# Towards automated content analysis of feedback: A multi-language study

Ikenna Osakwe[1], Guanliang Chen[1], Alex Whitelock-Wainwright[1], Dragan Gašević[1], Anderson Pinheiro Cavalcanti[2], and Rafael Ferreira Mello[2]

{richard.osakwe, guanliang.chen, alex.wainwright, dragan.gasevic}@monash.edu

apc@cin.ufpe.br, rafael.mello@ufrpe.br

[1]Monash University, [2]Universidade Federal Rural de Pernambuco

## ABSTRACT

Feedback is a crucial element of a student's learning process. It enables students to identify weaknesses and improve self-regulation. However, studies show this to be an area of great dissatisfaction in higher education. With ever-growing course participation numbers, delivering effective feedback is becoming an increasingly challenging task. Hence, this paper explores the use of automated content analysis to examine feedback provided by instructors for good feedback practices measured on *self*, *task*, *process*, and *self-regulation* levels. For this purpose, four binary XGBoost classifiers were trained and evaluated, one for each level of feedback. The results indicate effective classification performance on self, task, and process levels with accuracy values of 0.87, 0.82, and 0.69, respectively. Additionally, inter-language transferability of feedback features is measured using cross-language classification performance and feature importance analysis. Findings indicate a low generalizability of features between English and Portuguese feedback spaces.

## 1. INTRODUCTION

Despite widespread recognition of feedback's importance to learning [23, 29, 10], much of the current literature indicates a pervasiveness of low quality feedback in higher education [13]. Feedback quality is consistently rated one of the greatest causes of dissatisfaction for higher education students [9]. LA researchers are actively exploring automated feedback solutions that can enable instructors to efficiently identify and employ good feedback practices, and improve the speed of feedback delivery to students [15]. In that vein, several studies [17, 19, 28, 30] have examined the use of data mining methods to generate automated textual feedback. These analyses are often limited to domain specific areas such as computer programming or writing, or lack of grounding in educational theory. Much less work has gone into the exploration of automated domain-agnostic analysis to identify good feedback practices [4, 24]. Progress in such areas can enhance the instructor's ability to provide effective feedback comments and analyze features associated with good feedback practices for generalizable feedback generators. Therefore, this study aims to answer the following Research Questions (RQs):

1. To what extent can the automated analysis of feedback messages be used to identify good feedback practices?
   (a) How accurate are the predictions that are made about these feedback practices?
   (b) What are specific features of text that can be used to predict the use of good feedback practices?
2. How transferable are the identified feedback features to text written in different languages?

## 2. METHOD

### 2.1 Data

The dataset used in the current study consisted of feedback comments provided by instructors in Learning Analytics, Software Engineering, and Environmental Studies courses. A total of 2,092 observations were taken; 1,000 Portuguese records and 1,092 English records.

### 2.2 Coding Scheme

This study utilized Hattie and Timperley's [14] four levels of feedback due to its suitability for textual analysis due to its focus learning tasks, learning process, and self-regulation Cavalcanti et al. [4]. Hence, feedback examples were coded using Hattie and Timperley's [14] proposed four levels of feedback (see Table 1).

Feedback examples were coded by experts using instructions of Hattie and Timperley's [14] study. Each feedback record was examined by two expert coders separately. After this step, the differences between each pair of experts were compared. For the Portuguese feedback examples, the inter-rater agreement reached 72.2% with a Cohen's kappa (inter-rater ability considering chance [7]) of 0.44. The English feedback comments had inter-rater agreement of 63.8% and Cohen's kappa of 0.38. These measures met expectations for content analysis experimentation [20].

Table 1: Four levels of feedback identified by Hattie and Timperley [14]. Each level specifies different elements that the feedback is targeting and can be regarded as hierarchical, ranging from general comments made about the student themselves up to directives on how to improve self-regulation.

| **Level** | **Description** | **Example** |
|---|---|---|
| Feedback about the self (FS) | Personal evaluations about the learner | "You are a bright student" |
| Feedback about the task (FT) | How well tasks are understood or performed | "You need to include more about the Treaty of Versailles." |
| Feedback about the process (FP) | Processes needed to understand or perform tasks | "You need to edit this piece of writing by attending to the descriptors you have used so the reader is able to understand the nuances of your meaning." |
| Feedback about self-regulation (FR) | How to improve self-regulation | "You already know the key features of the opening of an argument. Check to see whether you have incorporated them in your first paragraph." |

Table 2: Number of instances for each class in the training and test datasets for each level of feedback.

| | | **Class 0** | **Class 1** | **Total** |
|---|---|---|---|---|
| **FS** | **Train** | 1149 (82.19%) | 249 (17.81%) | 1398 (70%) |
| | **Test** | 567 (82.17%) | 123 (17.83%) | 690 (30%) |
| **FT** | **Train** | 602 (43.06%) | 796 (56.94%) | 1398 (70%) |
| | **Test** | 297 (43.04%) | 393 (56.96%) | 690 (30%) |
| **FR** | **Train** | 1290 (92.27%) | 108 (7.73%) | 1398 (70%) |
| | **Test** | 637 (92.32%) | 53 (7.68%) | 690 (30%) |
| **FP** | **Train** | 808 (57.80%) | 590 (42.20%) | 1398 (70%) |
| | **Test** | 399 (57.83%) | 291 (42.17%) | 690 (30%) |

The annotation process led to a dataset with four sets of binary classes: class 0 if a feedback message did not belong to a particular level; class 1 if the feedback message belonged to the feedback level.

## 2.3 Feature Engineering

Feature extraction was informed by relevant studies [4, 24, 16]. The studies promote the use of linguistic features such as those developed in LIWC (Linguistic Inquiry and Word Count) [27] and Coh-Metrix [11] over traditional textual features such as lexical N-grams or Part-Of-Speech. According to Kovanović et al. [16], these features encourage overfitting by inflating the feature space. Additionally, these traditional features are data dependent and thus make it difficult to define the feature space beforehand [16]. Hence, we used feature sets that incorporated 86 LIWC [27] features, 78 Coh-Metrix [11] features, and two additional features, which are relevant to this content area — number of named entities and language of delivered feedback.

## 2.4 Analysis

### *2.4.1 Data Analysis and Pre-processing*

For the general classifier, feedback examples from both the English and Portuguese datasets were combined and split into 70% training and 30% test sets (Table 2). The training data suffered from class imbalances; particularly at the FS and FR levels.

### *2.4.2 Handling Class Imbalance*

Studies have shown class imbalances can have a negative impact on model prediction performance [26]. To alleviate the class imbalance problem, sampling algorithms are often employed to adjust the ratio of represented classes. SMOTE is a popular oversampling method that analyzes the data records in a two-dimensional vector space of given classes and generates data points as a linear combination of existing data points [5].

## 2.5 Model Selection and Evaluation — RQ1a

Decision tree ensembles are widely regarded classification algorithms that are well suited to feedback analysis [4, 24]. This is due to their white-box properties, easy interpretability, high accuracy and ability to identify important features in a dataset [4, 24, 6, 8].

This study employed a decision tree implementation called XGBoost [6]. XGBoost has been shown to outperform Random Forest on numerous classification tasks [22, 31]. The algorithm utilizes gradient boosting, which involves sequentially combining models (in this case, decision trees) that predict the residuals or errors of previous models at each iteration to improve overall accuracy [6]. XGBoost is ideal due to their superior accuracy and their implicit analysis of feature importance [6]. Four binary XGBoost classifiers were trained; one for each level of feedback.

### *2.5.1 Feature Analysis —- RQ1b*

The outputs of decision tree models can be analyzed with tools such as SHAP (SHapley Additive exPlanations) [18].Given an input of a machine learning model and data records, SHAP leverages the concept of Shapley values by measuring the average marginal contribution of a feature over all possible permutations. SHAP can diagnose the most impactful features using their SHAP value, which is the mean absolute contribution of each feature [18]. A higher SHAP value for a feature implies a greater importance compared to another feature.

### *2.5.2 Feature Transferability — RQ2*

To measure the transferability of features across languages, the dataset was split by language, creating Portuguese and English feedback datasets. Each of these datasets was split into training and test splits (70% training and 30% test set), and binary classifiers were trained and tuned, resulting in English feedback trained classifiers, and Portuguese feedback trained classifiers for each level of feedback, with the exception of the FR level. For the Portuguese feedback examples, the FR level had just eight positive instances out

Table 3: Performance of the classifiers trained to address research question RQ1 on the combined dataset involving both the English and Portuguese datasets. Legend: ACC – Accuracy; K – Cohen's kappa; F1 - F1 Score.

| | FS | | | FT | | | FR | | | FP | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Class Balancing | ACC | K | F1 | ACC | K | F1 | ACC | K | F1 | ACC | K | F1 |
| None | **0.88** | **0.52** | **0.58** | **0.82** | 0.64 | **0.83** | **0.92** | 0.00 | 0.00 | 0.68 | 0.33 | 0.57 |
| SMOTE | 0.87 | 0.51 | **0.58** | **0.82** | **0.65** | **0.83** | 0.91 | 0.04 | 0.07 | **0.69** | **0.35** | **0.59** |

of 1,000 records, which was not enough to train a machine learning algorithm [12]; hence, this level of feedback was excluded from all transferability analysis.

Once the English and Portuguese trained classifiers were developed, feature transferability was measured by i) *the inter-language prediction performance*: the prediction performance (measured by accuracy, F1 score and Cohen's $\kappa$) of the English trained classifier on the English test set was compared to the predictor performance on the Portuguese test set for the FS, FT, and FP levels of feedback. The same process was repeated for the Portuguese trained classifier; ii) *A comparison of significant features*: The most important features for the English and Portuguese trained classifiers are compared at the FS, FT and FP levels of feedback.

# 3. RESULTS AND DISCUSSION

The goal of this study was to examine how accurately one can model the feature space of good feedback practices, and how this feature space varies across languages. In that vein, four research questions were answered using novel statistical learning methodologies, with a view of promoting good feedback practices at scale.

## 3.1 Model Performance —- RQ1a

Research question RQ1 focused on investigating the extent to which the automated analysis of feedback messages can be used to identify good feedback practices. Four binary classifiers were developed using a variety of features (see 2.3) The best performing models were effective in identifying FS, FT and FP. While not a direct comparison due to the addition of the English feedback examples, the models achieved better results over those reported by Cavalcanti et al. [4]. The classifiers were able to improve accuracy by 0.07 and 0.05 for FT and FP, respectively and increase kappa values by 0.11, 0.35, and 0.06 for FS, FT, and, FP, respectively.

Similar to previous works [4], the FR classifier was not as effective in identifying instances. The model obtained a poor kappa of 0.06, which was likely caused by the model's poor ability of detecting positive cases of FR. Poor performance on this level was due to the significantly lower cases of positive instances as compared to the other levels of feedback.

## 3.2 Feature Analysis —- RQ1b

The focus of research question RQ1b was analyzing the most important textual features associated with the four levels of feedback. Hattie and Timperley [14] state that FS involves evaluations of the person, which are often a form of praise. The current findings add weight to this claim, as those features found to be most important in predicting the FS level were affective processes (particularly, positive emotions) and social processes, which align with the concept of praise. FS is often thought to be the least effective level of feedback [3, 14] and relatedly, the FS classifier had a negative association with discrepancy words; this might indicate FS comments have little actionable information or insight.

FT is sometimes referred to as corrective feedback and provides information on details related to task accomplishment such as correctness or behavior. Accordingly, this study found the predictors most associated with FT were those that related to the amount of information provided. Specifically, higher values of word counts, frequency of content words and minimum frequency of content —- all of which can be linked to greater information —- were positively correlated with observance of FT. Hattie and Timperley [14] suggest instructors not to rely solely on FT, but rather to view it as a process that moves the student to FP and FR. This theory is backed by the finding of strong negative association of causation words and FT; hence, FT comments were less likely to illustrate the causes of the student's failings, which is essential for the learner's self-regulation [14, 3, 21].

Compared to FT, FP is believed to promote a deeper understanding of learning as it enables the identification of relationships between resources and output, and the development of stronger cognitive processes. To achieve this, Balzer et al. [1] state FP should concern information about actual relations in the learning environment, relations which have been recognized by the learner, and relations between the learning environment and the learner's perceptions. Therefore, the value of FP comes from providing useful information on relationships. The findings of this study corroborate the theoretical views of FP. Amongst the most important features for FP were frequency of content words, adverbs, negative connectives and discrepancy words. These imply that FP comments were tied to providing new and corrective information. Other significant features can be tied back to relationships; including frequency of semicolons (semicolons are often used to link together ideas) and features associated with space and relativity.

According to Butler [3], one of the goals of FR should be to improve the student's ability to monitor current progress and use that information to form effective learning strategies. Accordingly, some of the most important predictors of FR were greater present and future focused processes.

## 3.3 Feature Transferability -— RQ2

To address research question RQ2, we studied inter-language classifier performance, and compared the most significant features for classifiers trained on different language feedback. Barbosa et al. [2] used similar linguistic features to those used in this project, such as LIWC and Coh-Metrix, to study cross-language classification of cognitive presence in online discussions, and found features to be independent of language; hence, we expected to find a moderate level of generalizability of feedback features across languages. However, our findings indicate a low transferability of feedback features. As seen on Table 5, the average accuracy differential on inter-language performance amounted to -0.06, -0.59, and -0.26; while the average kappa differential was approximately -0.50, -0.27, and -0.33 for FS, FT, and FP, respectively. Likewise, the Portuguese and English trained classifiers showed minimal overlap in their most important

Table 4: Top 10 important features are measured using SHAP and displayed from most to least important for FS, FT, FR and FP classifiers.

| FS | | | FT | | |
|---|---|---|---|---|---|
| **Variable** | **Description** | **SHAP** | **Variable** | **Description** | **SHAP** |
| liwc.Exclam | Freq. of exclamation marks | 1.02 | cm.WRDFRQa | Freq. of all words | 0.46 |
| liwc.posemo | Freq. of words with positive emotion | 0.73 | cm.WRDFRQc | Freq. of content words | 0.39 |
| liwc.you | Freq. of the word ”you” | 0.24 | cm.WRDFRQmc | Minimum freq. of content words | 0.34 |
| liwc.affect | Freq. of affective words | 0.20 | cm.DRNP | Noun phrase density | 0.10 |
| cm.SYNMEDlem | Minimal edit distance of lemmas | 0.20 | cm.DRAP | Adverbial phrase density | 0.10 |
| cm.WRDFRQc | Freq. of content words | 0.15 | liwc.SemiC | Freq. of semicolons | 0.08 |
| liwc.tentat | Freq. of tentative words | 0.15 | cm.DESWLsy | Mean word length | 0.07 |
| liwc.reward | Freq. of words associated with reward | 0.14 | liwc.adverb | Freq. of adverbs | 0.07 |
| liwc.informal | Freq. of informal words | 0.14 | liwc.social | Freq. of words related to social processes | 0.07 |
| cm.WRDPRP2 | Freq. of second person pronouns | 0.14 | liwc.article | Freq. of articles | 0.07 |
| **FS** | | | **FT** | | |
| **Variable** | **Description** | **SHAP** | **Variable** | **Description** | **SHAP** |
| cm.CRFNO1 | Noun overlap between adjunct sentences | 0.56 | liwc.SemiC | Freq. of semicolons | 0.39 |
| cm.WRDPRP3s | Freq. of third person pronouns | 0.50 | cm.LSASS1 | LSA measure of semantic coherence | 0.19 |
| cm.CRFSO1 | Word stem overlap between adjunct sentences | 0.43 | cm.CNCNeg | Freq. of negative connectives | 0.12 |
| cm.DRAP | Adverbial phrase density | 0.35 | liwc.adverb | Freq. of adverbs | 0.11 |
| cm.CRFCWOa | Content word overlap of all sentences | 0.25 | cm.DESWLltd | Standard deviation of average no. of letters/word | 0.09 |
| liwc.risk | Freq. of risk related words | 0.23 | liwc.space | Freq. of words related to space | 0.09 |
| liwc.differ | Freq. of words related to differentiation | 0.21 | liwc.verb | Freq. of verbs | 0.08 |
| liwc.focusfuture | Freq. of future focus words | 0.21 | liwc.shehe | Freq. of third person singular pronouns | 0.07 |
| liwc.focuspresent | Freq. of present focus words | 0.20 | cm.SYNLE | Mean no. of words before the main verb | 0.06 |
| liwc.affiliation | Freq. of affiliation words | 0.16 | liwc.discrep | Freq. of words associated with discrepancy | 0.06 |

Table 5: For RQ2 classifiers are exclusively trained on English (EN) and Portuguese (PT) feedback examples. Performance of each classifier is measured against EN and PT feedback examples. Legend: ACC – Accuracy; K – Cohen’s kappa; F1 - F1 Score.

| | | FS | | | FT | | | FP | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | ACC | K | F1 | ACC | K | F1 | ACC | K | F1 |
| EN Classifier | EN | 0.83 | **0.42** | **0.52** | **0.69** | **0.13** | **0.31** | **0.66** | **0.23** | **0.49** |
| | PT | **0.85** | 0.03 | 0.04 | 0.11 | 0.00 | 0.00 | 0.49 | -0.02 | 0.30 |
| PT Classifier | EN | 0.79 | 0.06 | 0.12 | 0.28 | 0.00 | 0.43 | 0.35 | 0.00 | 0.52 |
| | PT | **0.94** | **0.74** | **0.77** | **0.91** | **0.49** | **0.95** | **0.78** | **0.56** | **0.79** |

features across all levels of feedback.

One possible explanation for this finding might be the difference in courses represented in the English and Portuguese datasets. English feedback examples were primarily from STEM related courses, including Environmental Studies and Software Engineering, while Portuguese feedback examples had more of a mix, hailing from Biology and Literature courses. Hence, the different nature of represented courses might have influenced the transferability analysis.

Another explanation for the low transferability of features might be the cultural differences in communication. For instance, at the FS level of feedback, we observed greater association of friendship and social processes for the English feedback; i.e. English instructors might have displayed a greater level of familiarity with students. As an instructor can be viewed as an authority figure, this difference might be related to whether a culture is “horizontal”, and therefore emphasizes equality, or “vertical”, and emphasizes hierarchy [25]. The implications of this finding would indicate instructors will need to consider the cultural backgrounds of the learner while delivering feedback for improved efficacy.

# 4. CONCLUSION AND FUTURE RESEARCH

This study proposed four main contributions. First, this study explored how accurately a trained model can identify the presence of different feedback practices. The constructed classifiers, using primarily linguistic and psychological features, were effective in identifying the presence of FT, FP and FS levels of feedback and showed better performance than similar works in this content area; however, the FR classifier was marred by a lack of adequate data. The implications of these results provide a proof of concept for a tool that can automatically analyze and potentially diagnose the contents of an instructor’s feedback. This promotes the understanding and utilization of good feedback practices to improve their efficacy on learner adoption.

Another goal of this paper was to identify the prominent textual features of good feedback practices. Identified features were able to corroborate the findings of educational research on feedback theory. The presented findings can be further used to inspire the design of future automated feedback generators, e.g., intentionally including the prominent terms specific to different feedback practices when generating feedback.

Finally, this study conducted an analysis of the transferability of feedback features across languages. Feedback tools should be generalizable enough to cater to a variety of languages. By analyzing the transferability of feedback features across languages, this study aimed to enhance the global adaptability of current and future feedback tools. The findings indicate feedback features have low transferability between feedback examples delivered in English and Portuguese. However, a more expansive study is suggested, with a greater size and variety of feedback from different languages.

# References

[1] W. K. Balzer, M. E. Doherty, and R. O'Connor. Effects of cognitive feedback on performance. *Psychol. Bull.*, 106(3):410–433, 1989. ISSN 1939-1455, 0033-2909.

[2] G. Barbosa, R. Camelo, A. P. Cavalcanti, P. Miranda, R. F. Mello, V. Kovanović, and D. Gašević. Towards automatic cross-language classification of cognitive presence in online discussions. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*, LAK '20, pages 605–614, Frankfurt, Germany, Mar. 2020. ACM.

[3] D. L. Butler and P. H. Winne. Feedback and Self-Regulated Learning: A Theoretical Synthesis. *Review of Educational Research*, 65(3):245–281, Sept. 1995. ISSN 0034-6543, 1935-1046.

[4] A. P. Cavalcanti, A. Diego, R. F. Mello, K. Mangaroska, A. Nascimento, F. Freitas, and D. Gašević. How good is my feedback?: a content analysis of written feedback. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*, pages 428–437, Frankfurt Germany, Mar. 2020. ACM.

[5] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic Minority Oversampling Technique. *JAIR*, 16:321–357, June 2002. ISSN 1076-9757.

[6] T. Chen and C. Guestrin. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, San Francisco, California, USA, Aug. 2016. ACM.

[7] J. Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20 (1):37–46, Apr. 1960. ISSN 0013-1644, 1552-3888.

[8] D. Denisko and M. M. Hoffman. Classification and interaction in random forests. *Proceedings of the National Academy of Sciences*, 115(8):1690–1692, Feb. 2018. ISSN 0027-8424, 1091-6490.

[9] P. Ferguson. Student perceptions of quality feedback in teacher education. *Assess Eval High Educ*, 36(1):51–62, Jan. 2011. ISSN 0260-2938.

[10] C. Glover and E. Brown. Written Feedback for Students: too much, too detailed or too incomprehensible to be effective? *Bioscience Education*, 7(1):1–16, May 2006. ISSN null.

[11] A. C. Graesser, D. S. McNamara, M. M. Louwerse, and Z. Cai. Coh-Metrix: Analysis of text on cohesion and language. *Behav Res Methods Instrum Comput.*, 36(2): 193–202, May 2004. ISSN 0743-3808, 1532-5970.

[12] T. Hastie, R. Tibshirani, and J. H. Friedman. *The elements of statistical learning: data mining, inference, and prediction.* Springer series in statistics. Springer, New York, NY, 2nd ed edition, 2009.

[13] J. Hattie and M. Gan. Instruction based on feedback. *Handbook of research on learning and instruction*, pages 249–271, 2011.

[14] J. Hattie and H. Timperley. The Power of Feedback. *Review of Educational Research*, 77(1):81–112, Mar. 2007. ISSN 0034-6543.

[15] H. Keuning, J. Jeuring, and B. Heeren. A systematic literature review of automated feedback generation for programming exercises. *ACM Transactions on Computing Education (TOCE)*, 19(1):1–43, 2018.

[16] V. Kovanović, S. Joksimović, Z. Waters, D. Gašević, K. Kitto, M. Hatala, and G. Siemens. Towards automated content analysis of discussion transcripts: a cognitive presence case. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, LAK '16, pages 15–24, Edinburgh, United Kingdom, Apr. 2016. ACM.

[17] M. Liu, Y. Li, W. Xu, and L. Liu. Automated Essay Feedback Generation and Its Impact on Revision. *IEEE Trans. Learn. Technol.*, 10(4):502–513, Oct. 2017. ISSN 1939-1382.

[18] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee. Explainable AI for Trees: From Local Explanations to Global Understanding. May 2019.

[19] X. Ma, S. Wijewickrema, S. Zhou, Y. Zhou, Z. Mhammedi, S. O'Leary, and J. Bailey. Adversarial Generation of Real-time Feedback with Neural Networks for Simulation-based Training. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pages 3763–3769, Melbourne, Australia, Aug. 2017. International Joint Conferences on Artificial Intelligence Organization.

[20] K. A. Neuendorf. *The content analysis guidebook.* SAGE, Los Angeles, second edition edition, 2017.

[21] D. J. Nicol and D. Macfarlane-Dick. Formative assessment and self-regulated learning: a model and seven principles of good feedback practice. *Studies in Higher Education*, 31(2):199–218, Apr. 2006. ISSN 0307-5079, 1470-174X.

[22] B. Pan. Application of XGBoost algorithm in hourly PM2.5 concentration prediction. *IOP Conference Series: Earth and Environmental Science*, 113:012127, Feb. 2018. ISSN 1755-1315. Publisher: IOP Publishing.

[23] A. Parikh, K. McReelis, and B. Hodges. Student feedback in problem based learning: a survey of 103 final year students across five Ontario medical schools. *Med. Educ.*, 35(7):632–636, 2001. ISSN 1365-2923.

[24] A. Pinheiro Cavalcanti, R. Ferreira Leite de Mello, V. Rolim, M. Andre, F. Freitas, and D. Gasevic. An Analysis of the use of Good Feedback Practices in Online Learning Courses. In *2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT)*, pages 153–157, Maceió, Brazil, July 2019. IEEE.

[25] S. Shavitt, A. K. Lalwani, J. Zhang, and C. J. Torelli. The Horizontal/Vertical Distinction in Cross-Cultural Consumer Research. *Journal of Consumer Psychology*, 16(4):325–342, 2006. ISSN 1532-7663. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1207/s15327663jcp1604_3

[26] P.-N. Tan and others. *Introduction to data mining*. Pearson Education India, 2007.

[27] Y. R. Tausczik and J. W. Pennebaker. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *JLS*, 29(1):24–54, Mar. 2010. ISSN 0261-927X, 1552-6526.

[28] J. Villalón, P. Kearney, R. A. Calvo, and P. Reimann. Glosser: Enhanced Feedback for Student Writing Tasks. In *2008 Eighth IEEE International Conference on Advanced Learning Technologies*, pages 454–458, Santander, Cantabria, Spain, 2008. IEEE.

[29] M. R. Weaver. Do students value feedback? Student perceptions of tutors' written responses. *Assess Eval High Educ*, 31(3):379–394, June 2006. ISSN 0260-2938.

[30] S. Wijewickrema, X. Ma, P. Piromchai, R. Briggs, J. Bailey, G. Kennedy, and S. O'Leary. Providing Automated Real-Time Technical Feedback for Virtual Reality Based Surgical Training: Is the Simpler the Better? In *Artificial Intelligence in Education*, Lecture Notes in Computer Science, pages 584–598, Cham, 2018. Springer.

[31] Z. Xiao, Y. Wang, K. Fu, and F. Wu. Identifying Different Transportation Modes from Trajectory Data Using Tree-Based Ensemble Classifiers. *ISPRS International Journal of Geo-Information*, 6(2):57, Feb. 2017. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute.