International Journal of Artificial Intelligence in Education
https://doi.org/10.1007/s40593-019-00174-2

ARTICLE

Automated Analysis of Reflection in Writing: Validating Machine Learning Approaches

Thomas Daniel Ullmann[1]

Published online: 13 February 2019

Abstract

Reflective writing is an important educational practice to train reflective thinking. Currently, researchers must manually analyze these writings, limiting practice and research because the analysis is time and resource consuming. This study evaluates whether machine learning can be used to automate this manual analysis. The study investigates eight categories that are often used in models to assess reflective writing, and the evaluation is based on 76 student essays (5080 sentences) that are largely from third- and second-year health, business, and engineering students. To test the automated analysis of reflection in writings, machine learning models were built based on a random sample of 80% of the sentences. These models were then tested on the remaining 20% of the sentences. Overall, the standardized evaluation shows that five out of eight categories can be detected automatically with substantial or almost perfect reliability, while the other three categories can be detected with moderate reliability (Cohen's κ ranges between .53 and .85). The accuracies of the automated analysis were on average 10% lower than the accuracies of the manual analysis. These findings enable reflection analytics that is immediate and scalable.

Keywords Assessment and testing of learning outcomes · Automated content analysis · Educational data mining · Learning analytics · Machine learning · Modeling metacognitive skills · Reflection · Reflective writing · Text analysis

Introduction

Fostering reflective thinking (Boud et al. 1985; Dewey 1933; Mezirow 1991; Schön 1983) is an important educational practice recognized on national and international

✉ Thomas Daniel Ullmann
t.ullmann@open.ac.uk

[1] Institute of Educational Technology, The Open University, Walton Hall, Milton Keynes MK7 6AA, UK

levels. The worldwide operating Organisation for Economic Co-operation and Development (OECD) has endorsed reflective thinking as important for a "successful life and well-functioning society" (Rychen and Salganik 2005). Students worldwide are assessed regarding their reflection and evaluation skills via the Programme for International Student Assessment (PISA) test on reading literacy (OECD 2013). Examples for the importance of reflective thinking can be found on a national level, such as the quality code of the UK Quality Assurance Agency for Higher Education (QAA) that recommends all higher education providers establish a structured and supported process for learners to reflect upon their personal development (QAA 2012) or the U.S. Department of Education that puts reflective thinking on the national agenda as it recommends the creation of learning opportunities to promote reflection (Means et al. 2010). Students are frequently encouraged to maintain reflective journals or to engage in writing activities which aid their reflection on experiences.

Despite the importance of reflection for education, there is little research on artificial intelligence techniques to automate the analysis of writings regarding reflective thinking. Currently, the common educational practice to analyze and assess reflective writings is a manual process. Very often, they are analyzed by following the principles of the content analysis method (Krippendorff 2012), which is a manual, labor-intensive process (Birney 2012; Ip et al. 2012; Mena-Marcos et al. 2013; Poldner et al. 2014; Poom-Valickis and Mathews 2013; Prilla and Renner 2014; Wald et al. 2012). The typical content coding task of reflective writings entails the manual labeling of text units according to categories of a reflective writing model. The labeled text units are used to calculate scores that indicate the quality of reflection found in the writing (for example, see Clarkeburn and Kettula 2011; Mena-Marcos et al. 2013; Plack and Greenberg 2005; Poldner et al. 2014). This analysis is currently manual, as there is a lack of knowledge about automated methods specific to reflective writing.

There has been great success with the extensive research into automated text analysis methods for educational contexts (Dringus and Ellis 2005; Lin et al. 2009; Rosé et al. 2008). The areas of automatic essay assessment (Attali and Burstein 2006; Dikli 2006; Hearst 2000; Jordan 2014; Landauer 2003; Page 1968; Page and Paulus 1968; Rivers et al. 2014; Shermis 2014; Shermis and Burstein 2003; Wild et al. 2005; Wilson and Czik 2016), discourse analysis (Dascalu 2014; Dessus et al. 2009; Ferguson and Buckingham Shum 2011), and general natural language processing techniques that are prominently applied to the context of education (McCarthy and Boonthum-Denecke 2012).

Despite these research successes showing the potential of automated text analysis for education, there has been not much research regarding the automated text analytics of reflection. Furthermore, there has been some research using dictionary and rule-based approaches (see the following review section), but in general, not much is known about the extent to which automated methods can be used to reliably analyze texts regarding categories of reflective writing models.

This research replicates the aspect of the manual content analysis that assigns labels to text segments with automated means. This study is the first one to be based on a comprehensive model of reflective writing with the following eight categories that has been derived from 24 models of reflective writing: *reflection* (depth of reflection), *experience*, *feelings*, *personal belief*, *awareness of difficulties*, *perspective*, *lessons learned*, and future *intention* (breadth of reflection). The aim is to show that automated methods can be used to reliably draw inferences about the presence (and absence) of reflection in texts. This research further investigates the potential of supervised machine learning algorithms to reliably annotate text segments of writings according to categories of a reflective writing model.

Researching automated methods to analyze reflection is important because the manual analysis poses constraints to teaching and research and may hamper deeply personal reflections. The manual analysis of reflective writing is a time-consuming task, adding a cost that constrains the frequency and intensity of its pedagogical use. This cost also limits large-scale research explorations and research designs investigating immediacy of feedback. Automated methods do not have these constraints. Beyond saving time, automated methods may have benefits especially for learning to write reflectively because of the often sensitive nature of reflective writings. Students may feel a barrier to share their writings with a tutor. They may also feel more comfortable with an automated reflective writing aid to self-disclose those private thoughts relevant for their reflection. This assumption is backed by research that indicates self-disclosure is easier by using a computer (Richman et al. 1999).

The major highlights of this study are as follows:

a) The theoretical model of this study is based on categories that are common to many models that have been used to manually analyze reflective writing. Compared to previous research, this research tested machine learning on a comprehensive model and not on a particular model of reflection.
b) The empirical tests of the theoretical model show that the model categories can be annotated in a reliable way and the model components showed empirical validity. In other words, compared to previous research, this research shows that the model is reliable and valid.
c) The paper provides a comprehensive literature review of automated methods to analyze reflection in texts.
d) This research focuses on machine-learning-based approaches.
e) The evaluation was conducted in a standardized way for all data sets.
f) The evaluation shows that the theoretically derived model categories can be automatically analyzed in a reliable way using machine learning.
g) The evaluation suggests that there is a relationship between the manual and machine coding performance.
h) The inspection of the most important features of the machine learning algorithms confirmed several important features but also surfaced new features.

Model to Analyze Reflection in Writings

There is a debate about what exactly constitutes reflection [for example, see the definitions of Dewey 1933, Schön 1987, Boyd and Fales 1983, Boud et al. 1985, Mezirow 1991, and Moon 1999, 2006], but as of now, there is yet not a widely agreed definition of reflection. However, as discussed later in this work, some commonalities exist between the various theories of reflection. For the purpose of this paper, we define reflection as follows: *Reflective thinking is a conscious, goal-oriented cognitive process that seeks to learn solutions to personally important, often weakly defined, ambiguous* *problems of past and present experience and anticipated future situations that often involve thinking about the important elements of the experience, a critical analysis of the problem including the analysis of the thinkers' own personal perspective and feelings as well as the perspective of others in order to determine lessons-learned or future plans.*

All people do reflect on experience, but their reflective thinking skills may not be fully developed (Mann et al. 2007). Research indicates that reflective thinking can be taught and learned (Mann et al. 2007). An important educational practice to foster reflective thinking is the practice of reflective writing (Moon 2006; Thorpe 2004). Many higher education disciplines make use of reflective writing as part of their educational programs, such as teachers' pre-service training, early childhood education, business, physical therapy, literature, psychology (Dyment and O'Connell 2010), health profession (Mann et al. 2007), pharmacy (Wallman et al. 2008), language learning (Chang and Lin 2014; Lamy and Goodfellow 1999), and writing literacy (Yang 2010). Through writing, students capture their reflective thoughts to better understand their own experience.

Reflective writings are different from the types of writings students normally perform, such as essays, literature surveys, or reports. Although reflective thinking may have gone into these writings, the reflective thought process is usually not expressed or developed in these writings as this is not their primary purpose. These writings are different as they are centered on the personal thought process and how to learn from individual experiences. The writings are often very personal and can contain references to feelings. Also, authors are often self-critical. They may consider the perspectives of people who are important in their context or draw conclusions from contexts that are valued by the author. If the writing includes lessons learned or plans for the future, they are often very specific to the author's context. It is this constant challenging of routines to improve practice, making reflective thinking such an important asset. Thus, reflective writing is different from other types of academic writing. Students may not be familiar with it and therefore may find it difficult to engage with it, as the "rules" are different from what they have previously experienced. This lack of experience results in students not fully developing their reflective writing. For example, students often have problems writing in a descriptive/non-reflective way as they are used to describe events without deeply engaging their own thought process (Mann et al. 2007). An analysis of the writing is the first step toward improving reflective writing and thinking.

Common to all manual methods used to analyze writings regarding reflection is that they are based on a model (sometimes called a framework, assessment rubric, or coding scheme). This section outlines a comprehensive model of reflective writing used in this work as a realistically complex test case to automatically detect reflection.

For several decades, researchers have developed models to analyze reflective writings (early examples are Gore and Zeichner 1991; Hatton and Smith 1995; Richardson and Maltby 1995; Ross 1989; Sparks-Langer et al. 1990; Sparks-Langer and Colto 1991; Tsangaridou and O'Sullivan 1994). Reflective writing models often exhibit a quality of depth, of breadth, or both (Moon 2004). Many models that analyze the depth of reflection in writings define a scale with several levels usually with the lowest level characterized as showing no sign of reflection, i.e., a descriptive writing, while a writing on the highest level shows evidence of a deeply reflective writing. The quality of depth implies that reflection is hierarchical, with the highest level being the most desired outcome. Examples of depth models are the model of Wong et al. (1995) with the non-reflector, reflector, and critical reflector levels or the model of Lai and Calandra (2010) having the routine (non-reflective) level at the lowest level and the transformative (highly reflective) level at the highest level.

Breadth models are descriptive because they do not value one category over another one, as the depth models imply. For example, Wong et al. (1995) described their model to analyze reflective writings with categories, such as attending to feelings, validation, and outcome of reflection. Prilla and Renner (2014) analyzed reflection according to the categories such as describing an experience, mentioning and describing emotions, and challenging or supporting assumptions. Often, the unit of analysis (Weber 1990) of a depth model is the whole text, while the unit of analysis of breadth models are smaller parts of the texts. Several models describe a mapping from the breadth categories to the levels of reflection connecting both qualities (see Ullmann 2015a).

The breadth and depth dimensions for the model of reflection detection that are used for this research have been derived from the model descriptions of the following 24 models (Ullmann 2015a): the models of Sparks-Langer et al. (1990), Wong et al. (1995), McCollum (1997), Kember et al. (1999), Fund et al. (2002), Hamann (2002), Pee et al. (2002), Williams et al. (2002), Boenink et al. (2004), O'Connell and Dyment (2004), Plack and Greenberg (2005), Ballard (2006), Mansvelder-Longayroux (2006, 2007), Plack et al. (2007), Kember et al. (2008), Wallman et al. (2008), Chamoso and Cáceres (2009), Lai and Calandra (2010), Fischer et al. (2011), Birney (2012), Wald et al. (2012), Mena-Marcos et al. (2013), Poldner et al. (2014), and Prilla and Renner (2014). The reflective writing models described in those papers have been evaluated with the method of the manual content analysis, and they all reported inter-rater reliability scores, which gauge the degree to which coders can classify text units according to reflective writing model components. The reason for this selection is that a) content analysis is a common and principled approach to analyze and assess reflective writings and b) the information about reliability allows researchers to understand how well human raters can differentiate between the categories of the model, and it increases the confidence that the research can be replicated. All 24 models satisfy these criteria.

The synthesis of these models led to a model of common constituents, containing both qualities. We will refer to this model as the model for reflection detection. In total, the model consists of eight categories:

1. The depth dimension *reflection* was modeled as the binary category *reflective* vs. *non-reflective*. This is the common denominator of the depth models (Ullmann 2015a).

A synthesis of the breadth categories of the models derived seven categories (Ullmann 2015a). These are categories that are common components in many reflective writing models, and these seven breadth dimensions can be summarized as follows (see, Ullmann 2015a):

2. *Description of an experience:* Reflective writing models often contain a category that captured an experience of a writer. Often, this experience is the reason for the

writer to start a reflective writing. The description of the experience provides the context for the reflection (Ullmann 2015a).

3. *Feelings:* Many models contain a component that looked for expressions of emotions expressed in the writing. Feelings can be a key element of a reflective writing, as feelings can be the reason to start thinking reflectively (e.g., a feeling of puzzlement, uncertainty, surprise), and they can be the subject of the writing, reflecting the influence of feelings on our thought process (Ullmann 2015a).
4. *Personal belief:* Many models have searched for evidence of expressing personal beliefs as a component of reflective writing. A reflective thought is often of a personal nature, and a reflection is often about one's own perspectives and assumptions (Ullmann 2015a).
5. *Awareness of difficulties – critical stance:* All models contain a category that can be summarized as having an awareness of difficulties and problems and more generally a critical stance towards an experience or situation (Ullmann 2015a).
6. *Perspective:* Many models have described the importance of considering other perspectives. In the reflective writing, these models look for evidence of the description of the perspective of someone else, consideration of theory as an external perspective, or the social, historical, or ethical context (Ullmann 2015a).

Outcomes – Lessons learned and future intentions: Many of the models include a component that captures the outcomes of reflective writings. The models have exhibited two outcome dimensions:

7. *Lessons learned* – retrospective outcomes: These are outcomes that look back on what was learned. This could be a better understanding of the experience, new knowledge, behavior changes, changes to perception, and better self-awareness (Ullmann 2015a).
8. *Future intentions* – prospective outcomes: These are future potential outcomes that are yet not realized. Examples are intentions to do something or plans for the future (Ullmann 2015a).

We will refer to these categories throughout the text as to *Reflection*, *Experience*, *Feeling*, *Belief*, *Difficulty*, *Perspective*, *Learning*, and *Intention*. Overall, a synthesis of the categories of existing models derived common categories which formed the categories of the model for reflection detection. Each category of this model for reflection detection serves as a test case for the automated detection of reflection. The approach taken here is to evaluate the automated detection of reflection on a set of categories that are common to many models used in the context of the manual content analysis of reflective writing. An alternative approach is to test the automatability with an existing model. This, however, would have narrowed the scope under investigation to the characteristics of this specific model. The current approach has the benefit that the evaluation is about a model that represents cases that are commonly analyzed.

Automated Methods to Analyze Reflection

The review of the literature shows that the landscape of methods that have been used to automatically analyze reflective writing can be largely classified according to three approaches, namely the dictionary-based, the rule-based, and the machine-learning-based approaches. The main focus of this literature review is about the machine-learning-based approach. The other two approaches differ from the machine-learning-based approach insofar as they are expert-driven, meaning that experts explicitly define patterns to detect reflection. In the machine-learning-based approach, the algorithms define these patterns. The following literature review of methods to analyze text regarding the evidence of reflective thinking and related constructs extends the review of Ullmann (2015a).

Machine-Learning-Based Approach

There is a relatively small body of research that is concerned with machine-learning-based approaches to detect reflection in writings. These are approaches that use machine learning algorithms, especially supervised machine-learning algorithms for classification (Gupta and Lehal 2009; Sebastiani 2002) to classify text according to categories of reflective writing models. Compared to the rule-based approach that relies on the manual construction of rules and patterns, the machine-learning-based approach learns these patterns from example data automatically.

Research in this area is still fragmented. Currently, researchers investigate specific theoretical models that cover only some of the aspects of reflection. Furthermore, these models have not been tested regarding their validity and also often not regarding their reliability. These criteria should not be neglected as they are important to understand the quality of the model. To understand whether we can use machine learning for the analysis of writings regarding of reflection, we need to test a model that is valid and reliable and covers the most important characteristics of reflection.

Yusuff (2011) trained several machine learning models on a variety of text sources for his bachelor thesis. The texts of these sources were declared by the author as either reflective or not. Other dimensions of reflection have not been considered. The validity of this distinction was not determined, making it difficult to evaluate whether the texts have been reflective, and the inter-rater reliability of the coding was not reported, making it difficult to assess whether this research can be replicated by other researchers. Ullmann (2015a) reported the successful application of machine learning for the detection of reflection. As this paper is a substantial extension of that work, it is not further discussed. Cheng (2017) developed machine learning models to classify posts of an e-portfolio system according to categories of their self-developed A-S-E-R model in the context of L2 learning. The model proposes four elements of reflective writing, namely “experience”, “external influence”, “strategy application - analyze the effectiveness of language learning strategy”, and “analysis, reformulation and future application”. Each element consists of four levels capturing proficiency. Their model covers four out of the seven breadth dimensions of the model used in here, namely *Experience* (which is related to their category “experience”), *Perspective* (“external influence”), *Difficulties* (as part of the category “strategy application”) and future *Intentions* (part of the higher levels of the category “analysis, reformulation and future application”). They did not cover the categories *Beliefs* and *Feelings* and they did not explicitly specify the depth dimension of reflection. Their model has been derived from theory and adapted to the context of L2 language learning, but an empirical evaluation of the validity of the model is missing. Furthermore, the papers did not report inter-rater reliability. The model of Liu et al. (2018) has been inductively derived from an initial content analysis by experts. The model consists of two foci called "technical" and "personalistic beliefs" with each having three levels, namely "description", "analysis", and "critique". Their model covers three out of the seven breadth model categories of this paper, namely the categories *Experience* (which is related to their category "description"), *Belief* (which is part of the "personalistic" dimension), and *Difficulties* (which is part of the categories "analysis" and "critique"). Their model does not explicitly contain the depth of reflection dimension, and the model has not been empirically validated, but the authors report high inter-rater reliability between two coders. The model used in the research of Kovanović et al. (2018) covers two of the seven breadth dimensions of this paper's model, namely *Experience*, which in their model consists of the two categories "observation of own behavior" and "motive or effect of own behavior", and the category *Intention*, which is similar to the their category "indicating a goal of own behavior". Their model does not consider the depth dimension of reflection. They do not report any empirical evidence of the validity of their model, but they achieved high inter-rater reliability between two coders.

All models outlined have explicit links to the theory of reflective thinking. While not directly connected with reflection, there has been research about related concepts that use machine learning, such as cognitive presence (Corich 2011; Kovanovic et al. 2016; McKlin 2004) and argumentative knowledge construction (Dönmez et al. 2005; Rosé et al. 2008). As this research is only related, it is less relevant and is not further discussed.

The literature review shows that existing research is constrained to specific models of reflection. These theoretical models are less comprehensive compared to the theoretical model used in this paper and they miss important facets of reflection. Therefore, they do not allow us to assess whether all important dimensions of reflection can be assessed automatically with machine learning methods. Furthermore, the review shows that none of the models show any evidence for empirical validity. While many of the mentioned categories can be traced back to the theory or reflective thinking and thus have face validity, evidence of empirical validity would have strengthened the case for these models to actually measure reflective thinking in writings and not something else. Apart from two papers, the papers did not report inter-rater reliability values, making it difficult to assess whether their research can be replicated.

The research in this paper goes beyond the current state-of-the-art. Compared to previous research, this research is a) based on a general model derived from empirical research and b) includes empirical evidence about the reliability and validity of the model in the current context. Regarding the first point, the research model of this research has been generalized by many individual models. The model categories therefore stand for a set of categories that are commonly used to analyze reflective writing. The evaluation of the machine learning algorithms on these model categories therefore provides more generalized evidence about how well reflection can be analyzed with automated means. This approach is different from research that uses a particular model, such as the self-developed models by Liu et al. (2017) or the context-specific model of Cheng (2017; Cheng and Chau 2013) or the model of Kovanović et al. (2018). With regard to the second point, this research outlines both the performance of the machine learning algorithms as well as the performance of the manual coders, providing insights into the difficulty of the content analysis task, and provides empirical evidence about the validity of the model. Empirical evidence of validity for the model has not been reported in any of the previous works. Consequently, the test of the machine learning algorithms is based on a model that is a realistic test case of reflective writing models and is based on a model that is reliable and valid.

Besides model differences, current research into this area uses various machine learning algorithms ranging from latent semantic analysis (Cheng 2017) to Naïve Bayes and Support Vector Machines (Liu et al. 2018) and Random Forests (Kovanović et al. 2018). This variety suggests that all these algorithms are good candidate algorithms for the detection of reflection. Liu et al. observed that Naïve Bayes outperformed the SVM algorithms, suggesting that certain algorithms perform better on the same data set. We do not know which of the algorithms will perform best until we have tested several of them on the same data set.

Researchers also used different measures and measurement techniques to gauge the performance of the algorithms. They used the Cohen's κ (Cheng 2017; Kovanović et al. 2018) and the F_1-score (Liu et al. 2018) measures. Cohen's κ is often used in the educational area, as it is also a frequent measure of the inter-rater reliability between human coders. Using the same statistic for both, the human performance and the machine performance allows for a better comparison. The F_1-score is the harmonic mean of the statistics Precision and Recall. This is a measure often used in the area of Information Retrieval. However, both measures are not compatible, making their comparison difficult. Research differs with regard to the used performance statistic and in the methods to measure it, such as by using different splits of the test and training data set, various forms of cross-validation, and different numbers of class labels. All these factors make comparing research challenging.

We try to summarize the performance of the machine learning algorithms based on a mapping of the individual model categories to the categories of this paper's model for reflection detection. However, due to the outlined caveats, the outcome of the comparison may not be very informative. The performance of the category *Experience* has a reported Cohen's κ of 0.7 (Cheng 2017) and F_1-scores of 0.82 and 0.85 (Liu et al. 2018). The category "personalistic" dimension of Liu et al. (2018), which is similar to the category *Belief,* had F_1-scores ranging from 0.78 to 0.84. The categories "application of strategies" with a κ of 0.73 in the work of Cheng (2017) and "analysis" with an F1-score of 0.80 to 0.88 and "critique" with a score of 0.79 to 0.84 in the work of Liu et al. (2018) are similar to the category *Difficulties*. The category *Intention* can be found in the category "analysis, reformulation and future application" (κ of 0.6) of Cheng (2017). Kovanović et al. (2018) reported a Cohen's κ of 0.51 over all categories. The category with the highest error rate was the "goal" category (which has been mapped to the category *Intention*), followed by the categories "observation" and "motive" (*Experience*). Summarizing all this information, the categories *Experience* and *Difficulties* achieved the highest performance, followed by *Beliefs* and with some distance comes *Intention* at the last position. However, as the methods and measures varied between each paper and because of the limited amount of available research, the value of this ranking is limited.

Dictionary-Based Approach

In the context of automated methods, a dictionary often means a collection of words that are associated with a category. Usually, one or several experts define a set of words that represents a concept. A computer program can make use of these dictionaries to find occurrences of dictionary words in texts. The aim of dictionaries to analyze text is to convert text into numbers by counting the frequencies of dictionary words. This allows to quantify text and to use statistical methods to test the data. The focus of dictionary-based research is less about creating high-performing classifiers and most researchers do not consider this aspect, although there are exceptions (Ullmann 2017). Therefore, we cannot report about the performance for most of the paper cited in the following paragraphs. However, as this approach is relevant for the automated detection of reflection, this section provides an overview to show how widespread this approach has become. Benefits are that dictionaries can be set up quicker than rule-based or machine-learning-based approaches to test ideas of automatization. They can be important for rule-based approaches, which often use dictionaries in combination with rules, and they may even inform the creation of feature for machine learning.

One of the first examples of this approach was the *General Inquirer* (Stone and Hunt 1963). Other examples of this approach were the research in the scope of the *Textbank System* (Mergenthaler and Kächele 1991) and the *Linguistic Inquiry and Word Count* (LIWC) tool (Chung and Pennebaker 2012; Pennebaker and Francis 1996).

The following examples highlight the application of this approach in the context of the detection of reflection in texts. The dictionary-based approach has been researched in the context of multiple disciplines, such as in education (Bruno et al. 2011; Chang et al. 2012; Chang and Chou 2011; Gašević et al. 2014; Houston 2016; Kann and Högfeldt 2016; Lin et al. 2016; Ullmann 2011, 2015b, 2017; Ullmann et al. 2012, 2013), psychology (Fonagy et al. 1998; Mergenthaler 1996), and linguistics[1] (Birney 2012; Forbes 2011; Luk 2008; Olshtain and Kupferberg 1998; Reidsema and Mort 2009; Ryan 2011, 2012, 2014; Wharton 2012). Most of the research is based on English writings, but there is also research about Chinese (Chang et al. 2012; Chang and Chou 2011; Lin et al. 2016) and Swedish dictionaries (Kann and Högfeldt 2016). Consequently, there is an interdisciplinary interest in the analysis of reflection with dictionaries spanning several languages. As mentioned, there is a lack of research into the classification performance of these dictionaries. Ullmann (2017), however, devised a data-driven method based on large data sets to generate keywords that showed promising performance above the baseline accuracy for seven out of eight categories of the reflection detection model. The average Cohen's κ over all eight categories was 0.45. The highest value was 0.65 for *Experience* and the lowest value was 0.28 for *Perspective*. The Cohen's κ for *Reflection* was 0.59. These performances were achieved with a small number of words for each category. These words seemed to be a useful start to populate reflection specific dictionaries.

[1] The research summarised under the term linguistic approaches investigates reflection via defined sets of linguistic features. Texts are manually annotated according to these sets for further analysis. Often, these sets of features can be specified, which in principal allows their implementation either fully or in part via a dictionary-based or rule-based approach.

As with the rule-based approach, experts mainly drive the creation of the dictionaries (expert-driven approach). Often, a single researcher or a group of researchers determines which dictionaries are relevant for reflection and which words should belong to each dictionary based on the study of text examples. The machine-learning-based approach on the other side is data-driven. The algorithms learn from data which words are important and how these words must be connected to classify texts.

Rule-Based Approach

While dictionary-based approaches rely mostly on pattern matching the dictionary entries with the text, rule-based systems provide mechanisms that extend the capability of making inferences from texts. The core of a rule-based system is a set of rules to express knowledge about the domain. The logic expressed in these rules allows for formal reasoning over the knowledge base of rules. Thus, with the inference machine of a knowledge-based system, rules can be chained to deduce facts based on multiple conditions. This technique extends the expressiveness of the automated detector compared to the dictionary-based approach.

Compared to the dictionary-based approach, the rule-based approach is more recent in this domain. Research using this approach often combines natural language processing, dictionaries, and rules to create a text analysis pipeline that captures patterns of reflective writing, as defined by the expert modeler. Buckingham Shum et al. (2017; 2016) customized the Xerox Incremental Parser (Ait-Mokhtar et al. 2002), a general natural language parser, with custom generated dictionaries and rules in order to detect several categories of reflection. The categories of the model have been co-designed together with a practitioner (Buckingham Shum et al. 2017). Although the model described in the paper had several facets, the evaluation only tested whether the rule-based system can distinguish between reflective and unreflective sentences. This distinction may be similar to the depth category *Reflection* of the reflection detection model in this paper. Their best test result (second test) had a Cohen's κ of 0.43 (based on own calculation of the values presented in the confusion matrix of table 3 in Buckingham Shum et al. 2017), which was achieved after rule alterations based on the experiences with the first experiment and a rerun of the experiment on the same data set. The paper did not provide evidence for the validity of their concept nor did it report inter-rater reliability. Gibson et al. (2016) showed a rule-based system to analyze writings according to metacognition, which is a related concept of reflection. Their model defined four overarching categories with sub-categories. For each category, they created rules to find evidence of the categories in text using a combination of part-of-speech and dictionary words. In their evaluation, they combined all categories and tested whether the metacognitive activity was strong or weak. Their best test result (for strong authors) achieved a Cohen's κ of 0.48 (based on own calculation of the values presented in table 9 in Gibson et al. 2016). The paper did not report indicators of validity and reliability.

In the context of the analysis of writings according to facets of reflection, Ullmann et al. (2012) combined the dictionary-based approach with the rule-based approach. An inference machine reasoned over a set of rules that chained low-level rules with higher-level rules to derive facts that indicate reflection in writing. The descriptive results indicated a positive association between the predictions of the rule-based algorithm and the manual ratings of blog posts according to reflective categories, such as “description of an experience”, “personal experience”, “critical analysis”, “taking into account other perspectives”, “outcome”, “what next”, and “reflection”. The paper did not report any performance measures. The paper also did not report any evidence of the empirical validity of the theoretical model, but instead outlined the theoretical roots of each category, supporting the face validity of the model. The paper did not report the inter-rater reliability of the blog post coders.

Summary

Most research regarding automated methods to analyze writing about reflective thinking use the dictionary-based approach. There are other studies that use the rule-based systems and machine learning approaches. These three approaches have different capabilities in modeling text. The dictionary-based approach models dictionaries as word lists. Each word of this list belongs to the category expressed by the dictionary. Using this method of modeling text may result in lower accuracy, as, for example, it does not consider the polysemy of words. Words can have multiple meanings and therefore might express another concept than foreseen by the dictionary. In contrast, a rule-based approach has more capabilities to model text because it can use rules to disambiguate the meaning of words based on context information and provide better results. Machine learning has been highly successful to classify text (Hotho et al. 2005). It therefore appears as a promising approach to automatically analyze reflection in writing.

Research Questions

This study investigates whether machine learning algorithms can be used to reliably detect reflection in texts. The literature review showed that models that have been used for the manual analysis of writings according to reflection have two types of qualities, quality of depth and quality of breadth. The following two research questions consider both qualities:

1. Can machine learning reliably distinguish between *reflective* and *descriptive* (non-reflective) sentences?
2. Can machine learning reliably distinguish sentences per the presence or absence of categories that are common in reflective writings? The categories are the following: *description of an experience*, *feelings*, *personal beliefs*, *awareness of difficulties*, *perspective*, *lessons learned* and *future intentions*.

The following experiments have been constructed to answer these research questions. The experiments use a standardized process to evaluate the potential of machine learning to detect reflection in texts. This process ensures that all categories are assessed in the same way and that the results of the experiment are comparable.

Material and Methods

To generate the data set for each category of the reflection detection model, the researcher devised a standardized process[2] that was equally applied to all data sets. Based on a text collection of student writings, the texts were unitized, annotated, and split into eight times three data sets (for each of the eight categories of the reflection detection model exist three data set versions representing the three majority vote conditions outlined in the result section). Then, each data set was pre-processed and split into training and test data sets. These data sets served as inputs for the machine learning algorithms.

Text Collection

The text collection consisted of 77 student writings. Among them, 67 student writings came from the British Academic Written English (BAWE)[3] corpus (Gardner and Nesi 2013; Nesi and Gardner 2012). As most of the research about reflective writings was conducted in the context of academic writing, the BAWE corpus with its similar background seemed to be an appropriate choice. The BAWE corpus contains student essays, and some of them are responses to several reflective writing tasks (Nesi 2007). Relevance sampling (Krippendorff 2012) was used to retrieve 67 texts from the BAWE text collection and ten writings came from examples cited in the research literature. Relevance sampling was chosen over random sampling, as reflective texts are relatively rare (Ullmann et al. 2013); therefore, many texts without relevance would have entered the text selection if random sampling would have been applied. In total, 46 students wrote the 67 texts. Most of the essays were written by students of the health (20 students), business (9), engineering (9), tourism management (6), and linguistic (6) disciplines. They were mostly written by third- (23) and second-year students (21), followed by first-year (12) and postgraduate students (10), and for one text, the student level was unknown. Among the texts, 40 texts were awarded merit, 24 texts were awarded distinction, and for three texts, the grade was not known. In addition to these 67 texts, the text collection was extended by ten student writings that were cited in the literature of reflective writing (Korthagen and Vasalos 2005; Moon 2004, 2006; Wald et al. 2012) to add additional examples of reflective writings to the text collection.

Unitizing Text Collection

The related literature on the manual content analysis of reflective writings suggested that smaller units opposed to whole texts are more suitable to research the breadth

[2] Ullmann (2015a) contains a detailed description of the data set generation process.

[3] The BAWE corpus was developed at the Universities of Warwick, Reading, and Oxford Brookes under the directorship of Hilary Nesi and Sheena Gardner (formerly from the Centre for Applied Linguistics [previously called CELTE], Warwick), Paul Thompson (Department of Applied Linguistics, Reading) and Paul Wickens (Westminster Institute of Education, Oxford Brookes), with funding from ESRC (RES-000-23-0800). The corpus is available from the University of Oxford text archive: http://ota.ahds.ac.uk/headers/2539.xml.

quality of reflective writings (Bell et al. 2011; Fund et al. 2002; Hamann 2002; Plack and Greenberg 2005; Poldner et al. 2014; Wong et al. 1995). Therefore, the decision was made to choose single sentences as the unit of analysis. An added benefit of using sentences as the unit of analysis is that software can be used to automatically split texts into sentence units. The unit of analysis of sentences was also chosen for the depth category, although the levels of reflection are often assessed while considering the whole text (Fischer et al. 2011; Ip et al. 2012; Kember et al. 2008; Lai and Calandra 2010; O'Connell and Dyment 2004; Sumsion and Fleet 1996; Wald et al. 2012; Williams et al. 2002; Wong et al. 1995). The reason for this decision lies in the standardization of the experiment. Using the same unit of analysis for all categories simplifies their comparisons. The use of a smaller unit has the additional benefit that they can be aggregated on the level of the whole text. However, one of the drawbacks of using a sentence-based unit is that some of the meanings that stem from the wider context in which the sentence is embedded are not captured. Another drawback is that a sentence can consist of several meaningful parts and thus a smaller unit may be more useful.

A sentence splitter divided all texts of the collection (approximately 130,000 words) into sentences, and duplicated sentences and very short character strings were removed. Lastly, some of the sentences were used as qualifier questions for the coders, leaving a total of 5080 sentences (116,633 words).

Annotation

In the annotation step, all sentences were annotated with the categories of the reflection detection model. For this research, a crowd-worker platform[4] distributed the annotation task to thousands of workers, who received payment for their work. Their task was to rate sentences per eight questions. Each question represents an operationalization of one of the categories of the reflection detection model (see model section above). In other words, the operationalizations are indicators of the categories of the reflection detection model. For example, the operationalization "The writer describes an experience he or she had in the past" is indicative for the category *description of an experience* (see Table 1). Another example is the *outcome* category, which was covered with the two indicators "The writer has learned something", and "The writer intends to do something". The first indicator captures past outcomes by looking retrospectively back on outcomes, while the second indicator considers any future intent described by the writer. Table 1 contains the mapping between the categories of the model for reflection detection and the indicator questions that were used to capture the category. The words in parentheses are used as references to these indicators in the following text.

Table 2 shows several sentences from the text collection and their category label. These examples have been chosen from sentences of the data sets that have been agreed by all coders to represent the presence of a category. We chose two examples for each category. A sentence can have several labels.

[4] https://www.figure-eight.com/ (before it was http://www.crowdflower.com/)

Table 1 Indicators of the categories of the reflection detection model

Category	Indicator
1. Reflection	The sentence is descriptive … reflective (*Reflection*)
2. Description of an experience	The writer describes an experience he or she had in the past (*Experience*)
3. Feelings	The writer describes his or her feelings (*Feeling*)
4. Personal belief	The writer describes his or her beliefs (*Belief*)
5. Awareness of difficulties – critical stance	The writer recognizes difficulties/problems (*Difficulty*)
6. Perspective	The writer takes into account another perspective (*Perspective*)
7. Outcome – Lessons learned	The writer has learned something (*Learning*)
8. Outcome – Future intention	The writer intends to do something (*Intention*)

Pre-Processing

The same setup for pre-processing the data and to train and test the machine learning algorithms was applied to all data sets. The pre-processing step transformed the labeled data sets into data sets suitable for the machine learning algorithms. Important are the steps of feature construction and feature selection.[5] The choice made about features was to only use textual features in form of unigrams represented as a set of binary values (Sebastiani 2002). Although other representations are possible (Blake 2011; Brank et al. 2011), this research used a simple unigram set representation to estimate performance without using more complex features. The rationale was that if the machine learning models with simple features already show enough signal, then we can expect better results with more sophisticated features. The performance shown in the evaluation therefore represents a lower baseline that can be extended with feature engineering.

The extraction of the features from the texts produced many features. Feature selection aims at reducing the number of features to remove less informative features or features that introduce noise (Manning et al. 2008). There are many feature selection methods (Forman 2003; Mladenic 2011). In this study, we removed features such as punctuation, numbers, and spurious white space and all features that occurred fewer than ten times in the whole of the data set. The R text mining package tm (Feinerer and Hornik 2014; Feinerer et al. 2008) pre-processed the texts. After pre-processing, all data sets had the form of labeled feature vectors.

Training and Test Set

After the pre-processing of the data sets, they had the format required by the machine learning algorithms. We randomly divided each data set into a larger training data set (80%) that was used to train the machine learning algorithms and a smaller test data set (20%), which was used to assess the performance of the machine learning models derived from the training set. The test data set consists of novel/unseen instances, and this is a common best practice setup (Sebastiani 2002).

[5] Feature selection can also occur during the machine learning phase (Mladenic 2011). For more information, see the section about the features of the machine learning models.

Table 2 Annotated example sentences (two for each category)

Category	Examples
Reflection	I understood that, as sister, Jane needed to control the situation but I couldn't help wondering if a different approach would have brought about a preferable outcome.
	Yet within this profession it's not possible to work on my own and so it has helped me to try and improve my skills and confidence of working within a group and learning to listen to other people's opinions.
Experience	Because I had spent a lot of time with this class, I already had an idea of their abilities and I noticed that the same people always put their hands up.
	It kept coming back in my mind and over the next few days - I begun to think of the situation in lots of different ways.
Feeling	Although I was a bit anxious about the lesson, I kept on top of the feelings.
	But the rewarding feeling that all the work is worthwhile and the confidence that I'm better armed for the IT wars is priceless.
Belief	I must protect the interests of my participant, and guard them from harm.
	I decided this would be appropriate as I felt the actual content of the essay is more important than repeating myself in a summary.
Difficulty	Motivating the team was an interesting problem - how to encourage a group of first year students that instead of going down the pub, they should get together and discuss building a board game?
	There are still lots of areas that are greatly lacking, some for lack of resources, some because they should have been done earlier in the project and some I just hadn't used.
Perspective	Discussion with Jane revealed that she felt she received very little preparation or health promotion regarding breastfeeding within the antenatal period as the topic was discussed very briefly, there was no mention of the importance of skin to skin and its relation to successful breastfeeding -UNICEF, 11/11/04
	I decided to situate the interview in my living room; with no-one else was at home, hoping my subject would felt more comfortable.
Learning	Overall, I have learnt not just how to work with teams, but work well with teams; I have learnt how to manage a group of people, without alienating any of them.
	Having contact with the students on a daily basis, gave me some knowledge of other people's cultures.
Intention	I need to think about these things and form them into clear questions so that I can find out what I need to know and how to put it on paper in an acceptable way.
	By identifying and analysing my lack of self-confidence, and associated low self-esteem, I hope to be able to develop an action plan for my future practice in stage 2.

Class imbalance can be a problem for machine learning algorithms (Chawla et al. 2004). Several techniques exist to counterbalance this problem (Chawla 2005; Chawla et al. 2004; Menardi and Torelli 2012). Here, we use random oversampling on the training data as it has shown positive effects for data sets with class imbalance (Batista et al. 2004; Japkowicz and Stephen 2002). Random oversampling is a technique with which the minority class is randomly repeated until it matches the number of instances of the majority class. The test data set remained with the original class distribution to retrieve a realistic test performance.

We determined the best candidate model from the training data set with k-fold cross validation ($k = 10$) as the resampling technique (Kim 2009; Molinaro et al. 2005).

Resampling was also used to determine the tuning parameters of the machine learning algorithms.

Machine Learning Algorithms

Aggarwal and Zhai (2012) identified Support Vector Machines (SVM), Neural Networks, and Naïve Bayes classifiers as key text classification methods. Fernández-Delgado et al. (2014) found that Random Forests showed good performance on several data sets. They are therefore good-candidate machine learning algorithms to be evaluated on the problem of the automated detection of reflection in texts.

The R (R Core Team 2014) caret package developed by Kuhn et al. (2014) provided all machine learning algorithms used in this paper such as the implementations of the SVM (Hornik et al. 2006; Joachims 1998; Karatzoglou et al. 2004), Neural Networks (Venables and Ripley 2002), Random Forests (Breiman 2001; Liaw and Wiener 2002), and Naïve Bayes (Meyer et al. 2014; Weihs et al. 2005) algorithms.

Benchmarks

The performance of the machine learning algorithms was determined by comparing the prediction of the machine learning algorithm with the annotations generated by the coders. There exist several proposals about how to benchmark inter-rater reliability (Fleiss et al. 2004; Krippendorff 2012; Landis and Koch 1977; Stemler and Tsai 2008). A benchmark is a recommendation about acceptable levels of inter-rater reliability. For example, Landis and Koch (1977) defined Cohen's κ values below 0 as poor, between 0 and 0.20 as slight, between 0.21 and 0.4 as fair, between 0.41 and 0.6 as moderate, between 0.61 and 0.80 as substantial, and above 0.81 as almost perfect inter-rater reliability. Stemler and Tsai (2008) recommended for exploratory research a threshold of a Cohen's κ of 0.5. These thresholds provide guidance, but it is up to the research community to define acceptable levels for the context/practice in question. A high-stakes context requires stricter guidelines, while for a low-stakes context, a more lenient standard may suffice.

Results

Reliability and validity are important quality criteria of research (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education 2014). Therefore, the results section first shows that a) the annotation process of the data sets is reliable and b) that the model of reflection detection shows validity. Therefore, the machine learning algorithms are based on a theoretical model of common constituents of reflective writing models that has face validity and empirical validity and can be reliably annotated. The section afterward shows the performance estimates of the machine learning algorithms for each category of the reflection detection model. The last section inspects the most important features that the machine learning algorithms used to predict the class labels.

Reliability of the Manual Annotation

In the context of the manual content analysis, inter-rater reliability is usually calculated on an individual level and are usually calculated between two or three coders. In the context of crowdsourcing,[6] the inter-rater reliability is often calculated based on aggregated results. The underlying idea is that many (redundant) annotations can help compensate for noise, leading to high-quality annotations. A common aggregation strategy is using majority voting, which is a simple although not always optimal strategy (Li et al. 2013). This research uses several types of majority voting to aggregate votes. Also, the researcher must decide which type of majority voting best suits the research. Using a form of supermajority means applying a strict standard to the data compared to the simple majority, as only instances enter the data set where a supermajority of coders agreed. Simple majority voting needs less agreement than a supermajority voting; therefore, represents a less strict criterion. The benefit of a strict criterion is that it ensures to only include instances that have high agreement by many coders that represent the category. The downside of choosing a strict standard is that fewer instances meet the criterion, resulting in fewer instances to train the machine learning algorithms. This study considers simple majority and two-thirds and four-fifths supermajorities.

The process is as follows: To estimate the inter-rater reliability of the crowdsourced annotations, we randomly split the annotations for each unit of analysis into two groups. A form of majority voting (simple majority voting or a supermajority) determines the final annotation of the group. The inter-rater reliability is then calculated based on the aggregated ratings of both groups over all units. The random assignment to groups was repeated 50 times to compensate for grouping related effects. The reported inter-rater reliability of manually annotated data sets is the mean inter-rater reliability of these 50 random repetitions, and each group must consist of a minimum of four ratings. This criterion ensures that the voting is based on a group of coders and not only on the ratings of a single coder or only a few coders; consequently, the sample size is lower than the actual size of the training data set, as not all randomly sampled groups fulfill this criterion.

This approach allows to the various levels of inter-coder reliability to be compared with the corresponding performance of the machine learning algorithms to assess the order of magnitude of the performance difference and to estimate the machine learning performance from the manual performance.

Table 3 shows the estimates of the aggregated inter-rater reliability of the manual annotation of the sentences for each category of the reflection detection model. The table shows the accuracy and the Cohen's κ for three majorities, the two supermajorities (four-fifths and two-thirds majority), and the simple majority that were used to determine the class label of each sentence.

Cohen's κ, estimated from the aggregated ratings agreed by four-fifths of the ratings, is almost perfect for all but one category according to the benchmark of Landis and Koch (1977). The exception is the *Perspective* category, which is substantial. The estimates based on the two-thirds majority are all at least substantial. The same applies to the reliability estimates derived with the simple majority except for the *Perspective*

[6] For details, see Ullmann (2015a).

Table 3 Inter-rater reliability of the annotation task

Majority	Category	N	Accuracy	Cohens' κ
Four-fifths	Reflection	447	0.96	0.90
	Experience	624	0.99	0.98
	Feeling	454	0.98	0.94
	Belief	347	0.97	0.93
	Difficulty	477	0.98	0.96
	Perspective	336	0.96	0.78
	Learning	298	0.96	0.90
	Intention	724	0.99	0.95
Two-thirds	Reflection	851	0.89	0.74
	Experience	938	0.96	0.92
	Feeling	766	0.92	0.80
	Belief	682	0.89	0.78
	Difficulty	810	0.92	0.85
	Perspective	671	0.87	0.60
	Learning	623	0.86	0.68
	Intention	989	0.97	0.83
Simple	Reflection	1115	0.83	0.62
	Experience	1072	0.93	0.86
	Feeling	933	0.87	0.69
	Belief	887	0.83	0.66
	Difficulty	997	0.87	0.74
	Perspective	910	0.80	0.48
	Learning	852	0.80	0.55
	Intention	1094	0.94	0.72

and *Learning* categories, which are moderate. Overall, the measures show that the annotation process produced labels that the coding groups consistently agreed.

Validity of the Manual Annotated Data Set

The model for reflection detection postulates that the breadth dimensions of reflection are associated with the depth dimension. The analysis of validity uses the Fischer's exact test to investigate whether the breadth dimensions of reflection are independent of the depth dimension of reflection. The assumption is that each individual breadth category is associated with the category *Reflection*. In contrast, no relation between the depth and each breadth category would be a counterfactual against the model validity.

Table 4 shows the results of Fisher's exact test (R Core Team 2014) between the category *Reflection* and each breadth category of the model for reflection detection for three variations of the voting technique (four-fifths majority, two-thirds majority, and simple majority).

Table 4 Fisher's exact test of independence between Reflection and all breadth categories

Majority	Category	p	Odds ratio	Lower 95% CI	Upper 95% CI
Four-fifths	Experience	< .001	30.38	22.09	42.47
	Feeling	< .001	1391.42	724.42	2915.3
	Belief	< .001	360.3	189.58	764.52
	Difficulty	< .001	15.85	11.85	21.4
	Perspective	< .001	6.42	4.42	9.36
	Learning	< .001	79.49	52.22	124
	Intention	.04	1.47	1.01	2.13
Two-thirds	Experience	< .001	10.75	8.91	13.01
	Feeling	< .001	121.34	91.37	162.94
	Belief	< .001	43.39	32.89	58.14
	Difficulty	< .001	6.7	5.58	8.07
	Perspective	< .001	3.18	2.6	3.9
	Learning	< .001	17.56	14.2	21.77
	Intention	< .001	1.74	1.36	2.22
Simple	Experience	< .001	5.72	4.97	6.59
	Feeling	< .001	20.96	17.79	24.74
	Belief	< .001	10.67	9.07	12.6
	Difficulty	< .001	3.91	3.41	4.49
	Perspective	< .001	1.88	1.64	2.17
	Learning	< .001	5.93	5.15	6.83
	Intention	< .001	1.64	1.37	1.97

Table 4 shows that all breadth dimensions (from *Experience* to *Intention*) are likely not independent from Reflection ($p < 0.05$). Consequently, there is a high likelihood that there is a relation between the depth category and all breadth categories of reflection. These results corroborate the validity of the model for reflection detection. *Reflection* has the highest odds ratio to co-occur with the *Feeling*, *Belief*, *Learning*, *Experience*, and *Difficulty* categories and the lowest odds ratio with *Perspective* and *Intention* categories.

Machine Learning Training and Test Data Sets

Table 5 shows the amount (*N*) of training and test instances of the data sets for each category and each majority. The table shows the split of training and test instances by the two classes, present and absent. An instance (sentence) was either rated by the coders as being an example of the presence of a characteristic (e.g., is reflective) or absence (e.g., is descriptive/non-reflective). Also, 80% of the data were used for training and 20% were used for testing. The training and test data show the original class distribution. Also, the models were tested on the test data with the original class distribution.

Table 5 Training and testing instances

Majority	Category	Training instances			Test instances		
		N	Present	Absent	N	Present	Absent
Four-fifths	Reflection	2018	522	1496	504	130	374
	Experience	2701	1200	1501	675	300	375
	Feeling	2185	624	1561	545	155	390
	Belief	1900	916	984	473	228	245
	Difficulty	2187	1071	1116	546	267	279
	Perspective	1683	251	1432	420	62	358
	Learning	1563	544	1019	389	135	254
	Intention	2984	274	2710	745	68	677
Two-thirds	Reflection	2915	863	2052	728	215	513
	Experience	3367	1526	1841	841	381	460
	Feeling	3084	960	2124	770	240	530
	Belief	2949	1512	1437	736	377	359
	Difficulty	3039	1495	1544	759	373	386
	Perspective	2715	600	2115	678	150	528
	Learning	2690	930	1760	672	232	440
	Intention	3485	378	3107	870	94	776
Simple	Reflection	3671	1211	2460	917	302	615
	Experience	3765	1725	2040	941	431	510
	Feeling	3748	1246	2502	936	311	625
	Belief	3721	1943	1778	929	485	444
	Difficulty	3696	1820	1876	923	454	469
	Perspective	3655	1032	2623	913	258	655
	Learning	3659	2286	1205	914	343	571
	Intention	3799	493	3306	949	123	826

Performance of the Machine Learning Models

Table 6 shows the performance of the machine learning models assessed on the unseen test data for each category of the reflection detection model and for each majority condition. The category *Experience* had the highest performance over all three conditions. The Cohen's κ values were 0.85 (for the four-fifths condition), 0.78 (two-thirds), and 0.75 (simple majority). In all these three conditions, the best performing machine learning algorithm on the training data was the Random Forests algorithm trained either on the oversampled data set (overs.) or on the data set with the original class distribution (orig.). The category *Feeling* had the second highest Cohen's κ values, followed by *Reflection*, *Intention*, *Belief*, *Difficulty*, and *Learning*. The category with the lowest performance is *Perspective*, with Cohen's κ of 0.53 (four-fifths majority), 0.38 (two-thirds majority), and 0.30 (simple majority). The category with the highest accuracy

Table 6 Performance of machine learning algorithms

Majority	Category	ML model	N	Accuracy	Cohens' κ
Four-fifths	Reflection	Random Forests (overs.)	504	0.89	0.71
	Experience	Random Forests (overs.)	675	0.92	0.85
	Feeling	Random Forests (orig.)	545	0.90	0.74
	Belief	Neural Networks (orig.)	473	0.84	0.67
	Difficulty	Naive Bayes (overs.)	546	0.80	0.60
	Perspective	Naive Bayes (orig.)	420	0.88	0.53
	Learning	SVM (radial, overs.)	389	0.81	0.56
	Intention	Random Forests (overs.)	745	0.96	0.70
Two-thirds	Reflection	Naive Bayes (overs.)	728	0.84	0.62
	Experience	Random Forests (orig.)	841	0.89	0.78
	Feeling	Random Forests (overs.)	770	0.82	0.61
	Belief	SVM (poly, orig.)	736	0.78	0.57
	Difficulty	Random Forests (orig.)	759	0.79	0.58
	Perspective	Naive Bayes (overs.)	678	0.77	0.38
	Learning	Naive Bayes (orig.)	672	0.79	0.53
	Intention	Random Forests (overs.)	870	0.94	0.66
Simple	Reflection	Random Forests (overs.)	917	0.79	0.53
	Experience	Random Forests (orig.)	941	0.87	0.75
	Feeling	Random Forests (orig.)	936	0.80	0.53
	Belief	Random Forests (overs.)	929	0.74	0.48
	Difficulty	Random Forests (orig.)	923	0.73	0.45
	Perspective	Naive Bayes (orig.)	913	0.72	0.30
	Learning	Random Forests (overs.)	914	0.71	0.36
	Intention	SVM (linear, overs.)	949	0.87	0.47

over all three conditions is *Intention*, followed by *Experience*, *Feeling*, *Reflection*, *Belief*, *Difficulty*, *Learning*, and *Perspective*.

We can compare the accuracy with a baseline. The baseline accuracy is the accuracy of an algorithm that always predicts the class with the most instances as correct. The machine prediction should be better than this very simple algorithm that always predicts the majority class as true. The baseline accuracy can be calculated from the class distribution of the test data and is not shown here. All accuracies achieved by the machine learning algorithms are above the baseline accuracy except for the *Perspective* category for the simple and two-thirds majority.

We can compare the values with a baseline and with established benchmarks. For the four-fifths majority condition, the Cohen's κ values of all categories are above the 0.5 exploratory research threshold of Stemler and Tsai (2008). All categories have a Cohen's κ that is fair or better according to the benchmark of Landis and Koch (1977). Also, almost perfect is the *Experience* category in the four-fifths majority condition, followed by the categories with substantial reliability *Feeling*, *Reflection*, *Intention*, and *Belief*. *Difficulty*, *Learning*, and *Perspective* can be benchmarked as moderate.

Overall, we expect that the machine learning algorithms' performance to be below the performance of the manual coders. By averaging the performance values of all categories by each of the three conditions, the average machine performance accuracy was 8 % below the average manual annotation accuracy for the simple and two-thirds majority condition and 10 % below the average manual annotation accuracy for the four-fifths majority vote conditions. Cohen's κ was 18 points lower in the simple and two-thirds majority condition and 25 points lower for the four-fifths majority condition. In most cases, the Cohen's κ values of the machine learning models is one benchmark level lower than the level reported for the manual inter-rater reliability.

A simple linear regression was used to investigate the relation between machine learning performance and the rater performance. The inspection of the scatterplots (not shown here) between the machine learning and coder accuracy as well as machine learning Cohen's κ values and coder Cohen's κ values showed a strong positive correlation, which was confirmed with a significant Pearson's correlation (for accuracy: $r(22) = .84$, $p < .001$ and for Cohen's κ: $r(22) = .88$, $p < .001$). The linear regression indicated that the manual performance explained 71% of variance for accuracy ($F(1,22) = 54.07$, $p < .001$) and 77% of variance for Cohen's κ ($F(1,22) = 71.92$, $p < .001$). The results show a significant relationship between the performance of the machine learning algorithms and rater performance. The coefficient for the coder accuracy was 0.97 ($p < 0.001$) and for the coder Cohen's κ was 0.86 ($p < .001$).

Another interesting finding illustrated in Table 6 is that there is not a single best performing machine learning algorithm. The Random Forests algorithm achieved the highest performance for several categories but not in all categories. The highest performing models came also from the algorithms Naïve Bayes, Support Vector Machine variants (linear, radial and polynomial), and Neural Networks.

Features of the Machine Learning Models

The pre-processing section described how the texts were converted to features, in this case unigrams. The machine learning algorithms use these features for their classification. Many of the machine learning models do not use all the features as predictors. Based on the input data, they learn which of the input variables are important. We can inspect these variables to understand which features are more important and which are less important for the classification, providing insights about the inner workings of the otherwise difficult-to-examine machine learning algorithms. The inspection of these features can help to corroborate the validity of the machine learning models as often they intuitively make sense to an expert. Sometimes, the inspection reveals odd features, which may indicate errors when preparing the data set, for example, when the class information is accidentally included in the feature set. Other times, however, these features do not conform to human intuition but still achieve high performance.

The evaluative function to calculate the contribution of features to a model is different for each algorithm. Here, we use the functions provided by the R caret package developed by Kuhn et al. (2014). The variable importance for each unigram can be ranked and scaled with values ranging from 0 (no contribution) to 100 (contributes the most). While the variable importance can tell us about the features that contribute most to the classification, it cannot tell us about the direction of that contribution (e.g., is a feature used as a case in favor of a classification class or against it) as these features can be used in many ways to determine the response. Also, a highly important feature is not necessarily the sole differentiator between classes. The existence of a highly important feature in text only indicates that it is more likely that the instance belongs to the class or another class. Often, it is the combination of features that guarantees high performance, and Table 7 shows the most important features of the machine learning models.

To explain the feature presented in Table 7 in the context of reflective writing, they are interpreted based on the theory of reflection as well as the existing empirical studies. This interpretative context is also used to gauge the direction of the features, as we cannot derive this information from the variable importance measure. Due to space limitations and the large overlap of the features for the other two data set conditions, the discussion is only about the features of the four-fifths majority data set condition in Table 7.

Six of the features had a scaled feature importance of 10 and above in the Random Forests model of the reflective data set. This was the data set in which each sentence had a four-fifths majority of either being reflective or non-reflective/descriptive. The singular first-person pronouns “I”, “me”, and “my” have high levels of feature importance. Several other studies highlighted the importance of self-referential messages for reflection (Birney 2012; Ullmann 2015a, b; 2017; Ullmann et al. 2012), which is congruent with this finding. In the reflective writing literature, there is a debate whether a reflective account needs to be written from a first-person perspective (Moon 2006). Although a reflective account can be written without any self-references, they tend to be personal in nature. The word “have” can be indicative of the present perfect tense, which is used to express that something in the past still has importance for the present. Considering past experiences for current learning is an important facet of reflective thinking (see definition previously presented). The subordinate conjunction “that” is often used to provide additional information to information given by the main clause of the sentence. A reflection is also often a very detailed account of experience as the topic of the reflection is often complex. The conjunction “that” would be such a device to provide additional information. The word “that” can be part of a that-clause, for example, “I thought/feel/believe that”, which are that-clauses using verbs that refer to thinking processes. It can be used by the author to explicitly refer to mental processes, which is a sign of self-awareness and important for reflection. The verb “feel” is a thinking and sensing word. According to Birney (2012), verbs of that group highly correlate with reflection. She suggests that the combination of the first-person perspective together with thinking and sensing verbs (e.g., I feel) is a linguistic device that can be used to express self-awareness. Self-awareness is an important characteristic of reflection as it is much about expressing one’s own perspective, believes, and feelings.

The most important features of the category *Experience* are words indicating various verb tenses. The words “was”, “had”, and “were” indicate the past tense, the word “have” indicates the present perfect tense, the words “is”, and “are” indicate the present tense, and “will” indicates the future tense. A reflective writing is often a recount of a past experience. The author uses the past tense to express their experience. The description of an experience in a reflective writing is less about the presence or the future. The latter tenses, however, can indicate the absence of an experience. The other top features of *Experience* are the singular first-person pronouns “I” and “me”, as well

Table 7 Feature importance

Majority	Category	ML model	Unigrams	Top unigrams (scaled feature importance from 0 to 100)
Four-fifths	Ref.	Random Forests (overs.)	658	i (100), have (17), that (16), me (14), my (12), feel (11)
	Exp.	Random Forests (overs.)	889	i (100), was (84), had (35), is (35), we (30), me (22), were (17), will (16), are (13), my (13), have (10)
	Feel.	Random Forests (orig.)	711	i (100), feel (31), felt (14), me (12), my (12)
	Bel.	Neural Networks (orig.)	615	believe (100), feel (95), better (82), i (75), might (72), think (72), felt (63), good (63), used (62), people (61), own (60), seemed (58), asked (57), should (57), could (56), me (55), would (55), therefore (54), feelings (52), can (51)
	Diff.	Random Forests (orig.)	697	not (100), to (60), i (51), that (48), but (45), have (43), it (36), difficult (25), be (22), the (22), was (21), if (20), made (20), this (19), would (19), did (18), and (16), because (16), could (16), lack (16)
	Per.	Naive Bayes (orig.)	478	that (100), to (66), and (58), be (48), as (45), with (45), of (42), they (36), the (34), this (33), it (32), was (32), would (32), have (31), not (31), i (30), a (27), but (25), felt (25), is (24)
	Learn	SVM (radial, overs.)	468	i (100), that (67), to (48), my (47), and (44), have (41), this (38), in (37), the (36), of (35), me (29), a (26), it (24), more (22), on (19), about (17), experience (17), from (17), had (15), how (15)
	Int.	Random Forests (overs.)	957	will (100), i (92), to (20), my (19), was (16), need (12), next (11), improve (10)
Two-thirds	Ref.	Naive Bayes (overs.)	983	i (100), that (37), have (33), my (31), was (29), me (26), this (23), it (21), feel (19), to (18), is (17), not (17), about (15), had (15), would (15), could (14), felt (14), but (12), a (10), are (10)
	Exp.	Random Forests (orig.)	1136	i (100), was (88), had (39), is (34), were (21), we (18), are (17), me (17), my (16), will (14), did (11), have (10)
	Feel.	Random Forests (overs.)	1044	i (100), feel (24), my (21), me (13), felt (12), it (10)
	Bel.	SVM (poly, orig.)	980	i (100), that (47), this (44), it (43), to (43), have (40), my (39), not (32), a (30), be (28), feel (28), as (27), in (27), is (27), more (22), would (21), me (20), and (18), could (16), but (15)
	Diff.	Random Forests (orig.)	995	not (100), to (62), i (59), but (44), that (39), have (35), it (28), the (23), was (23), difficult (22), may (22), be (19), if (19), with (19), would (19), need (18), did (17), due (16), however (16), this (16)
	Per.	Naive Bayes (overs.)	866	that (100), and (84), to (79), not (46), as (45), be (43), of (43), with (39), they (38), have (37), i (36), this (36), would (35), it (33), a (29), in (29), the (28), was (27), but (25), more (25)
	Learn	Naive Bayes (orig.)	864	i (100), that (63), have (51), to (48), my (46), and (38), this (34), in (33), a (31), me (31), it (28), of (27), the (27), more (19), not (19), experience (17), as (16), on (16), feel (15), had (15)

Table 7 (continued)

Majority	Category	ML model	Unigrams	Top unigrams (scaled feature importance from 0 to 100)
	Int.	Random Forests (overs.)	1150	will (100), i (86), my (26), to (20), need (17), was (17), next (11), be (10)
Simple	Ref.	Random Forests (overs.)	1215	i (100), my (17), me (16), have (15), that (14), felt (11), it (11), feel (10)
	Exp.	Random Forests (orig.)	1253	i (100), was (86), had (38), is (35), we (21), were (19), will (19), me (18), my (16), are (15), have (11)
	Feel.	Random Forests (orig.)	1247	i (100), feel (27), my (17), me (16), felt (15), it (11)
	Bel.	Random Forests (overs.)	1236	i (100), that (49), feel (45), have (33), my (29), not (28), this (28), to (28), be (27), it (26), more (26), is (24), and (20), me (20), the (20), of (19), a (18), in (17), believe (16), could (16)
	Diff.	Random Forests (orig.)	1228	not (100), to (61), i (50), but (37), that (36), have (31), it (30), however (29), be (28), difficult (26), the (26), was (26), need (23), would (23), with (21), may (20), if (19), this (19), due (17), more (17)
	Per.	Naive Bayes (orig.)	1213	that (100), to (94), and (81), as (58), of (56), in (52), the (51), be (47), have (43), not (41), with (40), a (38), they (36), more (35), would (34), it (33), i (32), this (32), for (29), but (25)
	Learn	Random Forests (overs.)	1205	i (100), have (67), that (67), me (41), my (37), to (31), this (30), and (29), it (29), in (28), more (28), the (28), a (27), learnt (24), of (23), experience (21), is (19), as (18), with (18), for (17)
	Int.	SVM (linear, overs.)	1266	i (100), will (84), my (53), to (49), was (29), be (27), this (26), in (24), and (22), need (19), future (15), improve (14), me (14), on (13), try (13), use (13), what (13), would (13), how (12), more (12)

as the third-person pronoun “we”. These pronouns can be useful in describing the agent of the experience, which can be the person writing the account, and an experience with several actors of which the writer is one.

The words with the highest variable importance for the *Feeling* category consists of the first-person pronouns “I”, “me”, and “my” and the sensing word “feel” in its present tense form and past tense form. The phrases “I feel” or “I felt” are expressions that are often used to describe something that is not entirely in the known, something about which someone is not sure about, or something for which we do not hold a firm belief yet. Such intuitions that are often encapsulated in an expression related to emotions can be the reason to reflect about something to reach greater clarity. Supporting evidence for these features also comes from Birney (2012), who found a high relationship between these thinking and sensing words and reflection.

Many of the most important features of the category *Beliefs* are thinking and sensing words, such as “believe”, “feel”, “felt”, and “think”, and the noun “feelings”. The word “believe” directly addresses the category, and the other words in this category can be used to express beliefs and personal views. A writer can explicitly use these sensing words to express that something is a personal perception and not a fact. These words can also indicate that the writer is less definitive about something and does not yet accept something as a fact. This is related to the group of words expressing tentativeness, such as “seemed”, “should”, “could”, “can”, or “would”. Another class of words that has a high variable importance include the self-referential words “me”, and “own”. Reflective writings are often about one’s own personal beliefs. The transition adverb “therefore” can be used to express an addition such as a suggestion or conclusion. Birney (2012) examined causal reasoning and explanation resources, of which “therefore” is one, and found that they are important for reflective writings. In the context of writing about own beliefs, the word “better” can be used to indicate that something could have been better (is believed to be better). Currently, it less clear how the words “people”, “asked”, and “used”, which also have high levels of feature importance, fit into the context of *Beliefs*.

The top features of the category *Difficulties* use nouns indicating problems or discrepancies, such as the words “difficult”, or “lack”. Ullmann et al. (2012) noted a discrepancy annotator for reflection that among other words uses the word “lack”. The word “but” can be used to express a contrast, which can indicate a difficulty. Birney (2012) also found contrastive transition devices, such as “but”, important. The negation “not” is often used to describe that something did not happen or that there is a lack of something, also indicating a difficulty. The words “if”, “because”, and “that” (as in “given that”) can signal the premise part of an argument. These words comprise part of the premise annotator of Ullmann et al. (2012). As outlined, the subordinate conjunction “that” can be used to add extra information to the main clause of a sentence, which can be used to specify the exact nature or context of the problem or difficulty. The verb “be” is often used in the combinations of “to be”, “should be”, “need to be”, and “would be”. This verb is related to reality and existence and can be used to emphasise that something really happened and that something really exists, such as a problem. The verb “was” is the past tense form of “be” and thus focuses more on past aspects. The first-person pronouns “I” reflects that a reflection is often about personal problems or difficulties. The third-person pronoun “it” is frequently used to refer to something previously introduced, such as the problem that has been discussed. Similarly, the word “this” can be used to refer to something that has been previously mentioned. The verb “have” can be used to form the present perfect tense to indicate the importance of past events, such as a past problem, for the current situation. The verb “have” can be used in the form of “have to”, representing that something needs to be done. This feeling may come from the perception that something must be done to overcome a problem. Often “have” is used in combination with “could” to express that something “could have” been done differently, implying that something was not done in the best way possible because of a problem. The word “could” alone can signal uncertainty or a possibility, possibly signaling a difficulty. The word “would” can express something wished for or an imagined situation. In the context of expressing difficulties or problems, the expressed wish can indicate a target state that has not yet been reached and therefore implicitly indicate a problem. The verb “made” can be used to express that someone or something caused something, which in this context can describe a problem that has been caused. The function of the word “did” can be similar to “made”. Lastly, a link between the category and the words “to”, “the”, and “and” could not be established.

Regarding *Perspective*, the feature with the highest importance is the subordinate conjunction “that”. A writer can use “that” to add extra information to the main clause of the sentence. In the context of considering the perspectives of others, this extra information can be either details about that perspective or the source of the perspective. The third-person pronouns “they” and “it” can indicate the source of the perspective, which is in the first case a group of people and for the second case a single person. Conversely, the singular first-person pronoun “I” refers to the perspective of the writer and therefore can be a negative indicator for the presence of this category, i.e., indicating its absence. The word “but” can express a contrast, which can indicate a contrasting perspective. Similarly, Birney (2012) found that contrastive devices, such as “but”, can indicate “multiple perspectives”. The negation “not” can signal that something did not happen, which in the context of perspective can mean that something was different from one’s own perspective. The verb “felt” is a thinking and sensing word. In the context of *Perspective*, this verb can either express a feeling that the writer had about the perspective of someone else or that someone else had a feeling about the writer’s perspective. In contrast to Birney (2012), feeling and sensing words such as “felt” do not play a major role for the category “multiple perspectives”, a category that is closely associated with *Perspective*. The modal verb “would” can be used to express an imagined situation or something that does not necessarily need to be actual. In this context, the writer imagines something that adds another perspective to the train of thought. However, there was no clear connection between the category and the words “and”, “with”, “to”, “as”, “of”, “a”, “the”, “be”, “was”, “is”, and “have”.

The features with the highest importance for the data set *Learning* include the personal pronouns “I”, “my”, and “me”. This is congruent with the research of Birney (2012) that found a link between the personal voice and the evidence of learning. The personal pronoun “it” has been highlighted as important as well, but the importance currently cannot be explained. In the context of *Learning*, the word “experience” can specify a learning experience or the degree of experience of a person. The words “this” can be used to refer to something specific such as a specific learning experience or to refer back to something that has been previously mentioned. The subordinate conjunction “that” can signal additional information. In this context, this can be a clarification of the learning experience or other context information. The word “how” can be used to specify the way or manner of things. In this context, “how” can possibly signal a certain know-how gained by the writer. The word “more” can signal that something is now greater or better and can be used to describe an increase in learning. The words “have” and “had” have a high feature importance in the context of *Learning*. The word “have” can be part of phrases such as “I have learned a lot”, “I have a lot to learn”, “should have tried harder”, “could have done more”, expressing a statement about learning or lessons learned. The word “had” can refer to things that the writer had done, such as a recount of a learning experience. Currently, the connection between the words “of”, “about”, “in”, “on”, “from”, “to”, “and”, “the”, and “a” with *Learning* is unclear.

In the context of *Intention*, the personal pronouns “I” and “my” are important features to express future plans of the writer. The future tense indicator word “will” can indicate what the writer “will” try or do in future. Birney (2012) and Ullmann et al. also noted the importance of future tense words. The word “was” is the past tense form of “be”. “Was” refers to past events and therefore may be an indicator for the absence of an *Intention*. The temporal word “next” can indicate the “next” opportunity to do something. The verb “improve” signals areas that the writer wants to improve or areas that need improvement. The link between the word “to” and *Intention* cannot be explicitly established for now.

Discussion and Conclusions

The aim of this research was to determine whether machine learning algorithms can be used to reliably detect reflection in texts. This was tested on a comprehensive model of reflection that has been derived from theory, and its reliability and validity were confirmed with empirical evidence. The evaluation shows compelling evidence that machine learning can be used to analyze reflection in texts.

Reliability and Validity of the Model for Reflection Detection

The reflective writing literature showed that reflection is a multi-faceted construct that often describes two qualities, depth and breadth. Our model captured both dimensions with a total of eight categories that are the common categories of 24 reflective writing models. Previous research is based on models that are specific to the researchers' context and covered fewer categories, such as the model of Cheng (2017) with four categories, Liu et al. (2018) with three categories, and Kovanović et al. (2018) with two categories of the reflection detection model. The research using a rule-based approach suggested many model categories but empirically evaluated so far only one category. Compared to previous research, this paper tested the automatability of the analysis of reflection with a comprehensive reflective writing model.

The evaluation of the quality of our model showed that the theoretical model for reflection detection is reliable and valid, and these characteristics are two important criteria that indicate the quality of the data and the model. Previous research in this area did not report any evidence of the validity of their model, and only the research of Liu et al. (2018) and Kovanović et al. (2018) reported inter-rater reliability. In our research, the inter-rater reliability estimates showed that manual coders can reliably annotate sentences according to these eight categories. The evaluation also corroborated that the model for reflection detection is not only theoretical sound, but also showed evidence of empirical validity. Fisher's exact test strongly suggested that each of the breadth dimensions of reflection relate to the level of reflection category, agreeing with the theoretical model. Most categories related strongly to *Reflection*, such as *Feeling*, *Belief*, *Learning*, *Experience*, and *Difficulty*. *Perspective* and *Intention* relate as well, but their relations are weaker. These results are similar to the Spearman's rank correlations results shown in Ullmann (2015a). Ullmann (2015a) related the weaker relation between *Reflection* and *Intention* to the explanation that the concept of reflection may be more of a concept that entails looking back at past *Experiences* than generating *Intentions*, which is more of a forward-looking concept. Furthermore, considering other *Perspectives* is a concept that goes beyond one's own *Beliefs*. *Reflection* may be more associated with this inner self-perspective (see the high odds ratio with *Belief*) than the (outer) *Perspective* of others.

Reflection Detection Performance

To make the performance of the machine learning algorithm more comparable, the data set generation process and the machine learning training and testing were conducted in the same standardized manner for all categories of the reflection detection model. This approach was chosen as it allows a better comparison of the results over an unstandardized approach of individually optimized algorithms. The machine learning algorithms have been trained and tested on large data sets of initially 5080 sentences, which came from a corpus spanning several disciplines, academic years, and grades, aiding the robustness and generalizability of the results. The evaluation confirmed both research questions and showed that a) machine learning can reliably distinguish between sentences that are *reflective* and *descriptive* and b) that machine learning can reliably distinguish sentences according to the presence or absence of the categories *experience*, *feelings*, *personal belief*, *awareness of difficulties*, *perspective*, *lessons learned*, and *intention*. For all eight categories, the accuracy was above 80%, and Cohen's κ values have been benchmarked as substantial or almost perfect for all but two categories that had moderate inter-rater reliability for the data set generated with four-fifths majority. All Cohen's κ values were above the threshold for exploratory research.

The high performance of the machine learning classifiers on the category *Experience* is similar to the high performance reported in the research of Cheng (2017) and Liu et al. (2018) and is indicative of the error rate reported by Kovanović et al. (2018). According to the ranking of categories derived from previous research and reported in the literature review, the categories with the best performance after *Experience* are *Difficulty*, *Beliefs* and *Intention*. Our research showed a reversed trend of this order with a higher performance of *Intention* over *Beliefs* and *Difficulty*. Although the value of such rankings is currently limited, as outlined in the literature review, more research in this area may enable better conclusions about which categories of reflection are more difficult to detect. In the future research section, we provide suggestions towards better comparability of research results.

Our results showed a strong positive correlation between machine learning performance and coding performance. On average, the accuracy achieved by the machine learning models was eight to 10 % lower than the estimates from the manual annotation task. Cohen's κ was on average 18 to 25 points lower than the manual inter-rater reliability. Consequently, the machine performance was one benchmark level lower than the manual performance. Overall, these results strongly suggest that the reflection can be reliably detected with machine learning algorithms.

The results can guide the decision of researchers regarding the suitability of machine learning for their specific research context. The level of reliability depends on the stakes involved, with the expectation that research with high-stakes consequences should follow strict benchmark levels. Much research, however, is not high-stakes and therefore more lenient standards can apply. This research has shown that with the outlined machine learning configuration the reliability of the machine learning models will be likely one level below the manual coding performance. Consequently, by balancing some of the accuracy, texts can be automatically analyzed at scale, given that the levels of accuracy are acceptable. Automated analysis should not be used if the manual inter-rater reliability is already at the limits of what is acceptable. Until the research on automated methods to analyze texts regarding reflection reaches maturity, it is prudent to corroborate the quality of the results from the automated analysis with other information, for example by inspecting manually a sample of the results of the analysis.

Machine Learning Algorithms

The results also show that there has been not a single best performing machine learning algorithm for all model categories. It seems that the resulting data sets from the model categories have inherently different characteristics, possibly making them more suitable for a specific algorithm. We should be aware of this when testing with a few preferred algorithms (Cheng 2017; Kovanović et al. 2018; Liu et al. 2018). The key to creating high performing machine learning models may lie in the quality of the data sets, or the art of feature creation and selection, and in the selection of the machine learning algorithm.

Feature Importance

Overall, this investigation into the top features of the machine learning models shows that many of the features make intuitive sense in the context of their category, adding evidence to the validity of the machine learning algorithms. This section also highlighted possible features that have been found as relevant for reflection in other empirical work and showed many new features that can be important to express a reflective thought. This section also shows that some of the categories had top features, which were very general and hard to make sense of, such as the categories *Difficulty*, *Perspective*, and *Learning*. A potential reason for this can be that they had the lowest Cohen's κ values, that *Perspective* and *Learning* had the smallest amount of positive training classes, and/or that Naïve Bayes and SVM generated less interpretable features compared to the Random Forests algorithm or the Neural Networks.

Generally, the top features are distinct within each category, but there is also an overlap between words and categories, such as in the singular first-person pronoun "I" which is a top feature throughout all data sets or versions of the thinking and sensing word "feel" which is a feature of several categories. The questions that guided the annotation task asked to consider the perspective of the writer (see Table 1), possibly explaining the importance of this pronoun. The importance of expressing a personal stance agrees with most theoretical models of reflection. The word "feel" as part of thinking and sensing words is also important for several categories in Birney's empirically evaluated reflective writing model (2012).

Notable is that in this data-driven approach, the machine learning algorithms with their built-in feature selection methods choose features based on the data that maximize their performance. In contrast, the approaches chosen by Birney (2012) and Ullmann et al. (2012) are expert driven. Experts decide on a set of features and put these features to a test of their importance. The advantage of the data-driven approach is that the selection of features is not influenced by the decision of the experts. The automatically derived features can inform theory after careful interpretation and may inform expert-driven experiments. A disadvantage is that some of the features are difficult to interpret, diminishing their value in advancing theory. Notwithstanding, data-driven approaches add another perspective to the prevailing expert-driven approaches (Ullmann 2015b, 2017).

Another outcome of the inspection of the most important features concerns the generalizability of the results. Machine learning models that have been trained with a specific data set can be more easily transferred to other data sets if the model features are not specific to characteristics of the original data but specific to the construct in question. Examples of such irrelevant features can be discipline specific words or words that are specific for a writer. The results showed that the most important features are specific to the construct of reflection and not specific to the data. This suggests that the reported performance was less biased by characteristics that are specific to the data, which speaks for the generalizability of the results.

Limitations and Future Research

The results must be seen in their context. This study only examined the detection of reflection in English academic student writings, and other languages or contexts have not been explored. Academic student writings are the common case of reflective writing research. However, there are other text sources described in the reflective writing literature, such as blogs or transcripts of reflective conversations. It would be worth testing the detectability of reflection with other text sources.

For this study, the chosen unit of analysis was a sentence. Other units of analysis, such as the whole text, have not been explored. The sentence level has been chosen, as is frequently found as the choice of the unit of analysis in related research. Furthermore, working with a smaller unit of analysis allows for the aggregation of smaller units to larger units. However, to which degree larger units can be reliably detected must still be determined.

The reported performance is a lower limit of the potential performance of the machine learning algorithms, as the machine learning algorithms have not been individually optimized. The reason for this approach was that this research aimed at gauging comparable results which meant that a process was developed that was the same for all algorithms and data sets. Further research can evaluate the performance gains of individually optimized machine learning algorithms. Furthermore, this research showed that we can already achieve reflection detection with simple features, such as the unigrams that we have used in this research. Future research can evaluate other feature sets. Ideally, these features will be designed with performance in mind and with the aim to better understand reflection. Research can build on previous research, but important will be to find features that are specific to reflection (Kovanović et al. 2018; McNamara et al. 2015; Moschitti and Basili 2004).

In the state-of-the-art section of this paper, the author showed that researchers have investigated several methods to analyze writings regarding reflection and outlined high-level differences between these methods. Research would greatly benefit from a fine-grained understanding of the workings of each of these methods. Currently, however, a detailed comparative empirical evaluation of these approaches is not possible because of a lack of evaluation frameworks, model differences, and availability of tools and data sets. This paper made a start towards this aim with its ideas of using a standardized evaluation method, its proposal of reflective writing categories that are common to many models, its focus on model validity, and its reliability.

This research extends the method repertoire of the automated detection of reflection adding an extensive study of the less researched machine-learning-based approach. It showed that machine learning can be used to reliably classify text segments according to common categories of reflective writing. The benefit of this automated approach is that it allows to analyze reflective writings quicker, more frequently, and on larger scales, overcoming restrictions of the manual approach. For example, many of the questions researched with manual methods can now be analyzed automatically. This analysis can be based on the frequency counts of categories of reflection, but also on an aggregated text level (e.g. the whole text) using, for example, the mapping strategies from the breadth categories to the levels of reflection referenced in the literature review. The use of the same automated system for several studies can aid the comparability of the results as the system will rate texts always in the same way, which is not guaranteed with manual ratings. Furthermore, the automated analysis allows to repeatedly analyze large text collections. This will help to overcome limitations of current research that can be characterized as single point and single group studies with relatively small sample sizes.

Besides the automated analysis of writings regarding reflection we see great potential of this technology in the automated assessment of reflection. The automated analysis of reflection is a substantial part of the assessment of reflective writings, but assessment has a wider scope including areas such as feedback mechanisms and educational assessment quality standards. Further research will be necessary to better understand the usefulness of this technology for assessment. For example, automatically annotated writings can serve tutors as a second opinion potentially improving assessment accuracy (Winkler and Clemen 2004). Another example is systems that provide automatically feedback for students. Research started to investigate benefits of displaying annotations generated by a rule-based system in the context of an automated reflective writing analytics software (Gibson et al. 2017; Lucas et al. 2018). The approach of this paper can be used in a similar way to annotate texts regarding the categories of reflection. In the context of assessment and feedback, the requirements regarding the accuracy of the automated systems for reflection are high-stakes, which stresses the importance of research into the reliability of reflection detection. Besides reliability, construct validity will be another important criterion. As reflection is a complex construct an automated assessment system also has to reflect this complexity. This research has shown that the automated detection of reflection is possible for a wide range of facets of reflection, however, it has also shown that not all categories perform equally well, which calls for further research. This research focused especially on the two principles of reliability and validity because they are the two most important standards that help to understand the quality of the automated analysis. The other important principle is fairness, which is especially important in the context of assessment. Future research should explore group differences, such as gender, culture, or fluency in a language. This research would help to understand whether and to which extend specificities of the process to train and test the machine learning models, including the data sets produce a bias that may undermine the fairness of the results.

References

Aggarwal, C. C., & Zhai, C. (2012). A survey of text classification algorithms. *Mining text data*, 163–222. https://doi.org/10.1007/978-1-4614-3223-4_6.

Ait-Mokhtar, S., Chanod, J.-P., & Roux, C. (2002). Robustness beyond shallowness: Incremental deep parsing. *Natural Language Engineering, 8*(2–3). https://doi.org/10.1017/S1351324902002887.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, D.C.: American Educational Research Association.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V.2. *The Journal of Technology, Learning and Assessment, 4*(3). https://ejournals.bc.edu/ojs/index.php/jtla/article/view/1650. Accessed 5 Feb 2019.

Ballard, K. K. (2006). *Using Van Manen's model to assess levels of reflectivity among preservice physical education teachers*. Texas A&M University. Retrieved from http://hdl.handle.net/1969.1/4373. Accessed 5 Feb 2019.

Batista, G. E., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM Sigkdd Explorations Newsletter, 6*(1), 20–29.

Bell, A., Kelton, J., McDonagh, N., Mladenovic, R., & Morrison, K. (2011). A critical evaluation of the usefulness of a coding scheme to categorise levels of reflective thinking. *Assessment & Evaluation in Higher Education, 36*(7), 797–815. https://doi.org/10.1080/02602938.2010.488795.

Birney, R. (2012). *Reflective writing: Quantitative assessment and identification of linguistic features*. Waterford: Waterford Institute of Technology. Retrieved from http://repository.wit.ie/2658/. Accessed 29 Jan 2015.

Blake, C. (2011). Text mining. *Annual Review of Information Science and Technology, 45*(1), 121–155. https://doi.org/10.1002/aris.2011.1440450110.

Boenink, A. D., Oderwald, A. K., De Jonge, P., Van Tilburg, W., & Smal, J. A. (2004). Assessing student reflection in medical practice. The development of an observer-rated instrument: Reliability, validity and initial experiences. *Medical Education, 38*(4), 368–377. https://doi.org/10.1046/j.1365-2923.2004.01787.x.

Boud, D., Keogh, R., & Walker, D. (1985). *Reflection: Turning experience into learning*. Oxford: RoutledgeFalmer.

Boyd, E. M., & Fales, A. W. (1983). Reflective learning. *Journal of Humanistic Psychology, 23*(2), 99–117. https://doi.org/10.1177/0022167883232011.

Brank, J., Mladenic, D., & Grobelnik, M. (2011). Feature construction in text mining. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of machine learning* (pp. 397–401). Boston: Springer US. https://doi.org/10.1007/978-0-387-30164-8_303.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324.

Bruno, A., Galuppo, L., & Gilardi, S. (2011). Evaluating the reflexive practices in a learning experience. *European Journal of Psychology of Education, 26*, 527–543. https://doi.org/10.1007/s10212-011-0061-x.

Buckingham Shum, S., Sándor, Á., Goldsmith, R., Wang, X., Bass, R., & McWilliams, M. (2016). 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* (pp. 213–222). New York: ACM. https://doi.org/10.1145/2883851.2883955.

Buckingham Shum, S., Sándor, Á., Goldsmith, R., Bass, R., & McWilliams, M. (2017). Towards reflective writing analytics: Rationale, methodology and preliminary results. *Journal of Learning Analytics, 4*(1), 58–84. https://doi.org/10.18608/jla.2017.41.5.

Chamoso, J. M., & Cáceres, M. J. (2009). Analysis of the reflections of student-teachers of mathematics when working with learning portfolios in Spanish university classrooms. *Teaching and Teacher Education, 25*(1), 198–206. https://doi.org/10.1016/j.tate.2008.09.007.

Chang, C.-C., & Chou, P.-N. (2011). Effects of reflection category and reflection quality on learning outcomes during web-based portfolio assessment process: A case study of high school students in computer application courses. *The Turkish Online Journal of Educational Technology, 10*(3).

Chang, M.-M., & Lin, M.-C. (2014). The effect of reflective learning e-journals on reading comprehension and communication in language learning. *Computers & Education, 71*, 124–132. https://doi.org/10.1016/j.compedu.2013.09.023.

Chang, C.-C., Chen, C.-C., & Chen, Y.-H. (2012). Reflective behaviors under a web-based portfolio assessment environment for high school students in a computer course. *Computers & Education, 58*(1), 459–469. https://doi.org/10.1016/j.compedu.2011.08.023.

Chawla, N. V. (2005). Data mining for imbalanced datasets: An overview. In *Data mining and knowledge discovery handbook* (pp. 853–867). Springer. https://doi.org/10.1007/0-387-25465-X_40.

Chawla, N. V., Japkowicz, N., & Kotcz, A. (2004). Editorial: Special issue on learning from imbalanced data sets. *ACM Sigkdd Explorations Newsletter, 6*(1), 1–6.

Cheng, G. (2017). Towards an automatic classification system for supporting the development of critical reflective skills in L2 learning. *Australasian Journal of Educational Technology, 33*(4). https://doi.org/10.14742/ajet.3029.

Cheng, G., & Chau, J. (2013). An approach to identify levels of reflection using latent semantic analysis. In *2013 International Conference on IT Convergence and Security (ICITCS)* (pp. 1–3). https://doi.org/10.1109/ICITCS.2013.6717800.

Chung, C. K., & Pennebaker, J. W. (2012). Linguistic inquiry and word count (LIWC): Pronounced 'Luke' and other useful facts. In P. M. McCarthy & C. Boonthum (Eds.), *Applied natural language processing and content analysis: Advances in identification, investigation and resolution* (pp. 206–229). Hershey: Information Science Reference (an imprint of IGI Global).

Clarkeburn, H., & Kettula, K. (2011). Fairness and using reflective journals in assessment. *Teaching in Higher Education, 17*(4), 439–452. https://doi.org/10.1080/13562517.2011.641000.

Corich, S. P. (2011). *Automating the measurement of critical thinking in discussion forums*. Palmerston North: Massey University. Retrieved from http://hdl.handle.net/10179/2991. Accessed 3 Oct 2012.

Dascalu, M. (2014). Analyzing discourse and text complexity for learning and collaborating (Vol. 534). Cham: Springer International Publishing. Retrieved from https://doi.org/10.1007/978-3-319-03419-5

Dessus, P., Trausan-Matu, S., Van Rosmalen, P., & Wild, F. (2009). AIED 2009 workshops proceeedings volume 10: Natural language processing in support of learning: Metrics, Feedback and Connectivity.

Dewey, J. (1933). *How we think: A restatement of the relation of reflective thinking to the educative process*. Boston and New York: D.C. Heath and Company.

Dikli, S. (2006). An overview of automated scoring of essays. *The Journal of Technology, Learning and Assessment, 5*(1). https://ejournals.bc.edu/ojs/index.php/jtla/article/view/1640. Accessed 5 Feb 2019.

Dönmez, P., Rosé, C., Stegmann, K., Weinberger, A., & Fischer, F. (2005). Supporting CSCL with automatic corpus analysis technology. In *Proceedings of the 2005 conference on Computer support for collaborative learning: learning 2005: the next 10 years!* (pp. 125–134).

Dringus, L. P., & Ellis, T. (2005). Using data mining as a strategy for assessing asynchronous discussion forums. *Computers & Education, 45*(1), 141–160. https://doi.org/10.1016/j.compedu.2004.05.003.

Dyment, J. E., & O'Connell, T. S. (2010). The quality of reflection in student journals: A review of limiting and enabling factors. *Innovative Higher Education, 35*, 233–244. https://doi.org/10.1007/s10755-010-9143-y.

Feinerer, I., & Hornik, K. (2014). *tm: Text mining package*. Retrieved from http://CRAN.R-project.org/package=tm. Accessed 4 July 2014.

Feinerer, I., Hornik, K., & Meyer, D. (2008). Text mining infrastructure in R. *Journal of Statistical Software, 25*(5), 1–54.

Ferguson, R., & Buckingham Shum, S. (2011). Learning analytics to identify exploratory dialogue within synchronous text chat. In *Proceedings of the 1st International Conference on Learning Analytics and Knowledge* (pp. 99–103).

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research, 15*, 3133–3181.

Fischer, M. A., Haley, H.-L., Saarinen, C. L., & Chretien, K. C. (2011). Comparison of blogged and written reflections in two medicine clerkships. *Medical Education, 45*(2), 166–175. https://doi.org/10.1111/j.1365-2923.2010.03814.x.

Fleiss, J. L., Levin, B., & Paik, M. C. (2004). The measurement of interrater agreement. In *Statistical Methods for Rates and Proportions* (pp. 598–626). John Wiley & Sons, Inc. https://doi.org/10.1002/0471445428.ch18.

Fonagy, P., Target, M., Steele, H., & Steele, M. (1998). *Reflective-functioning manual, version 5.0, for application to adult attachment interviews*. London: University College London. Retrieved from http://mentalizacion.com.ar/images/notas/Reflective%20Functioning%20Manual.pdf. Accessed 6 March 2015.

Forbes, A. (2011). Evidence of learning in reflective practice: A case study of computer-assisted analysis of students' reflective blogs. *New Zealand Association for Cooperative Education 2011 Conference Proceedings*, 11–14.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research, 3*, 1289–1305.

Fund, Z., Court, D., & Kramarski, B. (2002). Construction and application of an evaluative tool to assess reflection in teacher-training courses. *Assessment & Evaluation in Higher Education, 27*(6), 485–499. https://doi.org/10.1080/0260293022000020264.

Gardner, S., & Nesi, H. (2013). A classification of genre families in university student writing. *Applied Linguistics, 34*(1), 25–52. https://doi.org/10.1093/applin/ams024.

Gašević, D., Mirriahi, N., & Dawson, S. (2014). Analytics of the effects of video use and instruction to support reflective learning. In *Proceedings of the fourth international conference on learning analytics and knowledge* (pp. 123–132). New York: ACM. https://doi.org/10.1145/2567574.2567590.

Gibson, A., Kitto, K., & Bruza, P. (2016). Towards the discovery of learner metacognition from reflective writing. *Journal of Learning Analytics, 3*(2), 22–36.

Gibson, A., Aitken, A., Sándor, Á., Buckingham Shum, S., Tsingos-Lucas, C., & Knight, S. (2017). Reflective writing analytics for actionable feedback. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference* (pp. 153–162). New York: ACM. https://doi.org/10.1145/3027385.3027436.

Gore, J. M., & Zeichner, K. M. (1991). Action research and reflective teaching in preservice teacher education: A case study from the United States. *Teaching and Teacher Education, 7*(2), 119–136. https://doi.org/10.1016/0742-051X(91)90022-H.

Gupta, V., & Lehal, G. S. (2009). A survey of text mining techniques and applications. *Journal of Emerging Technologies in Web Intelligence, 1*(1), 60–76.

Hamann, J. M. (2002). *Reflective practices and confluent educational perspectives: Three exploratory studies*. http://eric.ed.gov/?id=ED472393. Accessed 29 Jan 2015.

Hatton, N., & Smith, D. (1995). Reflection in teacher education: Towards definition and implementation. *Teaching and Teacher Education, 11*(1), 33–49. https://doi.org/10.1016/0742-051X(94)00012-U.

Hearst, M. A. (2000). The debate on automated essay grading. *IEEE Intelligent Systems and Their Applications, 15*(5), 22–37. https://doi.org/10.1109/5254.889104.

Hornik, K., Meyer, D., & Karatzoglou, A. (2006). Support vector machines in R. *Journal of Statistical Software, 15*(9), 1–28.

Hotho, A., Nürnberger, A., & Paaß, G. (2005). A brief survey of text mining. *LDV Forum, 20*, 19–62.

Houston, C. R. (2016). Do scaffolding tools improve reflective writing in professional portfolios? A content analysis of reflective writing in an advanced preparation program. *Action in Teacher Education, 38*(4), 399–409. https://doi.org/10.1080/01626620.2016.1226201.

Ip, W. Y., Lui, M. H., Chien, W. T., Lee, I. F., Lam, L. W., & Lee, D. (2012). Promoting self-reflection in clinical practice among Chinese nursing undergraduates in Hong Kong. *Contemporary Nurse, 41*(2), 253–262. https://doi.org/10.5172/conu.2012.41.2.253.

Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent Data Analysis, 6*(5), 429–449.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In C. Nédellec & C. Rouveirol (Eds.), *Machine learning: ECML-98* (pp. 137–142). Berlin Heidelberg: Springer. https://doi.org/10.1007/BFb0026683.

Jordan, S. (2014). *E-assessment for learning? Exploring the potential of computer-marked assessment and computer-generated feedback, from short-answer questions to assessment analytics*. The Open University. Retrieved from http://oro.open.ac.uk/41115/. Accessed 5 Feb 2019.

Kann, V., & Högfeldt, A.-K. (2016). Effects of a program integrating course for students of computer science and engineering. In *Proceedings of the 47th ACM technical symposium on computing science education* (pp. 510–515). New York: ACM. https://doi.org/10.1145/2839509.2844610.

Karatzoglou, A., Smola, A., Hornik, K., & Zeileis, A. (2004). Kernlab – An S4 package for kernel methods in R. *Journal of Statistical Software, 11*(9), 1–20.

Kember, D., Jones, A., Loke, A., McKay, J., Sinclair, K., Tse, H., et al. (1999). Determining the level of reflective thinking from students' written journals using a coding scheme based on the work of Mezirow. *International Journal of Lifelong Education, 18*(1), 18–30. https://doi.org/10.1080/026013799293928.

Kember, D., McKay, J., Sinclair, K., & Wong, F. K. Y. (2008). A four-category scheme for coding and assessing the level of reflection in written work. *Assessment & Evaluation in Higher Education, 33*, 369–379. https://doi.org/10.1080/02602930701293355.

Kim, J.-H. (2009). Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational Statistics & Data Analysis, 53*(11), 3735–3745. https://doi.org/10.1016/j.csda.2009.04.009.

Korthagen, F., & Vasalos, A. (2005). Levels in reflection: Core reflection as a means to enhance professional growth. *Teachers and Teaching: Theory and Practice, 11*, 47–71. https://doi.org/10.1080/1354060042000337093.

Kovanovic, V., Joksimović, S., Waters, Z., Gašević, D., Kitto, K., Hatala, M., & Siemens, G. (2016). Towards automated content analysis of discussion transcripts: A cognitive presence case. In *Proceedings of the sixth international conference on learning analytics & knowledge* (pp. 15–24). New York: ACM. https://doi.org/10.1145/2883851.2883950.

Kovanović, V., Joksimović, S., Mirriahi, N., Blaine, E., Gašević, D., Siemens, G., & Dawson, S. (2018). Understand students' self-reflections through learning analytics. In *Proceedings of the 8th international conference on learning analytics and knowledge* (pp. 389–398). New York: ACM. https://doi.org/10.1145/3170358.3170374.

Krippendorff, K. (2012). *Content analysis: An Introduction to its methodology* (3rd edn.). Thousand Oaks: Sage Publications, Inc.

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., et al. (2014). *caret: Classification and Regression Training*. Retrieved from http://CRAN.R-project.org/package=caret . Accessed 4 July 2014.

Lai, G., & Calandra, B. (2010). Examining the effects of computer-based scaffolds on novice teachers' reflective journal writing. *Etr&d-Educational Technology Research and Development, 58*(4), 421–437. https://doi.org/10.1007/s11423-009-9112-2.

Lamy, M.-N., & Goodfellow, R. (1999). 'Reflective Conversation'in the virtual language classroom. *Language Learning & Technology, 2*(2), 43–61.

Landauer, T. K. (2003). Automatic essay assessment. *Assessment in Education: Principles, Policy & Practice, 10*(3), 295–308. https://doi.org/10.1080/0969594032000148154.

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

Li, H., Yu, B., & Zhou, D. (2013). Error rate bounds in crowdsourcing models. In *ICML13 workshop: Machine learning meets crowdsourcing*. http://arxiv.org/abs/1307.2674. Accessed 12 Sept 2013.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News, 2*(3), 18–22.

Lin, F.-R., Hsieh, L.-S., & Chuang, F.-T. (2009). Discovering genres of online discussion threads via text mining. *Computers & Education, 52*(2), 481–495. https://doi.org/10.1016/j.compedu.2008.10.005.

Lin, C.-W., Lin, M.-J., Wen, C.-C., & Chu, S.-Y. (2016). A word-count approach to analyze linguistic patterns in the reflective writings of medical students. *Medical Education Online, 21*. https://doi.org/10.3402/meo.v21.29522.

Liu, Q., Zhang, S., Wang, Q., & Chen, W. (2017). Mining online discussion data for understanding teachers' reflective thinking. *IEEE Transactions on Learning Technologies, PP*(99), 1–1. https://doi.org/10.1109/TLT.2017.2708115.

Liu, Q., Zhang, S., Wang, Q., & Chen, W. (2018). Mining online discussion data for understanding teachers' reflective thinking. *IEEE Transactions on Learning Technologies, 11*(2), 243–254. https://doi.org/10.1109/TLT.2017.2708115.

Lucas, C., Gibson, A., & Buckingham Shum, S. (2018). Utilization of a novel online reflective learning tool for immediate formative feedback to assist pharmacy students' reflective writing skills. *American Journal of Pharmaceutical Education*, ajpe6800. https://doi.org/10.5688/ajpe6800.

Luk, J. (2008). Assessing teaching practicum reflections: Distinguishing discourse features of the "high" and "low" grade reports. *System, 36*(4), 624–641. https://doi.org/10.1016/j.system.2008.04.001.

Mann, K., Gordon, J., & MacLeod, A. (2007). Reflection and reflective practice in health professions education: A systematic review. *Advances in Health Sciences Education, 14*, 595–621. https://doi.org/10.1007/s10459-007-9090-2.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. New York: Cambridge University Press.

Mansvelder-Longayroux, D. D. (2006). *The learning portfolio as a tool for stimulating reflection by student teachers* (Doctoral thesis). ICLON, Leiden University Graduate School of Teaching, Leiden University. Retrieved from http://hdl.handle.net/1887/5430. Accessed 5 Feb 2019.

Mansvelder-Longayroux, D. D., Beijaard, D., & Verloop, N. (2007). The portfolio as a tool for stimulating reflection by student teachers. *Teaching and Teacher Education, 23*(1), 47–62.

McCarthy, P. M., & Boonthum-Denecke, C. (2012). *Applied natural language processing: Identification, investigation, and resolution*. Hershey: Information Science Reference.

McCollum, S. (1997). *Insights into the process of guiding reflection during an early field experience of preservice teachers*. Retrieved from http://hdl.handle.net/10919/30384. Accessed 5 Feb 2019.

McKlin, T. E. (2004). *Analyzing cognitive presence in online courses using an artificial neural network* (PhD Thesis). Atlanta: Georgia State University.

McNamara, D. S., Crossley, S. A., Roscoe, R. D., Allen, L. K., & Dai, J. (2015). A hierarchical classification approach to automated essay scoring. *Assessing Writing, 23*, 35–59. https://doi.org/10.1016/j.asw.2014.09.002.

Means, B., Toyama, Y., Murphy, R., Bakia, M., & Jones, K. (2010). *Evaluation of evidence-based practices in online learning: A meta-analysis and review of online learning studies* (monograph). https://www2.ed.gov/rschstat/eval/tech/evidence-based-practices/finalreport.pdf. Accessed 5 Feb 2019.

Mena-Marcos, J., García-Rodríguez, M.-L., & Tillema, H. (2013). Student teacher reflective writing: What does it reveal? *European Journal of Teacher Education, 36*(2), 147–163. https://doi.org/10.1080/02619768.2012.713933.

Menardi, G., & Torelli, N. (2012). Training and assessing classification rules with imbalanced data. *Data Mining and Knowledge Discovery, 28*(1), 92–122. https://doi.org/10.1007/s10618-012-0295-5.

Mergenthaler, E. (1996). Emotion–abstraction patterns in verbatim protocols: A new way of describing psychotherapeutic processes. *Journal of Consulting and Clinical Psychology, 64*(6), 1306–1315. https://doi.org/10.1037/0022-006X.64.6.1306.

Mergenthaler, E., & Kächele, H. (1991). University of Ulm: The Ulm Textbank research program. In L. E. Beutler & M. Crago (Eds.), *Psychotherapy research: An international review of programmatic studies* (pp. 219–225). Washington, DC: American Psychological Association. https://doi.org/10.1037/10092-025.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2014). *e1071: Misc functions of the Department of Statistics (e1071), TU Wien*. Retrieved from http://CRAN.R-project.org/package=e1071. Accessed 7 May 2014.

Mezirow, J. (1991). *Transformative dimensions of adult learning.* Jossey-Bass, 350 Sansome Street, San Francisco, CA 94104-1310 ($27.95).

Mladenic, D. (2011). Feature selection in text mining. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of machine learning* (pp. 406–410). Springer US. https://doi.org/10.1007/978-0-387-30164-8_307.

Molinaro, A. M., Simon, R., & Pfeiffer, R. M. (2005). Prediction error estimation: A comparison of resampling methods. *Bioinformatics, 21*(15), 3301–3307. https://doi.org/10.1093/bioinformatics/bti499.

Moon, J. A. (1999). *Reflection in learning & professional development*. London: Kogan Page.

Moon, J. A. (2004). *A handbook of reflective and experiential learning*. Oxford: RoutledgeFalmer.

Moon, J. A. (2006). *Learning journals: A handbook for reflective practice and professional development* (2nd edn.). London and New York: Routledge.

Moschitti, A., & Basili, R. (2004). Complex linguistic features for text classification: A comprehensive study. In *Advances in information retrieval* (pp. 181–196). Presented at the European Conference on Information Retrieval, Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-24752-4_14.

Nesi, H. (2007). The form, meaning and purpose of university level assessed reflective writing. In M. Edwardes (Ed.), *Proceedings of the BAAL Annual Conference*. London: Scitsiugnil Press. https://baal.org.uk/wp-content/uploads/2017/12/proceedings_07_full.pdf. Accessed 6 Feb 2019.

Nesi, H., & Gardner, S. (2012). *Genres across the disciplines: Student writing in higher education*. Cambridge: Cambridge University Press.

O'Connell, T. S., & Dyment, J. E. (2004). Journals of post secondary outdoor recreation students: The results of a content analysis. *Journal of Adventure Education & Outdoor Learning, 4*(2), 159–171. https://doi.org/10.1080/14729670485200511.

OECD. (2013). *PISA 2012 assessment and analytical framework.* Paris: OECD Publishing. https://doi.org/10.1787/9789264190511-en.

Olshtain, E., & Kupferberg, I. (1998). Reflective-narrative discourse of FL teachers exhibits professional knowledge. *Language Teaching Research, 2*(3), 185–202. https://doi.org/10.1177/136216889800200302.

Page, E. B. (1968). The use of the computer in analyzing student essays. *International Review of Education / Internationale Zeitschrift Für Erziehungswissenschaft / Revue Internationale de l'Education, 14*(2), 210–225.

Page, E. B., & Paulus, D. H. (1968). *The analysis of essays by computer. Final Report*. https://eric.ed.gov/?id=ED028633. Accessed 7 Oct 2012.

Pee, B., Woodman, T., Fry, H., & Davenport, E. S. (2002). Appraising and assessing reflection in students' writing on a structured worksheet. *Medical Education, 36*(6), 575–585. https://doi.org/10.1046/j.1365-2923.2002.01227.x.

Pennebaker, J. W., & Francis, M. E. (1996). Cognitive, emotional, and language processes in disclosure. *Cognition & Emotion, 10*(6), 601–626. https://doi.org/10.1080/026999396380079.

Plack, M. M., & Greenberg, L. (2005). The reflective practitioner: Reaching for excellence in practice. *Pediatrics, 116*(6), 1546–1552. https://doi.org/10.1542/peds.2005-0209.

Plack, M. M., Driscoll, M., Marquez, M., Cuppernull, L., Maring, J., & Greenberg, L. (2007). Assessing reflective writing on a pediatric clerkship by using a modified Bloom's taxonomy. *Ambulatory Pediatrics, 7*(4), 285–291. https://doi.org/10.1016/j.ambp.2007.04.006.

Poldner, E., Van der Schaaf, M., Simons, P. R.-J., Van Tartwijk, J., & Wijngaards, G. (2014). Assessing student teachers' reflective writing through quantitative content analysis. *European Journal of Teacher Education, 37*(3), 348–373. https://doi.org/10.1080/02619768.2014.892479.

Poom-Valickis, K., & Mathews, S. (2013). Reflecting others and own practice: An analysis of novice teachers' reflection skills. *Reflective Practice, 14*(3), 420–434. https://doi.org/10.1080/14623943.2013.767237.

Prilla, M., & Renner, B. (2014). Supporting collaborative reflection at work: A comparative case analysis. In *Proceedings of the 18th international conference on supporting group work* (pp. 182–193). New York: ACM Press. https://doi.org/10.1145/2660398.2660400.

QAA. (2012). *UK quality code for higher education. Part B: Assuring and enhancing academic quality. Chapter B3: Learning and teaching*. https://www.qaa.ac.uk/docs/qaa/quality-code/chapter-b3_-learning-and-teaching.pdf?sfvrsn=3500f781_8. Accessed 6 Feb 2019.

R Core Team. (2014). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. https://www.r-project.org/. Accessed 4 July 2014.

Reidsema, C., & Mort, P. (2009). Assessing reflective writing: Analysis of reflective writing in an engineering design course. *Journal of Academic Language and Learning, 3*(2), A117–A129.

Richardson, G., & Maltby, H. (1995). Reflection-on-practice: Enhancing student learning. *Journal of Advanced Nursing, 22*(2), 235–242. https://doi.org/10.1046/j.1365-2648.1995.22020235.x.

Richman, W. L., Kiesler, S., Weisband, S., & Drasgow, F. (1999). A meta-analytic study of social desirability distortion in computer-administered questionnaires, traditional questionnaires, and interviews. *Journal of Applied Psychology, 84*(5), 754–775. https://doi.org/10.1037/0021-9010.84.5.754.

Rivers, B. A., Whitelock, D., Richardson, J. T. E., Field, D., & Pulman, S. (2014). Functional, frustrating and full of potential: Learners' experiences of a prototype for automated essay feedback. In M. Kalz & E. Ras (Eds.), *Computer assisted assessment. Research into E-Assessment*. CAA 2014 (pp. 40–52). Springer International Publishing. https://doi.org/10.1007/978-3-319-08657-6_4.

Rosé, C. P., Wang, Y.-C., Cui, Y., Arguello, J., Stegmann, K., Weinberger, A., & Fischer, F. (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), 237–271. https://doi.org/10.1007/s11412-007-9034-0.

Ross, D. D. (1989). First steps in developing a reflective approach. *Journal of Teacher Education, 40*(2), 22–30. https://doi.org/10.1177/002248718904000205.

Ryan, M. (2011). Improving reflective writing in higher education: A social semiotic perspective. *Teaching in Higher Education, 16*(1), 99–111. https://doi.org/10.1080/13562517.2010.507311.

Ryan, M. (2012). Conceptualising and teaching discursive and performative reflection in higher education. *Studies in Continuing Education, 34*(2), 207–223. https://doi.org/10.1080/0158037X.2011.611799.

Ryan, M. (2014). Reflexive writers: Re-thinking writing development and assessment in schools. *Assessing Writing, 22*, 60–74. https://doi.org/10.1016/j.asw.2014.08.002.

Rychen, D. S., & Salganik, L. H. (2005). *The definition and selection of key competencies: Executive summary*. OECD. http://www.oecd.org/pisa/35070367.pdf. Accessed 6 Feb 2019.

Schön, D. A. (1983). *The reflective practitioner*. New York: Basic Books.

Schön, D. A. (1987). *Educating the reflective practitioner*. San Francisco: Jossey-Bass.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys, 34*(1), 1–47. https://doi.org/10.1145/505282.505283.

Shermis, M. D. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing, 20*, 53–76. https://doi.org/10.1016/j.asw.2013.04.001.

Shermis, M. D., & Burstein, J. C. (Eds.). (2003). *Automated essay scoring: A cross-disciplinary perspective*. Mahwah: Lawrence Erlbaum Associates, Inc.

Sparks-Langer, G. M., & Colto, A. B. (1991). Synthesis of research on teachers' reflective thinking. *Educational Leadership, 48*(6), 37–44.

Sparks-Langer, G. M., Simmons, J. M., Pasch, M., Colton, A., & Starko, A. (1990). Reflective pedagogical thinking: How can we promote it and measure it? *Journal of Teacher Education, 41*(5), 23–32. https://doi.org/10.1177/002248719004100504.

Stemler, S. E., & Tsai, J. (2008). Best practices in interrater reliability three common approaches. In J. Osborne (Ed.), *Best practices in quantitative methods* (pp. 29–49). Thousand Oaks: SAGE Publications, Inc. https://doi.org/10.4135/9781412995627.d5.

Stone, P. J., & Hunt, E. B. (1963). A computer approach to content analysis: Studies using the general inquirer system. In *Proceedings of the May 21–23, 1963, spring joint computer conference* (pp. 241–256). New York: ACM. https://doi.org/10.1145/1461551.1461583.

Sumsion, J., & Fleet, A. (1996). Reflection: Can we assess it? Should we assess it? *Assessment & Evaluation in Higher Education, 21*(2), 121–130. https://doi.org/10.1080/0260293960210202.

Thorpe, K. (2004). Reflective learning journals: From concept to practice. *Reflective Practice, 5*(3), 327–343. https://doi.org/10.1080/1462394042000270655.

Tsangaridou, N., & O'Sullivan, M. (1994). Using pedagogical reflective strategies to enhance reflection among pre service physical education teachers. *Journal of Teaching in Physical Education, 14*, 13–33. http://hdl.handle.net/10344/3200. Accessed 6 Feb 2019.

Ullmann, T. D. (2011). An architecture for the automated detection of textual indicators of reflection. In W. Reinhardt, T. D. Ullmann, P. Scott, V. Pammer, O. Conlan, & A. Berlanga (Eds.), *Proceedings of the 1st European workshop on awareness and reflection in learning networks* (pp. 138–151). Presented at the 6th European Conference on Technology Enhanced Learning: Towards Ubiquitous Learning 2011, Palermo, Italy: CEUR-WS.org. http://ceur-ws.org/Vol-790/paper14.pdf. Accessed 6 Feb 2019.

Ullmann, T. D. (2015a). *Automated detection of reflection in texts. A machine learning based approach* (PhD thesis). The Open University. Retrieved from http://oro.open.ac.uk/45402/. Accessed 6 Feb 2019.

Ullmann, T. D. (2015b). 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* (Vol. 1465, pp. 83–96). Presented at the 10th European Conference on Technology Enhanced Learning: Design for Teaching and Learning in a Networked World, Toledo, Spain: CEUR-WS.org. http://ceur-ws.org/Vol-1465/paper8.pdf. Accessed 6 Feb 2019.

Ullmann, T. D. (2017). Reflective writing analytics - empirically determined keywords of written reflection. In *In Proceedings of the 7th international conference on learning analytics & knowledge*. Vancouver: ACM.

Ullmann, T. D., Wild, F., & Scott, P. (2012). Comparing automatically detected reflective texts with human judgements. In A. Moore, V. Pammer, L. Pannese, M. Prilla, K. Rajagopal, W. Reinhardt, et al. (Eds.), *2nd workshop on awareness and reflection in technology-enhanced learning*. Presented at the 7th European Conference on Technology-Enhanced Learning, Saarbruecken, Germany: CEUR-WS.org. http://ceur-ws.org/Vol-931/paper8.pdf. Accessed 6 Feb 2019.

Ullmann, T. D., Wild, F., & Scott, P. (2013). Reflection - quantifying a rare good. In M. Kravcik, B. R. Krogstie, A. Moore, V. Pammer, L. Pannese, M. Prilla, et al. (Eds.), *Proceedings of the 3rd workshop on awareness and reflection in technology-enhanced learning* (pp. 29–40). Presented at the 8th European Conference on Technology Enhanced Learning: Scaling up learning for sustained impact, Paphos, Cyprus: CEUR-WS.org. http://ceur-ws.org/Vol-1103/paper2.pdf. Accessed 6 Feb 2019.

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S* (4th edn.). New York: Springer. https://www.springer.com/gp/book/9780387954578. Accessed 6 Feb 2019.

Wald, H. S., Borkan, J. M., Taylor, J. S., Anthony, D., & Reis, S. P. (2012). Fostering and evaluating reflective capacity in medical education: Developing the REFLECT rubric for assessing reflective writing. *Academic Medicine, 87*(1), 41–50. https://doi.org/10.1097/ACM.0b013e31823b55fa.

Wallman, A., Lindblad, A. K., Hall, S., Lundmark, A., & Ring, L. (2008). A categorization scheme for assessing pharmacy students' levels of reflection during internships. *American Journal of Pharmaceutical Education, 72*(1), 05.

Weber, R. P. (1990). *Basic content analysis* (2nd revised edition edition). Newbury Park: SAGE Publications, Inc.

Weihs, C., Ligges, U., Luebke, K., & Raabe, N. (2005). klaR analyzing German business cycles. In D. Baier, R. Decker, & L. Schmidt-Thieme (Eds.), *Data analysis and decision support* (pp. 335–343). Berlin: Springer-Verlag.

Wharton, S. (2012). Presenting a united front: Assessed reflective writing on a group experience. *Reflective Practice, 13*(4), 489–501. https://doi.org/10.1080/14623943.2012.670622.

Wild, F., Stahl, C., Stermsek, G., & Neumann, G. (2005). Parameters driving effectiveness of automated essay scoring with LSA. In *Proceedings of the 9th CAA Conference*. Presented at the Computer Assisted Assessment Conference, Loughborough: Loughborough University. https://dspace.lboro.ac.uk/2134/2008. Accessed 6 Feb 2019.

Williams, R. M., Wessel, J., Gemus, M., & Foster-Seargeant, E. (2002). Journal writing to promote reflection by physical therapy students during clinical placements. *Physiotherapy Theory & Practice, 18*(1), 5–15. https://doi.org/10.1080/095939802753570657.

Wilson, J., & Czik, A. (2016). Automated essay evaluation software in English language arts classrooms: Effects on teacher feedback, student motivation, and writing quality. *Computers & Education, 100*, 94–109. https://doi.org/10.1016/j.compedu.2016.05.004.

Winkler, R. L., & Clemen, R. T. (2004). Multiple experts vs. multiple methods: Combining correlation assessments. *Decision Analysis, 1*(3), 167–176. https://doi.org/10.1287/deca.1030.0008.

Wong, F. K., Kember, D., Chung, L. Y. F., & Yan, L. (1995). Assessing the level of student reflection from reflective journals. *Journal of Advanced Nursing, 22*(1), 48–57. https://doi.org/10.1046/j.1365-2648.1995.22010048.x.

Yang, Y.-F. (2010). Students' reflection on online self-correction and peer review to improve writing. *Computers & Education, 55*(3), 1202–1210. https://doi.org/10.1016/j.compedu.2010.05.017.

Yusuff, M. A. (2011). *Intelligent blogs for reflection* (BCS Computer Science Thesis). Leeds: University of Leeds. Retrieved from http://www.comp.leeds.ac.uk/cgi-bin/fyproj/reports/1011/Yusuff.pdf.gz. Accessed 11 Oct 2011.