

Automatic Assessment of Comment Quality in Active Video Watching

Negar MOHAMMADHASSAN*, Antonija MITROVIC, Kourosh NESHATIAN
& Jonathan DUNN

University of Canterbury, Christchurch, New Zealand

[*negar.mohammadhassan@pg.canterbury.ac.nz](mailto:negar.mohammadhassan@pg.canterbury.ac.nz)

Abstract: Active Video Watching (AVW-Space) is an online platform for video-based learning which supports engagement via note-taking and personalized nudges. In this paper, we focus on the quality of the comments students write. We propose two schemes for assessing the quality of comments. Then, we evaluate these schemes by computing the inter-coder agreement. We also evaluate various machine learning classifiers to automate the assessment of comments. The selected cost-sensitive classifier shows that the quality of comments can be assessed with high weighted-F1 scores. This study contributes to the automation of comment quality assessment and the development of personalized educational support for engagement in video-based learning through commenting.

Keywords: Video-based Learning, Learning Analytics, Applied Machine Learning, Text Classification.

1. Introduction

Educational videos have become a prevalent medium used in online learning. Learning through watching videos allows learners to acquire skills in various fields anywhere at their own pace. However, video watching can often be a passive activity due to the lack of direct interaction between students and teachers (Yousef et al., 2014). Thus, a crucial challenge in video-based learning (VBL) is providing support for engagement (Chatti et al., 2016).

AVW-Space is an online, controlled video-watching platform, which supports engagement via note-taking. Early studies with AVW-Space showed that only some students write comments, but those who do learn significantly more (Mitrovic et al., 2017a). In previous work, we used information about students' engagement to generate personalized nudges. Empirical studies showed that nudges are effective in supporting engagement: students who received nudges were more engaged and learnt more (Mitrovic et al., 2019). However, the initial nudges were too simplistic and focused mostly on comment writing and the types of comments.

In this paper, we present the analysis of comments quality as the primarily step for designing personalized nudges which focus on comment quality. Previous studies showed that some students write shallow comments, merely repeating what was said in the videos, without thinking deeply (Mitrovic et al., 2017b). Our goal is to develop an automatic way to assess the quality of comments as students write them, and generate personalized nudges to encourage students to write better quality comments. Towards this goal, we developed two quality schemes for comments and a Machine Learning (ML) classifier that can assess the quality of comments online. Our research questions are as follows:

1. How reliable are the quality schemes when used by human coders and machine learning models to classify comments based on their quality?
2. How does the ML model perform on data collected from different student populations and experimental settings?

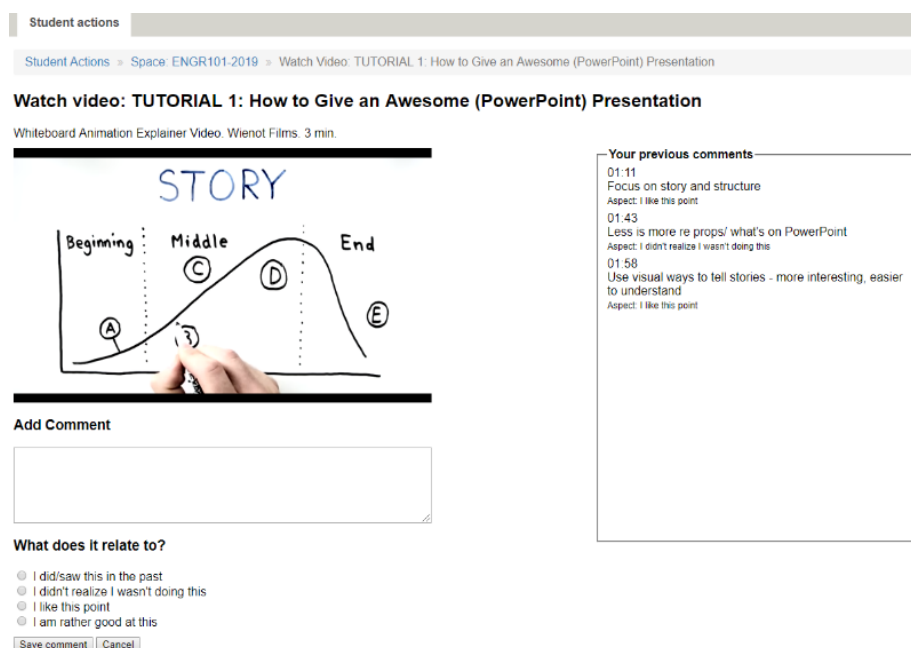
In Section 2, we introduce AVW-Space and previous studies on this platform, followed by an overview of related work on the assessment of written self-reports. After describing the datasets, we introduce and validate two quality schemes for the comments. Next, we investigate the performance of various ML models in assessing the quality of the comments. Finally, we reflect on the findings and propose potential improvements and applications.

2. AVW-Space

AVW-Space is an online, controlled VBL platform which was initially developed for teaching transferable skills. To create a new instance of AVW-Space, the teacher needs to select YouTube videos and specify micro-scaffolds for learning called aspects. Aspects are tags which direct students' attention to key points of the videos or prompt students to reflect on their knowledge and experience. Initially, students watch and comment on videos individually. To write a comment, the student can stop the video at any time, enter the text of the comment and select one of the aspects (Figure 1). In the second stage, the teacher selects the comments to be displayed anonymously to the class so that students can review and rate comments made by other students in the class.

Several studies have been conducted with an instance of AVW-Space focusing on presentation skills (Dimitrova et al., 2017; Hecking et al., 2017; Mitrovic et al., 2017a; Mitrovic et al., 2019; Taskin et al., 2019). The instance contained four tutorials on giving presentations. Students were instructed to watch tutorial videos first. The aspects for the tutorial videos are: "I am rather good at this", "I did/saw this in the past", "I didn't realize I wasn't doing it", and "I like this point". There were also four example videos of real presentations. The students were asked to critique those examples addressing the given aspects: "structure", "visual aids", "delivery" and "speech". Early studies with AVW-Space revealed that only students who wrote and rated comments improved their knowledge, while no increase in knowledge was observed for students who passively watched videos (Mitrovic et al., 2017a).

Based on the results from the initial studies, AVW-Space was enhanced by adding simple personalized nudges to foster engagement. The nudges provide hints to students based on their commenting behavior, encouraging students to write comments and providing examples of comments made by previous students. A study with the enhanced version of AVW-Space showed that students who received nudges wrote significantly more comments compared to the students who interacted with the original version of AVW-Space (Mitrovic et al., 2019). However, there were no nudges that provided feedback on the quality of comments, which is the motivation for the research presented in this paper. Although there have been several studies on comments in AVW-Space such as predicting whether a comment will receive a high number of ratings (Dimitrova et al., 2017) and identifying the pattern of vocabularies used by different types of the learners (Hecking et al., 2017), there has been no research on assessing the quality of comments.



Student actions

Student Actions > Space ENGR101-2019 > Watch Video: TUTORIAL 1: How to Give an Awesome (PowerPoint) Presentation

Watch video: TUTORIAL 1: How to Give an Awesome (PowerPoint) Presentation

Whiteboard Animation Explainer Video. Wienot Films. 3 min.

STORY

Beginning Middle End

A B C D E

Add Comment

What does it relate to?

- I did/saw this in the past
- I didn't realize I wasn't doing this
- I like this point
- I am rather good at this

Save comment Cancel

Your previous comments

01:11
Focus on story and structure
Aspect: I like this point

01:43
Less is more re props/ what's on PowerPoint
Aspect: I didn't realize I wasn't doing this

01:58
Use visual ways to tell stories - more interesting, easier to understand
Aspect: I like this point

Figure 1. AVW-Space commenting environment.

3. Analysis of Written Self-reports

Different forms of written self-reports analysis have been conducted to assess students' understanding of taught concepts. Preliminary research in this area focused on the textual analysis of students' essays (Carroll, 2007; McNamara, Crossley, & Roscoe, 2013), which are long, structured pieces of text compared to video comments. Text analysis has also been used for assessing answers to questions asked during or after teaching sessions (Arbogast & Montfort, 2016; Prevost et al., 2013). However, answers to pre-designed questions are less open-ended than comments in VBL. With the increased use of Massive Open Online Courses (MOOCs), there have been studies on online discussions (Crossley et al., 2015; Martín-Monje, Castrillo, & Mañana-Rodríguez, 2018) which are similar to comments in terms of length and flexibility in the context, but comments in AVW-Space are not conversational because students write them individually. The differences between comments and other types of self-reports underline the need for more dedicated research on comments in VBL platforms.

Many attempts have been made to investigate video annotations using linguistic features or ontologies. The focus of linguistic analysis is on syntactic and semantic components of textual data. A study on video annotations in the CLAS note-taking environment (Risko et al., 2012) proposed a scheme to identify the level of reflections in the annotations (Joksimović et al., 2019). This study used Coh-Matrix, a computational linguistics facility (Graesser et al., 2014), to derive linguistic properties to determine the depth of students' self-reflection. However, given the differences in educational environments, this scheme cannot be applied to other contexts. The study on predicting the social value of comments in AVW-Space (Dimitrova et al., 2017) also investigated the linguistic features of the comments using Linguistic Inquiry and Word Count (LIWC) (Tausczik & Pennebaker, 2010).

Content analysis, which clusters text into different conceptual categories, is used for comparing student's knowledge with the educational materials (Daems et al., 2014). In DiViDu (Hulsman, Harmsen, & Fabriek, 2009), four reflection categories (Observations, Motives, Effects and Goals) were suggested for written self-reflection of students on the recorded video of their communication. Then, the content of annotations in each category was analyzed manually. This reflection scheme has inspired further research on assessment of reflective notes by automated text analysis approaches (Joksimović et al., 2018; Martín-Monje et al., 2018). A study on AVW-Space comments used domain-specific concept ratio¹ to infer whether a comment is on-topic (Dimitrova et al., 2017). The pitfall of this measure is that for a comment which contains only one word from the domain ontology, the domain-specific ratio is equal to 1; such comments are treated as good comments while they do not illustrate critical thinking.

Network-text analysis (NTA) is a method in content analysis to extract relations between ontologies used by the learners. NTA models the text as a network of concepts (Popping, 2000). In JuxtaLearn (Daems et al., 2014), NTA was conducted on Science, Technology, Engineering, and Mathematics (STEM) video comments to assess students' conceptual change. NTA was also used to uncover the contextual pattern of comments in AVW-Space made by different types of learners (Hecking et al., 2017). This study showed that students with high self-regulatory skills wrote many comments containing most of the domain vocabulary, while students with weak learning skills wrote few comments using a small subset of the domain vocabulary. Overall, the content analysis of annotations is not sufficient for the quality assessment of comments since there could be high-quality comments having words from learners' own vocabularies.

Common approaches for automating the assessments of self-reports are rule-based, dictionary-based and machine learning methods (Ullmann, 2019). The rule-based approach utilizes rules defined by a human expert and a rule engine inferring information from the textual data. Defining rules manually and evaluating them could yet be labor-intensive tasks. In the dictionary-based approach, the frequency of words in various categories is analyzed. The accuracy of the dictionary-based approach depends on the selected dictionaries. However, machine learning approaches learn patterns from data automatically and recently have been frequently used for assessing reflective texts (Crossley et al., 2019; Liu et al., 2019).

¹ The number of words from the domain ontology appearing in the comment, divided by the total number of words in the comment

4. Materials and Methods

We used the data collected from the previous AVW-Space studies conducted in a first-year engineering course at the University of Canterbury in 2017, 2018 and 2019. This course used AVW-Space as an online resource for training students on presentation skills. We also used the data from a study conducted with postgraduate (PG) students (Mitrovic et al., 2016). All four studies used the same videos and aspects. The 2018 and 2019 studies shared the identical experiment design, with control and experimental groups. In addition to the standard AVW-Space features (aspects, videos and rating categories), the experimental group received nudges which encouraged students to write more comments using a variety of aspects (Mitrovic et al., 2019). However, the 2017 study (Mitrovic et al., 2017b) and the PG study did not include nudges. Table 1 presents the number of students who wrote comments and the number of comments on tutorial/example videos for each study.

Table 1. *Number of Students and Comments in the Studies*

	2017	2018	2019	PG
Participants	158	191	146	32
Comments on tutorial videos	670	1,144	1,101	346
Comments on example videos	575	687	660	368

4.1 Quality Schemes

To assess the quality of comments, it is necessary to define different categories of comments. There have been several frameworks proposed for students' reflective writing. A study on academic reflective essays identifies different types of reflection such as personal belief, lessons learned or future intentions (Ullmann, 2017; 2019). Another framework for assessing the depth of students' self-reflections on their performance groups the contents into observation, effect, or motivation and goal (Joksimović et al., 2019). A simulation environment for cross-cultural communications classifies the user's textual interactions with the system into different groups, such as statements on the situation and real-world stories (Dimitrova et al., 2013). However, a new labelling scheme for comments is needed for AVW-Space due to its special nature. Thus, we initially explored comments and their aspects, and proposed two quality schemes, one for comments on tutorial videos and another for those on example videos.

Table 2 presents the categories of comments on tutorial videos, with some examples from previous AVW-Space studies. This scheme contains five categories: (1) Affirmative, negative or off-topic, (2) Repeating, (3) Critical and analytical, (4) Self-reflective and (5) Self-regulating comments. Comments in categories 1 and 2 are pedagogically undesirable since they do not convey deep thinking about the videos. However, comments in category 3 show more critical thinking about the video, as learners elaborate on the video content. In category 4, learners reflect on their previous experience in relation to the video. Finally, students indicate a high level of learning in category 5 by planning how to improve their future presentations using the ideas covered in the videos. We also designed a quality scheme for comments on example videos (Table 3), which is similar to the scheme for tutorial videos with the exception of self-regulation and self-reflection categories. The reason for excluding those two categories is that the students were instructed to critique example videos using aspects like structure, visual aids, delivery and speech.

To evaluate the proposed quality schemes, we selected 167 comments via stratified sampling (110/57 comments on tutorial/example videos respectively) from 2018 and 2019 studies. Three expert coders labelled those comments independently. We used the ordinal Krippendorff's α (Krippendorff, 2010) to calculate inter-coder agreement, due to the number of coders and ordinal categories of the schemes. Krippendorff's α values were 0.78 and 0.69 for tutorial/example videos, respectively. Krippendorff suggests that the lowest acceptable value of α for inter-coder agreement is 0.66 (Krippendorff, 2010). After reviewing the comments on which the coders disagreed, we clarified the definitions in the schemes for further manual classification.

Table 2. *Quality Scheme for Comments on the Tutorial Videos*

Category	Definition
1. Affirmative, negative, off-topic	Comments which are irrelevant or merely affirmative/negative with no explanation. e.g. [Aspect: I did/saw this in the past] <i>“very helpful.”</i>
2. Repeating	Comments which only repeat the video content. e.g. [Aspect: I like this point] <i>“limit each slide to one key idea.”</i>
3. Critical and analytical	Comments which mention points that are implicitly covered in the video, or show critical thinking on the content of the video. e.g. [Aspect: I like this point] <i>“Presentations can be boring and long whereas stories are more enjoyable and can have clear direction if formulated properly.”</i>
4. Self-reflective	Comments in which the learner reflects on his/her behavior and previous experience or knowledge on giving presentations. e.g. [Aspect: I saw/did this in the past] <i>“My past speeches have had very interesting beginnings.”</i>
5. Self-regulating	Comments where the learner decides what they would do to improve themselves in future. e.g. [Aspect: I didn’t realize I wasn’t doing this] <i>“I will definitely be trying to smile more throughout my next presentation.”</i>

Table 3. *Quality scheme for comments on the example videos*

Category	Definition
1. Affirmative, negative	Comments which are irrelevant or merely affirmative/negative with no explanation. e.g. [Aspect: Visual aids] <i>“This was helpful.”</i>
2. Repeating	Comments that list or name good/bad practices in the presentations without explaining the effects and causes of the practice. e.g. [Aspect: Speech] <i>“End on a question.”</i>
3. Critical and analytical	Comments which criticize examples, explain the effect of a good/bad practice in the presentation or offer advice for improvement. e.g. [Aspect: Speech] <i>“Should give more meaning to the statistic by placing it in context.”</i>

4.2 Automation of Assessment using Machine Learning Approaches

To automate the assessment of comments quality using machine learning, we needed to convert comments to numerical features. We took a word count approach (Tausczik & Pennebaker, 2010) for extracting numerical features rather than full parsing, since these comments are not always grammatically correct. Therefore, the text is converted to numbers by counting the frequencies of words (or their stems) in dictionaries. We extracted 94 features for each comment using LIWC (Tausczik & Pennebaker, 2010). LIWC dictionaries are collected from various psychological constructs such as cognitive processes. Also, LIWC can calculate linguistic elements including pronouns, verbs and their tenses. As mentioned earlier, previous studies on reflective writing in various contexts also used LIWC features (Joksimović et al., 2018; Liu et al., 2019). In addition to LIWC features, we used the domain-specific ratio and unique domain-specific ratios² proposed in a previous study (Mitrovic et al., 2019) to consider the topic of comments. We also created a new binary feature indicating whether the used aspect is reflective (“I didn’t realize I was doing this”, “I am rather good at this” or “I did/saw this in the past”).

² The number of unique words from the domain ontology appearing in the comment, divided by the total number of words in the comment

We normalized all features to values between 0 and 1. To reduce the size of the feature set, we removed LIWC features that are not meaningful in this context, such as “biological” words or “personal concerns”. We also removed LIWC summary features such as “clout” and “authentic”, which are derived from primary features. Therefore, the LIWC features we selected are as follow:

word count, word per sentence, six-letter (or more) words, dictionary words, function words, pronouns, personal pronouns, I, we, you, she/he, they, impersonal pronouns, article, prepositions, auxiliary verbs, adverbs, conjunctions, negate, verbs, adjectives, comparative, interrogatives, number, quant, affect, positive emotions, negative emotions, social, cognitive processes, insight, causation, descriptive, tentative, certainty, differentiation, perceptual processes, see, hear, feel, ingest, drives, affiliation, achieve, power, reward, risk, focus on past (past tenses/adverbs), focus on present, focus on future, relativity, informal, swear words, net-speak, assent, non-fluencies, filler

To train different ML models and evaluate them, we merged the comments from 2018 and 2019 studies, since the experiment design was identical in those two studies. Overall, we had 1,347/2,245 comments on example/tutorial videos, which we randomly split into the training (80%) and test (20%) sets on the student level. That is, the students whose comments are in the training set have no comment in the test set. After labelling all comments manually using the quality schemes, we noticed that the data is imbalanced, as the numbers of comments in each category are different, as illustrated in Table 4 for the training set.

Table 4. *Categories Distribution in Tutorial and Example Comments in the Training Set*

Category	Tutorial comments					Example comments		
	1	2	3	4	5	1	2	3
Comments	48	924	395	387	41	27	668	320

To deal with the imbalanced data, we initially set class weights inversely proportional to the frequency of classes. Then, we examined traditional classifiers, such as support vector machine and decision trees. We also applied the ordinal classification (Frank & Hall, 2001) to consider the order for the quality categories. However, these classifiers did not perform well. Therefore, we took a cost-sensitive approach to define proper costs for different misclassifications regarding educational purposes. For example, misclassifying a comment in category 1 as 3 is worse than misclassifying a comment in class 3 as 5, because comments in category 1 show very limited engagement and therefore must be correctly detected and supported appropriately. Thus, the cost of misclassifying a comment in category 1 as 3 should be higher. We designed the cost matrices presented in Figure 2, which still consider the order of categories by increasing costs as the misclassification distance grows. In these matrices, misclassifications for categories 1 and 5 for tutorial videos and categories 1 and 3 for example videos have higher costs since the comments in category 1 need educational support the most. At the same time, a high-quality comment does not require pedagogical interventions. After many experiments with this approach, random-forest was selected as the well-performing base classifier. We refer to these classifiers as T_a and E_a , for comments on tutorial/example videos respectively.



Figure 2. Selected Cost Matrices for Comments on Tutorial (left) and Example Videos (right)

Table 5 reports the weighted mean of metrics for evaluating the performance of the cost-sensitive classifiers. F1-score is recommended for evaluating ML models on imbalanced data (Jeni, Cohn, & De La Torre, 2013). We also used the average cost and cost-saving to evaluate models in terms of costs. Cost-saving (equation 1) is the fraction by which the actual predictions reduce the costs in the

worst case of misclassification. We started with models E_a and T_a which were trained to identify each category from the corresponding quality scheme. That is, classifier T_a identifies comments belonging to one of the five categories in the quality scheme for the tutorial videos. The performance of classifiers T_a and E_a is not satisfactory. Therefore, we considered whether it is necessary to be able to predict each category of comments individually. For example, categories 4 and 5 of the tutorial comments show a very high reflection level; in such cases, a pedagogical intervention (in the form of a nudge) is not needed. When the student initially writes a comment belonging to one of those two categories, positive feedback would be encouraging, but providing positive feedback on each high-quality comment would not be appropriate for well-performing students. For that reason, we considered various combinations of quality categories and trained different classifiers. For comments on example videos, the only category requiring a nudge is category 1, so we grouped comments from categories 2 and 3 together; the resulting classifier E_b differentiates between two types of comments (category 1 versus the union of categories 2 and 3). For tutorial comments, we explored various groupings resulting in classifiers T_b to T_d .

$$cost_saving = 1 - \frac{cost\ of\ predictions}{Maximum_Cost} \quad (1)$$

As can be seen in Table 5, classifiers T_c , T_d and T_e have better performance than the initial model (T_a). Classifier T_d aligns well with the ICAP framework (Chi & Wylie, 2014): active learning (just repeating the lecture) is represented by categories 1 and 2; constructive learning (adding information that was not explicitly taught) is captured by categories 3, 4 and 5. However, having only 2 categories is not enough for capturing different behaviours and providing adequate support. Classifier T_c is similar to T_d , but it distinguishes between categories 1 and 2 to provide proper support. Classifier T_e groups comment into “off-topic/short affirmative or negative”, “commenting on the video” and “reflecting”. For comments on example videos, classifier E_b predicts whether a comment is describing the video and analyzing the strengths or weaknesses of the presentation in the video. E_b and T_e were selected as the best-performing classifiers.

Table 5. *The Performance of the Models on the Test Set*

Video	Model	TPR	FPR	Precision	F1-score	Avg. Cost	Cost-saving
Example	E_a : 1, 2, 3	0.71	0.22	0.75	0.71	3.79	0.85
	E_b: 1, 2+3	0.95	0.21	0.97	0.96	0.99	0.98
Tutorial	T_a : 1, 2, 3, 4, 5	0.72	0.18	0.72	0.68	3.53	0.877
	T_b : 1, 2, 3, 4+5	0.70	0.18	0.72	0.64	4.42	0.873
	T_c : 1, 2, 3+4+5	0.80	0.17	0.80	0.80	2.86	0.886
	T_d : 1+2, 3+4+5	0.74	0.26	0.80	0.73	3.10	0.877
	T_e: 1, 2+3, 4+5	0.84	0.15	0.86	0.84	2.08	0.881

Figure 3 presents the confusion matrices for the selected classifiers. Classifier T_e can identify comments in category 1 correctly. The only misclassifications in category 1 is for a comment saying “hello Jim” (Jim is the name of a character in a tutorial video). Besides, some of the very short comments in higher quality categories were misclassified as 1. For instance, “guideposts” should be classified as 2+3, but it is misclassified as 1. The F1-scores for classes 1, 2+3 and 4+5 are 0.72, 0.89 and 0.72 respectively. In classifier E_b , most of the comments in category 1 are classified correctly, but there are some misclassifications in class 2+3. The two misclassifications in category 1 are for comments that use domain-specific concepts to discuss the subject of the example rather than criticizing the presentation skills of the presenters. For example, “valid points” is misclassified as 2+3 since the comment includes “valid” and “points” which are two concepts from the domain ontology.

<i>Classified as</i>	$\begin{matrix} & 1 & 2+3 & 4+5 \\ \begin{matrix} actual = 1 \\ actual = 2 + 3 \\ actual = 4 + 5 \end{matrix} & \begin{pmatrix} 8 & 1 & 0 \\ 4 & 283 & 47 \\ 1 & 18 & 85 \end{pmatrix} \end{matrix}$	<i>Classified as</i>	$\begin{matrix} & 1 & 2+3 \\ \begin{matrix} actual = 1 \\ actual = 2 + 3 \end{matrix} & \begin{pmatrix} 7 & 2 \\ 13 & 310 \end{pmatrix} \end{matrix}$
----------------------	--	----------------------	---

Figure 3. Confusion Matrices for T_e (left) and E_b (right)

4.3 Generalizability of the Classifiers

We also evaluated the performance of the chosen models on unseen comments, from the 2017 and PG studies. The 2017 study was done with a similar population of students (the 2017 class of the same course), who have not received nudges. Therefore, we expected the performance of the classifiers to be similar to that on the test set. On the other hand, PG students usually have much stronger learning and metacognitive skills. Therefore, we expected that the performance of the classifiers on the PG data would be worse in comparison to the performance on the training/test set.

The performance of selected classifiers on the data from 2017 and PG studies is reported in Table 6. These classifiers perform differently for these datasets because the distributions of the combined categories are different from those in the 2018/2019 studies, as highlighted in Table 7. Postgraduate students wrote more reflective comments (categories 4 and 5) than the other groups of students. Also, the PG students made slightly more high-quality comments on example videos (categories 2 and 3) than first-year students. When looking at first-year students only, the 2017 set differs from the 2018/2019 data sets; the provision of nudges in 2018/2019 resulted in a higher number of reflective comments (Mitrovic et al., 2019).

Table 6. *The performance of merged categories models on 2017 and PG data*

Data	Model	TPR	FPR	Precision	F1-score	Avg. Cost	Cost-saving
2017	E_b : 1, 2+3	0.93	0.26	0.96	0.94	1.54	0.88
	T_e : 1, 2+3, 4+5	0.72	0.26	0.81	0.74	3.58	0.77
PG	E_b : 1, 2+3	0.96	0.85	0.96	0.96	2.08	0.82
	T_e : 1, 2+3, 4+5	0.68	0.30	0.70	0.69	4.61	0.79

Table 7. *Percentages of categories in different data sets*

Video	Categories	Training	Test	2017	PG
Example	1	2.66	2.72	5.91	2.16
	2+3	97.33	97.28	94.09	97.83
Tutorial	1	2.67	2.01	3.58	0.29
	2+3	73.48	74.72	79.40	58.95
	4+5	23.84	23.26	17.01	40.75

5. Conclusions and Future Work

In this research, we focused on the assessment of comment quality, as a starting point towards enhancing tailored support for engagement in AVW-Space. We proposed and evaluated two quality schemes for assessing comments in AVW-Space. We also automated the quality assessment of comments using a cost-sensitive approach. We tried combining categories to further improve the performance of the classifiers and to simplify the design of pedagogical interventions. Next, we selected the best-performing combinations as the automatic quality classifiers. The generalizability of these classifiers was also assessed by evaluating them on unseen data from two studies with different experimental setups. The performance of the classifiers was slightly lower for the 2017/PG datasets. However, there is still room for improving the performance of these classifiers for low-quality comments by trying

other cost-functions and a deeper feature-engineering. Furthermore, since the classifiers were trained and tested only on comments about giving presentations, the generalizability of this approach for other domains should be also investigated in future work.

The results of this study will enable us to design personalized nudges focusing on the quality of comments a student writes. For instance, when a student submits a comment in category 1, the system would provide an immediate nudge to encourage the learner to be more focused on the video content in his/her comment. When a student submits a comment in category 2 or 3, a nudge could suggest more elaboration or self-reflection. The system should also give positive feedback to the student when he/she writes a self-reflective or self-regulating comment.

Acknowledgments

AVW-Space has been developed in collaboration with Professor Vania Dimitrova and Dr Lydia Lau from the University of Leeds, and Dr Amali Weerasinghe. The authors are grateful to Jay Holland for technical support. We thank Professor Peter Gostomski and Dr Alfred Herritsch for their contribution in conducting the studies.

References

- Arbogast, C. A., & Montfort, D. (2016). Applying natural language processing techniques to an assessment of student conceptual understanding. *Proceedings of ASEE Annual Conference and Exposition*.
- Carroll, D. W. (2007). Patterns of student writing in a critical thinking course: A quantitative analysis. *Assessing Writing, 12*(3), 213–227.
- Chatti, M. A., Marinov, M., Sabov, O., Laksono, R., Sofyan, Z., Yousef, A. M. F., & Schroeder, U. (2016). Video annotation and analytics in CourseMapper. *Smart Learning Environments, 3*(1), 10.
- Chi, M. T. H., & Wylie, R. (2014). The ICAP Framework: Linking Cognitive Engagement to Active Learning Outcomes. *Educational Psychologist, 49*(4), 219–243.
- Crossley, S. A., Kim, M., Allen, L., & McNamara, D. (2019). Automated Summarization Evaluation (ASE) Using Natural Language Processing Tools. In S. Isotani, E. Millán, A. Ogan, P. Hastings, B. McLaren, & R. Luckin (Eds.), *Artificial Intelligence in Education* (pp. 84–95). Springer International Publishing.
- Crossley, S., McNamara, D., Baker, R., Wang, Y., Paquette, L., Barnes, T., & Bergner, Y. (2015). Language to Completion: Success in an Educational Data Mining Massive Open Online Class. *Proceedings of the 8th International Conference on Educational Data Mining*, (pp. 388-391).
- Daems, O., Erkens, M., Malzahn, N., & Hoppe, H. U. (2014). Using content analysis and domain ontologies to check learners' understanding of science concepts. *Computers in Education, 1*(2), 113–131.
- Dimitrova, V., Mitrovic, A., Piotrkowicz, A., Lau, L., & Weerasinghe, A. (2017). Using Learning Analytics to Devise Interactive Personalised Nudges for Active Video Watching. *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*, (pp. 22–31). Bratislava, Slovakia: ACM.
- Dimitrova, V., Steiner, C., Despotakis, D., Brna, P., Ascolese, A., Pannese, L., & Albert, D. (2013). Crowdsourcing for Evaluating a Simulated Learning Environment for Interpersonal Communication and Cultural Awareness. *Workshop on Culturally-aware Technology Enhanced Learning (CulTEL 2013)*.
- Frank, E., & Hall, M. (2001). A Simple Approach to Ordinal Classification. In L. De Raedt & P. Flach (Eds.), *Machine Learning: ECML 2001* (pp. 145–156). Springer Berlin Heidelberg.
- Graesser, A. C., McNamara, D., Cai, Z., Conley, M., Li, H., & Pennebaker, J. (2014). Coh-Metrix Measures Text Characteristics at Multiple Levels of Language and Discourse. *The Elementary School Journal, 115*(2), 210–229.
- Hecking, T., Dimitrova, V., Mitrovic, A., & Hoppe, H. U. (2017). Using Network-Text Analysis to Characterise Learner Engagement in Active Video Watching. *Proceedings of the 25th International Conference on Computers in Education*, (pp. 326–335). Asia-Pacific Society for Computers in Education.
- Hulsman, R. L., Harmsen, A. B., & Fabriek, M. (2009). Reflective teaching of medical communication skills with DiViDU: Assessing the level of student reflection on recorded consultations with simulated patients. *Patient Education and Counseling, 74*(2), 142–149.
- Jeni, L. A., Cohn, J. F., & De La Torre, F. (2013). Facing imbalanced data - recommendations for the use of performance metrics. In *2013 Humaine association conference on affective computing and intelligent interaction* (pp. 245-251). IEEE.

- Joksimović, S., Dowell, N., Gašević, D., Mirriahi, N., Dawson, S., & Graesser, A. C. (2019). Linguistic characteristics of reflective states in video annotations under different instructional conditions. *Computers in Human Behavior*, 96, 211-222.
- Krippendorff, K. (2010). Krippendorff's Alpha. In N. Salkind (Ed.), *Encyclopedia of Research Design*, 669-673. SAGE Publications.
- Liu, M., Shum, S. B., Mantzourani, E., & Lucas, C. (2019). Evaluating Machine Learning Approaches to Classify Pharmacy Students' Reflective Statements. In S. Isotani, E. Millán, A. Ogan, P. Hastings, B. McLaren, & R. Luckin (Eds.), *Proceedings of the 20th International Conference of Artificial Intelligence in Education* (pp. 220-230). Springer International Publishing.
- Martín-Monje, E., Castrillo, M. D., & Mañana-Rodríguez, J. (2018). Understanding online interaction in language MOOCs through learning analytics. *Computer Assisted Language Learning*, 31(3), 251-272.
- McNamara, D., Crossley, S., & Roscoe, R. (2013). Natural language processing in an intelligent writing strategy tutoring system. *Behavior Research Methods*, 45(2), 499-515.
- Mitrovic, A., Dimitrova, V., Weerasinghe, A., & Lau, L. (2016). Reflective Experiential Learning: Using Active Video Watching for Soft Skills Training (W. Chen, Ed.), *Proceedings of the 24th International Conference on Computers in Education* (pp. 192-201). Asia-Pacific Society for Computers in Education.
- Mitrovic, A., Dimitrova, V., Lau, L., Weerasinghe, A., & Mathews, M. (2017a). 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.), *Proceedings of the 18th International Conference of Artificial Intelligence in Education* (pp. 224-237). Springer International Publishing.
- Mitrovic, A., Gostomski, P., Herritsch, A., Dimitrova, V. (2017b) Improving presentation skills of first-year engineering students using Active Video Watching. In N. Huda, D. Inglis, N. Tse, & G. Town (Eds.), *Proceedings of the 28th Annual Conference of the Australasian Association for Engineering Education* (pp. 809-816).
- 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.), *Proceedings of the 20th International Conference of Artificial Intelligence in Education* (pp. 320-332). Springer, Cham.
- Popping, R. (2000). *Computer-assisted Text Analysis* (p. 97). Sage.
- Prevost, L. B., Haudek, K. C., Henry, E. N., Berry, M. C., & Urban-Lurain, M. (2013). Automated text analysis facilitates using written formative assessments for just-in-time teaching in large enrollment courses. *Proceedings of ASEE Annual Conference and Exposition* (pp. 23-26).
- Risko, E. F., Foulsham, T., Dawson, S., & Kingstone, A. (2012). The collaborative lecture annotation system (CLAS): A new TOOL for distributed learning. *IEEE Transactions on Learning Technologies*, 6(1), 4-13.
- Taskin, Y., Hecking, T., Hoppe, H. U., Dimitrova, V., & Mitrovic, A. (2019). Characterizing Comment Types and Levels of Engagement in Video-Based Learning as a Basis for Adaptive Nudging. In M. Scheffel, J. Broisin, V. Pammer-Schindler, A. Ioannou, & J. Schneider (Eds.), *Proceedings of European Conference on Technology-Enhanced Learning* (pp. 362-376). Springer International Publishing.
- Tausczik, Y. R., & Pennebaker, J. W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, 29(1), 24-54.
- Ullmann, T. D. (2017). Reflective writing analytics: Empirically determined keywords of written reflection. *Proceedings of 7th International Learning Analytics & Knowledge Conference*, (pp. 163-167). Vancouver, British Columbia, Canada: ACM.
- Ullmann, T. D. (2019). Automated Analysis of Reflection in Writing: Validating Machine Learning Approaches. *International Journal of Artificial Intelligence in Education*, 29(2), 217-257.
- Yousef, A. M. F., Chatti, M. A., & Schroeder, U. (2014). The state of video-based learning: A review and future perspectives. *International Journal of Advanced Life Sciences*, 6(3/4), 122-135.