques (bag-of-words, word2vec, or BERT), features extracted from the utterances (i.e., Jaccard similarity between the chat message and the game's text content, message length, and sentiment), and supplementary information about students (i.e., gender and prior knowledge).

Results on a dialogue corpus obtained from 44 middle-grade students' chat data in a collaborative game-based learning environment suggest that male students and students with low pre-test scores are likely to engage in disruptive talk more often during the learning process. Furthermore, proper facilitation is helpful in reducing disruptive talk among groups, which can be supported by our disruptive talk detection framework. The weak correlation between the disruptive talk and learning outcomes, which might have stemmed from either facilitators' or students' abilities to manage the disruptive situations, suggests a second use case for disruptive talk models; encouraging students to discuss with group members to manage disruptions. The findings on the automatic disruptive talk detection models indicate that LSTM-based disruptive talk detection accurately predicts disruptive talk with an accuracy rate of 0.92 and a recall rate of 0.75, suggesting that they offer potential for analyzing and supporting effective collaborative learning.

There are several promising directions for future work. It will be important to explore the framework with longer histories of messages since results have suggested that more context helps improve the models' predictive performance. Another promising direction is investigating approaches to enhancing the framework with contextualized word embeddings, such as transfer learning, by which pre-trained embeddings can be fine-tuned to better characterize the domain-specific nature of a corpus. In addition, it will be instructive to investigate an extended set of features that include student collaborative gameplay actions, which may provide additional contextual information to further improve disruptive talk detection performance.

ACKNOWLEDGMENTS

This research was supported by the National Science Foundation under Grants DRL-1561655, DRL-1561486, IIS-1839966, and SES-1840120. Any opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] Hua Ai, Marietta Sionti, Yi-Chia Wang, and Carolyn Penstein Rosé. 2010. Finding transactive contributions in whole group classroom discussions. In Proceedings of the 9th International Conference of the Learning Sciences - Volume 1 (ICLS '10). International Society of the Learning Sciences, 976–983.

[2] Steven Bird, Ewan Klein, and Edward Loper. 2009. Natural language processing with Python: analyzing text with the natural language toolkit. O'Reilly Media, Inc.

[3] Gian Barbosa, Raissa Camelo, Anderson Pinheiro Cavalcanti, Péricles Miranda, Rafael Ferreira Mello, Vitomir Kovanović, and Dragan Gašević. 2020. Towards automatic cross-language classification of cognitive presence in online discussions. In Proceedings of the Tenth International Conference on Learning Analytics & Knowledge (LAK '20). Association for Computing Machinery, New York, NY, USA, 605–614. https://doi.org/10.1145/3375462.3375496

[4] Marcela Borge, and Emma Mercier. 2019. Towards a micro-ecological approach to CSCL. International Journal of Computer-Supported Collaborative Learning 14, 2 (2019): 219-235.

[5] Dan Carpenter, Andrew Emerson, Bradford W. Mott, Asmalina Saleh, Krista D. Glazewski, Cindy E. Hmelo-Silver, and James C. Lester. 2020. Detecting off-task behavior from student dialogue in game-based collaborative learning. In Proceedings of the Twenty-First International Conference on Artificial Intelligence in Education, 55-66.

[6] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. Educational and psychological measurement 20, 1 (April 1960): 37-46. https://doi.org/10.1177/001316446002000104.

[7] Mihai Dascalu, Danielle S. McNamara, Stefan Trausan-Matu, and Laura K. Allen. 2018. Cohesion network analysis of CSCL participation. Behavior Research Methods 50, 2 (2018): 604-619.

[8] 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.

[9] Pierre Dillenbourg, Sanna Järvelä, and Frank Fischer. 2009. The evolution of research on computer-supported collaborative learning. In Technology-enhanced learning, 3-19. Springer, Dordrecht, https://doi.org/10.1007/978-1-4020-9827-7_1.

[10] Randi A. Engle, and Faith R. Conant. 2002. Guiding principles for fostering productive disciplinary engagement: Explaining an emergent argument in a community of learners classroom. Cognition and instruction 20, 4 (2002): 399-483.

[11] Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. arXiv preprint arXiv:1909.00512.

[12] Máverick Ferreira, Vitor Rolim, Rafael Ferreira Mello, Rafael Dueire Lins, Guanliang Chen, and Dragan Gašević. 2020. Towards automatic content analysis of social presence in transcripts of online discussions. In Proceedings of the Tenth International Conference on Learning Analytics & Knowledge (LAK '20). Association for Computing Machinery, New York, NY, USA, 141–150. https://doi.org/10.1145/3375462.3375495

[13] Ella L. F. Fu, Jan van Aalst, and Carol K. K. Chan. 2016. Toward a classification of discourse patterns in asynchronous online discussions. International Journal of Computer-Supported Collaborative Learning 11, 4 (2016): 441-478. https://doi.org/10.1007/s11412-016-9245-3

[14] Shay A. Geller, Nicholas Hoernle, Kobi Gal, Avi Segal, Amy X. Zhang, David Karger, Marc T. Facciotti, and Michele Igo. 2020. #Confused and beyond: detecting confusion in course forums using students' hashtags. In Proceedings of the Tenth International Conference on Learning Analytics & Knowledge (LAK '20). Association for Computing Machinery, New York, NY, USA, 589–594. https://doi.org/10.1145/3375462.3375485

[15] C.J. Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In Eighth International Conference on Weblogs and Social Media (ICWSM-14).

[16] Jane Guiller, and Alan Durndell. 2007. Students' linguistic behaviour in online discussion groups: Does gender matter?. Computers in Human Behavior 23, 5 (2007): 2240-2255. https://doi.org/10.1016/j.chb.2006.03.004

[17] Gahgene Gweon, Mahaveer Jain, John McDonough, Bhiksha Raj, and Carolyn P. Rosé. 2013. Measuring prevalence of other-oriented transactive contributions using an automated measure of speech style accommodation. International Journal of Computer-Supported Collaborative Learning 8, 2 (2013): 245-265. https://doi.org/10.1007/s11412-013-9172-5

[18] Cindy E. Hmelo-Silver. 2004. Problem-based learning: What and how do students learn?. Educational psychology review 16, 3 (2004): 235-266. https://doi.org/10.1023/B:EDPR.0000034022.16470.f3

[19] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. Neural Computation. 9, 8 (November 15, 1997): 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[20] Jeroen Janssen, Gijsbert Erkens, and Gellof Kanselaar. 2007. Visualization of agreement and discussion processes during computer-supported collaborative learning. Computers in Human Behavior 23, 3 (2007): 1105-1125.

[21] Karen A Jehn. 1997. A qualitative analysis of conflict types and dimensions in organizational groups. Administrative science quarterly 42, 3 (1997): 530-557. Accessed November 2, 2020. doi:10.2307/2393737.

[22] Heisawn Jeong, Cindy E. Hmelo-Silver, and Kihyun Jo. 2019. Ten years of computer-supported collaborative learning: A meta-analysis of CSCL in STEM education during 2005-2014. Educational Research Review 28 (2019): 100284.

[23] Kamran Kowsari, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes, and Donald Brown. 2019. Text classification algorithms: A survey. Information 10, 4 (2019): 150.

[24] Thomas K. Landauer, Darrell Laham, and Peter Foltz. 2003. Automated scoring and annotation of essays with the Intelligent Essay Assessor. In M. D. Shermis & J. Burstein (Eds.), Automated essay scoring: A cross-disciplinary perspective (87–112). Mahwah, NJ: Lawrence Erlbaum Associates.

[25] Dabae Lee, Yeol Huh, and Charles M. Reigeluth. 2015. Collaboration, intragroup conflict, and social skills in project-based learning. Instructional science 43, 5 (2015): 561-590.

[26] Jonna Malmberg, Sanna Järvelä, Hanna Järvenoja, Ernesto Panadero. 2015. Promoting socially shared regulation of learning in CSCL: Progress of socially shared regulation among high-and low-performing groups. Computers in Human Behavior 52 (2015): 562-572. https://doi.org/10.1016/j.chb.2015.03.082.

[27] Danielle S. McNamara, L. Allen, S. Crossley, Mihai Dascalu, and Cecile A. Perret. 2017. Natural language processing and learning analytics. Handbook of learning analytics (2017): 93-104.

[28] Danielle S. McNamara, Max M. Louwerse, Philip M. McCarthy, and Arthur C. Graesser. 2010. Coh-Metrix: Capturing linguistic features of cohesion. Discourse Processes 47, 4 (2010): 292-330.

[29] Danielle S. McNamara, Scott A. Crossley, and Rod Roscoe. 2013. Natural language processing in an intelligent writing strategy tutoring system. Behavior research methods 45, 2 (2013): 499-515.

[30] Neil Mercer. 2010. The analysis of classroom talk: Methods and methodologies. *British journal of educational psychology* 80, 1 (2010): 1-14.

[31] Neil Mercer. 1996. The quality of talk in children's collaborative activity in the classroom. Learning and instruction 6, 4 (1996): 359-377.

[32] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems, 3111-3119.

[33] Wookhee Min, Kyungjin Park, Joseph Wiggins, Bradford Mott, Eric Wiebe, Kristy Elizabeth Boyer, and James Lester. 2019. Predicting dialogue breakdown in conversational pedagogical agents with multimodal LSTMs. In International Conference on Artificial Intelligence in Education. Springer, Cham, 195-200.

[34] Jonas Mueller, and Aditya Thyagarajan. 2016. Siamese recurrent architectures for learning sentence similarity. In thirtieth AAAI conference on artificial intelligence 30, 1.

[35] Stefanos Nikiforos, Spyros Tzanavaris, and Katia-Lida Kermanidis. 2020. Virtual learning communities (VLCs) rethinking: Collaboration between learning communities. Education and Information Technologies (2020): 1-17.

[36] Suphakit Niwattanakul, Jatsada Singthongchai, Ekkachai Naenudorn, and Supachanun Wanapu. 2013. Using of Jaccard coefficient for keywords similarity. In Proceedings of the international multiconference of engineers and computer scientists 1, 6, 380-384.

[37] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 1532-1543.

[38] Lisa Hope Pelled, Kathleen M. Eisenhardt, and Katherine R. Xin. 1999. Exploring the black box: An analysis of work group diversity, conflict and performance. Administrative science quarterly 44, 1 (1999): 1-28. Accessed November 2, 2020. doi:10.2307/2667029.

[39] Chris Phielix, Frans J. Prins, and Paul A. Kirschner. 2010. Awareness of group performance in a CSCL-environment: Effects of peer feedback and reflection. *Computers in Human Behavior* 26, 2 (2010): 151-161.

[40] Christopher Potts. 2011. Sentiment symposium tutorial: Lexicons. http://sentiment.christopherpotts.net/lexicons.html (2011).

[41] Lauren B. Resnick, Sarah Michaels, and Catherine O'Connor. 2010. How (well structured) talk builds the mind. Innovations in educational psychology: Perspectives on learning, teaching and human development (2010): 163-194.

[42] Jürgen Schmidhuber. 2015. Deep learning in neural networks: An overview. Neural networks 61 (2015): 85-117. https://doi.org/10.1016/j.neunet.2014.09.003.

[43] Shree Krishna Subburaj, Angela E.B. Stewart, Arjun Ramesh Rao, and Sidney K. D'Mello. 2020. Multimodal, multiparty modeling of collaborative problem solving

performance. In Proceedings of the 2020 International Conference on Multimodal Interaction (ICMI '20). Association for Computing Machinery, New York, NY, USA, 423–432. https://doi.org/10.1145/3382507.3418877

[44] Chul Sung, Tejas Indulal Dhamecha, and Nirmal Mukhi. 2019. Improving short answer grading using Transformer-based pre-training. In International Conference on Artificial Intelligence in Education. Springer, Cham, 469-481. Lecture Notes in Computer Science, vol. 11625 https://doi.org/10.1007/978-3-030-23204-7_39.

[45] Md Arafat Sultan, Cristobal Salazar, and Tamara Sumner. 2016. Fast and easy short answer grading with high accuracy. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1070-1075.

[46] M. Tomai, M. E. Mebane, V. Rosa, and M. Benedetti. 2014. Can computer supported collaborative learning (CSCL) promote counter-stereotypical gender communication styles in male and female university students?. Procedia-Social and Behavioral Sciences 116 (2014): 4384-4392.

[47] Stefan Trausan-Matu, Mihai Dascalu, and Traian Rebedea. 2014. PolyCAFe-automatic support for the polyphonic analysis of CSCL chats. International Journal of Computer-Supported Collaborative Learning 9.2 (2014): 127-156.

[48] Viktor Wendel, and Johannes Konert. 2016. Multiplayer serious games. In Serious Games, Springer, Cham, 211-241. https://doi.org/10.1007/978-3-319-40612-1_8

[49] Yong Yu, Xiaosheng Si, Changhua Hu, and Jianxun Zhang. 2019. A review of recurrent neural networks: LSTM cells and network architectures. In Neural Computation 31, 7 (July 2019), 1235-1270. https://doi.org/10.1162/neco_a_01199