Kashihara, A. et al. (Eds.) (2024). Proceedings of the 32nd International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Exploring Explainable Artificial Intelligence in Active Video Watching

Raul Vincent LUMAPAS*, Antonija MITROVIC, Matthias GALSTER & Sanna MALINEN
University of Canterbury, New Zealand
*raulvincent.lumapas@pg.canterbury.ac.nz

Abstract: Active Video Watching supports engagement through scalable interventions, such as notetaking in the form of comments. Machine Learning is used to categorize comments based on their quality to provide personalized feedback to students. In previous work on AVW-Space, an online portal for active video watching, a machine learning model was trained using data from several studies on presentation skills. In this paper, we explore the effectiveness in assessing the comment quality of this model in Face-to-Face Meeting Communication skills in comparison to a model trained specifically for this soft skill. We used Explainable Artificial Intelligence to identify and compare the important features of the models. Results show the need for comment quality assessment models to be specific to the soft skill in question and show major differences between their important features, highlighting the necessity to create a model specific to a particular soft skill.

Keywords: Explainable Artificial Intelligence, Active Video Watching

1. Introduction

Increasing user engagement is a challenge in video-based learning (VBL). Studies show that students who watch videos passively achieve limited benefits in learning (Koedinger et al., 2015; Chi & Wiley, 2014). An effective way to support engagement in VBL is through active video watching (AVW). AVW-Space, a VBL platform (Mitrovic et al., 2019) supports engagement by nudging students to write good quality comments, as well as by reading comments written by peers. AVW-Space uses a Machine Learning (ML) classifier to analyze comments, trained using data collected from several studies on training presentation skills (Mitrovic et al., 2016, 2017, 2019). Mohammadhassan et al. (2020) proposed a quality scheme for comments, with category 1 being of the lowest quality and category 5 as the highest. Comments categories 1 and 2 are pedagogically undesirable as these comments do not convey deep thinking and reflection about the videos. The last three categories are considered high-quality, which convey critical thinking (category 3), reflection on past experiences (category 4), or self-improvement (category 5). It is important to build a robust ML model as nudges use comment quality to provide tailored feedback to students.

We investigated improvements of the ML models and utilized XAI to provide explanations of comment quality. This research looked at the performance of the current ML model (trained on presentation skills comments) when assessing comments written for a different soft skill: face-to-face communication skills. We trained a new model for the latter skill and investigated its performances as well as important features.

2. Improvements of the Face-to-Face Communications ML Model

The current ML model has been used in most AVW-Space studies, including studies on F2F communication skills (Mitrovic et al., 2023). We compared the comment categories produced by the ML model for 688 comments from the 2022 F2F communication skills study to the manual classifications of the same set of comments by two human raters. Despite the high F1-Score of .84 for the current model, the results show it often disagrees with both human raters. While the two human raters show a substantial level of agreement, with a Cohen's

Kappa of 0.732, the current model often disagrees with their classifications, resulting in a low Krippendorff's Alpha of 0.461, falling short of the acceptable alpha value of 0.66 (Krippendorff, 2004). This suggests the need to explore alternative ML models for F2F communication skills.

Using the comments from the three studies with 147 students on F2F communication skills in 2020-2022 (Mitrovic et al., 2023), we trained two new ML models. A total of 1,549 comments were divided into the training set (80% or 1,231 comments) and the testing set (20% or 308 comments). The first model, referred to as C_a, follows the merged categories (1, 2+3, 4+5) Mohammadhassan et al. (2020) used. The C_b model uses a "*1, 2, 3+4+5*" scheme. We compared the two models to the current model (Table 1). Classifier C_a, which uses the same merged categories but was trained using a larger data set, has the best F1-Score.

Table 1 shows the inter-rater agreement between the ML models and the human raters. The human coders often disagree with the current model. However, there is an increase in the agreement between the human coders and the new models, especially C_b. Although C_a and C_b fall short of the acceptable minimum Krippendorff's Alpha value, C_b significantly outperforms the current model and C_a. It is noted, however, that the low agreement is because the new models, C_a and C_b, were trained using a smaller number of comments. The C_b model with the lowest F1-Score among the three tested classifiers generated the highest value of Krippendorff alpha. This means that the actual results of that model produced comment qualities much closer to how expert raters would classify the comments. The higher inter-rater agreement shows a better performance of two models trained using F2F communication skills.

Table 1. *Inter-rater Agreement between the models and human coders*

	Current Model	**C_a**	**C_b**
F1-Score	.84	**.88**	.78
Krippendorff's Alpha	0.442	0.525	**0.623**
Average Pairwise Cohen's Kappa	0.478	0.524	0.628
Pairwise Cohen's Kappa – System & Rater 1	0.295	0.412	0.565
Pairwise Cohen's Kappa – System & Rater 2	0.407	0.482	0.587

3. Explaining the F2F Communications ML Model

We used the SHAP feature importance summary plot to show the top ten features used to classify comments. As shown in Figure 1, the summary plot for the current model shows that a single feature, reflective aspect, has a huge impact, downplaying other textual features. In comparison, this feature has lower importance than the *I* and *personal pronoun* features in the other two models. This means the presence of the personal pronoun "I" and other personal pronouns (such as me, he, she, we, etc.) has a high impact on the comment quality assessment. The summary plot also shows the balanced impact of features for quality categories 2+3 and 4+5 for model C_a, and clusters 2, and 3+4+5 for model C_b. It is worth noting that the impact is different for each quality category. For example, the *I* feature has a positive impact on category 4+5, which means that comments in this category contain the "*I"* word. Meanwhile, it has an adverse impact on cluster 2+3 since most comments of this category lack the word "*I."* It can also be observed that the features have a smaller impact on category 1 comments. This is because of the lack of textual features for category 1 comments.

We also compared the performance of models C_a and C_b. The feature importance summary plot shows that model C_b is significantly better than the other models. It also shows the importance of looking at how the ML model behaves. Despite C_b having a more balanced feature importance, there are still very dominant features (such as the *I, Personal Pronouns, Reflective Aspect,* and *Pronoun* features). Some issues might arise from these very dominant features. For example, comments that merely repeat the dialogue or content in the video are classified as category 2 comments. However, some of these dialogues or content in videos might contain the word *I*. Such comments like that might be misclassified as high-quality, given that the I feature has a high predictive impact in C_b. This shows that model C_b can be improved further but is nevertheless better than the other models.

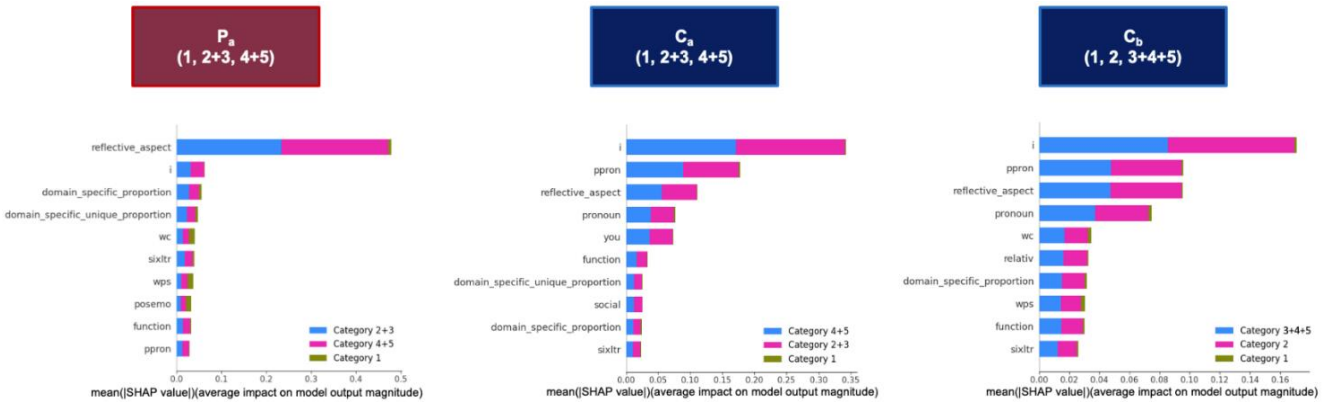

Figure 1. SHAP Feature Importance Summary Plot for the three ML models

4. Conclusions

We investigated the current ML model used in AVW-Space and identified the potential for improving it and for its use in other soft skills, such as F2F communication skills. Results show that the model performs poorly when used to classify comments in the F2F communication skills trainings, despite having a high F1-score. We then created two ML models (C_a and C_b) specifically trained using F2F communication skills comments. The comparison of the three ML models with the manual classifications of human raters shows that model C_b results in higher agreement with the human raters despite it having a much lower F1-score.

We analyzed the ML models using the recommended XAI tools from the question-driven design process and observed that model C_b is also more balanced in terms of feature importance. Both models trained using F2F communication skills comments use more features in comparison to the one trained using presentation skills and showed a significant improvement in comment quality assessment for Face-to-Face communication. Although the F2F communication skills models, particularly C_b, used more features, there is still a need to further investigate the other textual features when classifying a comment.

References

Chi, M. T. H., & Wylie, R. (2014). The ICAP Framework: Linking Cognitive Engagement to Active Learning Outcomes. Educational Psychologist, 49(4), 219–243.

Koedinger, K. R., Kim, J., Jia, J. Z., McLaughlin, E. A., & Bier, N. L. (2015). Learning is Not a Spectator Sport: Doing is Better than Watching for Learning from a MOOC. Proc. ACM Conf. Learning @ Scale (pp. 111–120).

Krippendorff, K. (2004). Content analysis: An introduction to its methodology. Sage.

Mitrovic, A., Dimitrova, V., Weerasinghe, A., Lau, L. (2016) Reflexive experiential learning: using active video watching for soft skills training. In: Chen, W. et al. (Eds.) Proc. 24th Int. Conf. Computers in Education (pp. 192-201). Asia-Pacific Society for Computers in Education

Mitrovic, A., Galster, M., Malinen, S., Holland, J., Musa, J. A., Mohammadhassan, N., & Lumapas, R. V. (2023). Effectiveness of Video-based Training for Face-to-face Communication Skills of Software Engineers: Evidence from a Three-year Study. ACM Transactions on Computing Education, 23(4), 1-25.

Mitrovic, A., Gordon, M., Piotrkowicz, A., & Dimitrova, V. (2019). Investigating the Effect of Adding Nudges to Increase Engagement in Active Video Watching. In Proc. Artificial Intelligence in Education (pp. 320–332). Springer International Publishing.

Mitrovic, A., Gostomski, P., Herritsch, A., Dimitrova, V. (2017) Improving presentation skills of first-year engineering students using Active Video Watching. In: N. Huda, D. Inglis, N. Tse, G. Town (Eds.) Proc. 28th Annual Conf. Australasian Association for Engineering Education (pp. 809-816).

Mohammadhassan, N., Mitrovic, A., Neshatian, K., Dunn, J. (2020) Automatic quality assessment of comments in active video watching using machine learning techniques. Proc. 28th Int. Conf. Computers in Education (pp. 1-10). Asia-Pacific Society for Computers in Education.