

Learner-centred Analytics of Feedback Content in Higher Education

JIONGHAO LIN, Centre for Learning Analytics, Monash University, Australia

WEI DAI, Centre for Learning Analytics, Monash University, Australia

LISA-ANGELIQUE LIM, University of Technology Sydney, Australia

YI-SHAN TSAI, Centre for Learning Analytics, Monash University, Australia

RAFAEL FERREIRA MELLO, CESAR School, Brazil and Centre for Learning Analytics, Monash University, Australia

HASSAN KHOSRAVI, The University of Queensland, Australia

DRAGAN GAŠEVIĆ, Centre for Learning Analytics, Monash University, Australia

GUANLIANG CHEN*, Centre for Learning Analytics, Monash University, Australia

Feedback is an effective way to assist students in achieving learning goals. The conceptualisation of feedback is gradually moving from *feedback as information* to *feedback as a learner-centred process*. To demonstrate feedback effectiveness, feedback as a learner-centred process should be designed to provide quality feedback content and promote student learning outcomes on the subsequent task. However, it remains unclear how instructors adopt the learner-centred feedback framework for feedback provision in the teaching practice. Thus, our study made use of a comprehensive learner-centred feedback framework to analyse feedback content and identify the characteristics of feedback content among student groups with different performance changes. Specifically, we collected the instructors' feedback on two consecutive assignments offered by an introductory to data science course at the postgraduate level. On the basis of the first assignment, we used the status of student grade changes (i.e., students whose performance increased and those whose performance did not increase on the second assignment) as the proxy of the student learning outcomes. Then, we engineered and extracted features from the feedback content on the first assignment using a learner-centred feedback framework and further examined the differences of these features between different groups of student learning outcomes. Lastly, we used the features to predict student learning outcomes by using widely-used machine learning models and provided the interpretation of predicted results by using the SHapley Additive exPlanations (SHAP) framework. We found that 1) most features from the feedback content presented significant differences between the groups of student learning outcomes, 2) the gradient boost tree model could effectively predict student learning outcomes, and 3) SHAP could transparently interpret the feature importance on predictions.

CCS Concepts: • **Applied computing** → **Education**; • **Computing methodologies** → *Natural language processing*.

Additional Key Words and Phrases: Feedback, Learning Analytics, Content Analysis, Interpretability

1 INTRODUCTION

Feedback has been acknowledged as an effective component in promoting students' learning in higher education [5, 11, 17, 41]. The conceptualisation of feedback in the existing literature has gradually shifted from feedback as information to feedback as a learner-centred process [41]. The earlier definition (i.e., feedback as information) of

*Corresponding Author

Authors' addresses: Jionghao Lin, jionghao.lin1@monash.edu, Centre for Learning Analytics, Monash University, Melbourne, Australia; Wei Dai, wei.dai1@monash.edu, Centre for Learning Analytics, Monash University, Melbourne, Australia; Lisa-Angelique Lim, lisa-angelique.lim@uts.edu.au, University of Technology Sydney, Sydney, Australia; Yi-Shan Tsai, yi-shan.tsai@monash.edu, Centre for Learning Analytics, Monash University, Melbourne, Australia; Rafael Ferreira Mello, rafaelmello@gmail.com, CESAR School, Recife, Brazil and Centre for Learning Analytics, Monash University, Melbourne, Australia; Hassan Khosravi, h.khosravi@uq.edu.au, The University of Queensland, St Lucia, QLD, Australia; Dragan Gašević, dragan.gasevic@monash.edu, Centre for Learning Analytics, Monash University, Melbourne, Australia; Guanliang Chen, guanliang.chen@monash.edu, Centre for Learning Analytics, Monash University, Melbourne, Australia.

effective feedback focuses on the information and timeliness [16], whereas the recent understanding (i.e., feedback as the learner-centred process) of feedback focuses more on the student's learning process [41]. Many previous studies found that feedback as a learner-centred process could lead to more effective feedback design to promote learners' achievements compared with the feedback as information [1, 2, 12]. To better understand the construction of learner-centred feedback, Dawson et al. [11] interviewed students and instructors about their experience on the feedback effectiveness. Students indicated that the quality of feedback content (e.g., actionable information for improvement on subsequent tasks) was the most effective aspect of feedback, whereas instructors indicated that feedback should be provided in the context of connected tasks (i.e., the feedback provided for the first task is connected with the subsequent task) [11]. These indications of effective feedback were further affirmed by a recent learner-centred feedback framework [35], which demonstrated a comprehensive set of effective feedback attributes (e.g., comments that provide actionable information for future performance) for textual feedback in higher education.

Effective feedback should deliver high-quality content and demonstrate the effect on student improvement such as improvement in learning outcomes and conceptual understanding [11, 35]. To understand how feedback content affects student improvement, researchers employed Learning Analytics (LA) approaches to analyse the correlation between feedback content and the students' learning outcomes. For example, Nicoll et al. [30] extracted textual features (e.g., N-grams) from feedback on the first assignment and measured student grade changes (e.g., grade increase and decrease) on the subsequent assignment as the proxy of the students' learning outcomes. Then, they [30] used LA approaches to analyse the correlation between the textual features and student grade changes. Though Nicoll et al. [30] demonstrated a method to understand the effect of feedback content on student learning outcomes, we argued that there was no explicit discussion between the textual features in their work [30] and the theoretically grounded attributes in the learner-centred feedback framework [35]. Additionally, the feedback effect (i.e., improvement in students' learning outcomes on the subsequent task) was not substantial since the connection between the two assignments was unclear in their work. Therefore, our study aimed to identify the textual features of feedback content based on a learner-centred framework and examine the use of these features among different groups categorised by students' subsequent learning outcomes on the consecutive assignment.

To facilitate the delivery of effective feedback, LA researchers also adopted machine learning models to automate the process of evaluating feedback quality [8, 30, 31]. Though previous works achieved promising model prediction performance [8, 30, 31], it is important to note that, beyond the prediction, interpreting how textual features contribute to the predicted result is also important as the interpretability of results could further enhance human trust in the analysis results [23, 24, 34]. The interpretation for the predicted results can be provided from the perspective of local and global level [18, 23]. The global interpretation can provide feature importance to the model prediction performance, whereas local interpretation can demonstrate feature effects on individual predictions [18]. Most previous LA studies on feedback demonstrated global interpretations on predicting effective feedback measures [8, 30, 31]. However, the local interpretation was under-explored in these works. According to [23], local interpretation can show the extent to which the features had positive or negative effects on individual predictions, which can enable human users to better understand why a model made a decision instead of blindly trusting the model's prediction output [23]. We deemed that the interpretation at the local level can allow a better understanding for instructors to decide which feedback comments should be modified. Therefore, our study aimed to advance existing knowledge by providing the interpretation of the predicted results at both global and local levels.

To this end, our study aimed to analyse textual features of feedback content based on a learner-centred feedback framework [35], examine the use of textual features in different groups based on students' learning outcomes on the

subsequent assignment, and interpret the prediction results of student grade changes. Formally, we aimed to answer the following **Research Questions**:

- **RQ 1** To what extent are features of feedback related to the student grade changes on the consecutive assignment?
- **RQ 2** To what extent do the features of feedback content contribute to the predictions of student grade changes?

Our study answered the above research questions by using a dataset containing feedback on two connective assignments in a higher education course. We analysed feedback content on the first assignment and observed student grade changes (i.e., increase or not increase) on the subsequent assignment, which is connected to the first assignment. To answer **RQ1**, we first employed the learner-centred feedback framework [35] to extract features about feedback content. Then, we examined the use of the features among different groups based on student grade changes on the subsequent assignment. Our results revealed that the use of some extracted features presented significant differences in different student groups regarding the student's grade changes. Additionally, the student performance on the first assignment was considered the most important indicator for differentiating improvement on the second connective assignment. For **RQ2**, the extracted features were further used as inputs for training machine learning models used in the previous works [7, 8, 30, 31] to predict students' grade changes on the subsequent assignment. Then, we used the interpretable framework **SHAP** (**SH**apley **A**dditive **eX**planations) [24] to provide the interpretation of predicted results. We found that the **GTB** (**G**radient **T**ree **B**oosting) model achieved better performance in predicting student grade changes compared to other selected models. The **SHAP** framework can transparently interpret the feature effect at global and local levels. For example, with the use of the **SHAP** framework, our study identified the positive emotional words in the feedback as one of the most significant features for model prediction performance (i.e., global-level interpretation) and also demonstrated that the high frequency of positive emotional words in feedback negatively correlated with the student grade increase on the subsequent assignment (i.e., local-level interpretation).

2 BACKGROUND

2.1 Conceptualisation of Feedback

Feedback has been widely acknowledged as a crucial part of students' learning achievements in higher education [11, 13, 17, 20, 32, 35, 39]. When examining the effectiveness of feedback, it is significant to understand the existing literature on feedback practice so as to identify potential indicators [39]. One popular definition of feedback by **Hattie and Timperley** [16] conceptualised effective feedback as the information which can help students to minimise the gap between current and expected performance [41]. They [16] proposed a well-known *Feedback Model* to characterise effective feedback into three types (i.e., *feeding back*, *forward* and *up*) and further into four-level focuses (i.e., *task*, *process*, *self-regulatory*, and *self* focus). The *Feedback Model* [16] places emphasis on the provision of task-relevant information. The recent literature on feedback has been gradually shifting to feedback as a learner-centred process (e.g., *Mark 2 model* [2], *Dialogic Triangle model* [42], *Feedback Literacy Framework* [6], and the *Typology of Feedback Impact* [17]) where students could make sense of the information on their work, and use feedback to further improve the quality of their future work [32, 41]. The reason behind this shift is that providing feedback as a learner-centred process is more beneficial to learners than as information [35, 41].

To better understand the design of the learner-centred feedback framework, **Ryan et al.** [35] reviewed the feedback publications from the past decade. They [35] summarised a comprehensive framework of learner-centred feedback, which can be organised into two main design layers: *context* and *artefact*. The *context* focuses on the feedback design related to the process of feedback provision (e.g., frequency and timeliness), while the *artefact* focuses on the attributes

of the feedback content. As the focus of this study is on the feedback content, we mainly focused on the *artefact* layer, which involved nine attributes as shown in Table 1.

Table 1. The description of *artefact* attributes based on the learner-centred feedback framework developed by Ryan et al. [35]

| Row ID | Artefact Attributes | Description |
|--------|--|--|
| Attr.1 | Strengthen teacher and learner relationships | builds social relationship with learners |
| Attr.2 | Encourage learner agency | influences the way where learners attend to the comment details |
| Attr.3 | Encourage positive learner affect | shows the emotional response and care on learners' feelings |
| Attr.4 | Highlight strengths of performance | acknowledges learners' strength in their works to enhance learners' self-efficacy |
| Attr.5 | Provide critiques about performance | demonstrates critiques on learners' works to develop learners' evaluative judgement |
| Attr.6 | Provide actionable information for future performance | helps learners achieve expected learning goals, develop learning strategies, and obtain improvement on the similar learning task |
| Attr.7 | Promote learner independence | rephrases comments as questions or suggestions rather than statements to encourage learners to think for themselves |
| Attr.8 | Usable for learners | rephrases comments in clear language, easy for learners to understand, relevant to the marking criteria, and explanatory to learners |
| Attr.9 | Invite dialogue about feedback | encourages learners to ask questions and seek help from others (e.g., family and peers) to have feedback dialogues |

2.2 Content Analysis on Feedback

Recent works have employed content analysis to examine textual feedback content [7, 8, 30, 31]. For example, Cavalcanti et al. [7] analysed the quality of 1,000 feedback collected from an online course provided by a Brazilian higher education institution. They [7] used the textual features (e.g., features from LIWC and Coh-Metrix) to train a Random Forest model to identify whether the feedback contained the seven principles of good feedback practices proposed by Nicol and Macfarlane-Dick [29], and their work [7] achieved 0.75 classification accuracy and 0.20 Cohen's κ on the testing dataset. Following the study [7], Cavalcanti et al. [8] claimed that the seven principles [29] were too general to annotate the written feedback. Instead, they further annotated the same feedback dataset from [7] based on the Hattie and Timperley [16] four feedback focuses. To automate the process of identifying four feedback focuses, Cavalcanti et al. [8] used the same textual features from [7] to train a Random Forest model and they [8] reached 0.75 and 0.32 for the averaged classification accuracy and Cohen's κ , respectively, on the testing dataset. Later on, Osakwe et al. [31] further extended this work [8] by using a more powerful machine learning model (i.e., Gradient Tree Boosting [9]) and reached 0.83 and 0.39 for the averaged accuracy and Cohen's κ , respectively.

Despite demonstrating the automated process of evaluating feedback quality based on the feedback practice (i.e., seven principles [29] and four level focuses [16]), these studies [7, 8, 31] draw limited conclusions on evaluating feedback quality based on the feedback effect (e.g., student performance changes on future tasks), which can shed light on the feedback design and automatic feedback generation [30]. To this end, Nicoll et al. [30] examined the correlation between the textual features and student grade changes on the subsequent tasks. They extracted textual features (e.g., N-grams and sentiment scores) from feedback and used these features to train a logistic regression model to predict students' grade changes (e.g., increase, no change, and decrease) on the subsequent task [30]. The model reached 0.79 and 0.76 for averaged classification accuracy and AUC, respectively [30]. Though Nicoll et al. [30] demonstrated a method for

Table 2. Students' demographics information

| Year | Semester | # Students | # Female | # Male | #Domestic | #International |
|--------------|----------|------------|-----------|------------|-----------|----------------|
| 2021 | S2 | 145 | 47 | 98 | 17 | 128 |
| 2022 | S1 | 127 | 36 | 91 | 10 | 117 |
| Total | - | 272 | 83 | 189 | 27 | 245 |

understanding the impact of feedback content, further insights are still missing regarding feedback practice. First, textual features (e.g., N-grams) in their work [30] were engineered from the linguistic perspective, which lacked a connection with the existing feedback literature. Additionally, the features (e.g., N-grams) used for training the machine learning model were dataset dependent, which might lead to overfitting issues [8, 19, 31]. Thirdly, the connection between the first and subsequent assignments in their work [30] was unclear. According to [17], feedback might not influence student improvement when the first and subsequent assignments are different. Lastly, the interpretation of feature importance in [30] did not demonstrate whether these features were positively or negatively correlated with the student grade changes, which might be hard for educational researchers to understand the impact of feedback content.

To advance existing literature, our study aimed to identify the features of feedback content given on the first assignment and investigate the use of these features in different groups categorised by the student grade changes on the subsequent connective assignment. Instead of using the data-dependent features, we aimed to select textual features in relation to the *artefact* attributes based on a comprehensive learner-centred feedback framework developed in [35]. Furthermore, we aimed to employ the widely-used interpretable framework by [24] to demonstrate feature importance on the model prediction and the features' correlation polarity (i.e., positive or negative) on the output.

3 METHOD

3.1 Data Preparation

Our study obtained ethical approval (ID:29874) from Monash University and collected instructors' rubric feedback from an introductory data science course taught in English at the postgraduate level. This course was offered for two semesters throughout the year, and students were required to work on several assignments (e.g., coding practice and writing a report) for each semester. The marks of all assignments accounted for 40% of the course, and each assignment was marked separately.

Inspired by the learner-centred feedback framework, i.e., effective feedback should be designed in a connective form (i.e., the comments of the current feedback can help students make improvements in the subsequent assignment) [11, 35], our study mainly investigated the feedback on the report assignment which required students to write a report proposal and a final report. To clarify the difference, we named the reported proposal **Assignment I** (counted 5% marks) and the final report **Assignment II** (counted 10% marks). Students were required to submit a report proposal for the assignment I to introduce a data science problem to be solved and describe relevant application background and types of business models. After students submitted assignment I, instructors marked the students' assignment I and provided feedback to help students work on assignment II. In assignment II, students could use the feedback from assignment I to complete the remaining parts of the report, such as the description of the data analysis. After students submitted assignment II, instructors graded students' work and provided corresponding feedback. Both assignments were assessed using the same rubric¹. However, as assignment I only required students to complete part of report writing, the instructor also used the subset of the rubric for assessing assignment I. Our study investigated the feedback

¹The marking rubric can be accessed by <https://github.com/jionghaolin/LAK2023>

content on assignment I (i.e., report proposal). We collected the feedback data from Semester 2 (Jul–Oct) in 2021 and Semester 1 (Feb–May) in 2022 since the assignment descriptions and rubric were the same for both semesters. In total, we had 288 student records collected by the learning management system. We filtered out 16 data records with two issues: 1) the records missed grades and feedback for the assignment I, and 2) duplicate records. Then, we obtained 272 records (83 female and 189 male). The students’ demographics in our dataset are presented in Table 2.

3.2 Representation of Students’ Grade Changes

Nicoll et al. [30] proposed a method to observe student grade changes by subtracting the students’ grades in the first task from the grades in the second task. Inspired by their work [30], we also measured student grade changes by subtracting the students’ grades in the assignment I from their grades in assignment II. The students’ assignments I and II were marked separately by the same instructor. According to the course assessment policy, each assignment was marked into five categories from low to high: Fail (**N**, scoring from 0-49), Pass (**P**, 50-59), Credit (**C**, 60-69), Distinction (**D**, 70-79), and High Distinction (**HD**, 80-100). We encoded five categories into a numerical scale (i.e., **N** = 0, **P** = 1, **C** = 2, **D** = 3, and **HD** = 4) to calculate the grade changes between two assignments. The positive values indicated students achieved performance increase on assignment II, and we encoded the records of the positive values as the Increase group. Whereas, the other calculated results were encoded as the Not Increase. However, it should be noted that 60 students were graded **HD** marks in both assignments I and II. We argue that maintaining the outstanding performance (i.e., **HD**) on the subsequent assignment is also an improvement but the performance increase could not be directly observed from the students’ subsequent performance change for these high-performing students. The feedback for these 60 high-performing students could be more concise (e.g., a line of comment, “*Well done!*”) compared with other students, which might influence the analysis results. Thus, we decided to filter out the feedback on these 60 students and planned to investigate them in future work. As a result, we obtained 212 records for analysis where 66 records were in the Increase and 146 in Not Increase group.

3.3 Mapping Artefact Level Features

To answer RQ1, we used a set of textual features extracted from feedback content to map against the *artefact* attributes of the learner-centred feedback framework by [35] as shown in Table 3. It should be noted that we did not find an automatic tool to identify the feedback attribute **Attr.9 Invite dialogue about feedback** so we decided to annotate this attribute manually. However, we did not observe the **Attr.9** in our feedback dataset because the invitation of dialogue feedback (i.e., weekly one-on-one consultation) was commonly sent by the instructors in an oral form during the class or by the forum in the learning management systems. Thus, we only analysed eight attributes in our study.

3.3.1 Linguistic Politeness. Expressing politeness in textual language can show respect and care for the interlocutors’ feelings [3], which are important to enhance the relationship between learners and instructors [21, 40] and design effective feedback [11, 20, 32]. Therefore, we deemed that linguistic politeness can be mapped to the **Attr.1** [17] “...development of healthy relationships between the teacher and the learner...”. To examine the politeness in feedback, our study employed the politeness tool² [43] (built on the Brown and Levinson [3] politeness theory) to extract 39 politeness strategies (e.g., Hedges strategy, “Adding a diagram *might* be better”) from the feedback. The description of these politeness strategies was documented in an electronic appendix, which is accessible via <https://github.com/jionghaolin/LAK2023>. The extracted politeness strategies were counted on their frequency from the feedback.

²<https://cran.r-project.org/web/packages/politeness/>

Table 3. Mapping the artefact attributes from Ryan et al. [35] learner-centred feedback framework with the textual features extracted by software and domain experts. **Agree**, **K**, and **Freq** represent **Agreement scores**, **Cohen’s κ** , and **Frequency**, respectively

| Row ID | Artefact Attributes | Textual Features | # Features | Method | Agree | K | Freq (%) |
|--------|---|-----------------------------------|------------|------------------------|-------|------|----------|
| Attr.1 | Strengthen teacher and learner relationships | Politeness Strategies | 39 | Politeness (R Package) | N/A | N/A | N/A |
| | | Relational Impact | 4 | | | | |
| Attr.2 | Encourage learner agency | Cognitive Impact | 10 | LIWC | N/A | N/A | N/A |
| Attr.3 | Encourage positive learner affect | Affective Impact | 5 | | | | |
| | | Self Focus | 1 | | 0.99 | 0.98 | 55.88% |
| Attr.4 | Highlight strengthen of performance | Task Focus (Positive Feedback) | 1 | | 0.98 | 0.96 | 58.46% |
| Attr.5 | Provide critiques about performance | Task Focus (Negative Feedback) | 1 | | 0.96 | 0.92 | 29.04% |
| Attr.6 | Provide actionable information for future performance | Task Focus (Non-corrective) | 1 | Manual Annotation | 0.93 | 0.79 | 77.94% |
| | | Process Focus | 1 | | 0.86 | 0.72 | 62.13% |
| | | Feeding Forward | 1 | | 0.91 | 0.78 | 88.60% |
| Attr.7 | Promote learner independence | Self-regulatory Focus | 1 | | 0.91 | 0.78 | 25.74% |
| Attr.8 | Usable for learners | Feeding Up | 1 | | 0.98 | 0.95 | 69.49% |
| | | Feeding Back | 1 | | 0.98 | 0.95 | 72.79% |
| | | Writing Metrics | 77 | | N/A | N/A | N/A |

3.3.2 Linguistic Inquiry and Word Count (LIWC). The LIWC dictionary has been widely used in content analysis of educational feedback to characterise written words into many psychological categories such as cognitive processes and emotions [7, 8, 13, 31], and these categories can reflect the writers’ psychological states [33]. As suggested by Derham et al. [13], they selected 19 categories from the LIWC dictionary, which could manifest the impact of feedback (i.e., relational, cognitive, and affective impacts [17]) on students. By scrutinising the description of these impacts, we found these impacts could potentially be mapped to the *artefact* attributes as shown in Table 3. Firstly, the *relational impact* might influence students’ engagement with the feedback (e.g., seeking feedback and making actions), which was related to the **Attr.1** [35]. Then, the *cognitive impact* might influence students’ thinking about how to process information, memorise the learning information, and form concepts [13, 17], which might be related to the **Attr.2** [35] “...and encouraging learners to engage in further independent study...”. Lastly, the *affective impact* might influence students’ affective states (e.g., happy and stress), which is related to the **Attr.3** [35] “...it may be beneficial for teachers to think about the potential affective impact...”. In line with the work of [13], our study also extracted 19 categories of features from the LIWC dictionary to represent relational (four categories), cognitive (ten categories), and affective (five LIWC categories) impacts. The examples of each category were shown in the electronic appendix, which is accessible via <https://github.com/jionghaolin/LAK2023>.

3.3.3 Feedback Model by Hattie and Timperley [16]. Inspired by the success of previous works [8, 31], it is promising to automate the process of identifying effective feedback practice (e.g., Hattie and Timperley four-level feedback focuses [16]). However, existing automatic tools for feedback analysis could not be applied to automate content analysis. Instead, we decided to manually annotate the feedback type and feedback focus for each instructor’s feedback from our dataset and aimed to build a classifier in future work to automatically identify the components of feedback models. During

the annotating process, we recruited two domain experts to annotate the entire feedback dataset by following the definition of feedback models introduced in [16]. The agreement score and Cohen’s κ were shown in Table 3, which demonstrated promising annotation results. The inconsistent cases were further resolved by a third feedback expert. We also demonstrated the frequency of each feedback model feature in Table 3.

By scrutinising the description of the learner-centred feedback framework [35], we deemed that the feedback models could be mapped to the *artefact* attributes. In terms of the four feedback focuses, we divided task-level focus into three sub-categories, i.e., positive feedback, negative feedback, and non-corrective comment (i.e., instruction-related information) as suggested in [13]. We posited that the positive feedback at the task focus to be related to the **Attr.4** [35] “...information which highlights what the learner has done well can be valuable...”, negative feedback at task focus to be related to the **Attr.5** [35] “...comments that provide critiques of a learner’s performance...”, and non-corrective comment at task focus to be related to the **Attr.6** “...to improve on similar task; to achieve learning outcomes...”. Then, the feedback on the process focus leads to more direct information about the learning process underlying a task [16], which is potentially related to the **Attr.6** [35]. Feedback on self-regulatory focus aims to promote students’ capabilities including self-monitoring, self-direction, and self-control to achieve learning goals, which is posited to be related to the **Attr.7** [35] “...phrase feedback comments as suggestions or questions rather than statements...”. Feedback on self focus involved the positive evaluation and affective components in the feedback, which is posited to be related to the **Attr.3** [35] “...to aim for positive emotional responses...”. In terms of feedback type, the feeding forward can inform students to determine the next steps in the subsequent tasks [16], which is assumed to be related to the **Attr.6** [35]. Then, the feeding up was used to clarify the goals and criteria of the assessment [14, 16], which was related to the **Attr.8** [35] “...the information should be relevant to the assessment criteria...”. The feeding back informs the learners’ progress towards the learning goals and responds to learners’ work [14, 16], which is posited to be related to the **Attr.8** [35] “...clearly related to particular aspects of the performance...”.

3.3.4 Coh-Metrix. To quantify the quality of feedback writing, we adopted the computational linguistic system Coh-Metrix, which was built by a set of metrics to calculate the complexity, cohesion, and readability of the written text [27]. The features extracted by Coh-Metrix have been widely used in the feedback content analysis to evaluate the quality of feedback writing [7, 8, 31]. By scrutinising the description of each metric in Coh-Metrix [27], we decided to select a subset of features from Coh-Metrix. According to the **Attr.8** [35], feedback should be easy to read, avoid the use of complex terms, and consider the detailed level of feedback. Firstly, to quantify the extent to which textual feedback can be easily read, previous work suggested that the traditional readability measures (e.g., Flesch Reading ease scores) and text cohesion-related metrics such as referential cohesion (i.e., measuring the writing cohesion level) in Coh-Metrix can be included [27]. Then, regarding the use of complex terms, as complex terms are rare in the students’ reading experience [27], we decided to use the metric word information (i.e., built upon the corpus of commonly used words [27]) to measure the use of complex terms. Thirdly, to measure the detailed level of feedback, we selected several metrics such as descriptive measures (i.e., overview of the feedback characteristics, for example, number of words per feedback) and lexical diversity (i.e., the richness of words) based on the Coh-Metrix handbook [27]. In total, we have 77 features which were detailed in the electronic appendix <https://github.com/jionghaolin/LAK2023>.

3.3.5 Learners’ Knowledge Level. Learners’ knowledge level might relate to the learners’ perception and sensemaking of the instructor feedback [17, 22]. For example, high-knowledge learners who obtain high performance on the assignment may engage more with feedback compared with low-knowledge learners [20, 39]. Therefore, our study decided to

include the learners' knowledge level in the data analysis and treated the students' performance on Assignment I as a proxy of their knowledge level.

3.4 Data Analysis

3.4.1 Statistical Analysis. To answer our **RQ1**, we calculated the mean values and standard deviation for each feature in different student grade change groups (i.e., Increase and Not Increase). Then, we examined the use of each feature among the different student grade change groups by using statistical tests. Regarding the Feedback Model features (e.g., *Process* and *Feeding Up*), we counted their appearance in a binary form (i.e., exist or not exist) for each feedback. Thus, we adopted the Chi-square test to examine the association between the Feedback Model features and student grade changes and used Cramer's Phi to measure the effect size [38]. The other features were generated in numerical values, so we adopted Mann-Whitney U test and used Rank-Biserial (i.e., r_{tb}) to measure the effect size [38].

3.4.2 Interpretation on Predictive Analysis. To answer **RQ2**, we first investigated the capability of machine learning models on our task, i.e., use the *artefact* features extracted from feedback on assignment I to predict whether students obtained Increase or Not Increase on assignment II. As discussed in Section 2.2, previous works demonstrated the effectiveness of applying the machine learning models (e.g., Logistic Regression [30], Random Forest [8] and Gradient Tree Boosting [31]) to the prediction tasks (e.g., feedback quality prediction [8, 31] and student performance prediction [30]). Inspired by these works [8, 30, 31], the current study also used these models to predict whether or not the student can achieve improvement (i.e., Increase or Not Increase) on the subsequent task (i.e., assignment II). **Logistic Regression (LR)** model is one of the widely used statistical models in education studies, which can make binary classification on the task [30]. **Random Forest (RF)** model is a type of ensemble learning method which combines many predictors to predict the results [8]. Each predictor learns the patterns from a random sample of the data records from the training dataset. The final prediction of the RF model is made by averaging the predictions of each individual predictor. **Gradient Tree Boosting (GTB)** model is also an ensemble machine learning method that makes the prediction based on many predictors [9]. The difference between the GTB and RF models is in the way of combining the internal predictors. The predictors of the RF model make the prediction independently and vote on the final prediction results. There might be many errors in some predictions during the training process. In contrast, the predictions of the GTB model are built in a sequential manner, and the prediction errors from the previous predictors can be fixed by the subsequent predictors.

Then, we adopted the well-established interpretable framework **SHAP (SHapley Additive exPlanations)** [24] to understand the contribution of features on model prediction performance. The SHAP framework was developed based on the game theory to quantify the features' contribution to the model's prediction [24]. According to [23, 24, 34], the SHAP framework shows the feature contribution not only at the global level (i.e., which features contributed most to the model prediction performance) but also at the local level (i.e., which features presented the positive and negative impact on individual prediction). It should be noted that the interpretability from SHAP does not indicate causality.

3.4.3 Study Setup. To evaluate the model prediction performance, we randomly split the dataset into *training set*, *validation set*, and *testing set* with the ratio 70%:10%:20%, respectively. The models were trained using the Python package `scikit-learn`³ and the models' hyper-parameters (e.g., *n_estimators*) were further tuned by applying grid search on

³<https://scikit-learn.org/>

Table 4. The comparison of selected features between the Increase group and the Not Increase group. Features marked with † were examined by Chi-square test and Cramer’s Phi effect size whereas the remaining features were examined by Mann-Whitney U test and Rank-Biserial (i.e., r_{rb}) effect size

| Row ID | Artefact Attributes | Feature Clusters | Features | Increase | | Not Increase | | Difference | |
|--------|--|------------------------------|---------------------------|----------|-------|--------------|-------|------------|-------|
| | | | | M | SD | M | SD | P-val | E.S |
| | | Knowledge Level | <i>1st Assgmt grades</i> | 1.5 | 1.03 | 3.16 | 1.10 | *** | 0.72 |
| Attr.1 | Strengthen teacher and learner relationships | Politeness [43] | <i>Negative.Emotion</i> | 0.74 | 0.79 | 0.36 | 0.64 | *** | -0.27 |
| | | | <i>Impersonal.Pronoun</i> | 1.98 | 2.18 | 1.40 | 2.28 | ** | -0.23 |
| | | | <i>Hedges</i> | 0.97 | 1.55 | 0.51 | 1.13 | ** | -0.19 |
| | | | <i>Give.Agency</i> | 0.24 | 0.50 | 0.08 | 0.26 | ** | -0.14 |
| | | Relational Impact (LIWC) | <i>affiliation</i> | 0.24 | 0.51 | 0.13 | 0.45 | * | -0.10 |
| Attr.2 | Encourage learner agency | Cognitive Impact (LIWC) | <i>interrog</i> | 1.69 | 2.11 | 0.98 | 1.95 | *** | -0.31 |
| | | | <i>cause</i> | 1.45 | 1.71 | 1.33 | 2.64 | * | -0.16 |
| | | | <i>QMark</i> | 0.58 | 1.05 | 0.37 | 1.02 | * | -0.13 |
| Attr.3 | Encourage positive learner affect | Affective Impact (LIWC) | <i>posemo</i> | 5.76 | 2.31 | 10.89 | 12.17 | *** | 0.48 |
| Attr.4 | Highlight strengths of performance | Feedback Model [16] | † <i>Task (Pos)</i> | 0.61 | 0.49 | 0.56 | 0.50 | N.S | 0.04 |
| Attr.5 | Provide critiques about performance | | † <i>Task (Neg)</i> | 0.44 | 0.50 | 0.25 | 0.43 | ** | 0.19 |
| Attr.6 | Provide actionable information | | † <i>Process</i> | 0.85 | 0.36 | 0.60 | 0.49 | *** | 0.25 |
| Attr.7 | Promote learner independence | | † <i>Self-regulate</i> | 0.38 | 0.49 | 0.24 | 0.43 | * | 0.14 |
| | | | † <i>Feeding Up</i> | 0.91 | 0.29 | 0.62 | 0.49 | *** | 0.29 |
| Attr.8 | Usable for learners | Writing Metrics (Coh-Metrix) | <i>DESWC</i> | 73.03 | 22.55 | 52.04 | 35.43 | *** | -0.48 |
| | | | <i>WRDADJ</i> | 24.86 | 9.41 | 16.99 | 12.50 | *** | -0.42 |
| | | | <i>DRAP</i> | 19.92 | 6.51 | 14.70 | 10.05 | *** | -0.41 |
| | | | <i>DESWLsyd</i> | 0.93 | 0.20 | 0.76 | 0.27 | *** | -0.41 |

Note: **M** = mean; **SD** = standard deviation; **P-val** = p-values; **E.S** = effect size; *1st Assgmt grades* = student grades on Assignment I; *posemo* = positive emotional words; *interrog* = interrogatives words; *cause* = causal language; *QMark* = question mark; *Task (Neg)* = negative feedback at task focus; *Task (Pos)* = Positive feedback at task focus; *DESWC* = number of words; *WRDADJ* = adjective incidence; *DRAP* = adverbial phrase incidence; *DESWLsyd* = standard deviation of the word length; In the columns of **P-val**, *** p<0.001; ** p<0.01; * p<0.05; N.S = not significant;

the *validation set*. Finally, all the models’ performances were evaluated on the *testing set* by using four representative metrics, i.e., F1-score, Area Under the Curve (AUC), Cohen’s κ , and classification accuracy.

4 RESULTS

4.1 Results on RQ1

We reported the statistical results based on the *artefact* attributes (shown in Table 4) and the results within each attribute were sorted based on their effect size⁴. We found that the values of feature *1st Assgmt grades* (i.e., student grades on Assignment I) in Increase group were significantly lower than the Not Increase group. Additionally, the effect size of the feature *1st Assgmt grades* presented a strong association [10] with the student grade changes.

For the **Attr.1** in Table 4, we investigated the linguistic politeness features and relational impact where the use of *Negative.Emotion*, *Impersonal.Pronoun*, *Hedges*, *Give.agency*, and *affiliation* were more frequent in the Increase group

⁴Due to the space limit, we presented the five most significant results for each *artefact* attribute.

than they were in the Not Increase. It should be noted that the *affiliation* feature is related to the sentences fostering a sense of encouragement (e.g., “It can **help** the analysis”), *Negative.Emotion* is related to the use of negative emotional words (e.g., “it is **difficult** to explain”), *Impersonal.Pronoun* is related to the non-person referents (e.g., “**That** is not a data science work”), *Hedges* is related to the indirect voice tone (e.g., “The focus **might** need to be narrowed down”), and *Give.Agency* is related to the sentences fostering a sense of suggestion (e.g., “It would be great if you can find a dummy dataset”). These significant results might indicate that though the comments were crafted with more negative emotional words in the Increase than Not Increase group, instructors also made more effort to maintain the relationship in Increase group by using expressions (i.e., *affiliation*, *Impersonal.Pronoun*, *Hedges*, and *Give.agency*).

For the **Attr.2** in Table 4, we investigated the cognitive impact features where the use of *interrog*, *cause* and *QMark* features in the Increase group were more frequent than those in the Not Increase. The use of *QMark* (e.g., “?”) and *interrog* (e.g., “**What** is the role of data scientist.”) were related to the questions in feedback and the significant results indicated that instructors might have posed more questions in their feedback to the Increase group than they did in feedback for the Not Increase group. Then, the *cause* feature could demonstrate comments expressed in the sense of causation (e.g., “**because** you need to address these in next assignment”) and the results indicated that the students might have received more comments contained *cause* in the Increase group than they did in the Not Increase.

For **Attr.3** in Table 4, we investigated the affective impact features where the use of *posemo* feature in the Not Increase group was more frequent than those in the Increase. The *posemo* feature is related to the adjective expressing positive emotion, which could capture most praise (e.g., “**Good** job!”, “**Excellent!**”) in feedback. Thus, the significant result might be the reason that the Not Increase students received more praise from instructors than the Increase since the students in the Not Increase had higher grades than those in the Increase on Assignment I.

In Table 4, four *artefact* attributes were captured by the Feedback Model. For the **Attr.4**, the *Task (Pos)* (i.e., positive feedback in task level focus) identified the feedback highlighted strengths of student performance (e.g., “**Good** topic with a clear discussion of project goals”). The significant difference was not found between the Increase and Not Increase groups. For the **Attr.5**, the *Task (Neg)* feature (i.e., negative feedback in task level focus) identified the feedback contained the instructor’s criticism on students’ submission (e.g., “**You did not** introduce the role of data science”). The significant results indicated that the use of *Task (Neg)* features was significantly correlated with the student grade change. For the **Attr.6**, the *Process* feature (i.e., the appearance of process level focus) identified the feedback with concrete instruction to improve on the subsequent task and achieve the expected goals (e.g., “It would be great if you add the flow chart presenting the overall structure of the project”). The significant differences indicated that the use of the *Process* features was significantly correlated with the student grade change. For the **Attr.7**, the *Self-regulate* (i.e., self-regulatory level focus) identified the feedback with the questions or statements to encourage learners to think more about improvement on their subsequent assignment. (e.g., “**What are** the data science roles?”). The significant differences indicated that the use of *Self-regulate* features was significantly correlated with the student grade change.

We investigated the **Attr.8** by using the feature *Feeding Up* and features related to Writing Metrics. Compared with the Not Increase group, Increase group had higher feature values of *Feeding Up*, *DESWC* (i.e., number of words per feedback, which indicated the informativeness level of the feedback [27]), *WRDADJ* (i.e., adjective incidence, which indicated the density level of using adjectives [27], “**Good** topic with **detailed** overview”), *DRAP* (i.e., the adverbial phrase incidence, which indicated the density level of using adverbial phrases [27], “**Describe** the data aspect **in the** goal”), and *DESWLsyd* (i.e., standard deviation of the word length, which indicated the level of text variation in terms of the word length [27]). The *Feeding Up* feature identifies feedback containing marking criteria or expected goals (e.g., “**You should clearly** describe the business benefits in your report” where the clear description of the business benefits

Table 5. The performance of **LR**, **RF**, and **GTB** in predicting student grade changes on assignment II.

| Model | Accuracy | F1-score | AUC | Cohen's κ |
|---------------------------------------|-------------|-------------|-------------|------------------|
| Logistic Regression (LR) | 0.81 | 0.78 | 0.77 | 0.56 |
| Random Forest (RF) | 0.80 | 0.76 | 0.75 | 0.51 |
| Gradient Tree Boosting (GTB) | 0.84 | 0.80 | 0.80 | 0.61 |

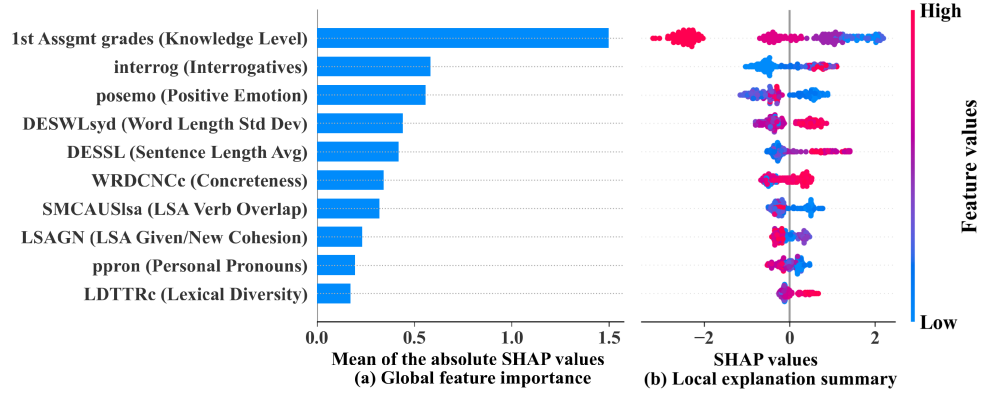


Fig. 1. Top 10 most significant features on the GTB model's prediction. Fig 1 (a) and (b) shared the same rank of feature importance. In Fig 1 (a), the bar chart ranked the features from the most important to the less important. The feature importance was measured by the mean of absolute SHAP values. In Fig 1 (b), the beeswarm plot demonstrated the distribution and direction of the feature effect at the instance level. Each line has an equal number of scatters, and the number of scatters in a line is the same as the number of instances in our training dataset. The position of each scatter in a line was determined by the SHAP value. The positive SHAP values are indicative of the Increase group, while negative are indicative of the Not Increase group. The colour of the points represents the feature values from low to high.

is in the marking criteria). The significant differences indicated that the use of *Feeding Up* features was significantly correlated with the student grade change. Regarding the writing metrics, the significant results of the features *DESWC*, *WRDADj*, *DRAP*, and *DESWLSyd* indicated that the students in Increase group might receive the feedback with more descriptive information compared with Not Increase.

4.2 Results on RQ2

The model evaluation results were shown in Table 5. We found that the performance of **GTB** model outperformed the other two models (i.e., **LR** and **RF**) across four metrics. Therefore, we decided to use the **GTB** model to analyse the contribution of *artefact* features to the model's prediction. By applying the SHAP framework to the trained GTB model, we presented the feature importance at the global level (shown in the bar chart Fig 1 (a)) and local level (shown in the beeswarm plot Fig 1 (b)). The features were sorted from the most significant to the less significant ones in both Fig 1. Due to the space limit, we only presented the top 10 most significant features. In Fig 1 (a), we observed that the *1st Assgmt grades* feature was considered the most significant feature by **GTB** model. Then, three significant features were extracted from the *LIWC* dictionary (i.e., *interrog*, *posemo*, and *ppron*) and six features calculated by *Coh-Mextrix* (i.e., *DESWLSyd* – standard deviation of the word length; *DESSL* – average length of sentences; *WRDCNCc* – concreteness level of the text; *SMCAUSIsa* – measures of verb overlap calculated by Latent Semantic Analysis (LSA);

LSAGN – LSA-based cohesion measure, which reflects the ratio of given and new information in the text; *LDTTRc* – lexical diversity ratio, measured by the number of unique words divided by the total words in a feedback comment.)

We further investigated the interpretation of feature importance at the local level, which presents the distribution of instances and direction of the feature effect. For example, in Fig 1 (b), the feature *1st Assgmt grades* presents a cluster of red scatters on the left side of the central axis, which indicates many instances with the high grades on assignment I were predicted as Not Increase. In contrast, the instances with low grades, concentrated on the right, tended to be predicted as Increase. The observed patterns of feature *1st Assgmt grades* demonstrated that the student grade on assignment I was a strong indicator to distinguish between Increase and Not Increase, which was in line with the result in Table 4. Similarly, we found that the high feature values of *interrog* concentrated on the right side, which indicates a positive correlation between the high feature values of *interrog* and the prediction of Increase. In contrast, most instances with low values of *interrog* concentrated on the left side, which indicates that the low values of *interrog* positively correlated with Not Increase. Then, we found that the high feature values of *posemo* concentrated on the left side. This result indicates a positive correlation with the prediction of Not Increase. The high value of *posemo* could be a strong predictor for the GTB model to predict Not Increase. However, the instances with low values of the *posemo* were dispersed from the positive SHAP values to the negative. It indicates *posemo* feature values may not have a clear correlation with either of the two groups – Increase or Not Increase. Additionally, we also observed that the low feature values of *DESWLsyd*, *WRDCNCc*, and *LDTTRc* converged towards the left, but high values of these features were dispersed. The results indicate that the low feature values of *DESWLsyd*, *WRDCNCc*, and *LDTTRc* positively correlated with the Not Increase.

5 DISCUSSION

5.1 Implications

Students with a lower grade on the first assignment were more likely to obtain a grade increase on the subsequent assignment. Unsurprisingly, the students with lower grades had more room to improve on the subsequent task compared to the students who had already obtained higher grades. Then, we observed many differences of *artefact* attributes between the group of Increase and Not Increase, which might shed light on the design and evaluation of the feedback based on the learner-centred framework.

Critiquing students' work as part of feedback is unavoidable, and crafting feedback comments in polite language for low knowledge students is recommended. Existing research also found that expressing feedback comments politely could help build rapport with students [28, 40] and also potentially enhance student performance [28], especially for the low-knowledge students who were more engaged with the polite comments [26]. Compared to the students in Not Increase, we observed that students in Increase group with lower grades on assignment I received more negative emotional words (e.g., “Your reference style is **incorrect**”) and also more polite elements such as encouraging student engagement and rewording the comments into suggestions (e.g., “It would be great if you add a flow chart”) instead of commands. Although more negative emotional words were used in the feedback, more polite elements were detected as well, which showed instructors trying to direct students to identify where to improve and meanwhile also gave the care to maintain the student and instructor relationship.

Positively framed feedback content might be a good indicator to distinguish between the student grade changes. Compared with the group of Not Increase, students in the Increase received fewer positive emotional words, which were related to comments encouraging positive affect and highlighting learners' strengths. This might be the reason

that students in the Increase group obtained lower grades on the first assignment than those in the Not Increase so the students in the Increase were less likely to receive the praise (e.g., “*Excellent Work!*”) from instructors [25]. Prior works also suggested that using praise for low knowledge students needs to be carefully delivered because the praise might undermine students’ own responsibility to identify their mistakes or errors [25]. Instead, using feedback to highlight the specific aspect of student strengths (e.g., “*Your report is well structured*”) is recommended as this positive feedback can affirm students’ understanding of success criteria, enhance student self-efficacy, maintain the student and teacher relationship, and mitigate the potential harshness of criticism.

To assist students’ improvements on the subsequent task, it is important to demonstrate comments in questions or suggestions. We noted that the questions in feedback mostly correlated with the use of the self-regulatory focus (e.g., “*What about other challenges?*”) and suggestions demonstrated the process focus (e.g., “*You should make sure that you have access to details about the datasets used here*”). The students in the Increase group received the feedback with more use of self-regulatory and process focus in feedback compared to those in the Not Increase. The existing literature found that the use of self-regulatory and process focus in feedback had a positive impact on promoting learner independence and grade improvement [15]. However, the self-regulatory focus was rarely provided in the feedback compared to the use of process focus (shown in Table 3). Thus, we recommend that instructors should make more effort in rephrasing statements into questions in the feedback.

Usable feedback comments should be both clear and detailed. Compared with the students in the Not Increase group, the students in the Increase group received more comments on reminding the success criteria of the assignment. The reason might be that the students in Increase group obtained lower grades than Not Increase on assignment I, which indicates the learning gap between the current and the expected grades was larger in the Increase group than the Not Increase group. Thus, instructors reminded more frequently of the success criteria for the low performance students to help them minimise the learning gap and make improvements on the subsequent assignment. Moreover, instructors provided more detailed feedback for the students in the Increase group. The reason might also be in relation to the students’ grades on the first assignments. As discussed before, low performance students received more comments on critiques, instructional questions and suggestions than the high performance students did, which led to more detailed feedback for the low performance students. Therefore, we recommend that instructors might consider reminding the low performance students of the successful goals and making an effort to elaborate detailed comments.

Lastly, our results demonstrated the effectiveness of the GTB model, which was in line with the results in [31]. To compensate for the GTB model’s low interpretability, our study employed the SHAP framework to interpret how the GTB model learned the patterns from the distribution of engineered features to predict student grade changes. Our results showed the potential of using the SHAP framework to deliver trustworthy analytic results. At the global level of interpretation, the SHAP framework identified some significant features corroborated what we observed when answering RQ1 such as *1st Assgmt grades* (student grade on the first assignment) and *posemo* (positive emotional words) showing significant differences between Increase and Not Increase. Then, the SHAP framework could present the direction (i.e., positive or negative) of the correlation between the significant feature values and the student grade changes at the local level, which provided transparent interpretations for the GTB model prediction. Since the features were built upon the feedback literature [35], we expected that the interpretations might help instructors observe the influence of using different feedback comments and further design their feedback on the basis of supporting students. As suggested in [34], both global and local level interpretation can help enhance the users’ trust since the human users are inclined to trust the predicted results if they can understand why the model makes the prediction. However, we

believe that future studies with instructors are needed to validate this claim. It is necessary to understand which exactly of the features are valued by the instructors and the extent to which the use of SHAP increases their trust.

5.2 Limitations and Future Work

We acknowledge that there are some limitations in the current study. *First*, our study was based on the feedback dataset from one course, which might not represent feedback in other courses. *Secondly*, some feedback content might be repeated since the same instructors might have given the same feedback to the students who encountered similar issues in their reports. *Thirdly*, we noted that the student's grade changes on the subsequent task might not fully demonstrate the feedback impact since many factors (e.g., peer discussions) might occur in the loop of feedback provision, which could in turn confound associations of feedback with the performance changes [17]. Future research should investigate the correlation between feedback content and other related factors (e.g., student learning strategies). *Lastly*, though the GTB model effectively predicted students' grade changes in our study, many falsely predicted cases still existed. The reasons for mistakenly predicted cases could be multi-faceted. For example, prediction accuracy could be improved by using more sophisticated machine learning models. Given the success of the deep neural network models in processing different tasks, it is worthwhile to investigate the value of applying deep neural networks to our task when we collect sufficient datasets for training such models. Another reason for falsely predicted cases might be related to the students' engagement with feedback. Since students might not always read feedback content, feedback might not be effective if students do not engage with it [4, 36]. Therefore, it is worthwhile to further incorporate students' engagement activities (e.g., students' posts in the forum [37]) before and after the feedback provision into the model training process.

Furthermore, an extension of our study would be to design a system that can automatically evaluate instructors' feedback based on the attributes from the learner-centred feedback framework and provide interpretable recommendations to the instructors about how feedback can be improved. For example, when the provided feedback lacks positive highlights of strength and contains many negative comments, the system can make a suggestion for the instructors to include positive emotional words to encourage students.

6 CONCLUSION

This paper illustrated how the learner-centred feedback framework could be used for analysing the feedback content by using the learning analytics approaches. We also demonstrated the potential of using machine learning models to predict student achievements on subsequent assignments and using the well-established framework SHAP to provide transparent interpretations of the predicted results. The implementations of the learner-centred analytics of feedback content in our study have important implications for the future design of automated feedback systems and the practice of feedback provision.

ACKNOWLEDGMENTS

This research was supported by the Australian Government through the Australian Research Council (DP220101209).

REFERENCES

- [1] Rola Ajjawi, David Boud, Michael Henderson, and Elizabeth Molloy. 2019. Improving feedback research in naturalistic settings. In *The Impact of Feedback in Higher Education*. Springer, 245–265.
- [2] David Boud and Elizabeth Molloy. 2013. Rethinking models of feedback for learning: the challenge of design. *Assessment & Evaluation in higher education* 38, 6 (2013), 698–712.
- [3] Penelope Brown and Stephen Levinson. 1987. *Politeness: some universals in language usage*. Cambridge University Press, Cambridge, UK.

- [4] David Carless. 2006. Differing perceptions in the feedback process. *Studies in higher education* 31, 2 (2006), 219–233.
- [5] David Carless. 2019. Feedback loops and the longer-term: towards feedback spirals. *Assessment & Evaluation in Higher Education* 44, 5 (2019), 705–714.
- [6] David Carless and David Boud. 2018. The development of student feedback literacy: enabling uptake of feedback. *Assessment & Evaluation in Higher Education* 43, 8 (2018), 1315–1325.
- [7] Anderson Pinheiro Cavalcanti, Rafael Ferreira Leite de Mello, Vitor Rolim, Máverick André, Fred Freitas, and Dragan Gašević. 2019. An analysis of the use of good feedback practices in online learning courses. In *2019 IEEE 19th ICALT*, Vol. 2161. IEEE, 153–157.
- [8] Anderson Pinheiro Cavalcanti, Arthur Diego, Rafael Ferreira Mello, Katerina Mangaroska, André Nascimento, Fred Freitas, and Dragan Gašević. 2020. How Good is My Feedback? A Content Analysis of Written Feedback. In *Proceedings of the LAK (LAK '20)*. ACM, New York, NY, USA, 428–437.
- [9] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 2016 Conference on KDD*. 785–794.
- [10] Jacob Cohen. 2016. A power primer. (2016).
- [11] Phillip Dawson, Michael Henderson, Paige Mahoney, Michael Phillips, Tracii Ryan, David Boud, and Elizabeth Molloy. 2019. What makes for effective feedback: staff and student perspectives. *Assessment & Evaluation in Higher Education* 44, 1 (2019), 25–36.
- [12] Phillip Dawson, Michael Henderson, Tracii Ryan, Paige Mahoney, David Boud, Michael Phillips, and Elizabeth Molloy. 2018. Technology and feedback design. *Learning, design, and technology* (2018).
- [13] Cathrine Derham, Kieran Balloo, and Naomi Winstone. 2021. The focus, function and framing of feedback information: linguistic and content analysis of in-text feedback comments. *Assessment & Evaluation in Higher Education* (2021), 1–14.
- [14] Douglas Fisher and Nancy Frey. 2009. Feed up, Back, Forward. *Educational Leadership* 67, 3 (2009), 20–25.
- [15] John Hattie. 2012. *Visible learning for teachers: Maximizing impact on learning*. Routledge.
- [16] John Hattie and Helen Timperley. 2007. The power of feedback. *Review of educational research* 77, 1 (2007), 81–112.
- [17] Michael Henderson, Rola Ajawi, David Boud, and Elizabeth Molloy. 2019. Identifying feedback that has impact. In *The impact of feedback in higher education*. Springer, 15–34.
- [18] Hassan Khosravi, Simon Buckingham Shum, Guanliang Chen, Cristina Conati, Yi-Shan Tsai, Judy Kay, Simon Knight, Roberto Martinez-Maldonado, Shazia Sadiq, and Dragan Gašević. 2022. Explainable artificial intelligence in education. *Computers and Education: Artificial Intelligence* (2022).
- [19] Vitomir Kovanović, Srećko Joksimović, Zak Waters, Dragan Gašević, Kirsty Kitto, Marek Hatala, and George Siemens. 2016. Towards automated content analysis of discussion transcripts: A cognitive presence case. In *Proceedings of the 6th LAK*. 15–24.
- [20] Lisa-Angelique Lim, Shane Dawson, Dragan Gašević, Srećko Joksimović, Abelardo Pardo, Anthea Fudge, and Sheridan Gentili. 2021. Students' perceptions of, and emotional responses to, personalised learning analytics-based feedback: an exploratory study of four courses. *Assessment & Evaluation in Higher Education* 46, 3 (2021), 339–359.
- [21] Jionghao Lin, Mladen Rakovic, David Lang, Dragan Gasevic, and Guanliang Chen. 2022. Exploring the Politeness of Instructional Strategies from Human-Human Online Tutoring Dialogues. In *LAK22: 12th International Learning Analytics and Knowledge Conference*. 282–293.
- [22] Anastasiya A Lipnevich, David AG Berg, and Jeffrey K Smith. 2016. Toward a model of student response to feedback. In *Handbook of human and social conditions in assessment*. Routledge, 169–185.
- [23] Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. 2020. From local explanations to global understanding with explainable AI for trees. *Nature machine intelligence* 2, 1 (2020), 56–67.
- [24] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Proceedings of the 31st NeurIPS* 30 (2017).
- [25] Effie Maclellan. 2005. Academic achievement: The role of praise in motivating students. *Active learning in higher education* 6, 3 (2005), 194–206.
- [26] Bruce M McLaren, Krista E DeLeeuw, and Richard E Mayer. 2011. A politeness effect in learning with web-based intelligent tutors. *International Journal of Human-Computer Studies* 69, 1-2 (2011), 70–79.
- [27] Danielle S McNamara, Arthur C Graesser, Philip M McCarthy, and Zhiqiang Cai. 2014. *Automated evaluation of text and discourse with Coh-Matrix*. Cambridge University Press.
- [28] Maria Mikheeva, Sascha Schneider, Maik Beege, and Günter Daniel Rey. 2019. Boundary conditions of the politeness effect in online mathematical learning. *Computers in Human Behavior* 92 (2019), 419–427.
- [29] David J Nicol and Debra Macfarlane-Dick. 2006. Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in higher education* 31, 2 (2006), 199–218.
- [30] Serena Nicoll, Kerrie Douglas, and Christopher Brinton. 2022. Giving Feedback on Feedback: An Assessment of Grader Feedback Construction on Student Performance. In *LAK22: 12th International Learning Analytics and Knowledge Conference*. 239–249.
- [31] Ikenna Osakwe, Guanliang Chen, Alex Whitelock-Wainwright, Dragan Gašević, Anderson Pinheiro Cavalcanti, and Rafael Ferreira Mello. 2022. Towards automated content analysis of educational feedback: A multi-language study. *Computers and Education: Artificial Intelligence* (2022).
- [32] Berry M O'Donovan, Birgit den Outer, Margaret Price, and Andy Lloyd. 2021. What makes good feedback good? *Studies in Higher Education* 46, 2 (2021), 318–329.
- [33] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. *The development and psychometric properties of LIWC2015*. Technical Report.
- [34] Filipe Dwan Pereira, Samuel C Fonseca, Elaine HT Oliveira, Alexandra I Cristea, Henrik Bellhäuser, Luiz Rodrigues, David BF Oliveira, Seiji Isotani, and Leandro SG Carvalho. 2021. Explaining Individual and Collective Programming Students' Behavior by Interpreting a Black-Box Predictive Model. *IEEE Access* 9 (2021), 117097–117119.

- [35] Tracii Ryan, Michael Henderson, Kris Ryan, and Gregor Kennedy. 2021. Designing learner-centred text-based feedback: a rapid review and qualitative synthesis. *Assessment & Evaluation in Higher Education* 46, 6 (2021), 894–912.
- [36] Shirley V Scott. 2014. Practising what we preach: towards a student-centred definition of feedback. *Teaching in Higher Education* 19, 1 (2014), 49–57.
- [37] Lele Sha, Mladen Raković, Jionghao Lin, Quanlong Guan, Alexander Whitelock-Wainwright, Dragan Gašević, and Guanliang Chen. 2022. Is the Latest the Greatest? A Comparative Study of Automatic Approaches for Classifying Educational Forum Posts. *IEEE TLT* (2022), 1–14.
- [38] Maciej Tomczak and Ewa Tomczak. 2014. The need to report effect size estimates revisited. An overview of some recommended measures of effect size. *Trends in sport sciences* 1, 21 (2014), 19–25.
- [39] Yi-Shan Tsai, Rafael Ferreira Mello, Jelena Jovanović, and Dragan Gašević. 2021. Student appreciation of data-driven feedback: A pilot study on OnTask. In *LAK21: 11th international learning analytics and knowledge conference*. 511–517.
- [40] Ning Wang, W Lewis Johnson, Richard E Mayer, Paola Rizzo, Erin Shaw, and Heather Collins. 2008. The politeness effect: Pedagogical agents and learning outcomes. *International journal of human-computer studies* 66, 2 (2008), 98–112.
- [41] Naomi Winstone, David Boud, Phillip Dawson, and Marion Heron. 2022. From feedback-as-information to feedback-as-process: a linguistic analysis of the feedback literature. *Assessment & Evaluation in Higher Education* 47, 2 (2022), 213–230.
- [42] Min Yang and David Carless. 2013. The feedback triangle and the enhancement of dialogic feedback processes. *Teaching in Higher Education* 18, 3 (2013), 285–297.
- [43] Michael Yeomans, Alejandro Kantor, and Dustin Tingley. 2018. The politeness Package: Detecting Politeness in Natural Language. *R Journal* 10, 2 (2018).