Disruptive Talk Detection in Multi-Party Dialogue within Collaborative Learning Environments with a Regularized User-Aware Network

Kyungjin Park[1], Hyunwoo Sohn[1], Wookhee Min[1], Bradford Mott[1], Krista Glazewski[2], Cindy E. Hmelo-Silver[2], and James Lester[1]
[1]Department of Computer Science, North Carolina State University
[2]Center for Research on Learning and Teaching, Indiana University Bloomington
[1]{kpark8, hsohn3, wmin, bwmott, lester}@ncsu.edu
[2]{glaze, chmelosi}@indiana.edu

Abstract

Accurate detection and appropriate handling of disruptive talk in multi-party dialogue is essential for users to achieve shared goals. In collaborative game-based learning environments, detecting and attending to disruptive talk holds significant potential since it can cause distraction and produce negative learning experiences for students. We present a novel attention-based user-aware neural architecture for disruptive talk detection that uses a sequence dropout-based regularization mechanism. The disruptive talk detection models are evaluated with multi-party dialogue collected from 72 middle school students who interacted with a collaborative game-based learning environment. Our proposed disruptive talk detection model significantly outperforms competitive baseline approaches and shows significant potential for helping to support effective collaborative learning experiences.

1 Introduction

Automatic analysis of dyadic dialogue utilizes a broad range of methods for intent recognition (Ahmadvand et al., 2019; Grau et al., 2004; Kim et al., 2010; Maraev et al., 2021). Compared to dyadic conversations, multi-party conversations are characterized by a high degree of complexity due to multi-way group interactions, thus, multi-party dialogue models should take into account group dynamics to reliably model phenomena. For example, previous research investigated giving less weight to participants whose convergence behaviors differ from the rest of the group (Rahimi and Litman, 2018) to examine which utterances should be clustered together (i.e., conversation threads) in multi-party dialogues (Mayfield et al., 2012; Tan et al., 2019).

In education, computer-supported collaborative learning environments promote social aspects of learning through the use of a variety of technological and constructive pedagogical strategies, including problem-based learning and inquiry learning (Dillenbourg et al., 2009; Hmelo-Silver, 2004; Jeong et al., 2019). Collaborative game-based learning environments often provide students with in-game chat features to help promote open discussion and negotiation among team members, facilitating the coordination of their in-game learning activities (Saleh et al., 2021). However, students are not always effective collaborators and may engage in improper communicative behavior, distracting from the group learning experience. The presence of negative socio-emotional engagement in collaborative learning environments can result in disruptive talk and can function as a barrier to the development of high-quality collaborative communication.

Previous work on detecting talk that can cause negative socio-emotional engagement (e.g., off-task behavior, bullying, disruptive talk) in collaborative learning environments investigated computational approaches using language models ranging from classic approaches (e.g., *n*-grams) and word embedding approaches (e.g., BERT). These language models have been combined with classic techniques (e.g., logistic regression, random forest) and deep learning techniques (e.g., long-short term memory networks) (Carpenter et al., 2020; Nikiforos et al., 2020; Park et al., 2021). However, the previous work either makes utterance-by-utterance predictions without taking

context into account or treats the entire multi-party conversation sequence as a continuous dialogue flow, despite the potential presence of multiple concurrent message threads with different topics.

In this paper, we propose a novel attention-based, regularized user-aware modeling approach for detecting disruptive talk in multi-party dialogue within a collaborative game-based learning environment. We investigate the use of target-user embeddings to help the prediction model determine the disruptiveness of the sequence more accurately with an additional user-specific network and attention mechanism. We also investigate a sequence-level dropout mechanism during training as a regularization technique that could help avoid overfitting possible diluted conversation sequences (i.e., presence of multiple threads in a sequence) in training data. Experimental results demonstrate that our attention-based, regularized user-aware model offers great potential for addressing disruptive talk detection in multi-party dialogues.

2 Related Work

Diverse prediction tasks have analyzed multi-party dialogue focusing on the asynchronous and entangled nature of group conversations, such as dialogue act classification using group thread history, and thread detection as well as cyberbullying and toxic message detection within group conversations (Anikina and Kruijff-Korbayova, 2019; Blackburn and Kwak, 2014; Ekiciler et al., 2021; Kim et al., 2012; Min et al., 2021; Tan et al., 2019).

Kim et al. (2012) investigated classic machine learning approaches for dialogue act classification, such as Naïve Bayes, support vector machines, and conditional random fields, along with contextual, structural, keyword, and dialogue interaction-based features of utterances for dialogue act classification in multi-party live chat datasets. As a sub-task of a disaster response mission knowledge extraction task, Anikina and Kruijff-Korbayova (2019) proposed a deep learning-based Divide&Merge architecture utilizing LSTM and CNN for predicting dialogue acts. Min et al., (2021) investigated the use of dialogue act prediction utilizing conditional random fields and ELMo contextualized word embeddings in multi-party team communication for providing adaptive team training support.

As multiple participants are involved in multi-party conversation, disentanglement of the conversation based on relevancy is another important task, which could enhance the conversational relevance rate of automated dialogue agents (Shamekhi et al., 2018) or improve summarization quality (Zhang and Cranshaw, 2018). Tan et al. (2019) proposed three LSTM-based context-aware thread detection architectures that automatically captures conversation threads in multi-party and multi-thread conversations, where the proposed model predicts which existing thread the current utterance belongs to (or whether it creates a new thread).

Another task that has received considerable attention in multi-party conversation is cyberbullying. The ability to detect bullying or toxic behavior is crucial to protecting users from cyberbullying. In particular, researchers are increasingly interested in toxic behavior in multiplayer games, such as multiplayer online battle arena (MOBA) games, where players compete against other teams in virtual online game environments (Kordyaka 2018). Blackburn and Kwak (2014) used random forest classifiers to detect toxic behavior in League of Legends using in-game performance, user reports, and chat data. The conversation data included 590,000 utterances, which were labeled via crowdsourcing on whether the conversation was toxic or not. Ekiciler et al. (2021) presented a linguistic analysis of gender-based toxic language usage in a Dota 2 chat dataset and investigated Naïve Bayes classifiers with three different Laplace smoothing parameters as an automatic approach for sexist toxic comment detection. A significant presence of gender discrimination in online games, mainly by young males and intense players, was revealed in their qualitative analysis.

Students' conversations can create disruption in collaborative learning environments, impeding collaborative learning processes. Recent research on bullying, off-task behavior, and disruptive talk in collaborative learning environments examined a range of word embedding techniques as well as a variety of classical machine learning and deep learning techniques (Carpenter et al., 2020; Nikiforos et al., 2020; Park et al., 2021). Nikiforos et al. (2020) explored the automatic detection of aggressive behavior (i.e., bullying) in two K-12 computer-supported collaborative learning environments. They used unigrams to represent words and examined machine learning approaches such as Naïve Bayes with Laplace smoothing,

decision tree classifiers, and feedforward neural networks. The prediction results suggest that approaches based on deep learning outperform other classical machine learning approaches. Carpenter et al. (2020) used dialogue analysis to identify if students' messages were on-task or off-task during collaborative game-based learning. To develop a model capable of reliably detecting off-task behavior, they investigated three different word embedding approaches (i.e., Word2Vec, ELMo, and BERT), various history lengths of previous utterances, and two deep learning and classical machine learning classifiers were trained on a feature set containing contextual information extracted from student chat messages. The empirical evaluations indicated that the LSTM-based off-task behavior detection model with BERT embeddings outperformed other baseline approaches. Park et al. (2021) presented an LSTM-based disruptive talk detection framework in a multi-party dialogue dataset from a collaborative game-based learning environment, utilizing features from chat messages, a range of linguistic features, gender, and pre-test scores. While this work has the potential to improve learning experiences by detecting disruptions within collaborative learning settings, they disregard the unique characteristics of multi-party dialogues. In our work, we improve predictive performance of disruptive talk detection models by incorporating an additional network that embeds the characteristics of the user of a target utterance and a sequence-level dropout mechanism.

3 Corpus

We next describe the collaborative game-based learning environment and its chat-interface, dataset collected from two field studies, and disruptive talk annotation process.

3.1 ECOJOURNEYS Collaborative Game-Based Learning Environment

ECOJOURNEYS is a collaborative game-based learning environment for middle school science education focused on ecosystems (Mott et al., 2019; Saleh et al., 2019) (Figure 1). Students visit a virtual island in the game-based learning environment and are tasked with determining what is causing a mysterious illness among the island's fish population. Students work in groups of four to solve the mystery within the game, where each student works on a different laptop and interacts

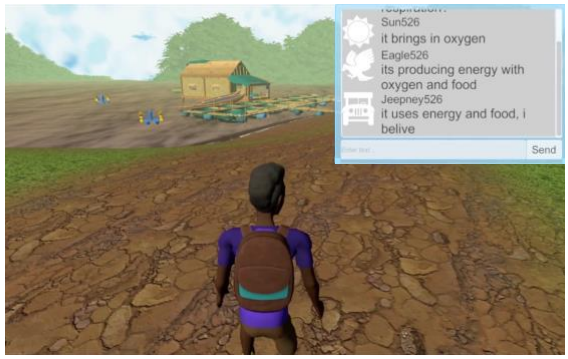

Figure 1: ECOJOURNEYS collaborative game-based learning environment and its in-game chat interface.

with peers in the virtual game environment. Individual students examine the fish illness during gameplay by collecting information and interacting with virtual characters. The virtual non-player characters serve as local experts, providing context for ecosystem concepts and the unfolding narrative (e.g., "Dissolved oxygen is a non-living component that animals and plants require to survive."). After investigating and gathering information, students meet at a virtual whiteboard within the game to share and categorize the information they have gathered and to discuss the most likely cause of the illness. Students are encouraged to exchange ideas, ask questions, and negotiate with their team members during the game's problem-solving activities using the in-game chat interface (Figure 1). This built-in chat system is accessible throughout the game. Each group is led by a facilitator, who is either a researcher or a teacher. The facilitator asks questions and encourages students to communicate with one another using the in-game chat interface. Facilitators can monitor and intervene on students' activities and conversations using an in-game screen, available only for facilitators, to guide students' learning. Facilitators can choose messages from a set of pre-written messages or write free-form messages using the in-game chat interface.

3.2 Dataset

The ECOJOURNEYS collaborative game-based learning environment was used in two classroom-based studies. Students were either in the sixth or seventh grade (11-13 years old) and played ECOJOURNEYS during six classroom periods. In total, 21 groups with 84 students (4 students per group) were involved in the two studies. From the 21 groups, the current work utilizes data from 18

groups consisting of 72 students (31 female and 41 male) who consented to the study and completed all the activities in the collaborative game-based learning environment. There are 9,236 chat messages available in the resulting dataset, with 2,440 messages from facilitators and 6,796 messages from students. We only consider the students' messages during the disruptive talk detection modeling working under the assumption that facilitators would not produce disruptive talk. On average, students in each group sent 382.4 messages (min = 89, max = 900, SD = 229.7).

3.3 Disruptive Talk Annotation

Adapted from prior work on disruptive talk analysis, the present work adopted a binary annotation scheme, *disruptive talk*, and *non-disruptive talk*, (Borge and Mercier, 2019). We labeled student utterances as disruptive talk if it had the potential to distract other group members from learning (e.g., "Um yea. yep, you can't work", "I WILL HAVE A MENTAL BREAKDOWN") and to interfere with deeper learning by interrupting the learning activity repeatedly (e.g., sending emojis multiple times). Otherwise, we labeled the utterance as non-disruptive talk.

Two human annotators labeled the students' chat-based dialogue collected during the study. Approximately 20% of the corpus was labeled by both annotators and an inter-rater agreement of 0.80 was achieved using Cohen's Kappa, indicating substantial agreement among the annotators (Cohen, 1960). All utterances labeled differently between the two annotators were discussed, and agreement was reached for certain situations without changing the high-level definition of disruptive talk we defined above. An example of those situations is when students exchange non-task-related messages, seemingly disruptive, before everyone is logged on and before starting the game, we agreed to label them as non-disruptive. A label was chosen for each utterance for which there was disagreement before proceeding with labeling the remainder of the corpus. Then, the remaining utterances were split in half and independently labeled by the annotators (approximately 40% each). The distribution of disruptive and non-disruptive utterances among the dataset was determined to be 1,864 (27.4%) and 4,932 (72.6%), respectively.

4 Method

4.1 Data Pre-Processing

The disruptive talk detection framework in our previous work utilizes linguistic features from student utterances and student attributes (i.e., gender and prior knowledge level) to determine how those features collectively contribute to prediction performance (Park et al., 2021). Here we keep all feature combinations from our previous work (i.e., sentence embedding, sentiment, Jaccard similarity between utterance and game text, gender, and pre-test scores) with an additional text cleaning pass that can be helpful for dealing with informal chat messages (Table 1).

We adopt a pre-trained BERT model, DistilBERT, a distilled version of BERT, that is a small, fast, and light Transformer model (Sanh et al., 2019). DistilBERT consists of 6 layers in the encoder with 40% fewer parameters than the BERT-base model and outputs 768-dimensional vectors for each word. We utilized a DistilBERT model that was trained on the Wikipedia dataset. For the sentence embedding, rather than taking the average of the embeddings of all the sentence words, we used the first token (i.e., [CLS]), a special token inserted in front of the input sentence in the BERT architecture, as it effectively represents what is in the input sentence and thus has been frequently used for BERT-based classification tasks (Devlin et al., 2018).

Approach	Example	Cleaned
Removed lengthening words[1]	"Hellllllllo"	"Hello"
Replaced slang[2]	"dis", "k"	"This", "Ok"
Spelling correction[3]	"who dat"	"Who that?"
Replaced abbreviated words	"don't", "can't"	"do not", "cannot"
Replaced emoji[4]	":-)"	"Happy"

Table 1: Text cleaning approaches.

[1] http://sentiment.christopherpotts.net/lexicons.html
[2] https://www.computerhope.com/jargon/c/chatslan.htm
[3] https://github.com/Azd325/gingerit
[4] https://pc.net/emoticons/

4.2 Attention-Based Regularized User-Aware Disruptive Talk Detection Modeling

When it comes to predicting disruptive talk based on the current message and a series of previous utterances, separately modeling the characteristics of the target user-specific utterances could be more effective than only utilizing messages from all group members equally; if a student makes a disruptive utterance, there is a higher chance that the same student will generate more disruption than the other group members. We propose an attention-based user-aware network that incorporates a target user-specific network that embeds the utterance histories of the target user as well as a separate network for modeling group-level utterances. We also apply the attention mechanism adapted from Bahdanau et al., (2014) to this output user representation and the hidden states of each time stamp to give weights to the group sequence output based on the user characteristics. An illustration of this attention-based user-aware network is shown in Figure 2

Suppose m_j^i is the feature embedding for the j^{th} message from user i in the n number of group utterance history, $Group_{Seq}$, including the current message.

$$Group_{Seq} = \{m_1^1, m_1^2, m_2^1, m_2^2, m_1^3, \cdots, m_j^i\}_{i=1,\cdots,4}$$

From this group sequence, we have user sequence $User_{seq}^i$ that only includes the utterances from user i.

$$User_{seq}^i = \{m_1^i, m_2^i, \cdots, m_j^i\}$$

The user network takes the utterance sequence, $User_{Seq}^i$, form the target-user only, then outputs a user embedding, $User_{emb}^i$.

$$User_{emb}^i = LSTM\left(User_{seq}^i\right)$$

We can get the attention score between the user embedding and the hidden state, h_t, at each LSTM time stamp of the group sequence.

$$\alpha_t = Softmax(User_{emb}^i \cdot h_t)_{t=1\ldots n}$$

Using this attention score, we can get the user-aware group sequence embedding.

$$Group_{emb} = \sum_{t=1}^{n} \alpha_t * h_t$$

Finally, we get the output probability by using a sigmoid function that takes as input $User_{emb}^i$ and $Group_{emb}$ via concatenation, then determine if the last utterance (e.g., m_j^2 in Figure 2) given to the model is an instance of disruptive talk with the threshold value of 0.5.

$$O = sigmoid(W_o * (User_{emb}^i + Group_{emb}))$$

We expect this attention-based user-aware approach will assist the model's inference on the target utterance by providing specific information about the target student characteristics embedded by the user-specific network, while simultaneously attending to the related utterances from the group sequence.

Additionally, we adopt a training approach that could better deal with the dynamics in multi-party dialogues. Different from dyad conversation, multiple conversation threads in the group conversation could make it difficult for the model to learn consistent and generalized aspects. We adopt a sequence dropout approach, which is one of the discourse perturbation methods used in (Koupaee et al., 2021), applied to the sequence inputs so that the model can learn different representations from the same context messages at every epoch as an approach to model regularization. For the given $Group_{Seq}$, we randomly drop utterances with a sequence dropout rate of r, range in (0, 0.5) excluding the target utterance. This sequence dropout rate can be fixed or can be randomly selected from the normal distribution. Figure 3 shows how this approach is applied during training.

We anticipate that by observing the same context message sequence from multiple dimensions, the disruptive talk detection model will learn generalized patterns by avoiding possible overfitting. Note that we do not drop anything from the user network with an assumption that utterances from the same user are consistent. It should be noted that this sequence dropout mechanism is different from the dropout technique commonly used for recurrent neural networks, which drops for the linear transformation of the inputs or the recurrent state, by dropping for the entire input at random time stamps. The sequence dropout is applied to the training data only for effective training through model regularization.

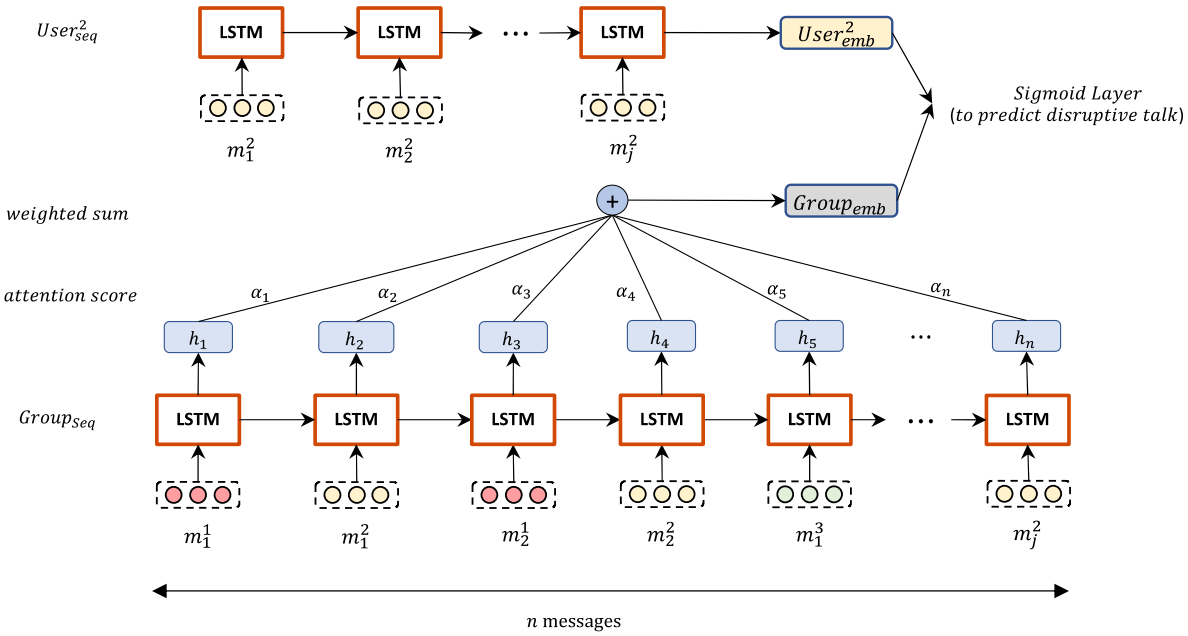

Figure 2: Proposed attention-based user-aware model. This figure illustrates what happens when the current message is from User 2.

4.3 Evaluation

We evaluate our modeling approaches in three steps. First, we compare our attention-based user-aware approach with our baseline model, which is based on a group sequence network with an attention mechanism (i.e., without the user-aware feature). This baseline modeling approach using LSTM-based disruptive talk detection model with DistillBERT as a sentence embedding approach, and 20 context messages, was adapted from our previous work (Park et al., 2021). Second, by comparing models with a fixed sequence dropout rates r from 0 to 0.5, and a model that adopts a random rate from normal distribution, we decide whether we would want to fix the sequence dropout rate of r or bring a complete randomness into the training phase. We did not raise the r over 0.5 (i.e., dropping 50% or more utterances every time) to avoid any possible data loss while training. All results are compared with the baseline model trained on full sequences-only, adopted from our previous work (Park et al., 2021). Furthermore, to account for the nature of randomness of sequence dropout approach, we run the models 5 times and average the results from each fold. Finally, we apply both the attention-based user-aware and the sequence dropout approaches to see that brings an additional performance enhancement.

We evaluate the performance of the disruptive talk detection models using the area under the receiver operating characteristic curve (AUC). AUC is one of the commonly used evaluation metrics for binary classification problems in machine learning, which represents the classification model's ability to separate between classes. The ROC curve shows the trade-off between true positive rate and false positive rate

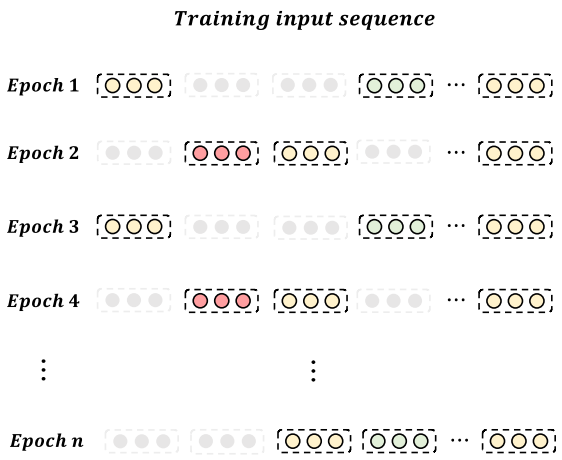

Figure 3: Sequence dropout training approach with a fixed rate of r. For the same training input sequence, the model drops r rate of inputs randomly. If r is chosen at random, a different number of inputs will be removed every epoch.

when varying the threshold values. An AUC of 1 indicates the classifier can perfectly discriminate between two classes, and 0.5 indicates the classifier cannot discriminate between two classes. We also evaluate the performance of the models using the area under the precision recall curve (PR-AUC) since AUC can give over-optimistic scores when the number of positive and negative classes are not balanced (Davis and Goadrich, 2006; Saito and Rehmsmeier, 2015). Like the ROC curve, the PR curve shows trade-off between precision (y-axis) and recall (x-axis) for different threshold values. It should be noted that when evaluating models based on the PR-AUC, it is essential to compare the performance with the PR-AUC of a no-skill classifier (i.e., Random chance), as the baseline performance varies depending on the task and the data distribution. To compare predictive performance, we report the average AUC and the average PR-AUC from cross-validation results.

We apply stratified group-level 10-fold cross-validation to avoid data leakage between training and testing data and retain the class distribution across folds. For each fold, we split the training data into a training and validation set to perform the early stopping based on the validation set. The distribution and the size of the validation set is the same as the test set. For all modeling approaches, we set the number of hidden units to 64, batch size to 32, and the number of epochs to 20, while using early stopping with a patience of 5 to avoid overfitting.

5 Result and Discussion

Table 2 shows evaluation results of the baseline model and the user-aware networks for disruptive talk detection. Our attention-based user-aware modeling approach outperforms the baseline modeling approach with respect to AUC (p=0.065), while it also brings improvement with respect to PR-AUC (p=0.161), where the statistical tests were conducted using the Friedman test, which is the non-parametric statistical test for multiple machine learning classifiers over multiple data sets, with a post-hoc analysis with the Wilcoxon signed rank test (Demšar, 2006). These results suggest that having the user-specific network was helpful for the model to identify whether the target utterance is disruptive or not. This might be because the user-specific network examines how the messages of target students have been developed without being affected by other student messages. The model obtains a clearer sense of the user's potential to be disruptive in a group conversation. In addition, it is possible that giving more weights to the hidden states that are more relevant to the target user embedding was effective to identify where to attend in the potentially noisy group sequence representation for the disruptive talk prediction of the target user.

Model	AUC	PR-AUC
No-Skill	0.5000	0.2466
Baseline	0.8292	0.5504
User-Aware	**0.8480**	**0.5691**

Table 2: Results of attention-based user-aware network (User-Aware). The best performance of each evaluation metric is marked in bold.

Table 3 shows the performance of sequence dropout approach (i.e., sequence dropout applied to a group-level network without a user-aware network) across the different sequence dropout rates and random choice. Except for $r = 0.1, 0.2$, all modeling approaches using different sequence dropout rates outperform the baseline with respect to AUC with a statistical significance ($p < 0.05$) when they were tested with the Wilcoxon signed rank test, while the model with random dropout rates applied perform the best. There were no significant differences in the performances among different dropout rates, except for the model using r is 0.1 or 0.2. These results might suggest that learning patterns from different sequence combinations were helpful for the disruptive talk detection model but dropping too few utterances would bring less significant enhancement to the performance. With respect to PR-AUC, the model

Dropout Rate (r)	AUC	PR-AUC
Baseline ($r = 0$)	0.8292	0.5504
Random (0, 0.5)	**0.8557***	**0.5649**
0.1	0.8413	0.5492
0.2	0.8426	0.5466
0.3	0.8507*	0.5517
0.4	0.8543*	0.5494
0.5	0.8498*	0.5526

Table 3: Sequence dropout approach across different dropout rate of r. The best performance of each evaluation metric is marked in bold, and * represents there is a statistically significant difference compared to the baseline.

Model	AUC	PR-AUC
Baseline	0.8292	0.5504
User-Aware	0.8480	0.5691
Sequence Drop	0.8557*	0.5649
SeqDrop+User-Aware	**0.8675***	**0.5991***

Table 4: Disruptive talk prediction results. The best performance of each evaluation metric is marked in bold, and * represents there is a statistically significant difference compared to the baseline.

with the random sequence dropout choice demonstrated improved performance compared to the other competitive modeling approaches, although the difference is not statistically significant when compared to the baseline model (p=0.138).

The performance enhancement with the sequence dropout training mechanism suggests that the conversation sequences may have contained noise due to the presence of multiple conversation threads, and that the model had some trouble determining how to extract the essential parts of the conversation sequences that could help with disruptive talk predictions. The model was given the opportunity to learn multiple variants of utterances from the same sequence because of the random dropping of a different subset of sequences at each training epoch. It is possible that this method could help achieve improved predictive performance by regularizing the disruptive talk detection models to effectively deal with noisy conversation data.

Finally, we compare the baseline model with the combined model, which utilizes both the attention-based user-aware and sequence dropout approaches. We compare the performance of this combined model with the models from the previous phases. Here, we adopt the random choice for the sequence dropout rate since it yielded higher performance with respect to both AUC and PR-AUC than the ones with the fixed sequence dropout rate. Results in Table 4 shows that our proposed disruptive talk detection model combining both User-Aware and Sequence dropout approaches. Our proposed regularized user-aware networks significantly outperform the baseline approach for both evaluation metrics (p<0.01 for AUC and p=0.09 for PR-AUC) with an alpha of 0.1. It also outperforms the models using each of the two proposed mechanisms: user-aware only (p<0.01 for AUC and p=0.06 for PR-AUC) and sequence dropout only (p=0.08 for AUC and p=0.16 for PR-AUC). These results suggest that the combined approach brings a synergetic effect to disruptive talk detection prediction. We observed from our repeated experiments (i.e., 5 executions) for all models using sequence dropout during training that the coefficient of variations (i.e., standard deviation / mean) of all approaches are less than 1, which is considered to be low variance between the values. This might suggest that the models were reliably trained even with randomness that resulted from dropping for a different set of utterances in dialogue sequences in each run.

Lastly, we note potential limitations of our research. Because of the nature of stratified group-level sampling where the sampling procedure must take into account both the label distribution and the group index, it is not possible to apply the exact same distribution across different folds, which could result in large performance variations between folds. In addition, while our proposed modeling approach demonstrated a promising result in our testbed collaborative game-based learning environment, the proposed model could be evaluated with other computer-supported collaborative learning environments to demonstrate generalizability of the technique.

6 Conclusion

Multi-party dialogue modeling poses significant challenges because of the complexity driven by group dynamics characterized in multi-party conversations. Detecting disruptive talk in collaborative game-based learning environments is crucial to support high-quality collaborative learning. We have presented a novel deep learning-based disruptive talk detection model that incorporates a user-aware attention network and a random sequence dropout training mechanism, where the model utilizing both approaches significantly outperform the baseline approaches. The proposed model shows significant promise for addressing key challenges in multi-party dialogue prediction. In the future, it will be important to test our model's capability with multi-party dialogue corpora from other computer-supported collaborative learning environments to test the generalizability of our model. It will also be important to implement the disruptive talk detection model in a real-time setting and investigate how it informs adaptive support for collaborative student learning.

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

Ali Ahmadvand, Jason Ingyu Choi, and Eugene Agichtein. 2019. Contextual dialogue act classification for open-domain conversational agents. *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval*, pages 1273-1276.

Sergio Grau, Emilio Sanchis, Maria Jose Castro, and David Vilar. 2004. Dialogue act classification using a Bayesian approach." In *9th Conference Speech and Computer*.

Su Nam Kim, Lawrence Cavedon, and Timothy Baldwin. 2012. Classifying dialogue acts in multi-party live chats. In *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation*, pages 463-472.

Wookhee Min, Randall Spain, Jason D. Saville, Bradford Mott, Keith Brawner, Joan Johnston, and James Lester. 2021. Multidimensional team communication modeling for adaptive team training: A hybrid deep learning and graphical modeling framework. In *International Conference on Artificial Intelligence in Education*, pages 293-305.

Vladislav Maraev, Bill Noble, Chiara Mazzocconi, and Christine Howes. 2021 Dialogue act classification is a laughing matter. In *Proceedings of the 25th Workshop on the Semantics and Pragmatics of Dialogue*.

Zahra Rahimi and Diane Litman. 2018. Weighting model based on group dynamics to measure convergence in multi-party dialogue. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 385–390, Melbourne, Australia. Association for Computational Linguistics.

Elijah Mayfield, David Adamson, and Carolyn Penstein Rosé. 2012. Hierarchical Conversation Structure Prediction in Multi-Party Chat. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 60–69, Seoul, South Korea. Association for Computational Linguistics.

Ming Tan, Dakuo Wang, Yupeng Gao, Haoyu Wang, Saloni Potdar, Xiaoxiao Guo, Shiyu Chang, and Mo Yu. 2019. Context-aware conversation thread detection in multi-party chat. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6456-6461. 2019.

Pierre Dillenbourg, Sanna Järvelä, and Frank Fischer. 2009. The evolution of research on computer-supported collaborative learning. In *Technology-enhanced learning*, pages 3-19. Springer, Dordrecht.

Cindy E. Hmelo-Silver. 2004. Problem-based learning: What and how do students learn? *Educational psychology review*, 16(3): 235-266.

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.

Asmalina Saleh, Chen Feng, Haesol Bae, Cindy E. Hmelo-Silver, K. Glazewski, Seung Lee, Bradford Mott, and James Lester. 2021. Negotiating accountability and epistemic stances in middle-school collaborative discourse. In *Proceedings of the International Conference on Computer-Supported Collaborative Learning*.

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 *International Conference on Artificial Intelligence in Education*, pages. 55-66. Springer, Cham.

Stefanos Nikiforos, Spyros Tzanavaris, and Katia-Lida Kermanidis. 2020. Virtual learning communities (VLCs) rethinking: influence on behavior modification—bullying detection through machine learning and natural language processing. *Journal of Computers in Education* 7(4): 531-551.

Kyungjin Park, Hyunwoo Sohn, Bradford Mott, Wookhee Min, Asmalina Saleh, Krista Glazewski, Cindy Hmelo-Silver, and James 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*, pages 405-415.

Tatiana Anikina and Ivana Kruijff-Korbayová. 2019. Dialogue act classification in team communication for robot assisted disaster response. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 399-410.

Jeremy Blackburn and Haewoon Kwak. 2014. STFU NOOB! predicting crowdsourced decisions on toxic behavior in online games. In *Proceedings of the*

23rd international conference on World wide web, pages 877-888.

Aslı Ekiciler, İmran Ahioğlu, Nihan Yıldırım, İpekİlkkaracan Ajas, and Tolga Kaya. 2021. The bullying game: Sexism based toxic language analysis on online games chat logs by text mining. In *Conference on Gender Studies and Sexuality*.

Ameneh Shamekhi, Q. Vera Liao, Dakuo Wang, Rachel KE Bellamy, and Thomas Erickson. 2018. Face Value? Exploring the effects of embodiment for a group facilitation agent. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1-13. 2018.

Amy X. Zhang, and Justin Cranshaw. 2018. Making sense of group chat through collaborative tagging and summarization. *Proceedings of the ACM on Human-Computer Interaction* 2(CSCW): 1-27.

Bastian Kordyaka. 2018. Digital Poison Approaching a theory of toxic behavior in MOBA games. In *International Conference on Information Systems*.

Bradford Mott, Robert Taylor, Seung Lee, Jonathan Rowe, Asmalina Saleh, Krista Glazewski, Cindy Hmelo-Silver, and James Lester. 2019. Designing and developing interactive narratives for collaborative problem-based learning. *Proceedings of the Twelfth International Conference on Interactive Digital Storytelling*, pages 86-100, Snowbird, Utah.

Asmalina Saleh, Cindy Hmelo-Silver, Krista Glazewski, Bradford Mott, Yuxin Chen, Jonathan Rowe, and James Lester. 2019. Collaborative Inquiry Play: A Design Case to Frame Integration of Collaborative Problem Solving with Story-Centric Games. *Information and Learning Sciences*, 120(9): 547-566.

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

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1): 37-46.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

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*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Mahnaz Koupaee, Greg Durrett, Nathanael Chambers, and Niranjan Balasubramanian. 2021 Don't let discourse confine your model: Sequence perturbations for improved event language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 599-604.

Jesse Davis and Mark Goadrich. 2006. The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233-240.

Takaya Saito and Marc Rehmsmeier. 2015. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, 10(3): e0118432.

Janez Demšar. 2006. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, 7:1–30.