Evaluating an Automated Analysis using Machine Learning and Natural Language Processing Approaches to Classify Computer Science Students' Reflective Writing

*Author1[1], Author2 [1], Author3[1], and Author4[2]
[1]Address 1
[1*]Email id
[2]Address 2
[2]Email id

* Corresponding Author

Abstract. Reflection writing is a common practice in higher education. However, manual analysis of written reflections is time-consuming. This study presents an automated analysis of reflective writing to analyze reflective writing in CS education based on conceptual Reflective Writing Framework (RWF) and application of natural language processing and machine learning algorithm. This paper investigates two groups of features extraction (n-grams and PoS n-grams) and random forest (RF) algorithm that utilize such features to detect the presence or absence of the seven indicators (description of an experience, understandings, feelings, reasoning, perspective, new learning, and future action). The automated analysis of reflective writing is evaluated based on 74 CS student essays (1113 sentences) that are from the final year project reports in CS's students. Results showed the seven indicators can be reliably distinguished by their features and these indicators can be used in an automated reflective writing analysis for determining the level of students' reflective writing. Finally, we consider the implications of how the conceptualization of refection quality and providing individualized learning support to students in order to help them develop reflective skills.

Keywords: Reflection Assessment, Machine Learning, Natural Language Processing, Reflective Writing, Refection, Computer Science education

1 Introduction

Reflection has been used in higher education to support students to become thoughtful practitioners by enabling them to extract knowledge from their experiences [1-3] and can support metacognition [1]. The disciplines of higher

education have been including reflection in some of their programs, for instance, teachers' pre-service training [4], medicine [5], management [6], and Computer Science (CS) [7-9]. The focus of this research is on reflective writing in CS education. In terms of this, Fekete [9] stated that 'reflection is worth encouraging, for its indirect effect on the technical skills and knowledge which are our ultimate purpose in teaching Computer Science' (p.144). Reflection improves students' awareness of how to learn from situations, e.g., how to deal with a sequence of steps required to reach a certain goal or how to identify the roots of a problem rather than concentrate on their feelings about the problem [10].

Despite the widespread use of reflective writing approaches for assessment, leading to the support of personalised learning [11], assessing reflective writing remains a challenge [12-16]. Reflection assessment is labour-intensive when manual content analysis [17] is applied to the task. Such assessment is employed to understand how students reflect and to support their reflective practice [18]. Of course, the fact that such assessment is so labour-intensive has led to the idea that automated approaches might potentially have a role in achieving it.

Automated assessment methods which assess writing based on evaluating its reflective content have generally used natural language processing (NLP) [11, 12, 14, 16, 19] to automate the utilization of a reflection framework. This research aims to explore the automated assessment of reflective writing for CS education. This field represents a significant challenge, due to the limited research which has taken place with respect to automated methods for analysing reflective writing.

Little research has been undertaken on the automated assessment of reflective writing. This present research aims to evaluate a system that undertakes the automatic assessment of reflective writing for CS education using advanced methods of natural language processing (NLP). This paper aims to (a) determine empirically each indicator of reflective writing features by examining different linguistic groups (unigram (word), bigram (word), trigram (word), unigram (PoS), bigram (PoS) and trigram (PoS)) and (b) build and evaluate a machine learning approach for binary classification for reflective writing in CS. The findings shed light on the structure of CS students' reflections and the first attempt to develop an automated reflective writing assessment using advanced NLP and machine learning techniques to allow personalized reflection. The research question for this study is as follows**:** What are the linguistic features which can be found in CS students' reflective writings which indicate the presence of each of the reflection indicators?

2 Literature review

2.1 The Importance of Reflection in CS Education

Reflection is commonly described as evidence of understanding of one's experiences of the situation used to take action in the future [2, 20]. In CS education, various activities necessitate the application of variations on the common reflection processes, such as judgment, evaluation, reasoning, problem-solving, and memorizing [10, 21-23].
Chng study indicated that it is necessary to teach problem-solving and reasoning skills in the course of CS education to improve students' awareness of how to learn from a situation they are presented with [10]. Hazzan & Tomayko also showed the importance of reflection in CS to support the student in the complexity in the development of software systems, which requires the developer to improve their understanding of their mental processes [23], and as this can be achieved by applying a reflection approach, it teaches developers how to think effectively. For these reasons, reflective writing is important in CS education.

2.2 Methods to Analyze Reflection

The reflective writing frameworks can develop different indicators in different fields (see appendix 1), such as medical education or teacher preparation [24-26], and along similar lines [27-35].

There is variation in the ranges of characteristics covered by each indicator of each framework, with some frameworks tending to combine two or more indicators into one [36] or to divide what is generally one indicator into multiple sub-indicators [25, 35]. For example, Moallem's framework focuses on the writers/students' perspectives by using indicators such as explore; imagine alternatives; and gain exposure (to a variety of interpretive considerations in dialogue with others). On the other hand, Mamede and Schmid [36] use only one of their indicators to represent perspective.
In terms of CS, [37] proposed seven indicators (description of an experience, understandings, feelings, reasoning, perspective, new learning, and future action) of the conceptual RWF and the framework was empirically evaluated [38]. These indicators are described in the following section.
The manual analysis of refection is time-consuming [18]. Of course, the fact that such assessment is so labour-intensive has led to an interest in using advanced methods of analyzing reflective writing [11, 12, 14, 16].

2.3 Automatic Method to Analyze the Reflection

The existing approaches to automatic reflection analysis can be classified into keyword-based and machine learning-based approaches [11, 12, 39]. The

keyword-based methods depend on locating specific keywords in the input text as indicators of reflection, using a keyword matching process. The presence/absence or frequency of the keywords can be used to analyse input text using the keyword-based approach [15, 40, 41]. Further, machine learning-based frameworks use existing classification algorithms to find patterns associated with each indicator at the training stage and then classify 'unseen' input texts using these mined patterns [11, 12, 14, 16].

Ullmann proposed a data-driven keyword-based technique for automatic reflective writing classification [11, 12, 39]. The datasets used were constructed from the British Academic Writing English (BAWE) corpus — from the health, engineering, and business fields. This majority voting system raised the reliability yielded by Cohen's κ from 62% to 92%.

Lin et al. used a Chinese version of Linguistic Inquiry and Word Count (LIWC) [41], a text processing tool that characterizes the words that are utilized in association with different psychological and cognitive processes [42] —to identify the functions of the words used in reflective writings. The results showed that the words designated as indicators were the ones most frequently in the analyzed narrative.

Ullmann proposed a data-driven keyword-based technique for automatic reflective writing classification which identified various indicators. Eight datasets were constructed using the BAWE corpus [11, 12, 14, 16]. The sentences were processed with the same annotation procedure using majority voting as given by Ullmann. However, Ullmann differs in the way in which keywords were derived for a framework of reflection that contains various indicators: reflection, experience, feelings, belief, difficulty, perspective, learning, and intention. The approach used for determining the keywords was based on log-likelihood, as discussed in [43].

Cui et al. proposed a framework, based on the LIWC list, for identifying the most important words and phrases in terms of identifying each indicator in the proposed reflection framework[15]. A total of 27 dental students' reflections (using six reflective statement types) over four years were employed to identify the required features [15].

The LIWC dictionaries are not specific to one word that means a word can to more than one group, such as the word 'died' appears in several categories in past tense words, verbs, and death. Chung & Pennebaker pointed out that 'NLP approaches will outperform LIWC on many classification tasks' in comparison to machine learning algorithms [42]. A recent study by Liu et al. stated that 'the use of only LIWC emotional features is insufficient to detect the depth of the Feeling Factor' (p.12) [44].

3 Methodology

3.1 Dataset

The dataset consists of sample texts that were collected from the projects of CS undergraduate students in the UK university. The dataset used in this study – of 174 different reflective writing documents – was employed for use by coders for the annotation. The data were collected from 174 third- and fourth-year CS student projects. These had been undertaken in the course of the academic years 2013 through 2016 at the author's university and anonymized before analysis.

The unit of analysis was taken sentence of reflection as indicators can occur across sentences as evidence of reflection [45]. The dataset was coded by three coders until the reliability was stable at an acceptable agreement for three coders; using Cohen's κ for each of the seven binary indicators (presence or absent); as suggested by [46]: poor agreement results are indicated by a Cohen's κ below 0, slight agreement results are in the range [0-0.2], fair agreement results are in the range [0.21-0.4], moderate agreement results are in the range [0.41-0.6], substantial agreement results are in the range [0.61-0.8] and almost perfect results are represented by a Cohen's κ of above 0.8. The agreement between coders was calculated as follows: when all the coders agreed on the same sentence this was considered agreement, but when even just one coder did not agree on a particular sentence, this was classified as disagreement.

Applying the conceptual RWF of 1113 sentences for the CS dataset, the agreement between coders was calculated as the seven indicators achieved kappa statistic values of between 0.46 to 0.75, and this range means moderate to the substantial agreement [46].

3.2 Manually Analysis of Reflection' Indicators (Content Analysis)

Reflection's indicators are used to assess the presence of each reflection's indicator based on the framework of Alrashidi et al. [37, 38]. The conceptual RWF is described in detail with text examples.

The Description of an experience indicator occurs when the writer describes the experience with no interpretation.

An example of reflective writing in CS is when the writer constructs superficial descriptions of situations. For example, 'It was my responsibility to ensure that all aspects of the project were progressing as expected and ensure that everyone was working towards the overall goal.'

The understanding indicator is encountered when attempts are made to reach an understanding of a concept or topic and/or an understanding related to personal experience.

For example, from the CS dataset, when the students show an abstract understanding of situations; 'Until this year, I had no understanding of many image processing technologies, such as background subtractors and camera

calibration'

The Feelings indicator occurs when the writer has identified their thoughts, feelings, and/or behaviours.

For example, from the CS dataset, when the students show evidence of the reflective level indicator of feelings, in terms of emotions and thoughts; as the following sentence,

'Group meetings had a good atmosphere and members of the group felt comfortable sharing and discussing their ideas and proposals'

The Reasoning indicator emerges when in-depth analysis is made which leads to a significant conclusion – i.e., a deeper understanding of the experience. The reasoning indicator signifies that the writer has made an effort to explain the experience in question.

For example, from the CS dataset, where students show that they recognized issues by providing explanations and/or excuses for what happened; the examples of sentences may contain reasoning indicators.

'As a result of this, I focused more on the design of the individual system components, and in discussing as well as researching the theoretical foundations behind the fleet scheduling problems'

The Perspective indicator occurs when the writer shows awareness of alternative perspectives. For example, from the CS dataset, where the students show awareness of their and/or others' perspectives; these examples of sentences may contain reflection indicators other than this particular one.

'Given how I had completed my share of the work and there was not enough time to finish and polish our software due to the slow progress of the work from some of the team members, I decided to offer my time and skills towards their tasks, also, because the overall outcome of the project would have affected me too.'

The New Learning indicator occurs when the writer describes what they have learned from experience.

For example, from the CS dataset, where the students describe what has been learned; these examples of sentences may contain reflection indicators other than this particular one.

'I have developed skills in using new tools, such as the MIRToolbox [4] extension to MATLAB, and have gained knowledge of various technical methods such as Principal Component Analysis and how different clustering algorithms work.'

The Future action indicator suggests that the writer would, given the same circumstances again, intentionally do something differently or that they would plan their actions based on the new understanding that has resulted from considering and reviewing the original experience.

Here are examples, from the CS dataset, where students show an awareness of the outcome of their action and a consequent change in perception.

'If I had been in a leadership role, I would have had to greatly improve my leadership skills in terms of organising others and being assertive, as I believe that, currently, if were assigned the leader role I would not be able to lead a team

effectively.'

unit of analysis was taken sentence of reflection as indicators can occur across sentences as evidence of reflection [45]. The dataset was coded by three coders until the reliability was stable at an acceptable agreement for three coders; using Cohen's κ for each of the seven binary indicators (presence or absent); as suggested by [46] poor agreement results are indicated by a Cohen's κ below 0, slight agreement results are in the range [0-0.2], fair agreement results are in the range [0.21-0.4], moderate agreement results are in the range [0.41-0.6], substantial agreement results are in the range [0.61-0.8] and almost perfect results are represented by a Cohen's κ of above 0.8. The agreement between coders was calculated as follows: when all the coders agreed on the same sentence this was considered agreement, but when even just one coder did not agree on a particular sentence, this was classified as disagreement.

Applying the conceptual RWF of 1113 sentences for the CS dataset, the agreement between coders was calculated as the seven indicators achieved kappa statistic values of between 0.46 to 0.75, and this range means moderate to the substantial agreement [46].

3.3 Proposed Framework

The proposed automated RWF applying, the n-gram is used as it can encompass the features that are commonly used in NLP which are particularly relevant for the classification of the seven indicators. According to Jurafsky and Martin [47], the n-gram model supports the processing of important kinds of features commonly encountered in speech and language processing in general [48]. However, one of the limitations of the n-gram model is the fact that the method is ignorant of the grammatical nature of the text. Recently, a Part of Speech (PoS) n-gram model was used extensively for text classification [49, 50].

A recent study by Liu et al. [44] stated that 'further writing analytics development is needed, particularly for the feelings factor, which does not appear to be well covered by the variables we have considered here – what writing analytic features might prove to be more reliable indicators of feelings if they are denoted as important to the quality of a reflection?'(p.12). in order to capture the feelings indicator, there is a need to capture the linguistic feature as the word and phrase as well the grammatical combination that exhibited in texts, n-grams (uni-gram, bi-gram and tri-gram) and PoS n-grams (uni-gram, bi-gram and tri-gram) are often utilized; this is because these methods, together, exhibit diversity in capturing the associations of the related indicators (e.g., feeling, perspective).

In figure 1, to extract the relevant features from the input text, after the text has been preprocessed, the text is tagged with its part-of-speech. Then, a set of features are extracted that allows for the analysis of the influence of these features on the indicator classification and hence the results of the automated RWF processing. Two groups of features are extracted, which are, n-grams (uni-gram, bi-gram and tri-gram) and PoS n-grams (uni-gram, bi-gram and tri-gram). In table

1, illustrates the classification steps as input text until measuring the automated classification vs the manual coding.

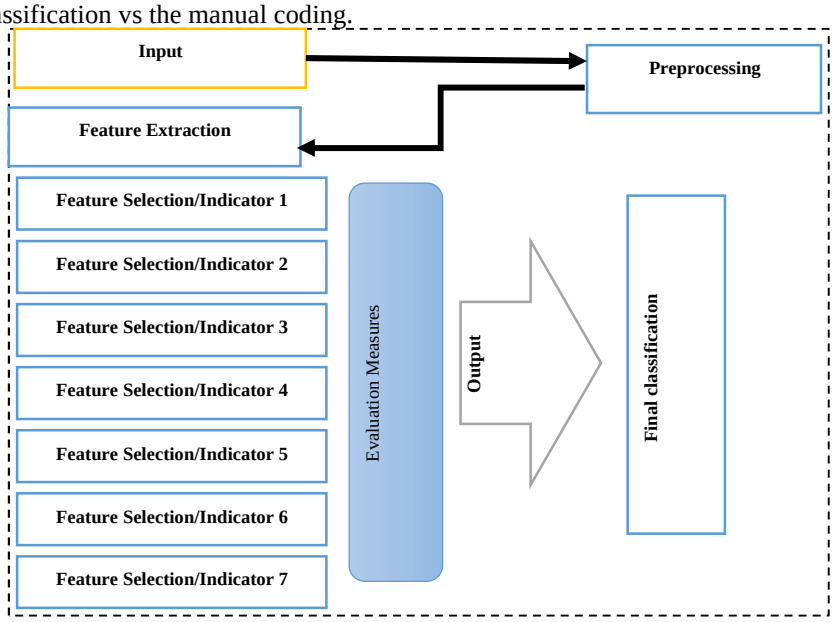

Figure 1: The automated reflective writing framework

Table.1. The classification steps

Component	*Description*
Inputs	Text, (coding the seven indicators for training-only)
Ground truth	Manual coding
Measurements	Accuracy and Cohen's κ
Comparison	Performance of classifier (manual coding vs automated coding)

Random forest is an ensemble classification technique that provides low-bias, low-variance performance. it also allows for feature inspection by building multiple decision trees using random subsets of features on bootstrapped samples [51]. Previous research has employed the RF algorithm as the best available for classification — as compared to the rest of the algorithms [11, 14, 16]. This study applies the RF algorithm.

Cohen's κ is often used concerning the manual annotation in the educational area to measure IRR (as between human assessors), while the F-measure and accuracy are the most common measures used in automatic reflective writing assessment systems [16]— in order to assess the performance of the machine learning algorithms applied.

In this study the automated RWF, a binary classification process is

implemented for each of the predetermined indicators. For each indicator, the inputs are classified into the classes, 0 or 1 (absence/presence of the indicator in the text) in the breadth-based form of the automated RWF task. For each input text, a feature vector based on the extracted and reduced features is created.

4 Results

Feature selection produces a set of features that can, potentially, be selected — one for each indicator. This selected set of features is sub-grouped based on the feature type. These are unigram (word), bigram (word), trigram (word), unigram (PoS), bigram (PoS) and trigram (PoS).

4.1 Description of an experience

The description of an experience indicator is often positively associated with some PoS type features (unigram and bigrams); Using analysis of the dataset, it was discovered that the first person pronouns ‘I’, ‘my’, ‘me’, and the third-person pronoun ‘it’ have a positive effect in terms of this indicator. This means that the presence of first-person pronouns and third-person pronouns is associated with inputs that include the text of a purely descriptive nature. Similarly, Birney [52] and Ullmann [11] reported that the first and third-person pronouns can be found in the description of an experience category. Accordingly, ‘have’, ‘has’, and ‘had’ can be seen to be associated with this indicator, in the present, past, and perfect forms. This finding is congruent with those of the researches of Birney [52], Ullmann [11], Ryan [15], and Jung and Wise [16]; these all found a link between the past tense form and the description of an experience indicator.

4.2 Understanding

In the dataset, the word ‘expect’ was appeared only in the non-understanding examples (i.e., text items that were as not complying with the understanding indicator by the expert), such as ‘this ended up being a lot of work which I was not expecting due to the number of modules’, ‘…would work as expected and performed the required role’. and ‘I learned how to deal with multiple deadlines in a way where you can achieve the best expected results’. Besides, the understanding indicator is characterized negatively by phrases such as ‘it would’, ‘for example’, and ‘gained further knowledge’. In the dataset, there are text items involving such phrases which can be characterized as non-understanding, such as, ‘I also gained further knowledge in people skills; due to the situation’, and ‘For example, instead of a plain jar file’. Finally, the understanding indicator is characterized positively by the phrase ‘in conclusion’. This appears in examples of text complying with the understanding indicator in the dataset, such as ‘In conclusion, I have gained much

from this project,…'.

4.3 Feelings

Using analysis of the dataset, it was discovered that the bigrams of words (there being 353 of these) form the majority of the complete list of features. In particular, there are the bigrams which include the sensing and thinking verbs that refer to mental processes, e.g., 'I think', ' I feel', 'I believe', 'I find', I assume' and 'I realise' — in their present tense and past tense forms. There are many instances of words/bigrams in the dataset which referred to negative feelings: such as 'struggle', 'difficulty', 'negative', 'unfortunately', 'quite difficult', and 'conflict'. However, there were also quite a few words/phrases which referred to positive feelings: such as 'satisfy', 'excellent', 'good', 'appreciation', 'happy', and 'proud'. Such intuitions, which are often encapsulated in expressions linked to feeling or thinking, can be a justification for considering that a particular text is reflecting on something in order to gain greater clarity.

The subordinating conjunction 'that' is often used. It can also be used in sentences that answer the question 'why' and in these cases the phrase to which the 'that' refers can be used to determine the context of the problem being resolved whereas the preceding verb gives an indication of the mental state involved. Here are some examples from the dataset: word tri-grams 'I believe that', 'I felt that', 'so that we', and 'mean that you'. In addition, the PoS tri-grams, pronoun + verb + conjunction covers a great many of these cases.

Supporting evidence for the nature of these features as described above can be found in Birney [52], Ullmann [11], Ryan [53] and Cui et al. [15]. Feelings are linked to the use of pronouns: singular (I, my, me), plural (we) (as stated by [11]) and sensing and thinking verbs (as stated by [52] and [53]).

4.4 Reasoning

From the dataset, it was discovered that different bigrams of words also included in these selected features have different effects: for example, the bigram 'because of' has a negative effect in terms of identifying input with a reasoning label, while the bigram 'role within' has a positive impact. In the dataset, there are some items which include the phrase 'role within' that are classified positively concerning the reasoning label, such as 'adopting a different role within the group would have led to a fundamentally different position for myself,'. Other items (other than the one given above) which include the phrase 'because of' are classified as negative for the reasoning indicator, such as '….this sprint cycle was hard to adhere to because of time being focused…..'.

The best indicators were words unigram and bigrams specifying the presence of causal links, such as 'hence', 'result', 'due to', 'the fact', 'as this', 'such as', 'as my', 'and hence', 'this cause', and 'result of'. These words and phrases evidenced

that students were using causal links in their reflective writing, so adding explanation (Birney [52]; Ryan [53]).

These results are consistent with the literature, that is with Birney [52], Ullmann [11], Ryan [53] and Cui et al. [15] and Shum et al. [54]. In terms of the use of causal words in reflective writing, Birney [52]; Ryan [53] indicated that such causal words can help students to provide explanations of their actions. Ullmann [11] also showed empirically that automated analysis detected 'because' as a signal of the premise part of an argument. Shum et al. [54] indicated that the ability expressions evidence reflective writing that can describe the students' capabilities.

4.5 Perspective

From the dataset, the word 'question' is a feature that was selected concerning the perspective indicator. It has a positive influence in terms of identifying input text items that comply with the perspective indicator. This is the case also with the word 'would' and some phrases involving this: 'would also', and 'would have', among others. Adverbs in general also have a positive influence. In the dataset, there are examples of the use of such constructs as, 'If I could do anything differently, I would have chosen a clearer project title sooner, ….', '….Given my intense interest in this topic, I really hope I have the opportunity to go back and answer those questions', and 'Being able to discern helpful materials from the outset would have saved a considerable amount of time throughout the project'.

Expressions of ability may use modal verbs and phrases. The students are usually using the adjective 'able' to express their ability to do something. Employing temporal links such as 'now' in combination with expressions of ability can be used to express the acquisition or possession (now) of authority, skill, or other means of doing something. And expressions of ability can be found in association with this indicator, e.g., 'I was able', 'would be able', 'enable me to', 'I become', 'have allowed me', 'possible to', 'this ensures that', and 'I am able'. This implies that adjectives used in expressions of ability can also be found in association with this indicator.

Finally, this pattern of findings (as discussed above) is consistent, in technological terms, with Ullmann [11] results, and in theoretical terms, with Birney [52]; Ryan [53]. Ullmann [11], in particular, found that the automated detection of reflective writing concerning the perspective indicator showed evidence of the use of first- and third-person pronouns.

4.6 Future Action

From the dataset, auxiliary verbs such as 'would' can be used to talk about the past, and about the future in the past, or something desired but not actual at present, or about an imagined situation in general. Also, the word 'will' can be

used to describe something that is to take place in the future, though it can refer to the past or future depending on the context.

In the dataset, the texts 'In the future, I would definitely take a more proactive role in ensuring the health of the group as a whole' and 'If I were to complete this research project again ……' can have a positive future action label assigned to them based on the words and phrases.

These findings relating to the future action indicator are aligned with those of Birney [52], Ullmann [11], Cui et al. [15] and Jung and Wise [16]. These all found that the use of the future tense and first-person pronouns evidence a consideration, by the student, of their future actions as moderated/modified by the lessons gained from the experience being described.

4.7 New learning

From the dataset, The verb 'learn' can be used to refer to what the student has learned from an experience. However, the word was found in different tenses and parts of speech forms, such as 'I have learned', 'the learning outcome', 'managed to learn', 'author has learned', 'lesson I learned' 'was learning through', and 'learned about java'. Additionally, the phrases, 'teach me' and 'taught me' can be used to refer to gain in knowledge.

The first-person singular pronouns, 'I', 'my', and 'me' had a positive influence on the detection of the learning indicator, which means that the presence of the first-person pronouns is associated with inputs that can be labelled with the learning indicator. Similarly, Birney [52], Ullmann [11] and Jung and Wise [16] found empirical evidence that the use of first-person pronouns can provide evidence of learning.

Accordingly, it can be concluded that the learning indicator is evidenced by a great many different linguistic features. In contrast, other researchers, Birney [52], Ullmann [11], and Jung and Wise [16] only found a few linguistic features involved with this indicator.

Evaluation Results

Table 2. Performance of Random Forest Algorithm for each Indicator

Indicator	accuracy	Kappa
Description of an experience	0.80	0.43
Understanding	0.84	0.17
Feelings	0.75	0.51
Reasoning	0.75	0.51
Perspective	0.85	0.35
Future action	0.96	0.53
New learning	0.93	0.67

As given in Table 2, using random forest can improve the performance of n-grams and POS-n-grams features that showed straightforwardly to be effective in various text classification models. The classification model for reflective indicators showed slight to substantial performance (κ = 0.17 to 0.67). Future action and New learning classifiers had the highest accuracy (0.96 and 0.93) and the best performance (κ = 0.67 and 0.53), while Feelings and Reasoning classifiers performed moderate kappa by 0.51 aligned with the accuracy performance ($\geqq$ 0.75). the Perspective and Understanding classifiers had a high accuracy above 0.80, but the kappa had slight to fair (0.17 and 0.35). Accordingly, combining different types of features in the proposed framework improves the results. Different features need to experiment with a different field and with different inputs.

5 Discussion

RQ: What are the linguistic features which can be found in CS students' reflective writings which indicate the presence of each of the findings reflection indicators?

The two types of linguistic features have been used for predicting/detecting the reflection indicators. These features were shown to be effective in predicting the presence or absence of the pre-determined indicators (i.e., description of an experience, understanding, feelings, reasoning, perspective, new learning, and future action).

Compared to the state-of-the-art, our automated RWF captured a wide range of features and obtained good results using features that had not previously been tested in relation to any similar task. [14] used three types of features, n-grams, LIWC, and Coh-matrix, [11] used uni-grams only while [16] used features that were extracted based on LIWC only.

It is worth noting that our findings relating to the indicators are well aligned with the theoretical and technical literature concerning reflective writing. These findings have shown that many different linguistic features are useful for detecting indicators and that therefore analyzing these features may lead to a greater understanding of the nature of the various levels of reflective writing and their characteristics. And these findings/features were indeed examined here in order to demonstrate the kind of value this kind of study can have in terms of highlighting potential areas for future investigation.

The features identified by the system for each indicator are generally different from those for any other indicator, but there is also some overlap of linguistic resources and terms with respect to the indicators. For example, the singular first-person pronoun is one of the top features for all the indicators. It was also found that verbs such as thinking and sensing were important for several indicators. The findings relating to first-person pronouns and the thinking and sensing words were investigated empirically as important features of reflection; this investigation was

based on the theoretical literature of Birney [52], and the empirical evaluation incorporated in the reflective writing and technical literature analyses of Ullmann [11] and Shum et al. [16]. These findings with respect to the thinking and sensing words and phrases suggest that such can be used to foster students' reflections which are conditional upon and require thinking [55]. Regarding the description of an experience indicator, the first-person pronoun and the third person pronoun are important for identifying this indicator. These pronouns can be used to describe who was involved in the experience.

Further, both the understanding and the feelings indicators exhibit similarities in terms of the use of linguistic features such as the subordinating conjunction, 'that', the auxiliary verbs (to be, have), adjectives, the past tense, and the thinking and sensing words (understand, think, feel, believe, relies on, etc.). These findings align theoretically with both Birney [52] and Ryan [54] who focused on manual analysis, and technically (in terms of features for machine learning) with Kovanović et al. [14], Ullmann [11], and Jung and Wise [16].

The set of linguistic features identified as important for predicting/detecting the reasoning indicator is a new and very useful result as no evidence concerning the linguistic features which are linked with this indicator has previously been provided in the literature. The features relevant to the reasoning indicator are found to overlap with those important for the other indicators. For instance, there is an overlapping in terms of the use of adverbs between the reasoning, description of an experience, perspective, and new learning indicators. Similarities were also found between the features important to the reasoning indicator and those used for the detection of the perspective indicator, both were strongly predicted via the adjective 'able' (i.e., an expression of ability) and some verbs (i.e., thinking and sensing words). This may suggest that the students, in texts of this kind, express their ability to do something or think of something. In relation to the reasoning indicator, the use of causal links (i.e., because, due to, as a result, etc.) and temporal connectives (i.e., first, second, during, later, etc.) can refer to specific things and help to link-up the various things that have happened in a situation. This suggests that a description of a sequence of events can imply action over time. In her reflective writing model, Ryan [53] stated that causal links can help students to provide explanations of their actions.

The perspective and future action indicators were associated with similar linguistic features, primarily the use of the term 'would', and the future tense. The recognition of these features here is aligned theoretically with Birney [52] and technically with Ullmann [11] findings that using the future tense is important in relation to both these indicators.

The auxiliary verbs were the most important features for detecting the description of an experience indicator. Employing auxiliary verbs in this context is often done to refer to the past or the future of the current experience. Such verbs are usually used with the main verb in order to show that verb's tense. The auxiliary verbs were somewhat important for most indicators — other than the reasoning indicator. The encountering of this situation may further justify the findings (about the perspective and future action indicators) made by Ullmann

[11] concerning the words, ‘was’, ‘is’, and ‘have’.

Regarding the new learning indicator, the adverb feature played a significant role in the detection of this indicator. In this context, adverbs can often be used to describe how an action has been performed or to allow the student to convey emotions relating to their internal dialogue which relate to how an action affected them, or to describe feelings such as ‘socially’. In addition, the adverb ‘how’ is often employed to describe the way a thing is done, to ask about the extent or degree of something, or sometimes to express a strong feeling about the extent of something. Ryan [53] described the use of ‘adverbial groups to show reason’ in her description of her reflective writing framework — in terms of linguistic evidence (p.105). She found empirical evidence that when students were writing about their evaluation of a particular topic, they used adverbs frequently and as necessary. This is in contrast to Birney [52], who argued that adverbs are not important for the detection of the learning indicator. On the other hand, Ullmann [11], in terms of his framework, indicated that the adverb ‘how’ was frequently used to show evidence of learning.

The use of n-grams and PoS (n-grams) showed great potential for making intuitive sense of reflective writings. These kinds of the feature are linguistically based and so refer to well-established linguistics concepts. The features we mostly used were ones that have been widely seen in literature: the first-person pronoun, the sensing and thinking verbs and phrases (e.g., believe, see, feel), the causal links and phrases (i.e., because of, as a result, due to), and future and past words and phrases. The use of PoS tagging allowed differentiation of reflection indicators via the consideration of the syntactic relationships between the words and phrases in the sentences [13, 14, 16].

Random forest models showed a good performance in the majority of the indicator as it was outperformed other approaches of reflection classification tests, according to earlier research Kovanović et al. [14], Ullmann [11] and Cui et al. [15].

6 Conclusion

This paper makes two significant contributions. Firstly, we developed the automated reflective writing classification for students based on Alrashidi et al [37, 38]. The classification model for reflective indicators of random forest reached an accuracy of 0.96 and substantial performance (κ =0.67), which is regarded as a moderate level of agreement. The application of n-grams and PoS n-grams features shows considerable potential for understanding students' reflective writings, which are constructed using well-established linguistic for psychological processes.

Secondly, our study provides an evaluation of the linguistic features of the seven reflection indicators based on Alrashidi et al [37, 38]. The features we

mostly used were ones that have been widely seen in literature: the first-person pronoun, the sensing and thinking verbs and phrases (e.g., believe, see, feel), the causal links and phrases (i.e., because of, as a result, due to), and future and past words and phrases. The use of PoS tagging allowed differentiation of reflection indicators via the consideration of the syntactic relationships between the words and phrases in the sentences [13, 14, 16].

Lastly, our findings demonstrated some advantages of using the reflection in a specific context that captured a different kinds of linguistic features for each indicator for CS students reflective writing. In the future, we will focus on examining more linguistic features in order to capture a specific semantic for each reflection indicator. we also will focus on developing our tool by using advanced techniques such as data mining.

References

1. Schön, D.A., *Educating the reflective practitioner: Toward a new design for teaching and learning in the professions*. 1987: Jossey-Bass.
2. Boud, D., R. Keogh, and D. Walker, *Reflection:Turning experience into learning*. 1985: Routledge.
3. Moon, J.A., *Learning journals: A handbook for academics, students and professional development*. 1999: Routledge.
4. Cohen-Sayag, E. and D. Fischl, *Reflective Writing in Pre-Service Teachers' Teaching: What Does It Promote?* Australian Journal of Teacher Education, 2012. **37**(10): p. 2.
5. Wald, H.S., et al., *Fostering and evaluating reflective capacity in medical education: developing the REFLECT rubric for assessing reflective writing.* Academic Medicine, 2012. **87**(1): p. 41-50.
6. Betts, J., *Theology, therapy or picket line? What's the 'good'of reflective practice in management education?* Reflective Practice, 2004. **5**(2): p. 239-251.
7. George, S.E., *Learning and the reflective journal in computer science.* Australian Computer Science Communications, 2002. **24**(1): p. 77-86.
8. Demmans Epp, C., G. Akcayir, and K. Phirangee, *Think twice: exploring the effect of reflective practices with peer review on reflective writing and writing quality in computer-science education.* Reflective Practice, 2019. **20**(4): p. 533-547.
9. Fekete, A., et al., *Supporting reflection in introductory computer science.* ACM SIGCSE Bulletin, 2000. **32**(1): p. 144-148.
10. Chng, S.I., *Incorporating reflection into computing classes: models and challenges.* Reflective Practice, 2018. **19**(3): p. 358-375.
11. Ullmann, T.D., *Automated analysis of reflection in writing: Validating machine learning approaches.* International Journal of Artificial Intelligence in Education, 2019. **29**(2): p. 217-257.
12. Ullmann, T.D., *Automated detection of reflection in texts. A machine learning based approach*. 2015, The Open University.
13. Gibson, A., et al. *Reflective writing analytics for actionable feedback.* in *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*. 2017. ACM.

14. Kovanović, V., et al. *Understand students' self-reflections through learning analytics*. in *Proceedings of the 8th international conference on learning analytics and knowledge*. 2018.
15. Cui, Y., A.F. Wise, and K.L. Allen, *Developing reflection analytics for health professions education: A multi-dimensional framework to align critical concepts with data features.* Computers in Human Behavior, 2019. **100**: p. 305-324.
16. Jung, Y. and A.F. Wise. *How and how well do students reflect? multi-dimensional automated reflection assessment in health professions education*. in *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*. 2020.
17. Krippendorff, K.H., *Content Analysis - 3rd Edition : an Introduction to Its Methodology*. 2013, Thousand Oaks: SAGE Publications, Inc.
18. Koole, S., et al., *Factors confounding the assessment of reflection: a critical review.* BMC Medical Education, 2011. **11**(1): p. 104.
19. Liu, M., et al. *Evaluating Machine Learning Approaches to Classify Pharmacy Students' Reflective Statements.* in *International Conference on Artificial Intelligence in Education*. 2019. Springer.
20. Dewey, J., *A restatement of the relation of reflective thinking to the educative process.* 1933: DC Heath.
21. Schraw, G. and R.S. Dennison, *Assessing metacognitive awareness.* Contemporary educational psychology, 1994. **19**(4): p. 460-475.
22. Stone, J.A. and E.M. Madigan, *Integrating reflective writing in CS/IS.* ACM SIGCSE Bulletin, 2007. **39**(2): p. 42-45.
23. Hazzan, O. and J.E. Tomayko, *Reflection and abstraction in learning software engineering's human aspects.* Computer, 2005. **38**(6): p. 39-45.
24. Surbeck, E., E.P. Han, and J. Moyer, *Assessing reflective responses in journals.* Educational Leadership, 1991. **48**(6): p. 25-27.
25. Moallem, M., *Reflection as a Means of Developing Expertise in Problem Solving, Decision Making, and Complex Thinking of Designers.* 1998.
26. Mamede, S. and H.G. Schmidt, *The structure of reflective practice in medicine.* Medical Education, 2004. **38**(12): p. 1302-1308.
27. Plack, M.M. and L. Greenberg, *The reflective practitioner: reaching for excellence in practice.* Pediatrics, 2005. **116**(6): p. 1546-1552.
28. Greiman, B. and H. Covington, *Reflective thinking and journal writing: Examining student teachers' perceptions of preferred reflective modality, journal writing outcomes, and journal structure.* Career and Technical Education Research, 2007. **32**(2): p. 115-139.
29. Crawford, P.A., S.K. Roberts, and R. Hickmann, *Nurturing early childhood teachers as leaders: Long-term professional development.* Dimensions of early childhood, 2010. **38**(3): p. 31.
30. Black, P.E. and D. Plowright, *A multi-dimensional model of reflective learning for professional development.* Reflective Practice, 2010. **11**(2): p. 245-258.
31. Kwon, K. and D.H. Jonassen, *The influence of reflective self-explanations on problem-solving performance.* Journal of Educational Computing Research, 2011. **44**(3): p. 247-263.
32. Ellmers, G., *The graphic design project: employing structured and critical reflection to guide student learning.* Communication Design, 2015. **3**(1): p. 62-79.
33. Amador, J.M., et al., *Preparing preservice teachers to become self-reflective of their technology integration practices*, in *Pre-Service and In-Service Teacher Education: Concepts, Methodologies, Tools, and Applications*. 2019, IGI Global. p. 1298-1325.

34. Antonio, R.P., *Developing Students' Reflective Thinking Skills in a Metacognitive and Argument-Driven Learning Environment.* International Journal of Research in Education and Science, 2020. **6**(3): p. 467-483.
35. Noer, S., P. Gunowibowo, and M. Triana. *Development of guided discovery learning to improve students reflective thinking ability and self learning*. in *Journal of Physics: Conference Series*. 2020. IOP Publishing.
36. Babb, J., R. Hoda, and J. Nørbjerg, *Embedding reflection and learning into agile software development.* IEEE software, 2014. **31**(4): p. 51-57.
37. Alrashidi, H., et al. *A Framework for Assessing Reflective Writing Produced Within the Context of Computer Science Education*. in *In Proceedings of the 10th International Learning Analytics & Knowledge Conference (LAK20)*. 2020. Frankfurt, Germany.
38. Alrashidi, H., et al. *Educators' Validation on a Reflective Writing Framework (RWF) for Assessing Reflective Writing in Computer Science Education*. 2020. Cham: Springer International Publishing.
39. Ullmann, T.D. *Reflective writing analytics: empirically determined keywords of written reflection*. in *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*. 2017. ACM.
40. Ullmann, T.D., *An architecture for the automated detection of textual indicators of reflection*, in *1st European Workshop on Awareness and Reflection in Learning Networks*. 2011: Palermo, Italy. p. 138-151.
41. Lin, C.-W., et al., *A word-count approach to analyze linguistic patterns in the reflective writings of medical students.* Medical education online, 2016. **21**(1): p. 29522.
42. Pennebaker, J.W. and C.K. Chung, *Expressive writing: Connections to physical and mental health.* Oxford handbook of health psychology, 2011: p. 417-437.
43. Ullmann, T.D. *Keywords of written reflection-a comparison between reflective and descriptive datasets*. in *CEUR Workshop Proceedings*. 2015.
44. Liu, M., K. Kitto, and S. Buckingham Shum, *Combining factor analysis with writing analytics for the formative assessment of written reflection.* Computers in Human Behavior, 2021. **120**: p. 106733.
45. Moon, J.A., *Reflection in learning and professional development: Theory and practice*. 2013: Routledge.
46. Landis, J.R. and G.G. Koch, *The measurement of observer agreement for categorical data.* biometrics, 1977: p. 159-174.
47. Jurafsky, D. and J.H. Martin, *Speech and Language Processing: An introduction to speech recognition, computational linguistics and natural language processing.* Upper Saddle River, NJ: Prentice Hall, 2008.
48. Ogada, K., W. Mwangi, and W. Cheruiyot, *N-gram based text categorization method for improved data mining.* Journal of Information Engineering and Applications, 2015. **5**(8): p. 35-43.
49. Schulman, A. and S. Barbosa. *Text Genre Classification Using only Parts of Speech*. in *2018 International Conference on Computational Science and Computational Intelligence (CSCI)*. 2018. IEEE.
50. Fang, A.C. and J. Cao, *Part-of-Speech Tags and ICE Text Classification*, in *Text Genres and Registers: The Computation of Linguistic Features*. 2015, Springer. p. 71-82.
51. Breiman, L., *Classification and regression trees*. 2017: Routledge.
52. Birney, R., *Reflective writing: Quantitative assessment and identification of linguistic features*. 2012, Waterford Institute of Technology.
53. Ryan, M., *Improving reflective writing in higher education: A social semiotic perspective.* Teaching in Higher Education, 2011. **16**(1): p. 99-111.

54. Shum, S.B., et al. *Reflecting on reflective writing analytics: Assessment challenges and iterative evaluation of a prototype tool*. in *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*. 2016. ACM.
55. Boud, D., R. Keogh, and D. Walker, *Reflection: Turning experience into learning*. 2013: Routledge.

Appendix 1

Author	*Reference Framework*	*Indicators*	*Dictionary/ List*	*Field*	*Results*
(Ullmann, 2011)	NA	Focuses only on actually reflective text	Manual list	NA	Examples only
(Bruno et al., 2011)	NA	Focuses only on actually reflective text	Manual mental words list	NA	NA
(Ullmann, 2015b)	Based on 24 Frameworks	Experience, personal, feelings, critical stance, perspective, outcome	Data-driven	Multiple	NA
(Lin et al., 2016)	(Gibbs, 1988)	Description, feelings, evaluation, analysis, conclusion, and action plan	LIWC list	Medical	NA
(Ullmann, 2017)	Based on 24 Frameworks	Reflection, description of an experience, feelings, beliefs, difficulties, perspective, and outcome (lessons learned and future intentions)	Data-driven	Multiple	0.78
(Cui et al., 2019)	(Gibbs, 1988)	Description, analysis, feelings, perspective, evaluation, and outcome	LIWC list	Dental	NA