Understand students' self-reflections through learning analytics

Understand students' self-reflections through learning analytics

Author one*
Author one department
Author one institution
Author one city, Author one state
Author one country
author.one.email@institution

Author two
Author two department
Author two institution
Author two city, Author two state
Author two country
author.two.email@institution

Author three
Author three department
Author three institution
Author three city, Author three state
Author three country
author.three.email@institution

Author four
Author four department
Author four institution
Author four city, Author four state
Author four country
author.four.email@institution

Author five
Author five department
Author five institution
Author five city, Author five state
Author five country
author.five.email@institution

Author six
Author six department
Author six institution
Author six city, Author six state
Author six country
author.six.email@institution

Author seven
Author seven department
Author seven institution
Author seven city, Author seven state
Author seven country
author.seven.email@institution

ABSTRACT

Reflective writing has been widely recognized as one of the most effective activities for fostering students' reflective and critical thinking. The analysis of students' reflective writings has been the focus of many research studies. However, to date this has been typically a very labor-intensive manual process involving content analysis of student writings. With recent advancements in the field of learning analytics, there have been several attempts to use text analytics to examine student reflective writings. This paper presents the results of a study examining the use of theoretically-sound linguistic indicators of different psychological processes for the development of an analytics system for assessment of reflective writing. More precisely, we developed a random-forest classification system using linguistic indicators provided by the LIWC and Coh-Metrix tools. We also examined what particular indicators are representative of the different types of student reflective writings.

CCS CONCEPTS

• **Information systems** → **Clustering and classification**; • **Applied computing** → **E-learning**; **Distance learning**;

*Corresponding author

LAK'18, March 5–9 2018, Sydney, NSW, Australia

ACM ISBN 978-1-4503-4190-5/16/04…$15.00
https://doi.org/10.1145/2883851.2883950

KEYWORDS

learning analytics, text mining, self-reflections, online learning

ACM Reference format:
Author one, Author two, Author three, Author four, Author five, Author six, and Author seven. 2018. Understand students' self-reflections through learning analytics. In *Proceedings of 8th International Learning Analytics and Knowledge (LAK) Conference, Sydney, NSW, Australia, March 5–9 2018 (LAK'18),* 10 pages.
https://doi.org/10.1145/2883851.2883950

1 INTRODUCTION

An important characteristic of modern education is the focus on developing student higher cognitive skills and critical thinking. In this regard, some of the most fundamental learning activities relate to the use of (self-)reflection. The act of reflection is widely considered to be the essence of thinking process [22]. Reflection also represents an integral part of student self-regulation, and is essential for metacognitive adaptation of study approaches and goals [69]. The benefits of reflection are well recognized in contemporary educational practice [62, 63].

Over the years, there have been many approaches developed for fostering student reflection [cf. 62]. Among the different strategies used, reflective writing represents one of the most popular methods. An approach that is widely used for triggering the process of self-reflection that is necessary for metacognitive regulation. Not surprisingly, the assessment of student reflective writings has been the focus of many studies. These studies have largely employed more manual and labour intensive content-analysis methods to evaluate student reflective writings [41]. With the recent advancements in the field of learning analytics, there have been some attempts

to use more automated methods. With learning analytics increasingly being used as an approach to promote student awareness and regulation of their learning activities [28], developing automated processes to assess and understand the content of student self-reflections is an important undertaking. For example, the work of Ullmann [62] provides early evidence how analytics methods can be used to understand reflection as expressed in student essays.

This present study examined the use of automated text analytics methods for assessing the content of students' reflections on their learning activities. More precisely, we developed a learning analytics system for assessing reflection as expressed in student video annotations of their musical performances and examined what characteristics of their written language is most predictive of the different types of reflection.

2 BACKGROUND WORK

2.1 Reflection and Self-Regulated Learning

Across a range of disciplines, self-reflection is a key skill and strategy for students to cultivate as they enhance their higher order thinking skills and prepare for professional practice [19]. Reflection, or reflective practice, provides students with the opportunity to develop autonomy and confidence in their learning as they establish learning goals and take ownership of their learning strategies [14]. The higher education environment provides an opportune time for students to learn how to think independently, comment critically, and reflect on their learning [21] as they build the self-monitoring and self-regulating skills needed to be life-long learners [34]. Particularly, in clinical [36] or performance based disciplines [43] where students can watch a video recording of themselves demonstrating a particular skill, self-reflection exercises and assessment tasks can promote student self-assessment of their performance and identify areas of improvement. Reflective journals [14] and video annotation tools [3, 35] have shown to be effective tools for supporting students' self reflection and establishment of learning goals.

Despite the promotion of self-reflection skills through scaffolded activities and assessment, the depth of students' reflection and the progression from potentially superficial levels of reflection (e.g. descriptive) to more higher-order and goal-oriented [8, 35] requires an analysis of the specificity or type of statements made. An examination of the type of reflection in varying pedagogical or instructional conditions, can help identify students who are struggling to grasp higher levels of reflective thought (e.g. establishing goals) and those who may be largely focused on describing their skill or performance rather than critiquing it further. Reflection that involves goal-setting is much more challenging [58] and strategies are required to support students mastery of it. For example, ××× [4], examined the effect of students' experience with reflective tasks and the instructional conditions (graded vs ungraded activity) on the level or specificity of students' reflective statements. Their study concluded that prior experience with reflection along with continual grades and formative feedback on their reflective tasks encourages a greater amount of higher order critical reflection (e.g. goal-orientated or analysis of their motive or effect of their performance). Hence, scaffolding and an early introduction of reflective practice in the curriculum is needed for raising the depth and complexity of student reflection [19].

2.2 Automated analysis of self-reflections

While analysis of student self-reflections provides important insights into the development of students' higher order thinking, it is for the most part, very time-consuming manual process [61, 64, 65]. In most cases, it involves the quantitative content analysis [41] of student writings using a pre-defined coding scheme that focuses on identifying word indicators of the different facets of reflections [61]. Broadly speaking, different content analysis approaches exploit the underlying differences in the distributions of different linguistic categories between reflective and non-reflective statements and texts [63]. The majority of prior work has focused on the analysis of student essays and journal writings, with an emphasis on the depth of student reflection expressed (e.g., no reflection, simple reflection, and critical reflection) [66] . Not surprisingly, (self-)reflection in student writings was found to be substantially less frequent than desired [66], primarily on the descriptive [33] and shallow [56] levels.

Given the potential of computational methods for understanding student self-reflections, there have been several attempts to develop automated systems for assessment of student writings, including self-reflective texts. According to Ullmann [64], the existing automated content analysis systems can be divided into three broad and overlapping groups based on the adopted methodology:

1) Dictionary-based approaches [e.g., 15, 18, 46, 47, 61, 62],
2) Rule-based approaches [e.g., 30, 61, 65], and
3) Machine learning approaches [e.g., 1, 2, 5, 17, 46, 47, 62].

These three general approaches are also often combined. For example, Ullmann [61] proposed a system for identification of reflection in student essays using the combination of predefined dictionaries, regular expressions and rule-based analytics. Ullmann [61] also used synonym expansions to extend the list of words associated with reflective writings and provide more generalizable and stable performance. Similarly, Ullmann et al. [65] developed a rule-based system for reflection analysis in students' blog postings using WordNet [25], Linguistic Inquiry and Word Count (LIWC) tool [59], Stanford NLP parser [45], and synonym database. Using a custom-built vocabulary of the important keywords and focusing on the type of pronouns used (e.g., first person singular, third person plural) Ullmann et al. [65] devised a set of rules for identification of the different elements of reflective writings. A similar approach based on LIWC [59] and Coh-Metrix [31, 49] has been utilized by ××× [2] for the identification of students' level of critical thinking as expressed in discussion forum postings.

A further common approach to analyzing student writings is based on the use of natural language processing (NLP) methods. This is often applied in combination with different machine learning algorithms. The simplest NLP methods use frequencies of *N-grams* (i.e., word sequences of length N) as classification features [e.g., 1, 62]. For example, Ullmann [62] used N-grams as features for binary classification of 5,081 student reflection sentences and elements of reflective writings (i.e., experience, feelings, personal, critical stance, perspective, outcome), reporting classification accuracy as Cohen's κ range of .49–.83, depending on the particular coding category. Similarly, Gibson et al. [30] used part-of-speech (POS) tagging to match students' writings to the common POS phrases indicative of student's metacognitive activities, while Latent Semantic Analysis

(LSA) has been used by Cheng [17] to understand reflection in English language learners. Likewise, the Gibson and Kitto [29] NLP-based approach utilized the TF-IDF scoring [40], Latent Dirichlet Allocation (LDA) [10], and different keyword-based metrics for identification of the level of subjectivity and affectivity in students' reflective writings. Finally, × × × [1] used POS-tagging, Name-entity tagging, and syntactic dependency parsing (via Stanford NLP toolkit [45]) to build a classification system for examining students' levels of critical thinking.

3 RESEARCH QUESTIONS

While there has been a substantial amount of research on automated assessment of student reflective writings, the primary domain of analysis were long, complex texts, such as essays, blogs, or journals, in which students were expected to exercise reflective and critical thinking. As reflection is typically represented in just a small part of the written text, a large part of the existing research focused on the identification of different parts of written text that represent different types and facets of reflection. This was typically achieved through a combination of custom-built keyword and phrase matching mechanism, or by a data-driven NLP indicators, such as N-grams, that were chosen depending on the specifics of a particular study context. Hence, there are concerns regarding the external validity in the literature published up to date, with regards to what are the highly predictive – and psychologically sound – indicators of (self-)reflection in student writings and how they can be used to develop analytics systems for reflection assessment. As such, the research questions addressed in this study are

> Research Question 1:
> *What are the linguistic indicators of self-reflection, as captured in students' writings?*
>
> Research Question 2:
> *Can the identified indicators of self-reflection be used to develop an automated system for assessment of students' self-reflection?*

To address these questions, we used psychologically-sound and well-established linguistic measures of different psychological processes (e.g., affective, cognitive, social, biological) provided by the widely used LIWC [59] and Coh-Metrix [31, 49] tools in addition to the widely used N-grams, in order to develop an automated classification system for reflection assessment. To make the identification of the relevant reflection indicators more precise, we focused on analyzing short self-reflective writings rather than longer (e.g., essays or blogs) texts which typically contain much lower proportion of reflective writing. In particular, we examined students' self-reflection in their short annotations of the video recordings of their own musical performances.

4 METHOD

4.1 Study data

4.1.1 Study setting. The dataset in the present study is the same dataset that was used in the study described in [3]. The data comes from the four undergraduate courses in performing arts discipline offered in the 2012/2013 academic year at a large research-intensive public university in Canada. Course 1 and Course 2 were offered in the Fall 2012 semester while Course 3 and Course 4 were offered in the Winter 2013 semester. In all four courses, students were providing self-reflections on the video recordings of their own musical performances. In Course 1, the recordings were of students' group performances while in the other three courses, video recordings were of students' individual performances (Table 1). In addition, in Course 1 and Course 4, the creation of self-reflections was optional activity, while it was a course requirement in Course 3 and Course 4 and part of student assessment. In total, there were 77 different students across the four courses, with some students taking more than one course.

Table 1: Description of included courses and coded units of analysis

Course	Recording type	CLAS required	Enrolled students	Coded analysis units
Course 1	Group	No	31	145 (3.27%)
Course 2	Individual	Yes	40	1393 (31.44%)
Course 3	Individual	Yes	28	2457 (55.46%)
Course 4	Individual	No	20	435 (9.82%)
		Total:	119 (77[1])	4,430 (100%)

[1] Unique number of students.

To create their self-reflections, students used Collaborative Lecture Annotation System (CLAS) [50, 55], which is a software tool that enables students to annotate video materials, which are in this case videos of their art performances. In terms of the functionality, CLAS enables students to create *time-stamped annotations*, which are associated with a particular part of a video, and *general annotations* which are not associated with any part of the video and used to create general comment or summary of the video. Both time-stamped and general annotations can be either private or public, with the latter providing the opportunity for student collaboration and peer feedback.

4.1.2 Content analysis. After students' self-reflections were collected, the quantitative content analysis [41] was undertaken to categorize each student reflection using the coding scheme adapted

Table 2: Description of coding categories

Category	Definition	Example
`Observation`	Student indicates what they observed about their own behavior, but does not indicate why the behavior occurred.	"I still continue to have problems making eye contact..."
`Motive`	Student indicates what they observed and why it occurred.	"...being up there made me insecure and nervous, which led to my eyes dropping frequently..."
`Goal`	Student indicates what they will do next time or what they need to work on.	"What I really want to avoid is ending up just mirroring everything."

Table 3: Distribution of coding categories in test and train data sets

	Dataset		
Category	Train (75%)	Test (25%)	All
`Observation`	1,135 (34.17%)	382 (34.48%)	1,517 (34.24%)
`Goal`	1,848 (55.63%)	625 (56.41%)	2,473 (55.82%)
`Motive`	174 (5.24%)	56 (5.05%)	230 (5.19%)
`Other`	165 (4.97%)	45 (4.06%)	210 (4.74%)
Total:	3,322 (100%)	1,108 (100%)	4,430 (100%)

from Hulsman et al. [35]. Originally, Hulsman et al. [35] define four types of reflections based on the specificity of goals observed in them: (1) *observations* of own behavior (`Observation`), (2) *motive or effect* of own behavior (`Motive`), (3) *asking for feedback* for improvement (`Feedback`), and (4) *indicating a goal* of own behavior (`Goal`). Given that in our case reflective task was an individual learning activity, we omitted the *asking for feedback* category, resulting in the three different coding categories. The description and representative examples of each of the categories is given in Table 2.

As each annotation can potentially contain several reflections, the unit of analysis was a sentence segment, which was in most cases a complete subordinate or dependent clause. In total, 971 annotations which consisted of 3,324 individual sentences were coded by two coders, resulting in 4,430 coded units of analysis. Both coders went through the same training process and coded smaller sub-samples of data until Cohen's κ above 0.75 was reached. The distribution of different codes is shown in Table 3. We see that the majority of units were coded as either goal indications (55.92%) or observations (34.24%), while motivation was far less frequent, occurring in only 5.19% of the analysis units. Finally, we also included the category *Other* to code units that did not contain expression of any of the three reflection types, and it was used to code 4.74% of the analysis units.

4.2 Training and test data preparation

As the first step in our analysis process, we first split the data into training and test datasets (75% and 25% of the whole corpus, respectively), as commonly done in the machine learning [32, 54]. The model development and parameter tuning are done using the training set, while the final evaluation of model's performance is done on test set. By doing this, we prevent for overestimating the model performance which will occur if we estimated the model accuracy on the same data on which we learned the model parameters [32]. In total, training and test datasets contained 3,322 and 1,108 instances, respectively (Table 3). It should be noted that training and test datasets are created in a *stratified* manner, which means that the original proportions of coding categories (i.e., `Observation`, `Goal`, `Motive`, and `Other`) is preserved in both subsets (Table 3).

4.3 Feature extraction

In order to develop a classification system for student reflections, we extracted several different types of features. The extracted features were heavily based on the existing work in educational text and discourse analysis [e.g., 1, 2, 5, 6, 20, 23, 24, 31, 37, 49, 57, 60, 67], including the features which are strongly theory-driven and empirically validated. In total, we extracted 503 different features which we describe in the reminder of this section.

4.3.1 N-grams. As commonly done in text classification systems, we extracted basic *N-grams* features (i.e., unigrams, bigrams, and trigrams) from the training data (i.e., 75% of the whole corpus). Prior to N-gram extraction, we first removed *stopwords*, which are the highly frequent words in English (e.g., a, the, be, can, have) that do carry useful information for classification purposes [54]. Given that the use of N-grams results in inflation of the feature space and overfitting of the training data, we extracted only top 100 unigrams, bigrams, and trigrams to keep the size of the feature space limited and less prone to overfitting. The top ten most frequent unigrams, bigrams, and trigrams (Table 4) are about the quality of student performances and students' needs, goals, and feelings which could be used to gauge the type of student reflection. As expected, we also see a sharp decline in N-gram frequencies as `N` increases.

After we extracted a set of 300 N-gram features from the training set, we extracted the same set of N-grams features from the test set (i.e., the remaining 25% of the whole corpus that were not included in the training set). Therefore, the definition of the feature space only depends on the training data, while the test data is completely put aside and used only for the final validation of the classifier performance.

4.3.2 LIWC features. In addition to N-gram features, following the work of × × × [2], we used the Linguistic Inquiry and Word Count (LIWC) tool [59] to extract a large set of linguistic measures which are indicative of a large set of biological and psychological processes (e.g., perceptual, cognitive, affective, social), as well as different topics (e.g., work, achievement, personal, leisure, time) and linguistic categories (e.g., nouns, verbs, adjectives). The previous work [2] indicated that LIWC measures can be successfully used within learning analytics systems to uncover important psychological processes behind student behavior observed in trace data logs. In the current study, we used the 2015 version of the LIWC tool which provides the total of 93 empirically validated linguistic measures [cf. 59], including four high-level measures: (1) analytical thinking, (2) social status, leadership, and confidence, (3) authenticity, and (4) emotional tone.

Table 4: Top 10 unigrams, bigrams, and trigrams from the training data

Unigram	Freq.	Bigram	Freq.	Trigram	Freq.
need	383	left hand	112	practice front mirror	17
conducting	279	eye contact	71	use left hand	14
think	248	need work	55	third goal would	11
music	239	make sure	54	make eye contact	10
really	200	front mirror	36	second goal would	10
hand	182	goal would	32	first goal would	10
practice	181	feel like	30	three critical goals	8
ensemble	171	beat pattern	30	critical goals improvement	8
work	170	right hand	29	really need work	8
beat	161	also need	26	influence sound moment	7

4.3.3 Coh-Metrix features. In addition to LIWC, similarly to ×× × [2], we also used the Coh-Metrix tool [31, 49], which is a text analytics tool designed to measure different aspects of writing cohesion. Coh-Metrix provides 109 different measures of text cohesion (i.e., referential, causal, co-reference, temporal, spatial, and structural cohesion), several measures of text complexity and readability, and measures of linguistic category use.

Coh-Metrix has been extensively used in masny studies in the domain of collaborative learning to assess student outcomes [23], online discourse [24, 68], development of social ties [37–39], quality of student essays [7, 48], and learning resources [31]. Coh-Metrix has also been successfully used in learning analytics systems for assessing student-produced writings, such as student discussion messages [2, 67]. Given the goal of understanding the processes driving student self-reflections, Coh-Metrix provides a valuable set of empirical measures that can be used to understand the characteristics of each of the types of student reflections.

4.3.4 Context features. Given that several units of analysis can be present in a single sentence, we also included a single binary feature `first_in_sentence` which captures whether a particular unit of analysis is the first (or the only) unit in a given sentence. We hypothesized that students' observations would more often be first in a sequence of annotations given their sensemaking nature.

4.4 Data preprocessing

After feature extraction, we addressed the problem of class imbalance, as visible in Table 2. Following the approach suggested by ××× [2], we used the Synthetic Minority Oversampling Technique (SMOTE) [13, 16], which is a popular method for addressing the class imbalance problem. The SMOTE algorithm works by constructing additional synthetic data points as a linear combination of the existing data points. To process an existing data point X in an n-dimensional feature space $X = (x_1, x_2, x_3, ...x_n)$ using the SMOTE algorithm:

- Find K nearest neighbors of X (in our case, K=5) belonging to the same minority class.
- Select at random one of those K nearest neighbors (called Y),
- Generate a new synthetic data point as a random linear combination of X and Y:

$$Z = X + c * Y$$

where c is a random number between 0 and 1.

To increase the size of the minority class by N times, each minority-class data point would be processed N times. In contrast, to increase the size of the minority class by less than 100%, first a subset of the original data points was selected and then each of those data points would be processed exactly once. Figure 1 illustrates the application of SMOTE algorithm in our training set. The size of `Other` category was increased 11-fold (from 165 to 1815), and the size of `Motive` category was increased 10-fold (from 174 to 1740). In contrast, the size of the `Observation` category was increased only for 60% (from 1135 to 1816) by first selecting 60% of the original data points which were then processed by the SMOTE algorithm. At the end, the class imbalance problem was significantly reduced,

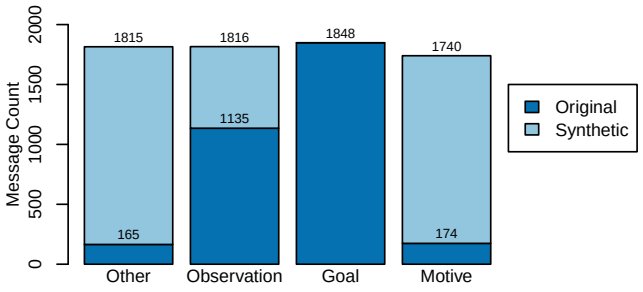

Figure 1: SMOTE preprocessing for class balancing.

which should increase the overall performance of the classification system.

Finally, we removed three extracted features that had the same value for all training instances, which effectively made them useless for our classification problem. Those three features were all LIWC metrics: (1) `family`: capturing family-related topics, (2) `filler`: representing the use of filler words (e.g., um, uh, ah, like, okay), and (3) `Quote`: concerning the use of quotation marks.

4.5 Model Selection and Evaluation

To develop a classification system for self-reflections, we used random forests [12], which are widely-used ensemble classification technique. Random forests combine a large number of decision-trees and bootstrap sampling to provide low-bias low-variance classification method [12]. A large study by Fernández-Delgado et al. [26] compared performance of 179 different classification techniques on 121 different datasets identified random forests along with Gaussian kernel Support Vector Machines (SVMs) as the state-of-the-art classification techniques.

A random forests classifier is an ensemble of a large number of decision trees (controlled by the `ntree` parameter) and the final classification decision is obtained by a simple majority voting mechanism across the whole ensemble [12]. An important characteristic of random forests is that each decision tree is constructed on a different bootstrap sample (i.e., a sub-sample with repetitions of the same size as original) of the training data, and evaluated on the data points that were not included in the bootstrap sample. Moreover, each tree is constructed using only a random subset of the available features (the size of feature subset is controlled by the `mtry` parameter) without tree pruning [12].

Random forest classification enables the assessment of the importance of the different classification features, by looking how often and how early each feature occurs in the decision tree ensemble. While there are many concrete measures of feature importance [44], one of the most widely used measures of feature importance is the Mean Decrease Gini (MDG) index which measures the reduction of the Gini impurity in the resulting decision sub-trees. In this manner, the MDG index assesses how useful a given feature is for separating data instances among different classes. For a classification feature F_i, MDG is calculated as the average decrease in the Gini impurity across all decision tree nodes where feature F_i was used.

As previously stated, random forest classifiers require specification of the two configuration parameters: (1) `ntree`: the number

of trees in ensemble, and (2) `mtry`: the number of random features used by each tree). The number of trees in the ensemble should be sufficiently large so that the performance of the classifier is stabilized [51] while the number of features used by each tree should be carefully optimized to balance bias-variance tradeoff [32]. According to Oshiro et al. [51], ensembles of 64–128 trees are recommended to balance between the processing time, memory usage, and classification accuracy. This recommendation is aligned with our previous implementations of random forests [2] where the classification performance stabilized from around 100–150 trees. Still, given the relatively small size of our training set (7219 instances), the processing time and memory constraints were less critical so we decided to use 500 trees in the ensemble (i.e., `ntree = 500`). Finally, to optimize `mtry` parameter, we used ten repetitions of 10-fold cross-validation to examined 19 candidate values: 2, 3, 4, 6, 8, 11, 15, 20, 27, 36, 48, 65, 87, 156, 209, 279, 373, and 500. The actual parameter values were generated by the `caret` package and its default grid search strategy.

4.6 Implementations

The implementation of the classifier was done in the Python and R programming languages and by using several software packages and libraries:

(1) The extraction of N-grams was done using `NLTK` library [9] for Python programming language,
(2) The extraction of psychological indicators was done with LIWC 2015 tool [53, 59],
(3) The extraction of text coherence measures was performed with Coh-Metrix toolkit [31, 49],
(4) Stratified sub-sampling of test and train data was done through `scikit-learn` [52] machine learning library for Python programming language,
(5) The development of a random forest classifier was done using `randomForest` R package [44], and finally,
(6) The model training, selection, and validation was performed with `caret` R package [27].

4.7 Limitations and future work

The major limitation of the adopted approach is that the collected data are from the same domain (i.e. performing arts) and thus, might not be representative of a broader range of student self-reflections across different disciplines. As such, one direction for our future work will be to examine the performance of the developed classification scheme on datasets from different study domains. Moreover, there are several potentially useful classification features which have not been included in the design of our system. For example, it is likely that the inclusion of the codes from student's previous reflections as classification features would provide important additional information that would substantially increase the classification accuracy. For example, if a student wrote an observation reflection, then it is more likely that his following reflection would be goal or motive reflection. However, at the moment, each annotation is categorized in isolation from all other reflections made by a student, which likely reduces the classifier's performance. In this regard, the use of structured classification approach, such as one employed by [67] is an important direction for the future work.

Table 5: Random forest parameter tuning results

mtry	Accuracy (SD)	Cohen's κ (SD)	mtry	Accuracy (SD)	Cohen's κ (SD)
2	.81 (.01)	.75 (.02)	48	.88 (.01)	.85 (.01)
3	.84 (.01)	.79 (.02)	65	.88 (.01)	.84 (.01)
4	.86 (.01)	.81 (.02)	87	.88 (.01)	.84 (.02)
6	.87 (.01)	.83 (.01)	116	.89 (.01)	.85 (.02)
8	.88 (.01)	.84 (.01)	156	.88 (.01)	.84 (.02)
11	.88 (.01)	.84 (.01)	209	.88 (.01)	.84 (.02)
15	.88 (.01)	.84 (.01)	279	.88 (.01)	.84 (.02)
20	.88 (.01)	.84 (.01)	373	.87 (.01)	.83 (.02)
27	.88 (.01)	.84 (.01)	500	.86 (.01)	.82 (.02)
36	.88 (.01)	.85 (.01)			
Min:	.81	.75	Range:	.07	.10
Max:	.89	.85	Mean:	.87	.83

5 RESULTS

5.1 Model training and evaluation

Figure 2 and Table 5 show the results of random forest model optimization via cross-validation. The best performance of .89 classification accuracy (SD = .01) and Cohen's κ of .85 (SD = .02) was achieved with 116 features per decision tree on the training dataset. The difference between the worst- and best-performing model was 0.07 in classification accuracy and .10 κ which confirms the importance of parameter optimization and tuning on the final model performance (Table 5). The performance of the random forest model on the complete training set using the optimal `mtry` value is shown on Figure 3. We can see that the performance of the classifier stabilized with around 100 decision trees, indicating that 500 trees selected was more than enough to ensure good classifier performance. The average out-of-bag (OOB) error rate was .12, suggesting only 12% of the data points being misclassified in the training set. As expected, the error rates for the two most resampled classes (i.e., `Other` and `Motive`) were the lowest, while the highest error rate was observed for `Observation` category which was not resampled.

After developing the random classifier on the training data, we validated its performance on the holdout test data (25% of the whole

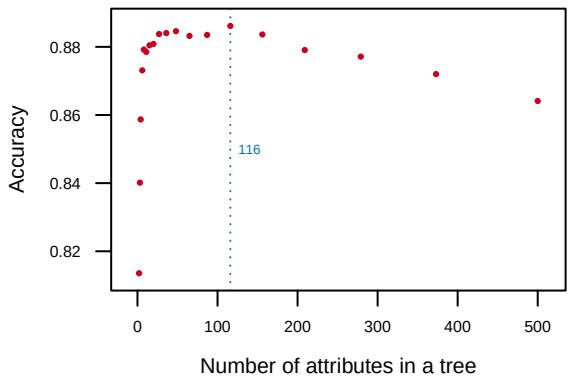

Figure 2: Random forest parameter tuning results.

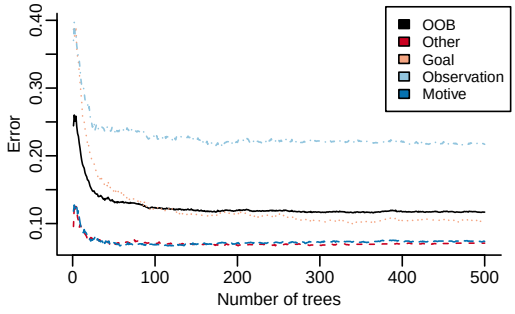

Figure 3: Best random forest configuration performance.

dataset). Our random forest classifier achieved .75 classification accuracy (95% CI[0.72, 0.77]) and Cohen's κ of 0.51 which is considered "Moderate" accuracy above the pure chance level [42]. The confusion matrix for the test data is shown in Table 7. We see that error rate for the `Goal` category is the lowest, followed by the moderate error rate for the `Observation` category. In contrast, we see that the `Other` and `Motive` categories were mostly misclassified as belonging to two former large categories.

Finally, to examine the value of the SMOTE preprocessing, we examined the confusion matrix of the random forest model developed using the original training and test datasets. The optimal `mtry` value was 500 by which the classifier obtained .73 (SD = .02) classification accuracy and Cohen's κ of .48 (SD = .04). Further validation of the classifier performance on the holdout test data showed .74 classification accuracy (95% CI[.72, .77]) and Cohen's κ of 0.50 which was slightly lower than the classifier performance obtained after the SMOTE pre-processing.

5.2 Feature importance analysis

In addition to assessing the classification accuracy, we also examined the contribution of different features to random forest performance. Table 8 provides the summary of feature MDG scores, while Figure 4 shows MDG scores for all 500 classification features.

Table 6: Train data confusion matrix for the final model

	Predicted				
Actual	Other	Observation	Goal	Motive	Error rate
Other	1686	80	49	0	.07
Observation	5	1655	177	11	.10
Goal	6	**373**	1422	15	.22
Motive	1	59	68	1612	.07

Table 7: Test data confusion matrix for the final model

	Predicted				
Actual	Other	Observation	Goal	Motive	Error rate
Other	9	**9**	**27**	0	.80
Observation	1	250	**131**	0	.34
Goal	1	**59**	564	1	.10
Motive	0	**30**	**25**	1	.98

We see a wide spread in MDG scores; 50% of features obtained an MDG score below 1.06 and 75% of features obtained an MDG score below 15.34. In contrast, certain features obtained much higher MDG scores, with the maximum MDG score of 219.94.

The detailed analysis of top twenty most important classification features is given in Table 9. While 146 classification features had above average MDG scores, given the space limitations, we focused our analysis on top twenty. We see that the most important classification feature was the LIWC category of perceptual words (`liwc.see`). In addition the use of past-oriented words (`liwc.focuspast`), punctuation, causal words, passive voice, and connectives were among the most important classification features.

Among the Coh-Metrix features, the most important were the ratio of causal particles to causal verbs (`cm.SMCAUSr`), use of agentless passive voice (`cm.DRPVAL`), use of nouns (`cm.WRDNOUN`) and noun phrases (`cm.DRNP`), use of connectives (`cm.CNCCaus`), causal verbs (`cm.SMCAUSv`), intentional cohesion of the the text (`cm.SMINTEr`), and number of words before main vebs in sentences (`cm.SYNLE`). The `Motive` reflections had the highest number of causal particles to causal verbs and words before the main verbs, indicating the complex language structure used to describe student motivation. Similarly, the use of agentless passive voice and connectives was strongly associated with the `Motive` category that also exhibited the highest intentional cohesion. In contrast, highest numbers of nouns and noun phrases were associated with the `Other` category, whereas causal verbs were most strongly associated with `Goal` category.

The most important LIWC features were related to students use of perceptual words (`liwc.see`) which were most strongly associated with the `Observation` and `Goal` categories and the least with the `Other` category. The `Observation` and `Motive` reflections also had a strong focus on the past events (`liwc.focuspast`), whereas `Goal` reflections did not. Somewhat unexpectedly, words related to

Table 8: Summary of classification feature importance

Min.	Q1	Median	Mean	Q3	Max.
0.00	0.12	1.06	10.82	15.34	219.94

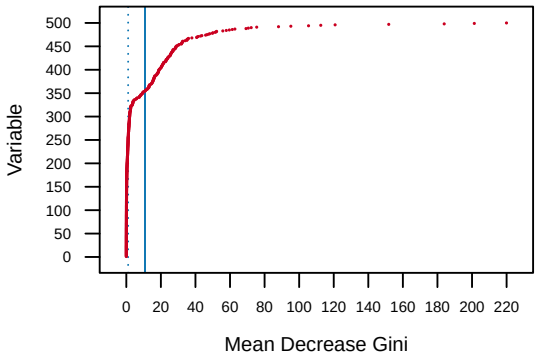

Figure 4: Feature importance by Mean Decrease Gini (MDG) measure. Dotted blue line shows median MDG score (1.06), while solid blue line shows average MDG score (10.81).

Table 9: Twenty most important features and their mean scores for messages in different phases of cognitive presence

			Coding category			
Feature	Description	MDG[1]	Other	Observation	Goal	Motive
liwc.see	Perceptual processes: seeing (e.g., view, saw, seen)	219.94	1.01 (3.35)	1.64 (4.43)	1.62 (4.57)	1.55 (3.29)
cm.SMCAUSr	Situation model: ratio of casual particles to causal verbs	201.33	0.12 (0.32)	0.13 (0.34)	0.11 (0.30)	0.37 (0.47)
cm.DRPVAL	Syntactic pattern density: agentless passive voice density, incidence	183.90	2.46 (12.11)	3.02 (17.57)	2.03 (13.36)	4.66 (20.52)
liwc.focuspast	Time orientation: focus towards past (e.g., ago, did, talked)	151.83	1.46 (3.38)	4.57 (6.34)	0.80 (2.81)	4.80 (6.66)
cm.WRDNOUN	Word information: noun incidence	120.80	252.12 (212.42)	186.47 (99.92)	208.13 (127.82)	194.34 (95.04)
liwc.ingest	Biological processes: ingestion (e.g., dish, eat, pizza)	112.50	0.58 (4.03)	0.30 (1.68)	0.33 (2.21)	0.39 (1.66)
cm.CNCCaus	Connectives: causal connectives, incidence	105.42	28.27 (58.70)	20.86 (42.82)	28.57 (47.68)	42.91 (50.47)
trust.ensemble	Frequency of "trust ensemble" bigram	95.18	0.00 (0.07)	0.00 (0.04)	0.00 (0.05)	0.00 (0.00)
cm.SMINTEr	Intentional cohesion: ratio of intentional particles to intentional actions/events	88.05	0.30 (0.66)	0.23 (0.49)	0.36 (0.59)	0.47 (0.68)
liwc.Period	Punctuation: use of full stop	75.44	9.17 (13.13)	6.38 (6.93)	7.14 (8.60)	5.45 (6.08)
cm.DRNP	Syntactic pattern density: incidence score of noun phrases	72.26	390.98 (208.55)	319.76 (115.17)	306.22 (133.34)	314.47 (105.34)
chamber.music	Frequency of "chamber music" bigram	70.59	0.00 (0.07)	0.00 (0.08)	0.00 (0.03)	0.01 (0.09)
liwc.AllPunc	Punctuation: all (e.g., periods, commas, question marks)	69.26	19.22 (19.05)	13.02 (9.85)	14.06 (12.39)	11.66 (8.12)
liwc.cause	Cognitive processes: causality (e.g., because, effect)	62.82	2.02 (4.45)	1.85 (4.32)	2.03 (4.66)	3.98 (5.33)
cm.SMCAUSv	Situational model: incidence score of causal verbs	61.34	33.60 (47.91)	46.08 (59.57)	65.65 (76.14)	44.28 (49.62)
liwc.insight	Cognitive processes: insight (e.g., think, know)	59.62	2.98 (5.02)	3.64 (5.34)	2.36 (5.34)	3.50 (4.56)
cm.SYNLE	Syntactic complexity: mean number of words before the main verb in the main clause	57.85	2.36 (3.68)	2.65 (3.08)	1.78 (2.60)	3.53 (4.16)
liwc.home	Personal concerns: home (e.g., kitchen, landlord)	55.83	0.10 (1.19)	0.08 (0.85)	0.05 (0.59)	0.04 (0.49)
liwc.Analytic	Summary measures: the measure of formal, logical, and hierarchical thinking processes	52.24	60.02 (37.26)	54.60 (35.94)	70.94 (33.05)	57.36 (34.69)
liwc.percept	Perceptual processes: all (e.g., look, heard, feeling)	51.97	4.20 (7.29)	4.85 (6.80)	4.69 (7.31)	5.04 (5.67)

[1] Mean decrease Gini impurity index.

biological ingestion processes (liwc.ingest) were strongly predictive of reflections in the Other category. The same category was also most strongly associated with personal concerns (liwc.home), the use of full stops (liwc.Period), and punctuation in general (liwc.AllPunc), and least associated with words describing perceptual processes (liwc.see and liwc.percept). On the other hand, the Goal reflections were the most analytic (liwc.Analytic), while the Motive reflections contained most perceptual (liwc.perceipt) and causal (liwc.cause) words

With regards to the contextual feature first_in_sentence, it did not show in the list of the top twenty features (Table 9). Upon a more detailed inspection, we found that first_in_sentence was the 28th most valuable classification feature, with an MDG score of 43.72, which is also substantially above the average MDG score of 10.81 or median MDG score of 1.06. The closer examination revealed that segments in Other category were most likely to be at the start of the sentence (or a complete sentence) ($Mean = 1.94, SD = 0.23$), followed by Observation ($Mean = 1.79, SD = 0.41$), Goal ($Mean = 1.72, SD = 0.45$), and finally Motive ($Mean = 1.67, SD = 0.47$) [1].

6 DISCUSSION

The classification results on the testing dataset showed that the use of N-grams and LIWC and Coh-Metrix features provides a good basis for the development of an automated self-reflection classification system. Cohen's κ of 0.51 represents a moderate level of agreement above the change level [42]. These results are promising and showing the potential of our approach. The results also indicate the significant benefits of classifier parameter tuning, given the substantial variation in the classifier's performance on the training dataset (Table 5). Table 5 indicates that 7% of the classification accuracy and .10 Cohen's κ can be solely attributed to the optimization of the mtry parameter (i.e., the number of attributes used in each tree of the forest). The most directly comparable results are by Ullmann [62] who reported slightly higher Cohen's κ values (.49–.83), albeit on a different, binary classification problem with different coding categories.

A further contribution from the study is the examination of the important classification features. While SVMs provided the best performance in most experiments by Ullmann [62], we opted for more interpretable classification methods which can be used to improve conceptual understanding of students' self-reflection. Our results showed that a small subset of highly predictive indicators can be used to distinguish between the different types of reflective statements (Table 9). In particular, several of the indicators that capture different linguistic structures (e.g., agentless passive voice density, syntactic pattern density, connectives) were identified as some of the best predictors of student self-reflection. Hence, in our future work, we will also examine the inclusion of syntactic dependency features, such as the ones used by $\times\times\times$ [1].

The important classification indicators (Table 9) indicate they are for the well aligned with the previous research on student (self-)reflection. Both Observation and Motive showed the strong use of past-oriented words, which is not surprising given that both categories relate to the descriptions of previous events (i.e., their past

[1] first_in_sentence was coded as: Yes=2, No=1

performance currently watched) and already identified by Ullmann [64] and Ullmann [63]. However, it is likely that the use of video recordings as a media to facilitate reflection had an impact on the use of perceptual words. If students were reflecting based on their memories of the past events, it is likely that they will use less perceptual words. As such, it seems important to provide students with not only instructional scaffolds, but also resources and materials for reflection in a format that will best promote (self-)reflection and critical thinking development.

We also see a strong use of words describing cognitive process of insight in the `Observation` and `Motive` categories. This is not surprising, given that reflection is one of the most effective approaches to fostering students' higher order thinking skills which is conditioned upon inquiry and insightful thinking [11]. The `Motive` category was also associated with higher intentional cohesion and causality, which is well aligned with the properties of the Motive-Effect type of reflective statements that capture student intentions and outcomes of particular actions. Our results also provide more detailed insights into the particular syntactic structures used to express motives and effects of student actions. We see that `Motive` category is associated with more agentless passive statements, higher use of connectives, higher ratio of causal particles to causal verbs, and higher complexity of verb phrases. In contrast, the `Goal` category was characterized by a higher use of causal verbs and a more formal, logical, and hierarchical thinking processes. This implies that `Goal` statements were generally expressed using causal, yet simpler linguistic structures (active language, simple causal statements), whereas the `Motive` category was characterized by a more complex language (i.e., more complex verb phrases, more passive expressions, more causal particles). Finally, we also see a unique profile of non-reflective statements (the `Other` category), which were characterized by a higher focus on personal topics and less driven by the perceptual processes. We also see a more frequent use of punctuation, which is likely caused by the use of emoticons in the non-reflective messages. Interestingly, on a linguistic level, we found a higher use of nouns, which requires a further study that we will conduct in our future work.

7 CONCLUSIONS

The contributions of this paper are twofold. First, we developed a classification system for categorization of students' reflections in accordance with the coding scheme by Hulsman et al. [35] which provides a moderate accuracy (accuracy of 89% and Cohen's κ of .51) over the chance level. The use of LIWC and Coh-Metrix features shows a great potential for understanding students' reflective writings, which are based on well-established linguistic metrics of different psychological processes. Second, our study provides a detailed evaluation of the linguistic indicators of the different types of student reflection. Interestingly, the most significant predictor was the use of perceptual words (e.g., seen, view, saw) and the complexity of causal expressions (i.e., the ratio of causal particles to the causal verbs). We also found that basic N-gram features provided less value than highly theorized linguistic metrics from LIWC and Coh-Metrics analysis tools. Finally, our results also showed some benefits of utilizing the reflection context, which was in our case captured by a single variable that indicated the relative position of the statement in a sentence. As such, in our future work, we will focus on providing more a detailed operationalization of the annotation context. We will also examine the use of the system for provision of the real-time feedback to students, which is one of the most promising uses of learning analytics [28].

REFERENCES

[1] ×××. 2014. Blinded for peer review.
[2] ×××. 2016. Blinded for peer review.
[3] ×××. 2017. Blinded for peer review. (2017).
[4] ×××. 2017. Blinded for peer review. (2017).
[5] ×××. 2017. *Blinded for peer review.* Ph.D. Dissertation.
[6] ×××. 2017. Blinded for peer review.
[7] Laura K. Allen, Erica L. Snow, and Danielle S. McNamara. 2014. The long and winding road: Investigating the differential writing patterns of high and low skilled writers. In *Proceedings of the 7th International Conference on Educational Data Mining*. http://educationaldatamining.org/EDM2014/uploads/procs2014/short%20papers/304_EDM-2014-Short.pdf
[8] John Bain, Roy Ballantyne, Jan Packer, and Colleen Mills. 1999. Using Journal Writing to Enhance Student Teachers' Reflectivity During Field Experience Placements. *Teachers and Teaching* 5, 1 (1999), 51–73. https://doi.org/10.1080/1354060990050104
[9] Steven Bird. 2006. NLTK: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*. Association for Computational Linguistics, 69–72. https://doi.org/10.3115/1118108.1118117
[10] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* 3 (2003), 993–1022.
[11] David Boud, Rosemary Keogh, and David Walker. 2013. *Reflection: Turning Experience Into Learning*. Routledge, London, New York.
[12] Leo Breiman. 2001. Random Forests. *Machine Learning* 45, 1 (2001), 5–32. https://doi.org/10.1023/A:1010933404324
[13] Chumphol Bunkhumpornpat, Krung Sinapiromsaran, and Chidchanok Lursinsap. 2009. Safe-Level-SMOTE: Safe-Level-Synthetic Minority Over-Sampling TEchnique for Handling the Class Imbalanced Problem. In *Advances in Knowledge Discovery and Data Mining (Lecture Notes in Computer Science)*. Springer, Berlin, Heidelberg, 475–482. https://doi.org/10.1007/978-3-642-01307-2_43
[14] Gemma Carey, Scott Harrison, and Rachael Dwyer. 2017. Encouraging reflective practice in conservatoire students: a pathway to autonomous learning? *Music Education Research* 19, 1 (2017), 99–110. https://doi.org/10.1080/14613808.2016.1238060
[15] C.-C. Chang, C.-C Chen, and Y.-H. Chen. 2012. Reflective behaviors under a web-based portfolio assessment environment for high school students in a computer course. *Computers & Education* 58, 1 (2012), 459–469.
[16] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* (2002), 321–357. https://www.jair.org/media/953/live-953-2037-jair.pdf
[17] Gary Cheng. 2017. Towards an automatic classification system for supporting the development of critical reflective skills in L2 learning. *Australasian Journal of Educational Technology* 33, 4 (2017). https://doi.org/10.14742/ajet.3029
[18] Stephen Corich, Kinshuk Hunt, and Lynn Hunt. 2012. Computerised Content Analysis for Measuring Critical Thinking within Discussion Forums. *Journal of e-Learning and Knowledge Society* 2, 1 (2012). http://www.je-lks.org/ojs/index.php/Je-LKS_EN/article/view/700
[19] Debra Coulson and Marina Harvey. 2013. Scaffolding student reflection for experience-based learning: a framework. *Teaching in Higher Education* 18, 4 (2013), 401–413. https://doi.org/10.1080/13562517.2012.752726
[20] Scott Crossley, Rod Roscoe, and Danielle S. McNamara. 2014. What Is Successful Writing? An Investigation into the Multiple Ways Writers Can Write Successful Essays. *Written Communication* 31, 2 (2014), 184–214. https://doi.org/10.1177/0741088314526354
[21] Ryan Daniel. 2001. Self-assessment in performance. *British Journal of Music Education* 18, 03 (2001). https://doi.org/10.1017/S0265051701000316
[22] John Dewey. 1933. *How we think. A restatement of the relation of reflective thinking to the educative process.* D. C. Heath, Boston, MA.
[23] Nia Dowell, Oleksandra Skrypnyk, Srećko Joksimović, Arthur C. Graesser, Shane Dawson, Dragan Gašević, Pieter de Vries, Thieme Hennis, and Vitomir Kovanović. 2015. Modeling Learners' Social Centrality and Performance through Language and Discourse. In *Proceedings of the 8th International Conference on Educational Data Mining (EDM 2015)*. Madrid, Spain. http://www.educationaldatamining.org/EDM2015/proceedings/full250-257.pdf
[24] Nia M.M. Dowell, Christopher Brooks, Vitomir Kovanović, Srećko Joksimović, and Dragan Gašević. 2017. The Changing patterns of MOOC discourse. In *Proceedings of the Fourth (2017) ACM Conference on Learning @ Scale (L@S '17)*. ACM, New York, NY, 283–286. https://doi.org/10.1145/3051457.3054005

[25] Ingo Feinerer, Kurt Hornik, and Mike Wallace. 2013. Package 'wordnet'. (2013). ftp://140.247.115.226/CRAN/web/packages/wordnet/wordnet.pdf

[26] Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinani Amorim. 2014. Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? *Journal of Machine Learning Research* 15 (2014), 3133–3181. http://jmlr.org/papers/v15/delgado14a.html

[27] Max Kuhn. Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, and Can Candan. 2015. *caret: Classification and Regression Training*. http://CRAN.R-project.org/package=caret R package version 6.0-58.

[28] Dragan Gašević, Shane Dawson, and George Siemens. 2015. Let's not forget: Learning analytics are about learning. *TechTrends* 59, 1 (2015), 64–71. https://doi.org/10.1007/s11528-014-0822-x

[29] Andrew Gibson and Kirsty Kitto. 2015. Analysing Reflective Text for Learning Analytics: An Approach Using Anomaly Recontextualisation. In *Proceedings of the Fifth International Conference on Learning Analytics and Knowledge (LAK '15)*. ACM, New York, NY, USA, 275–279. https://doi.org/10.1145/2723576.2723635

[30] Andrew Gibson, Kirsty Kitto, and Peter Bruza. 2016. Towards the Discovery of Learner Metacognition From Reflective Writing. *Journal of Learning Analytics* 3, 2 (2016), 22–36. https://doi.org/10.18608/jla.2016.32.3

[31] Arthur C. Graesser, Danielle S. McNamara, and Jonna M. Kulikowich. 2011. Coh-Metrix Providing Multilevel Analyses of Text Characteristics. *Educational Researcher* 40, 5 (2011), 223–234. https://doi.org/10.3102/0013189X11413260

[32] Trevor J Hastie, Robert J , Tibshirani, and Jerome H , Friedman. 2013. *The elements of statistical learning*. Springer.

[33] N. Hatton and D. Smith. 1995. Reflection in teacher education: Towards definition and implementation. *Teaching and Teacher Education* 11, 1 (1995), 33–49.

[34] Eleanor Hawe and Helen Dixon. 2016. Assessment for learning: a catalyst for student self-regulation. *Assessment & Evaluation in Higher Education* (2016), 1–12. https://doi.org/10.1080/02602938.2016.1236360

[35] R. L. Hulsman, A. B. Harmsen, and M. Fabriek. 2009. Reflective teaching of medical communication skills with DiViDU: Assessing the level of student reflection on recorded consultations with simulated patients. *Patient Education and Counseling* 74, 2 (2009), 142–149. https://doi.org/10.1016/j.pec.2008.10.009

[36] Robert L. Hulsman and Jane van der Vloodt. 2015. Self-evaluation and peer-feedback of medical students' communication skills using a web-based video annotation system. Exploring content and specificity. *Patient Education and Counseling* 98, 3 (2015), 356–363. https://doi.org/10.1016/j.pec.2014.11.007

[37] Srećko Joksimović, Nia Dowell, Oleksandra Skrypnyk, Vitomir Kovanović, Dragan Gašević, Shane Dawson, and Arthur C. Graesser. 2015. Exploring the Accumulation of Social Capital in cMOOC Through Language and Discourse. *Submitted* (2015).

[38] Srećko Joksimović, Nia Dowell, Oleksandra Skrypnyk, Vitomir Kovanović, Dragan Gašević, Shane Dawson, and Arthur C. Graesser. 2015. How Do You Connect?: Analysis of Social Capital Accumulation in Connectivist MOOCs. In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge (LAK '15)*. ACM, New York, NY, USA, 64–68. https://doi.org/10.1145/2723576.2723604

[39] Srećko Joksimović, Vitomir Kovanović, Jelena Jovanović, Amal Zouaq, Dragan Gašević, and Marek Hatala. 2015. What Do cMOOC Participants Talk About in Social Media?: A Topic Analysis of Discourse in a cMOOC. In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*. Poughkeepsie, NY, 156–165. https://doi.org/10.1145/2723576.2723609

[40] Anne Kao and Stephen R Poteet. 2007. *Natural language processing and text mining*. Springer, London.

[41] Klaus H. Krippendorff. 2003. *Content Analysis: An Introduction to Its Methodology*. Sage Publications, Thousand Oaks, CA.

[42] J. R. Landis and G. G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33, 1 (1977), 159–174. https://doi.org/10.2307/2529310

[43] Äli Leijen, Ineke Lam, Liesbeth Wildschut, P. Robert-Jan Simons, and Wilfried Admiraal. 2009. Streaming video to enhance students' reflection in dance education. *Computers & Education* 52, 1 (2009), 169–176. https://doi.org/10.1016/j.compedu.2008.07.010

[44] Andy Liaw and Matthew Wiener. 2002. Classification and Regression by randomForest. *R News* 2, 3 (2002), 18–22. http://CRAN.R-project.org/doc/Rnews/

[45] Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. 55–60.

[46] Tom McKlin. 2004. *Analyzing Cognitive Presence in Online Courses Using an Artificial Neural Network*. Ph.D. Dissertation. Georgia State University, College of Education, Atlanta, GA, United States.

[47] Tom McKlin, SW Harmon, William Evans, and MG Jones. 2002. Cognitive presence in web-based learning: A content analysis of students' online discussions. In *IT Forum*, Vol. 60.

[48] Danielle S. McNamara, Scott Crossley, and Philip M. McCarthy. 2009. Linguistic Features of Writing Quality. *Written Communication* (2009). https://doi.org/10.1177/0741088309351547

[49] Danielle S. McNamara, Arthur C. Graesser, Philip M. McCarthy, and Zhiqiang Cai. 2014. *Automated Evaluation of Text and Discourse with Coh-Metrix*. Cambridge University Press.

[50] Negin Mirriahi and Shane Dawson. 2013. The Pairing of Lecture Recording Data with Assessment Scores: A Method of Discovering Pedagogical Impact. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge (LAK '13)*. ACM, New York, NY, USA, 180–184. https://doi.org/10.1145/2460296.2460331

[51] Thais Mayumi Oshiro, Pedro Santoro Perez, and José Augusto Baranauskas. 2012. How Many Trees in a Random Forest?. In *Machine Learning and Data Mining in Pattern Recognition (Lecture Notes in Computer Science)*. Springer, Berlin, Heidelberg, 154–168. https://doi.org/10.1007/978-3-642-31537-4_13

[52] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12, Oct (2011), 2825–2830. http://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf

[53] J. W. Pennebaker, C. K. Chung, M. Ireland, A. Gonzales, and R. J. Booth. 2007. *The development and psychometric properties of LIWC2007*. http://hdl.handle.net/2152/31333

[54] Anand Rajaraman and Jeffrey David Ullman. 2011. *Mining of Massive Datasets*. Cambridge University Press, Cambridge, UK.

[55] E. F. Risko, T. Foulsham, S. Dawson, and A. Kingstone. 2013. The Collaborative Lecture Annotation System (CLAS): A New TOOL for Distributed Learning. *IEEE Transactions on Learning Technologies* 6, 1 (2013), 4–13. https://doi.org/10.1109/TLT.2012.15

[56] D. D. Ross. 1989. First steps in developing a reflective approach. *Journal of Teacher Education* 40, 2 (1989), 22–30.

[57] Carolyn Rosé, Yi-Chia Wang, Yue Cui, Jaime Arguello, Karsten Stegmann, Armin Weinberger, and Frank Fischer. 2008. Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning. *International Journal of Computer-Supported Collaborative Learning* 3, 3 (2008), 237–271. https://doi.org/10.1007/s11412-007-9034-0

[58] Mary Ryan. 2013. The pedagogical balancing act: teaching reflection in higher education. *Teaching in Higher Education* 18, 2 (2013), 144–155. https://doi.org/10.1080/13562517.2012.694104

[59] Yla R. Tausczik and James W. Pennebaker. 2010. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology* 29, 1 (2010), 24–54. https://doi.org/10.1177/0261927X09351676

[60] Abinash Tripathy, Ankit Agrawal, and Santanu Kumar Rath. 2016. Classification of sentiment reviews using n-gram machine learning approach. *Expert Systems with Applications* 57 (2016), 117–126. https://doi.org/10.1016/j.eswa.2016.03.028

[61] Thomas Daniel Ullmann. 2011. An Architecture for the Automated Detection of Textual Indicators of Reflection. In *1st European Workshop on Awareness and Reflection in Learning Networks held in conjunction with the 6th European Conference on Technology Enhanced Learning: Towards Ubiquitous Learning 2011*. http://oro.open.ac.uk/33406/

[62] Thomas Daniel Ullmann. 2015. *Automated detection of reflection in texts. A machine learning based approach*. PhD thesis. The Open University.

[63] Thomas Daniel Ullmann. 2015. Keywords of written reflection-a comparison between reflective and descriptive datasets. In *Proceedings of the 5th Workshop on Awareness and Reflection in Technology Enhanced Learning*. 83–96. http://oro.open.ac.uk/id/eprint/44590

[64] Thomas Daniel Ullmann. 2017. Reflective Writing Analytics: Empirically Determined Keywords of Written Reflection. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference (LAK '17)*. ACM, New York, NY, USA, 163–167. https://doi.org/10.1145/3027385.3027394

[65] Thomas Daniel Ullmann, Fridolin Wild, and Peter Scott. 2013. Comparing automatically detected reflective texts with human judgements. In *2nd Workshop on Awareness and Reflection in Technology- Enhanced Learning*. 101–116. http://oro.open.ac.uk/id/eprint/37830

[66] Thomas Daniel Ullmann, Fridolin Wild, and Peter Scott. 2013. Reflection - quantifying a rare good. In *Proceedings of the 3rd Workshop on Awareness and Reflection in Technology Enhanced Learning*. 29–40. http://oro.open.ac.uk/id/eprint/40065

[67] Zak Waters, Vitomir Kovanović, Kirsty Kitto, and Dragan Gašević. 2015. Structure matters: Adoption of structured classification approach in the context of cognitive presence classification. In *Information Retrieval Technology*, Guido Zuccon, Shlomo Geva, Hideo Joho, Falk Scholer, Aixin Sun, and Peng Zhang (Eds.). Number 9460 in Lecture Notes in Computer Science. Springer International Publishing, 227–238. https://link.springer.com/chapter/10.1007/978-3-319-28940-3_18

[68] Jaebong Yoo and Jihie Kim. 2013. Can Online Discussion Participation Predict Group Project Performance? Investigating the Roles of Linguistic Features and Participation Patterns. *International Journal of Artificial Intelligence in Education* 24, 1 (2013), 8–32. https://doi.org/10.1007/s40593-013-0010-8

[69] Barry J. Zimmerman. 2002. Becoming a Self-Regulated Learner: An Overview. *Theory Into Practice* 41, 2 (2002), 64–70.