Shih, JL. et al. (Eds.) (2023). Proceedings of the 31st International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Student Engagement Detection: Case Study on Using Peer-to-Peer Emotion Comparison with Context Regularization

Geyu LIN, Manas GUPTA, Cheryl Sze Yin WONG & Huayun ZHANG
*Institute for Infocomm Research (I^2R), Agency for Science Technology and Research (A*STAR), Singapore*
{lin_geyu, manas_gupta, cheryl_wong, zhang_huayun}@i2r.a-star.edu.sg

Abstract: This paper describes a method to automatically assess participants' engagement in online education. Similar to emotion recognition, student engagement can be subjective. Hence, it is challenging to obtain large-enough and consistent ground-truth engagement labels for automatic student engagement. We propose an unsupervised method that could detect abnormal engagement states using peer-to-peer emotion correlation analysis in different modalities. Without any human engagement labeling, this zero-shot method accurately pinpoints the abnormal student engagements in our experiment. Modality-dependent engagement prediction also suggests possible distractions on the student's device.

Keywords: Student Engagement, Virtual Learning, Multimodality

1. Introduction

Market for online education keeps expanding in recent years with the development of hardware, such as internet infrastructures, personal digital devices, etc. and the progress of software, such as online education tools and online education theories. Coronavirus sped up this procedure when schools all over the world had to shut down their physical classes and move classes online (Zheng et al., 2021). Many remote education programs outlast the pandemic. Schools and students prefer online learning to some extent because of the flexibility and convenience it offers.

Due to the intrinsic difference between real classroom and online classroom, both teachers and students are facing new challenges for online education. Without the proper classroom management and discipline restriction, students may find it hard to resist the temptation of food, drink, and other entertainment within their reach. Without face-to-face interaction with their students, teachers may not be able to gauge the effectiveness of varying teaching strategies on engaging the students. Many teachers reported they lose confidence in their ability to meet students' needs. Hence, the use of automatic student engagement detection (Baker et al, 2004) can help teachers manage the online classroom and identify the students who are disengaged. Automatically monitoring and diagnosing student engagement variation could become a teacher assistant for quality online education.

However, training machine learning models for accurate student engagement detection would require accurate ground-truth engagement labels (Khan et al., 2022). The accurate ground-truth labels could be hard to obtain due to subjective perspectives and differing cultures (Ocumpaugh et al. 2014). Hence, an unsupervised method to detect abnormal engagement states is proposed in this paper.

We examine the feasibility of analyzing students' engagement during the online classes by using learning context regularization to find out the out-of-sync student using correlation of emotion prediction of webcam video with learning contexts' emotion prediction.

2. Problem

Student engagement is correlated with the student emotion states and it is reflected by the student's behaviour response in the class and his/her academic improvement before and after the learning procedure. Multimodal data collected in virtual classroom, including individual performance in class quiz, average time answering the questions, activities in the classroom, what the student is watching, what the student is listening, eye contact, facial and body movement, can be collected and carefully labelled to build deep learning models detecting student engagement.

A fundamental challenge to deep learning models is the reliance on supervised learning and by extension good data - data that is large-enough and consistently labelled. Small, and inconsistent data often lead to serious performance degradation in real-world. While this is not unique to engagement detection, the lack of a widely accepted measurement for student engagement level makes it even more difficult to collect and label high quality data for engagement detection task, in addition to individual difference between students and subjective bias of human labelers.

In this study, an alternative way is adopted for detecting out-of-sync students. Instead of training deep learning model using well labelled engagement data, peer-to-peer emotion comparisons are conducted in different modalities to pinpoint abnormal engagement states in the student group.

3. Methodology

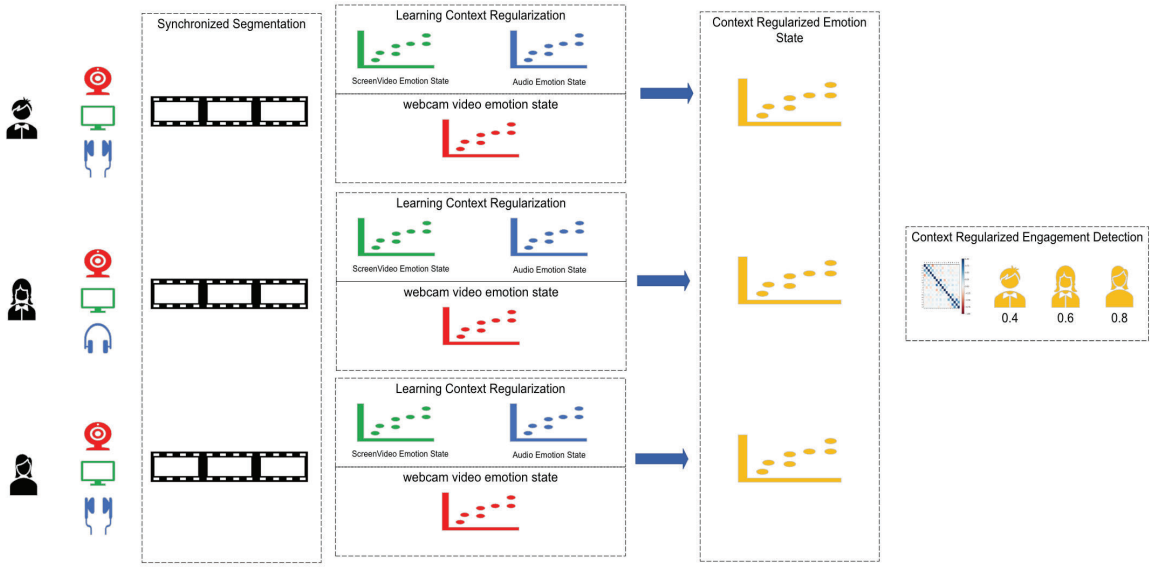

Figure 1. Block Diagram for Proposed Method

The proposed framework for detecting possible disengagement consists of three components, namely synchronized segmentation to determine the learning context, emotional states based on multiple modalities of data and the concept of peer-to-peer correlation in the same context. This is being elaborated in the following sections.

3.1 Synchronized Segmentation

Experiments are conducted in asynchronous settings. Students attend the virtual class at their own convenience using a pre-recorded lecturing video. By attending the virtual class, they agree that their learning procedures, including their webcam recording, their screen

activities, and audio output on their learning devices, will be saved in a multi-stream video file and it will be used for research purposes after anonymization. Student video must be segmented for engagement variation analysis along with time. Two synchronization methods are adopted for student video segmentation.

3.1.1 Fixed-Window-Length Segmentation

Student videos are synchronized by the instant starting time of lecturing video and chopped without overlapping. All student videos are segmented into fixed 10-sencond clips without considering the continuance of the lecturing content.

3.1.2 Lecturer-Speech-Content Based Segmentation

The lecturer's speech in the pre-recorded video is transcribed using openai/whisper-large-v2 (Alec et al., 2022) speech recognition model. This model was trained on 680K hours of multilingual and multi-task transcribed speech data collected from the web. While this model does produce highly accurate transcription for lecturer's speech, the corresponding timestamps it provided are insufficient and sometimes inaccurate. To obtain accurate timestamps for student video segmentation, another HMM-ANN (Dahl et al., 2012) (Hidden-Markov Model – Deep Neural Network) hybrid model was adopted for force-alignment between the audio stream in student video and the whisper transcription of the lecturing video. Student videos were then segmented into varying-length clips according to the sentence-end punctuation in the whisper transcription. Doing in this way, student videos are synchronized by lecturer-speech-content and student video clips are associated with relatively complete lecturer sentences. The average length of student video clips is around 10 seconds.

3.2 Multimodal Emotion States

In an online learning environment, there can be multiple modalities of data that can be captured, such as the learner behavior through video and learning context through video and audio, which can be converted into text. Hence, we leverage on the Emolysis toolkit (Ghosh et al., 2023) for video and text emotion recognition. The video modality model used for prediction is the High-Speed face Emotion (HSE) model (Savchenko et al., 2022), architecture based on EfficientNet (Tan & Le, 2018), pre-trained on AffectNet (Mollahosseini et al., 2019). The text modality model used for prediction is a pre-trained RoBERTa module (Liu et al., 2019) followed by a three-layer Deep Neural Network (DNN) that maps the latent RoBERTa features to label space. The RoBERTa is fine-tuned and DNN is trained with the CMU-MOSEI dataset (Zadeh et al., 2018).

The models predict the valence, arousal, as well as the emotion states, namely fear, anger, joy, sad, disgust, surprise, trust, anticipation and none. In our analysis, the various emotion states may not be relevant for student engagement, thus we focus on the valence and arousal values. One of the challenges in multimodality data is the variance in prediction frequency for different modalities of data. In our case, it would be the webcam video, screen video and audio (transcript to text). Hence, the calculated average values in each clip obtained using synchronized segmentation would be utilized.

3.3 Peer-to-Peer Correlation and Engagement in Context

Learning context is an important factor to consider when analyzing the correlation of webcam video emotion prediction between students. Students might have different reaction times as the lecture flows, hence analyzing the emotion correlations of students directly might not work well.

The learning contexts between different students should be highly correlated because they should be watching the same video. With this fact, we can pick out students that are, a) did not open the video, b) watching different video and c) not listening to the video by only analyzing the correlation of the context. By adding an additional modality of webcam video, we can identify the students that were not focusing on the lecture through making valence and arousal scores on each student.

In our paper, we have experimented with two different types of learning contexts. First, the screen recording of the lecture itself. Second, the text content generated from the audio of the video. These two factors serve as regularizations when we evaluate the correlation between students to pick out outliers from students.

We make prediction for the arousal and valence score of the context, then take an average of webcam videos' emotion prediction and context emotion prediction to get final arousal and valence score to perform correlation evaluation. With learning context regularization on webcam videos' emotion prediction, we can identify out-of-sync student easily using correlation on the combined arousal and valence score between students.

4. Experiments

4.1 Multimodal Data Collection

We use the OBS Studio open-source recording tool to collect multimodal data. The tool allows us to collect the webcam of the student, the video of the lecturer, the audio coming from the lecturer and the audio generated by the student for each recording session. The various modalities are synced by default. We extract the relevant modalities from the raw data and process them individually for running our experiments.

4.2 Pilot Data Analysis

We created two outliers for evaluation to represent different types of distractions that the student may have to illustrate the feasibility of evaluating engagement level in an online learning setting.

4.2.1 Student Watching Unrelated Content

Participant id003 does not open the lecture video on screen. Hence id003 should be a representative of watching unrelated content during lecture. This participant should be identified as an outlier when using the screen video and webcam video.

4.2.2 Student Not Focusing on The Lecture

Participant id002 is a representation for students that are having the lecture on screen but are sleeping or looking around. This participant should be identified as an obvious outlier using screen video and audio.

4.3 Result

4.3.1 Pairwise Correlation Coefficient (PCC) Matrix

PCC matrix consists of the pairwise correlation coefficient (PCC) of each of the participants. The (i, j)-th entry of the PCC matrix represents the PCC of prediction score of participants i and participant j . We use heatmap to visualize the matrix, the color varies from red to

green, representing the PCC from lower to higher values. Hence, if a row or column for a participant consists of many red grids, this participant should be identified as an outlier.

4.3.2 Engagement Score

We assign an **engagement score** to each participant's arousal and valence prediction using the following method*:*

a) Suppose we have n participants, each participant have l clips*.*

b) Duplicate participants i *'s series* $(n-1)$ times*.*

c) Concatenate all other participants' series *except participant* i

d) Calculate the PCC of two series obtained from step (b) and step (c). Assign the PCC as the engagement score for participants i

The engagement score is a summarized score about how a participant's prediction emotion score correlates with others. This is used as a unified measure to represent how a student is engaged during the lecture.

4.3.3 Regularization Factor

The regularization factor refers to the factor that is highly correlated under the experiments setting. The regularization factors can simply be learning contexts, as discussed in section 3.3. We have two regularization factors*:*

a) Screen video emotion score.

b) Audio modality emotion score.

4.3.4 Overall Score Calculation

Given the emotion score prediction (arousal, valence) from video (student and lecturer), audio and text modality. We apply average on student video arousal with each of the regularization factor and both regularization factors, then analyze the engagement score and PCC matrix to see whether we can identify the outlier from calculation. Also, we compare two segmentation techniques, a) sentence segmentation, b) uniform 10s segmentation, for webcam video and audio modalities analysis and webcam video, screen video and audio modalities analysis to *verify how good regularization influences the outlier identification.*

4.3.5 Screen Video with webcam video

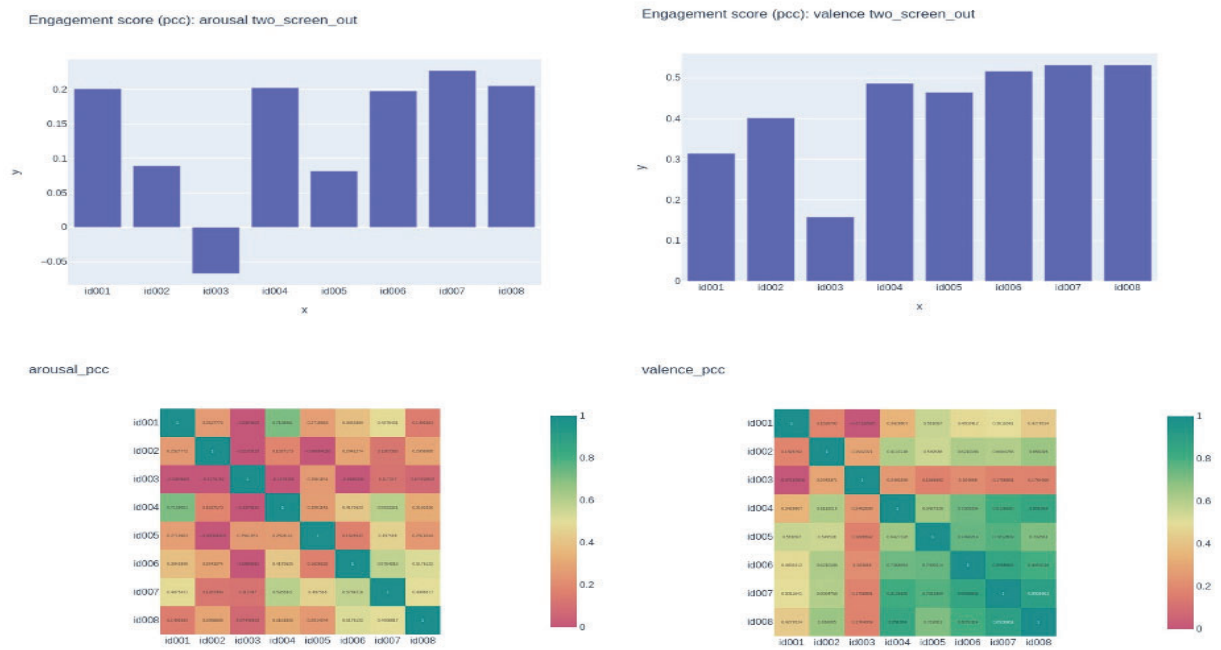

Figure 2. PCC matrix and engagement score (screen + webcam) using content-based segmentation

The above are the engagement score and PCC matrix of arousal and valence for screen video and webcam video. In this section, we examine the result using screen video as regularization factor.

From the confusion matrix of valence, we can observe that the id003 is the most obvious outliner because id003 did not open the screen of instructed video. The emotion prediction of screen video will differ for id003 compared to other participants. However, from the confusion matrix of arousal, it might not be obvious that id003 is an outlier. With the help of engagement score, we can observe that id003 is a clear outlier with negative correlation with others. This shows the importance of providing a consolidated score for better interpretation.

Hence, using screen video and webcam video, we can identify outlier described in section 4.2 by considering the PCC matrix and the engagement score of valence and arousal.

4.3.6 Audio With Webcam Video

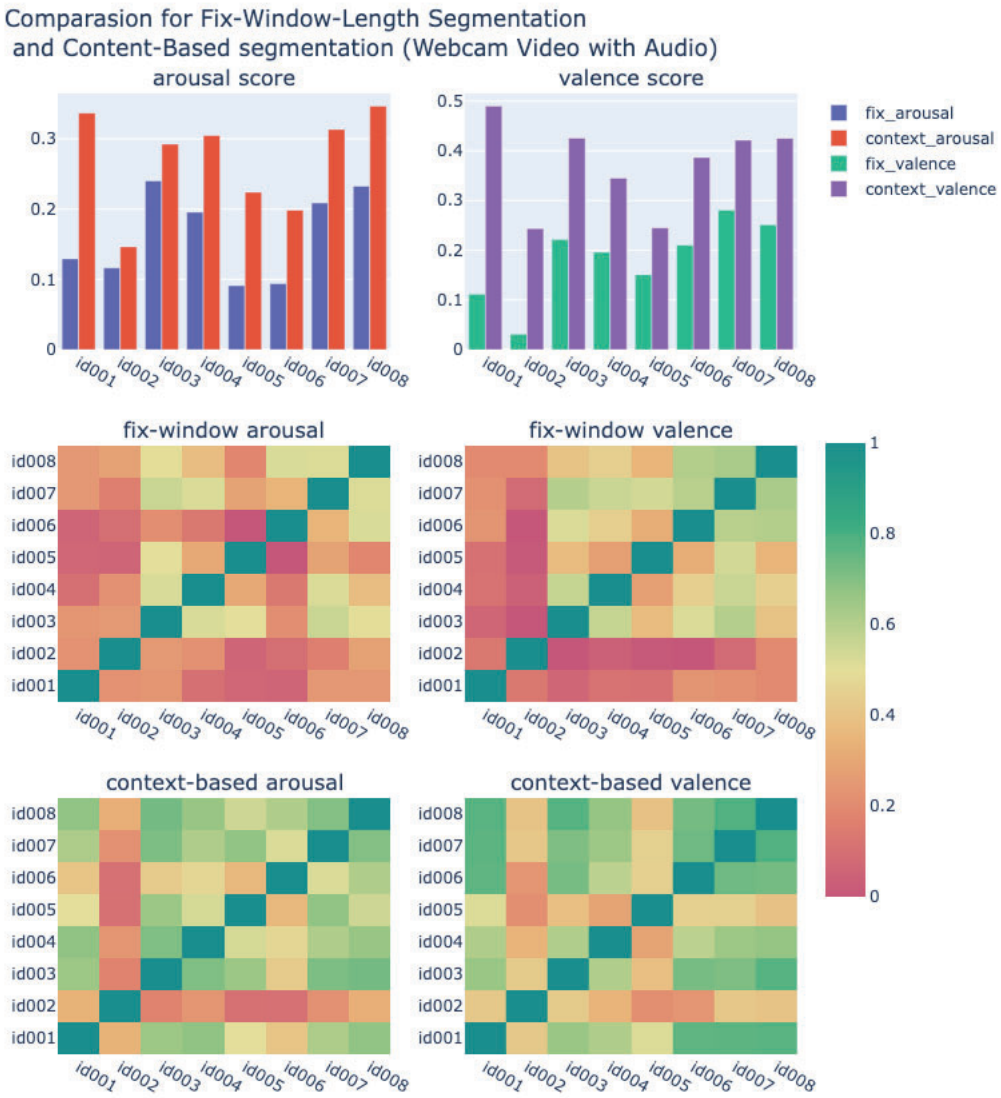

Figure 3. Results for Audio with Webcam Video

The above shows the engagement score result and PCC matrix of webcam video with audio emotion as regularization factor using fix-window-length segmentation. We can observe that the outlier is not obvious in this case, because the uniform 10s segmentation cannot have a good regularization compared to sentence level segmentation due to the incomplete sentence information. Some sentence might be trimmed, such that the model cannot predict the emotion of the learning context accurately.

With content-based segmentation, we can clearly identify id002 as an outlier from both PCC matrix and engagement score of both arousal and valence. Hence, this suggests that with a good regularization factor, we can identify the outlier by comparing the correlation between different students and using only webcam video of student's face and audio emotion prediction.

4.3.7 Screen Video, Webcam Video and Audio

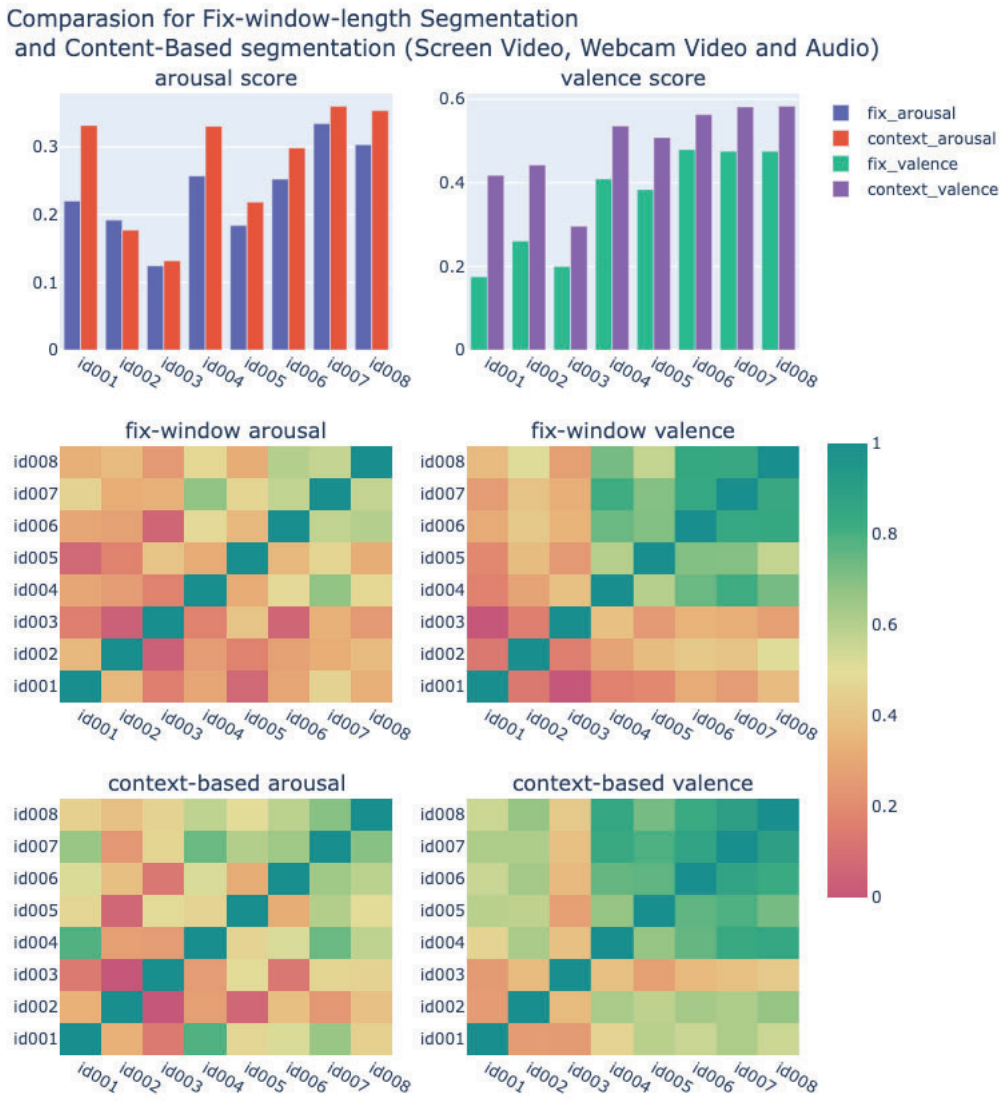

Figure 4: Results for Audio with Webcam Video and Screen Video

From the engagement score and PCC matrix generated from fixed-window-length segmentation, it's hard to conclude the outlier because the engagement score are comparable for arousal and id001 has lower engagement score for valence than both id002 and id003. Also, from the PCC matrix, we might conclude that id001 is also an outlier besides id002 and id003.

From the result of combing three modality with content-based segmentation, we can differentiate both outlier (id002 and id003) from engagement score of arousal and valence. Even though id001 seems also have outlier behavior from engagement score of valences, the outlier behavior of arousal prediction is much more obvious. We still can identify id002 and id003 as outlier by PCC matrix of both arousal and valence. This proves that the context regularization helps identify different kinds of distractions that students may have during online learning sessions.

5. Conclusion

In the paper, we demonstrate two feasibilities of analyzing engagement under online learning settings, by analyzing the context regularized arousal and valence prediction as follows:

a) Record the screen video and webcam video together, the screen video emotion prediction served as a regularization factor to help to identify student that is less engaged who might be watching other unrelated content (e.g., id003) or not listening carefully (e.g., id002).

b) With only the students' face, the audio emotion prediction output served as an effective regularization factor because audio modality should be uniform unless student is listening to something else. Furthermore, we could identify students who are not listening carefully through a regularized score between webcam video emotion prediction and audio emotion prediction.

The content-based segmentation provides a more reliable context regularization due to the content dependency on the text modality. Using fixed-window-length segmentation, the text will be trimmed randomly. This will result in the unreliable prediction on text modality. Hence with content-based segmentation creates a more reliable context regularization when we evaluate the correlation between students.

Acknowledgements

We would like to thank Siti Maryam Binte Ahmad Subaidi for coordinating the data collection and post processing, Chen Hao, Jeremy Wong, Lily Hoang Huong Giang and Tan Chin Tuan (alphabetical order) for helpful discussion of this case study. We would also like to thank the Artificial Intelligence, Analytics Informatics (AI3) and Institute for Infocomm Research (I2R), A*STAR for supporting this work.

References

Alec Radford, Jong Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever (2022), Robust Speech Recognition via Large-Scale Weak Supervision. Technical Report, OpenAI.

Baker, R. S., Corbett, A. T., Koedinger, K. R., & Wagner, A. Z. (2004, April). Off-task behavior in the cognitive tutor classroom: When students" game the system". In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 383-390).

Dahl G.E., Yu Dong, Deng Li, and Acero A. (2012), Context-Dependent Pretrained Deep Neural Networks for Large Vocabulary Speech Recognition, IEEE Trans. Audio, Speech, and Language Proceeding, 33-42.

Ghosh, S., Cai, Z., Gupta, P., Sharma, G., Dhall, A., Hayat, M., & Gedeon, T. (2023). Emolysis: A Multimodal Open-Source Group Emotion Analysis and Visualization Toolkit. arXiv preprint arXiv:2305.05255.

Khan, S. S., Abedi, A., & Colella, T. (2022). Inconsistencies in the Definition and Annotation of Student Engagement in Virtual Learning Datasets: A Critical Review. arXiv preprint arXiv:2208.04548.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). ROBERTA: A robustly optimized BERT pretraining approach. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1907.11692

Ocumpaugh, J., Baker, R., Gowda, S., Heffernan, N., & Heffernan, C. (2014). Population validity for educational data mining models: A case study in affect detection. British Journal of Educational Technology, 45(3), 487-501.

Mollahosseini, A., Hasani, B., & Mahoor, M. H. (2019). AffectNet: A database for facial expression, Valence, and arousal computing in the wild. IEEE Transactions on Affective Computing, 10(1), 18–31. https://doi.org/10.1109/taffc.2017.2740923

Savchenko, A. V., Savchenko, L. V., & Makarov, I. (2022). Classifying emotions and engagement in online learning based on a single facial expression recognition neural network. IEEE

Transactions on Affective Computing, 13(4), 2132–2143. https://doi.org/10.1109/taffc.2022.3188390

Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. International Conference on Machine Learning, 6105–6114. http://proceedings.mlr.press/v97/tan19a/tan19a.pdf

Zadeh, A. B., Liang, P. P., Poria, S., Wang, Z., & Morency, L. (2018). Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). https://doi.org/10.18653/v1/p18-1208

Zheng, M., Bender, D. & Lyon, C. Online learning during COVID-19 produced equivalent or better student course performance as compared with pre-pandemic: empirical evidence from a school-wide comparative study. BMC Med Educ 21, 495 (2021). https://doi.org/10.1186/s12909-021-02909-z