Assessment of Comment Quality in Active Video Watching using Deep Learning

Chantelle BALDWIN* **& Antonija MITROVIĆ**
University of Canterbury, New Zealand
chantelle.baldwin@pg.canterbury.ac.nz

Abstract: This study investigates the effectiveness of deep learning (DL) methods for classifying student comments based on quality within the AVW-Space video-watching platform. Given the limitations of existing machine learning (ML) techniques, this research explores whether DL models can improve classification performance. The study compared six DL models (BERT, RoBERTa, ALBERT, ELECTRA, GPT, and GPT-2) by training them on a dataset of 13,440 student comments. The results show that RoBERTa outperforms all models, demonstrating precision, recall, and F1-score improvements. Fine-tuning experiments led to an optimised RoBERTa model. We examine methods to address class imbalance, with weighted loss functions and random undersampling proving ineffective. This study contributes to the automation of comment assessment and supports personalised educational experiences.

Keywords: LLMs, Deep learning, Classifiers, Learning analytics, Video-based Learning, Automated Evaluation

1. Introduction and Related Work

Using video media for learning has become prevalent. Video-based learning (VBL) has unique features supporting various learning styles and has the potential to lead to better learning outcomes (Zhang et al., 2006). Recent VBL systems also facilitate self-regulated learning (Dimitrova & Mitrovic, 2022), but the lack of direct interaction with teachers can lead to passive learning (Yousef et al., 2014). Active Video Watching (AVW) mitigates inefficient learning and encourages engagement in video-based teaching.

AVW-Space is a controlled video-based learning platform that enables teachers to create customised spaces by selecting YouTube videos and specifying aspects to direct students' attention (Mitrovic et al., 2016). These aspects prompt students to reflect on their knowledge and experiences. Students can pause the video to leave comments at any point, and teachers may select comments to be anonymously shared for peer review, enabling classmates to rate and respond. Continuous enhancements have optimised AVW-Space for learning and engagement (Dimitrova & Mitrovic, 2022; Mitrovic, Dimitrova, et al., 2017; Mitrovic et al., 2016, 2019; Mohammadhassan et al., 2020). This study focuses on comments from a presentation skills module, which contained four tutorial videos on delivering presentations and four actual recordings of presentations.

Students who reviewed and wrote comments improved their knowledge, while those who passively watched did not (Mitrovic, Dimitrova, et al., 2017). Initial studies demonstrated the benefits of behaviour-based hints designed to encourage commenting and offer exemplar responses, known as personalised nudges, leading to increased student comments (Mitrovic et al., 2019). AVW-Space was enhanced to deliver personalised nudges; however, students did not yet receive real-time feedback on comment quality.

In response, Mohammadhassan et al. (2020) explored whether machine learning (ML) could assess comment quality, introducing a five-class scheme specific to tutorial videos. Class 1 (Affirmative, negative, off-topic) contains irrelevant or purely affirmative/negative remarks without explanation. While Class 2 (Repeating) consists of comments restating video content, both are considered low quality. The remaining classes represent higher quality. Class 3 (Critical and analytical) reflects critical thinking, while Class 4 (Self-reflective) connects

to the learner's prior experiences or behaviours. In Class 5 comments, students write about how to improve in future. Weighted classes and cost-sensitive error handling addressed the dataset's class imbalance. Misclassifying Class 3 as Class 2 is less severe than misclassifying Class 1 as Class 5, as the latter represents the most significant gap in engagement and thus incurs the highest penalty. A weighted random forest classifier achieved an F1-score of 0.68 and an average cost of 3.53. Performance was improved by merging classes, though this reduced the specificity of the feedback.

Reflective writing classification is challenging due to its unstructured nature and varied vocabulary. Literature shows deep learning (DL) methods often outperform traditional ML. Wang et al. (2024) achieved an F1-score of 0.699 using BERT, improving to 0.7474 with GPT-4-generated data and contrastive learning to stabilise training. Wulff et al. (2023) classified teachers' reflections with BERT, but faced tokenisation issues due to the niche domain. Li et al. (2023) found that BERT outperformed ML for reflective writing (F1-score: 0.7217).

This study investigates whether DL can outperform existing ML methods for AVW-Space comment classification. We evaluate transformer-based models; BERT (Devlin et al., 2019), ALBERT (Lan et al., 2020), RoBERTa (Liu et al., 2019), ELECTRA (Clark et al., 2020), GPT (Radford et al., 2018), and GPT-2 (Radford et al., 2019). Experiments assess hyperparameter tuning (dropout, hidden layers, batch size, epochs, learning rate), and testing strategies for handling class imbalance. Findings will inform NLP-driven educational tools and deepen understanding of DL's effectiveness for educational text classification.

2. Proposed Method

Mohammadhassan et al. (2020) used 2,245 manually classified comments, collected from 2016 to 2019, to develop an ML-based method. They split the data into 80% training and 20% testing. Since that study, more data is available for training and testing the model. We experimented with 13,440 comments written by first-year engineering students and collected from 2018 to 2023 (Dimitrova & Mitrovic, 2022; Mitrovic et al., 2019; Mohammadhassan et al., 2020, 2022). The studies were approved by the Human Research Ethics Committee (HREC) of the University of Canterbury.

We split the data randomly into a training, validation and testing set of 70%, 15% and 15%, respectively. The class distribution of the comments is uneven. There are 7,599 comments in Class 2 (Repeating), accounting for 57% of the dataset. Class 3 (Critical and analytical) has 2,813 comments, making up 21%. Class 4 (Self-reflective) has 2,253 comments, 17% of the dataset. Class 1 and Class 5 contain the smallest number of comments, both holding 3% of the total collected comments, 341 and 334 comments, respectively.

2.1 Model Comparison

The DL models selected for comparison were BERT, ALBERT, RoBERTa, ELECTRA, GPT, and GPT-2. This selection provides a broad overview of the performance of two leading architectural frameworks, BERT-based and GPT-based models. We selected these models over larger, more recent architectures due to practical and ethical considerations. Fine-tuning and deploying larger-scale models exceeded available computational resources. Also, AVW-Space requires near-instantaneous comment evaluation, favouring smaller models with faster inference. Additionally, performing evaluations locally safeguards student privacy by avoiding transmission of sensitive data to external entities. Each model is publicly available via Hugging Face (bert-base-uncased, albert-base-v2, roberta-base, google/electra-base-discriminator, openai-community/openai-gpt, and openai-community/gpt2) (Hugging Face, 2025).

We used the F1-score, precision, recall and average cost to compare with the results obtained from the ML benchmark testing. The goal of the evaluation was to help identify the model that best captures the complex nature of student comments, ultimately improving the accuracy of the comment classification system in AVW-Space. Model performance determined the selection of a single model for the remaining tests.

2.2 Fine-tuning and Weighted Training

Fine-tuning a model to find the optimal hyperparameters involved training models using different parameter combinations to maximise performance while ensuring generalisability to unseen data. We tested dropout rates of 0.1, 0.2, 0.3, 0.4, and 0.5; batch sizes of 16, 32, and 64; training for 2 to 6 epochs; learning rates of 1e-5, 2e-5, and 5e-5; and hidden layer counts ranging from 2 to 12. The evaluation covered 2,046 configurations, each trained and validated on the same dataset split for comparability. To address variability in DL training, the six top-performing configurations from the initial tests underwent four additional training runs. Subsequent experiments utilised the best-performing configuration.

To address class imbalance, we modified the default cross-entropy loss using focal loss for hard-to-classify examples. We applied inverse frequency weighting through both weighted cross-entropy and alpha-balanced focal loss. In addition, random undersampling balanced all classes to 334 samples to create an even distribution of comment classification.

3. Results

We tested six models to assess whether a DL approach would outperform the existing ML method to classify the quality of student comments. The selected models are BERT, ALBERT, RoBERTa, ELECTRA, GPT, and GPT-2. The BERT-based models were trained with a batch size of 64, while GPT models used 32 due to higher computational costs. Preliminary tests showed validation loss typically increased after four epochs, suggesting overfitting; thus, all the DL model training consisted of four epochs. Table 1 presents the aggregated precision, recall, F1-score and average cost obtained for each of the DL models.

Table 1. *Recall, precision, F1-score and average cost of different models*

Model	Recall	Precision	F1-Score	Average Cost
RoBERTa	**0.740**	**0.735**	**0.734**	**3.418**
ALBERT	0.738	0.734	0.732	3.513
GPT	0.731	0.728	0.728	3.658
BERT	0.724	0.722	0.723	3.738
GPT-2	0.729	0.721	0.712	3.823
ELECTRA	0.718	0.693	0.704	3.943
Retrained ML	0.719	0.706	0.692	4.030

The retrained ML model achieved a recall of 0.719, a precision of 0.706, and an F1-score of 0.692. All the DL models outperformed the retrained ML model across F1-score and average cost. RoBERTa achieved the highest performance across all metrics: a recall of 0.740, precision of 0.735, F1-score of 0.734, and the lowest average cost of 3.418. Overall, while the model performs well, further fine-tuning could improve its performance further.

ALBERT showed promising results with a recall of 0.738, precision of 0.734, and an F1-score of 0.732. ALBERT's compact architecture provides advantages in terms of computational efficiency, making it a good option for resource-constrained environments. However, RoBERTa remains the best-performing model due to its superior balance of accuracy and efficiency, which justifies its selection for further experimentation. Our findings were consistent with previous findings, which demonstrate that RoBERTa outperforms other DL alternatives (Sy et al., 2024, p. 20).

Additionally, GPT models obtained lower performance metrics than both RoBERTa and ALBERT. GPT models are generally less suited to tasks requiring fine-tuning for specific downstream applications, as their pre-training objective is broad in scope. This generality could have led to the suboptimal performance compared to models like RoBERTa, which are trained on a diverse vocabulary and further optimised for targeted tasks. However, it is important to note that the GPT models used a smaller batch size of 32 for training. The smaller batch size could have led to overfitting and reduced its ability to generalise

3.1 Fine-Tuning and Weighted Training

We tested 2,046 hyperparameter configurations for fine-tuning RoBERTa on student comment classification. Of these, 72 outperformed the default RoBERTa model (F1-score: 0.734), with the highest achieving 0.746, a marginal improvement of 0.012.

To ensure consistency, we trained the top six configurations three additional times. Results showed slight improvements over the default model but revealed that fine-tuning does not always guarantee better performance. The best performing configuration (dropout of 0.2, 10 hidden layers, batch size of 16, 6 epochs, and learning rate of 5e-5) achieved the highest F1-score of 0.747, precision of 0.749, and recall of 0.753. The configuration reduced the average cost to 3.29 from the RoBERTa default of 3.42, and therefore served as the configuration for further testing. This configuration also improved performance across all classes (Figure 1). Class 1 achieved an F1-score of 0.789, precision of 0.811 and recall of 0.768. Class 2 maintained strong performance with an F1-score of 0.829, precision of 0.794, and recall of 0.868, comparable to the retrained ML model.

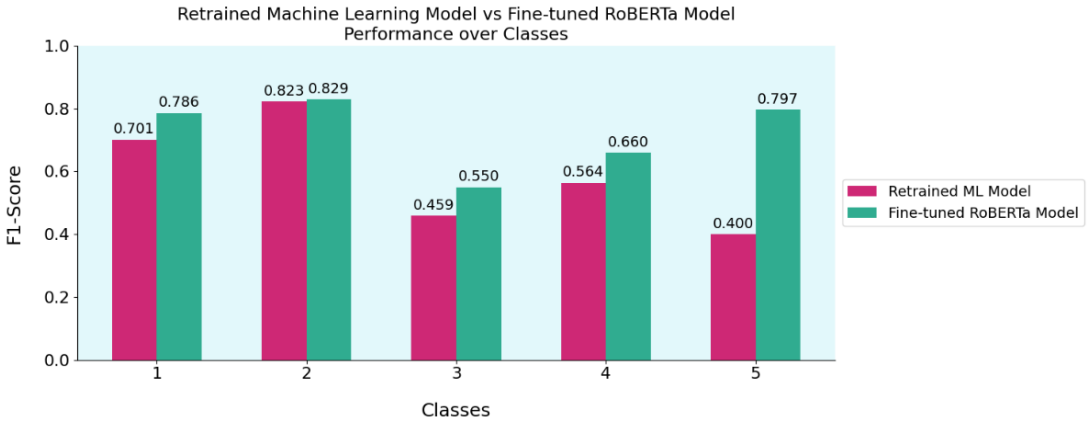

Figure 1. Retrained ML method versus the fine-tuned RoBERTa model performance

Class 3 continued to underperform with an F1-score of 0.550, precision of 0.577, and recall of 0.525, possibly due to the factual nature of comments, which made it difficult to differentiate from Class 2. For example, one comment classified correctly as Class 2, with the aspect "I did/saw this in the past," stated, "Only using meaningful numbers and statistics that enhance your point,". However, another comment with the same aspect, "Making the audience imagine works as a good opener," was misclassified as Class 2 instead of Class 3. The Class 3 classification was dependent on the content of the video. Incorporating video transcripts might improve classification performance but risks reducing generalisability.

Class 4 also improved with an F1-score of 0.660, a precision of 0.777, and a recall of 0.573. The lower recall compared to precision indicates that the model has a conservative approach to classifying into Class 4, often misclassifying as Class 2 comments. An example is a Class 4 comment with the aspect "I am rather good at this," stated, "Show you are actually interested in your topic," but was misclassified as Class 2. The example comment highlights the challenge of distinguishing self-reflective content when the reflective components are contained in the aspect rather than the comment.

Class 5 saw the most significant gain, achieving an F1-score of 0.797, a precision of 0.731, and a recall of 0.875, marking a 0.397 increase in F1-score over the retrained ML model. The higher cost of misclassifying Class 5 comments suggests that the average cost reduction was primarily due to the improved classification of these comments.

Although the aggregated F1-score presents marginal improvement from the RoBERTa model, there is a decrease in average cost and improvement in performance across all classes. The difference in each class's performance highlights the limitations of evaluating multiclassification models such as these using aggregated metrics.

The top fine-tuned model performed well for minority Class 1 and 5 but struggled with Class 3 and 4, which were harder to distinguish from the majority Class 2. To address the

class imbalance, weighted cross-entropy, focal loss, alpha-balanced focal loss, and random undersampling were tested. The default cross-entropy achieved the strongest performance with an F1-score of 0.747, while focal loss produced a comparable result with an F1-score of 0.735. Random undersampling performed considerably worse, with an F1-score of 0.624, reflecting the loss of training data. Weighted cross-entropy and alpha-balanced focal loss produced F1-scores of 0.669 and 0.674, respectively. Although these approaches improved recall for Classes 3 and 4, they also increased false positives. The models more readily classified comments into Classes 3 and 4, but did so inaccurately, resulting in decreased F1-score. Overall, the alternative methods reduced overall performance, suggesting that more targeted strategies are required to address the imbalance.

4. Discussion and Conclusions

AVW-Space enables students to leave comments that are automatically classified into quality classes using a weighted random forest classifier developed in 2020 (Mohammadhassan et al., 2020). Our goal was to determine whether DL techniques outperformed the existing ML method and whether fine-tuning and weighted techniques could optimise performance.

Training of the 2020 ML model utilised a smaller dataset. To ensure a fair comparison, we retrained it using the expanded dataset, yielding marginal improvements in the F1-score. We then evaluated six transformer-based DL models: BERT, RoBERTa, ALBERT, ELECTRA, GPT, and GPT-2. All DL F1-scores outperformed the retrained ML model, with RoBERTa emerging as the best performer. Our findings confirm RoBERTa's superior performance over other DL and ML approaches, aligning with prior research (Mohammadhassan et al., 2020; Sy et al., 2024). Fine-tuning RoBERTa by adjusting hyperparameters (dropout rate, hidden layers, batch size, number of epochs, and learning rate) further refined its performance. Weighted testing used the configuration that obtained the highest F1 score.

A limitation of our study is the imbalanced dataset, with most comments falling into Class 2. Differentiating Class 3 and 4 from Class 2 proved challenging. The factual nature of Class 3 comments and the importance of the aspect in some Class 4 comments likely contributed to this issue. Attempts to address class imbalance using alternative loss functions showed mixed results. Focal loss performed similarly to standard cross-entropy, while weighted cross-entropy and alpha-balanced focal loss reduced overall accuracy: although they increased classification into Class 3 and 4 over Class 2, these predictions were often incorrect. Undersampling also decreased performance, as the benefit of a balanced class distribution did not compensate for the information lost.

As AVW-Space evolves with new features, the nature and quality of comments will change. Future research should consider synthetic data generation or selective subsampling of recent comments to adapt to these evolving standards. Additionally, the fine-tuned model is specialised for comments on presentation skills, meaning its performance may not generalise to other topics. Further testing is needed to assess its effectiveness outside the trained domain. Also, aggregated performance metrics obscured class performance variations, suggesting that alternative evaluation methods, such as class-wise thresholds or tailored metrics, could provide a better assessment of model performance. Also, annotator bias remains a concern, emphasising the need for standardised labelling to enhance classification consistency. Incorporating video transcripts might improve classification performance but could also reduce generalisability.

Overall, this research identified a fine-tuned RoBERTa model as the best-performing DL approach, achieving an increased performance across precision, recall, F1-score and average cost. The fine-tuned RoBERTa model obtained an improved F1-score across all classes, specifically presenting vast improvements for Class 5 comments. These results highlight the effectiveness of DL techniques in enhancing comment quality classification within AVW-Space. Future work will focus on refining classification strategies, improving evaluation metrics, and developing efficient alternatives to support deployment.

Acknowledgments

We thank the members of the Intelligent Tutoring Group for their support

References

Clark, K., Luong, M.-T., Le, Q. V., & Manning, C. D. (2020). ELECTRA: Pre-training text encoders as discriminators rather than generators. International Conference on Learning Representations.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), Proc. Conf. the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol 1, (pp. 4171–4186). Association for Computational Linguistics.

Dimitrova, V., & Mitrovic, A. (2022). Choice Architecture for Nudges to Support Constructive Learning in Active Video Watching. Int. Journal of Artificial Intelligence in Education, 32(4), 892–930.

Dimitrova, V., Mitrovic, A., Piotrkowicz, A., Lau, L., & Weerasinghe, A. (2017). Using Learning Analytics to Devise Interactive Personalised Nudges for Active Video Watching. 25th Conference on User Modeling, Adaptation and Personalization, 22–31.

Hugging Face. (2025). Hugging Face Hub Documentation. https://huggingface.co/docs/hub/index

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). ALBERT: A Lite BERT for Self-Supervised Learning of Language Representations. Proc. 8th Int. Conf. Learning Representations.

Li, Y., Raković, M., Dai, W., Lin, J., Khosravi, H., Galbraith, K., Lyons, K., Gašević, D., & Chen, G. (2023). Are deeper reflectors better goal-setters? AI-empowered analytics of reflective writing in pharmaceutical education. Computers and Education: Artificial Intelligence, 5, 100157.

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-1907.

Mitrovic, A., Dimitrova, V., Lau, L., Weerasinghe, A., & Mathews, M. (2017). Supporting Constructive Video-Based Learning: Requirements Elicitation from Exploratory Studies. In E. André, R. Baker, X. Hu, Ma. M. T. Rodrigo, & B. du Boulay (Eds.), Artificial Intelligence in Education (pp. 224–237). Springer International Publishing.

Mitrovic, A., Dimitrova, V., Weerasinghe, A., & Lau, L. (2016). Reflexive Experiential Learning: Using Active Video Watching for Soft Skills Training. In W. Chen & others (Eds.), Proc. 24th Int. Conf. on Computers in Education (pp. 192–201). Asia-Pacific Society for Computers in Education.

Mitrovic, A., Gordon, M., Piotrkowicz, A., & Dimitrova, V. (2019). Investigating the Effect of Adding Nudges to Increase Engagement in Active Video Watching. In S. Isotani, E. Millán, A. Ogan, P. Hastings, B. McLaren, & R. Luckin (Eds.), Artificial Intelligence in Education (pp. 320–332).

Mohammadhassan, N., Mitrovic, A., & Neshatian, K. (2022). Investigating the effect of nudges for improving comment quality in active video watching. Computers & Education, 176, 104340.

Mohammadhassan, N., Mitrovic, A., Neshatian, K., & Dunn, J. (2020). Automatic assessment of comment quality in active video watching. In H.-J. So, Ma. M. Rodrigo, J. Mason, A. Mitrovic, D. Bodemer, W. Chen, Z.-H. Chen, B. Flanagan, M. Jansen, R. Nkambou, & L. Wu (Eds.), 28th International Conference on Computers in Education, (pp. 1–10). APSCE.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. 1–12.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., & others. (2019). Language models are unsupervised multitask learners. OpenAI Blog, 1(8), 9.

Sy, C. Y., Maceda, L. L., Canon, M. J. P., & Flores, N. M. (2024). Beyond BERT: Exploring the Efficacy of RoBERTa and ALBERT in Supervised Multiclass Text Classification. Int. Journal of Advanced Computer Science & Applications, 15(3).

Wang, Z., Hsu, C.-Y., Horikoshi, I., Li, H., Majumdar, R., & Ogata, H. (2024). Classifying Self-Reflection Notes: Automation Approaches for GOAL System. In A. Kashihara & others (Eds.),32nd Int. Conf. Computers in Education. APSCE.

Wulff, P., Mientus, L., Nowak, A., & Borowski, A. (2023). Utilizing a Pretrained Language Model (BERT) to Classify Preservice Physics Teachers' Written Reflections. Artificial Intelligence in Education, 33, 439–466.

Yousef, A. M. F., Chatti, M., & Schroeder, U. (2014). The State of Video-Based Learning: A Review and Future Perspectives. International Journal on Advances in Life Sciences, 6, 122–135.

Zhang, D., Zhou, L., Briggs, R. O., & Nunamaker, J. F. (2006). Instructional video in e-learning: Assessing the impact of interactive video on learning effectiveness. Information & Management, 43(1), 15–27.