Education and Information Technologies (2023) 28:2961–2999
https://doi.org/10.1007/s10639-022-11254-7

Applications of deep language models for reflective writings

Jan Nehyba[1] · Michal Štefánik[2]

Received: 26 January 2022 / Accepted: 20 July 2022 / Published online: 5 September 2022

Abstract

Social sciences expose many cognitively complex, highly qualified, or fuzzy problems, whose resolution relies primarily on expert judgement rather than automated systems. One of such instances that we study in this work is a reflection analysis in the writings of student teachers. We share a hands-on experience on how these challenges can be successfully tackled in data collection for machine learning. Based on the novel deep learning architectures pre-trained for a general language understanding, we can reach an accuracy ranging from 76.56–79.37% on low-confidence samples to 97.56–100% on high confidence cases. We open-source all our resources and models, and use the models to analyse previously ungrounded hypotheses on reflection of university students. Our work provides a toolset for objective measurements of reflection in higher education writings, applicable in more than 100 other languages worldwide with a loss in accuracy measured between 0–4.2% Thanks to the outstanding accuracy of the deep models, the presented toolset allows for previously unavailable applications, such as providing semi-automated student feedback or measuring an effect of systematic changes in the educational process via the students' response.

Keywords Deep learning · Natural language processing · Reflection dataset · Reflection classification · Analyses of reflective journals · Generalized linear mixed models

✉ Jan Nehyba
nehyba@ped.muni.cz

Michal Štefánik
stefanik.m@mail.muni.cz

[1] Faculty of Education, Masaryk University, Poříčí 7/9, 639 00 Brno, Czech Republic

[2] Faculty of Informatics, Masaryk University, Botanická 68a, 602 00 Brno, Czech Republic

1 Introduction

The recent years have observed a significant rise in accuracy of machine learning methods that has moved the possibilities of automation. This trend has been particularly significant in an area of Natural Language Processing (NLP), where, thanks to the models pre-trained for a general language understanding (GLU) (Devlin et al., 2019), machines are able to outperform a performance of human annotators on tasks such as Question Answering (Lan et al., 2020) or paraphrasing (Liu et al., 2019), while on other, more complex tasks such as Machine Translation (Bahdanau et al., 2014) or cross-lingual classification (Conneau & Lample, 2019), these models have reached the before-unseen qualitative levels, which for the first time in history, makes such algorithms applicable in these use-cases without expensive expert supervision.

However, in specialized applications of social sciences, an applicability of the novel NLP technologies needs to tackle specific obstacles which we believe might be a reason for their lagging widespread deployment (Zawacki-Richter et al., 2019).

Our article focuses on a specific area of application of deep learning methods[1] in the education context, which is the identification of reflective categories in the written text, specifically for student teachers. One of the main tools of reflective practice is reflective writing, a ubiquitous tool for professional development (Section 2) whose effectiveness is evidenced by many Authors (Section 2.1). However, learning reflective writing turns out to be a complex topic and faces many challenges. For example, acquiring the skill of reflective writing requires consistent, individualized feedback, which is difficult for many students (Section 2.2). The application of machine learning methods offers solutions to some of these challenges, as faced by the use of reflective writing in the university environment. Although this may seem like a completely new connection, a significant branch of work combines various automatic analyzes and reflective writing (Section 2.3). However, these applications face some problems that we believe are characteristic for the applications of technology in social sciences.

In the case of reflection identification, we experience the problem in a form of low inner-coherence and cross-annotator coherence (Section 3.3). Such situation can be a consequence of the fact that the inner models of the problem among annotators are not compatible, or even when the model for a selected annotator is not itself consistent.

A statistical model trained on such inconsistently-annotated data has to inevitably choose between fitting a subjective, yet either the most consistent, or over-represented scheme of the specific annotator, or not being able to properly fit annotators variability at all (a situation also refered to as *underfitting*).

[1] We use the terms "machine learning" and "deep learning" algorithms. Machine learning commonly refers to a broader set of algorithms with the ability to adapt its parameters to given data automatically. Deep learning is a subset of machine learning identifying models composed of multiple layers of simple decision-making units that can together accurately model more complex, non-linear problems. Deep learning models are also referred to as "artificial neural networks".

We find that the obstacles might lay in an insufficiently distinctive definition of the reflection categories, leading to a complicated collection of consistent data. We learn to address this hypothesis in our methodology, as overviewed in Section 3.2, and formulate the problem in attempt to *maximize* the annotators consistency. Instead of the classic approaches of ensembling annotators voting or selecting only the samples with a high confidence that turn out insufficient for the situation, we avoid the non-representative samples identified by very high inconsistency, and let the annotators discuss and eventually agree on the samples where they choose its category inconsistently.

We observe that the change in data collection methodology provides a major gain in accuracy of all final models. For example, in case of a Random Forrest, the best-performing non-neural classifier, the gain represents 26.65% of accuracy. We describe a process of data collection more precisely on Section 3.3 and we publish an anonymised version of the resulting dataset freely available (Section 3.2).

We follow with defining a reflectivity extraction as a sentence classification problem in Section 3.4. We distinguish two essential groups of models that we experiment with that we denote as *shallow classifiers*, introduced in Section 3.4.1 and *deep classifiers*, introduced in Section 3.4.2. In case of the deep classifiers, we additionally perform experiments in cross-lingual settings: the classifier is trained on reflectivity identification in a selected language, but evaluated in a different language. In English-Czech cross-lingual settings, we demonstrate that our multilingual models might be applied in any other of more than 100 of its pre-trained languages, listed in Conneau et al. (2020), (see Appendix 1).

We dedicate a significant amount of attention to a seamless reproducibility of our results, to which we provide detailed instructions in the project repository[2] (Section 3.5).

In Section 3.5, we proceed with an application of the trained models to investigate some interesting and unanswered exploratory hypotheses selected from the related literature. These investigate the relation of the amount of reflection in the reflective journals to student performance (Section 3.5.1) and a development of types of reflectivity in student journals in time (Section 3.5.2).

Applying our tools for reflection identification, we perform two exploratory analyses. The first one focuses on a relation of reflective writing to a perceived performance of the students by mentors, while the latter one is concerned with a temporal development of student writing with respect to the contained reflective entities.

These analyses also demonstrate how our models can be used in further research; sources and outputs of all our analyses are publicly available (Section 3.5) and can be reproduced with a single click.

[2] See the instructions in project repository on https://github.com/EduMUNI/reflection-classification

2 Reflective practice and reflective writing of student teachers

Reflective practice is commonly defined as a process of learning through and from experience towards gaining new insights of self and practice for future acts (Boud et al., 1985). It is a strategy for professional development, presenting a firm footing in most teacher-education programs (Cochran-Smith, 2005).

Reflective practice is an important concept in education, but it is difficult to do and equally difficult to teach (Finlay, 2008). One of the most important tools for reflective practice of student teachers is reflective writing. Reflective practice using writing is documented by a number of authors (e.g. Loughran & Corrigan, 1995; Mena-Marcos et al., 2013; Moon, 2006; Bolton, 2010).

General forms of reflective writing often include a composition of reflective or learning journals (Hatton & Smith, 1995; Bain et al., 2002; Ukrop et al., 2019), portfolios (Zeichner & Wray, 2001; Darling, 2001), blogs (Stiler & Philleo, 2003) and learning networks (Cardenas, 2014). These forms have some basic characteristics, which Moon (2006) summarized into three categories: More or less structured journals, individual, dialogue or collaborative journals and printed or electronic journals (or different forms such as audio or videotape).

In our study, we frame a term of reflective writing as unstructured, individual, and electronic reflective journals, which record students' thoughts and reactions during their teaching practice (Lee, 2008). In this scope, reflective journals are used as supporting material for group discussion in reflective seminars. These seminars are concerned with a group reflection focused on the practice of the students. This approach is similar to Tan (2013) who integrates reflective writing with reflective dialogue and uses reflective writing as an initiator of group discussion supervised by a university tutor.

2.1 The efficiency of reflective writings

The reflective writing practice of student teachers has several essential benefits. For instance, Krol (1996) emphasizes that it is an "approach that fosters reflection and is an effective source of dialogue between student and teacher" (p. 1); or that reflective writing can serve as a self-assessment to evaluate own personal epistemology (Hume, 2009; Lee, 2008). The latter states that reflective writing can also provide an opportunity for peer discussion/assessment of the students (Hedlund, 1989; LaBoskey, 1994; Colton & Sparks-Langer, 1993). Reflective writing also helps to provide an insight into how students structure their experience during the reflection (Bean & Stevens, 2002; Maloney & Campbell-Evans, 2002; Wallin & Adawi, 2018). Overall, this approach helps to support student teachers in the development of their teacher identities and provides instructive insights into the student teachers' teaching and learning (Whipp et al., 1997).

Despite these results, some voices argue that the efficiency of reflective writing is ambiguous. For instance, Alger (2006) notes that many of the reflective activities "have the potential to encourage reflection, but there is little research evidence to show that, as a result of engaging in these reflective activities, teachers develop a reflective disposition or stance to their teaching" (p. 289). Mena-Marcos et al. (2013) submit the statement that "deliberate reflection can support the construction of professional knowledge, yet this occurs rarely" (p. 147). Many of the studies in higher education do not consider developmental psychology and that reflective judgment of the students is also influenced by age (cf. King & Kitchener, 2004).

Systematic literature reviews can be a guide in this discussion. In these studies, the authors found that reflective writing appears to be an effective tool for supporting reflective practice and teaching performance. Lindroth (2015) supports this by documenting an efficiency of reflective writing in the context of student teachers, while Dyment and O'Connell (2011) find reflective writing to be also efficient in an environment of higher education.

Despite some ambiguities, the vast amount of previous research pointed out the relation between reflective practice and developing reflective thinking or competence in the students. However, the open research issue is the relationship between reflective practice and the real performance of teaching. When there is a measurement between reflection and performance, it often happens through self-assessment questionnaires (Fallon et al., 2003) or observations (Cattaneo & Motta, 2020). For example, Cohen-Sayag and Fischl (2012) submit the claim based on their research that "the result indicated that the link between levels of student teaching and levels of reflective writing is not clear" (p. 33). The intensive reflective writing improved the levels of reflection in one group but did not improve their teaching acts correspondingly. Still, student teachers that reached a threshold of critical levels of reflection also improved their teaching acts. Based on these theoretical findings, we construct one of our exploratory hypotheses (Section 3.5.1), denoting significantly higher perceived performance of students by their mentors in a group having reflective proportion in writing over a selected threshold, as compared to a group of students below this threshold.

2.2 Low level of reflection and fostering of reflective writing

Previous findings point to a suspicion that for many beginning teachers, it might be challenging to exhibit higher levels of reflection in their writings. (Ryken and Hamel, 2016, p. 31) point out that "consistent finding in the research on teacher reflection is that higher levels of reflection are rarely observed among teacher candidates (Klein, 2008; Larrivee & Cooper, 2006; Lee, 2005; Mena-Marcos et al., 2013; Pedro, 2005; Shoffner, 2008; Ward & McCotter, 2004)". The situation seems independent of the form of reflection produced; Lepp et al. (2020) describe that written or video journals by teacher candidates have the same most common level of reflection which is the descriptive level.

Additionally, it has been reported that students are prone to a devaluation of the understanding of reflective practice that justifies their own assumptions (Loughran, 2007). This situation is referred to as retroreflection (Kolb, 2014); such a reflection is rarely melted into action. Even if a student performs reactive action, it is carried out mechanically, without reflection, and without long-term effects.

These difficulties in mastering reflection may be related to the fact that it is challenging for students and novice teachers to apply this way of thinking from a developmental point of view (cf. King & Kitchener, 2004).

From a previous point of view, it seems important to foster reflective practice by a mentor, worksheet or other scaffolding tools. Houston (2016) found that "a small number of studies do indicate that scaffolding tools are an effective means to help students with reflective writing, but little research has been done in this area (Arrastia et al., 2014; Fox & White, 2010; Larrivee & Cooper, 2006; Wilcox, 1996)". Hanafi (2019); Pasternak and Rigoni (2015) also showed that training in reflective writings is important for fostering a reflective stance of the candidate teachers. Some authors provide exercises for developing the competence of reflective writing. For example, students analyze three different texts with different reflective writing levels, from the descriptive level to the critical level of reflection. The task is to underline the sentences that seem to be associated with reflection, followed by a joint discussion (cf. Moo, 2006; Hanafi, 2019).

Part of the support given by a mentor or instructors at the university is to provide direct feedback on written reflection. Spalding et al. (2002) considered that "individualized and personalized feedback for students instructors were most important in helping them grow". From this perspective, an ability to automatically assess students' reflective writings seems particularly crucial, as many teacher training programmes in universities have a number of students not feasible for fully personalised feedback. Here, machine learning algorithms or deep learning algorithms can provide, or ease a way to deliver necessary, personalized feedback.

In our study, students do not get feedback on their reflective writings; based on this fact and theoretical findings, we expect that the number of categories (as an indicator of the level of reflection) in reflective journals does not change among the submissions of reflective journals in different time (hypothesis described in Section 3.5.2).

2.3 Machine learning approaches in an assessment of reflective journals

The traditional way of assessment of reflective journals often has a form of qualitative or quantitative content analysis (Fox et al., 2019; Lepp et al., 2020; Mena-Marcos et al., 2013). These assessments are usually performed manually in the form of classification of selected segments of reflective writing. Common objectives of the assessment are research analysis of reflective writing or using assessment as formative feedback for students; As mentioned, in order to develop reflective competency or skills in the writing, essential individualized and personalized feedback for students is necessary (Spalding et al., 2002, cf.).

The first attempts to automate the evaluation of reflective writing began to appear with the transition of NLP techniques to Social Sciences. Initially, we observe the evaluation techniques mainly based on a dictionary-based approach (e.g. Bruno et al., 2011; Chang et al., 2012; Chou and Chang, 2011; Cui et al., 2019) or rule-based approach (e.g. Gibson et al., 2016; Shum et al., 2017). A more detailed overview of some of the listed research can be found in Ullmann (2019).

2.3.1 Literature review: Methodology and overview

The following text overviews works on the automatic assessment of reflective journals based on machine learning algorithms. The literature was retrieved using the following methodology: based on the keywords "reflective writing" and "machine learning", we performed a search in the Web of Science and Google Scholar databases (WOS = 18 + GS = 100 studies). The time frame for collecting the studies was from 2000 to 2020. By an inspection of title and abstracts, we eliminated 98 studies with an outcome of 20 studies. We critically read these articles and subsequently removed 11 studies that do not meet the established criteria enumerated below, resulting in 9 closely related studies.

These studies meet the following criteria: (1) the article focuses on the topic of reflective writing; (2) the article contains applications of automatic recognition of reflective categories (Classification Tasks) using machine learning algorithms, including deep learning algorithms; (3) annotations are made by human raters. Exclusion criteria were (1) the article does not focus on reflective writing, but rather, for example, only on the issue of academic writings (García-Gorrostieta et al., 2018); 2) automatic recognition is implemented using non-adaptive techniques, i.e., the techniques that are not able to adjust automatically to given the data set, such as dictionary-based or rule-based approaches; (3) dataset annotation is not performed by humans, but is, for example, automatically generated (Beaumont & Al-Shaghdari, 2019).

Our overview, also summarized in Table 4 in the Appendix, shows that most studies have used traditional machine learning algorithms for automatic recognition and prediction, with the exception of Carpenter et al. (2020) and Ullmann (2019), using deep learning techniques. The best-performing algorithms for most of these studies are Random Forests (4 studies) and Naïve Bayes (3 studies). The prediction performance is commonly reported in measures of Accuracy, F1-score, Precision, and Recall. The success of the algorithms ranges between Accuracy = .68–.96, Precision = .52–.96, F1-score = .41–87.

Annotation schemes for reflection classification are unique to each study, and we do not identify any efforts for mutual comparability at the time. In some cases, categories were created based on qualitative research for a specific context (e.g. Carpenter et al., 2020); in the others, the categories are theoretically derived (e.g. Jung & Wise, 2020). We denote two subcategories in these meta-schemes, one focusing on the depth of reflection and the other on the breadth of reflection (cf. Ullmann, 2019). Within the subcategory of the breadth of reflection, the most common category is description (4 direct occurrences), feelings (3 direct occurrences), and analysis (3 direct occurrences).

Inter-rater reliability is represented by standard variables such as Cohen's κ (4 studies), Krippendorff's unweighted α (2 studies), and Intraclass correlation (2 studies). Datasets of reflection most often come from a university background, but we can also find teachers or students from secondary school as an author of reflection.

The size of annotated datasets varies from 301 to 10,000 sentences (M = 2,954.56, Mdn = 1,966, SD = 2,943.61). The most common technique used for selecting the test dataset for benchmarking the proposed classification method is standard cross-validation or the traditional split of 80% to 20% or another ratio.

Despite the fact that the annotation schemes are so diverse, in our work, we choose to follow annotation meta-scheme of Ullmann (2019), which is based on an analysis of 24 models of reflective writing. Meta-schema contains 8 categories that are connected to reflection: *Reflection*, *Description of an Experience*, *Feelings*, *Personal belief*, *Awareness of difficulties*, *Perspective*, *Outcome – Lessons learned*, *Outcome – Future intention*.[3] We identify the benefits of this scheme in its justified selection of diverse categories, based on a systematic analysis, providing an easier comprehension by annotators. An important factor is also an ability to compare to the best-performing models of Ullmann (2019), that provides enough technical specifics to be reproduced.

3 Methodology

This section describes the background of our research question and hypotheses (Section 3.1), followed by a chronological overview of the steps taken to address them (Section 3.2). We proceed with a thorough description of the data collection process (Section 3.3). Eventually, we describe the technical aspects of reflection identification in context of NLP (Section 3.4) and the statistical methods we use to assess our hypotheses (Section 3.5).

3.1 Research question and hypotheses

This section summarizes the motivation and main objectives of our study in a form of a main research question and hypotheses.

Based on the literature review on assessing reflective journals through machine learning algorithms, we identify some shared shortcomings of the peer studies; Regardless of the complexity of the problem, studies primarily proceed with elementary, or not directly relevant text representations. Subsequently, a prediction is performed by mostly linear classifiers, or neural architectures with no structural relation to the problem.

We do not identify incremental qualitative progress among the studies; arguably, this can be due to limited reproducibility options. Except for Ullmann (2019), which states to annotate reflection in originally-public BAVE corpus, the data sources of the related studies are not publicly available. Additionally, no studies share the same meta-scheme of reflection categories that would allow respective qualitative comparison. Only a few studies are proceeded by a practical application, where we identify only a work of Knight et al. (2020) and ReflectR proceeding of Ullmann (2019), although relevant literature provides a wide variety of theories requiring support in data.

[3] A detailed description of each category with examples can be found in our full annotation manual: https://github.com/EduMUNI/reflection-classification/blob/master/data/annotation_manual.pdf

We acknowledge the semantic complexity of the problem and focus primarily on an application of deep language models of Transformers (Vaswani et al., 2017), but we critically compare their quality to a performance of thoroughly-tuned traditional approaches, reported to work the best in the literature. Our central research question (RQ) addresses the qualitative aspect of the result:

RQ: ***What quality can we reach in automated classification of reflection in written journals of student teachers?***

For the reasons mentioned in (Section 2.3.1), we choose an annotation scheme of Ullmann (2019). For convenience, we shorten the original names of the categories in this scheme, acquiring the following annotated categories: *Reflection*, *Experience*, *Feelings*, *Beliefs*, *Difficulties*, *Perspective*, *Learning*, *Intention*. The only customization of the scheme proposed by Ullmann (2019) is an addition of the *Others* category, referring to the sentences that do not correspond to any category of reflection. Thanks to this category, our models can also distinguish non-reflective sentences, allowing us to use machine learning models to evaluate the relative ratio of reflection in text.

We use the automated classification in a set of exploratory analyses to support selected hypotheses based on the selected theoretical studies of reflection.

The first analysis is based on a claim stating that only reaching a critical level of reflection is connected with better results in their teaching acts (Cohen-Sayag & Fischl, 2012) as most student teachers reach only a descriptive level of reflection (Lepp et al., 2020) (see Section 2.1). We pose a Hypothesis 1 (H1) to confirm this.

H1: ***There exists a "critical reflection" threshold such that, we can see a statistically-significant difference in the performance of the students.***

To estimate a quality of the teaching acts, we use student teachers' evaluation through a questionnaire completed by student mentors from practice. Thanks to the ability to identify reflection with a satisfiable quality, we are able to relate the statistics of journals with the performance of their authors in a corresponding time, reported from their mentors.

The second hypothesis is related to the statement that only individual and personalized supporting in writing reflection leads to better results in reflective writings (Spalding et al., 2002, p. 1393). We set a hypothesis 2 (H2) to instead reflect the expected state, i.e. that authors improve their reflection over time without support.

H2: ***The number of categories in reflective journals depends on the ordering of submission of reflective journals.***

In our settings, some student teachers collected a sequence of journals, amid which, they did not receive any personalized feedback to their writings. Hence, if the students require feedback to reflect more, we expect no significant changes in a number of categories over time and no significant trend depending on the ordering would be seen.

3.2 Methodology overview

This section brings an overview of the steps covered in our research aimed to answer the central research question and the exploratory hypotheses. These range from data collection to machine learning evaluation and statistical tests based on classified reflection. Each of the steps will be more thoroughly described in the subsequent sections.

Reflective journals collection We collected 1,070 reflective journals from Czech student teachers. Out of these, 300 reflective journals were randomly selected, while prioritizing the journals of the same authors.

Data annotation – first round Sampled journals were loaded to Doccano - annotation software (Nakayama et al., 2018). Six annotators were trained to identify reflectivity in the journals. The scheme of reflective categories follows the one of Ullmann (2019). This scheme was operationalised to a detailed manual for annotators (available on Github[4]). To reassure consistency, each journal was annotated by three different annotators.

Data annotation – consistency evaluation We evaluated inter-rater reliability in each pair of annotators and other qualitative analyses. We quantitatively concluded a quality of the collected data to be insufficient for further use.

Data annotation – second round From the annotated set of the first round, we have picked 7,128 sentences with high likeliness of belonging to a consistent reflective label – see below for details. The methodology was simplified to an annotation of full sentences with a single label. In cases where annotators did not agree on the category of the sentence, annotators were requested to discuss and agree on a single label.

Automated reflectivity identification – evaluation We have trained the selected machine learning classifiers on 90% of such annotated sentenced and evaluated them on a distinct 5% held-out set. The remaining 5% was used for hyperparameter tuning. We selected several classic classifier algorithms for language processing based on the literature (Section 2.3): Random Forest, Naïve Bayes, Support Vector Machine, Logistic Regression, as well as some well-performing deep models, supporting multilingual classification: BERT (Devlin et al., 2019) and XLNet (Conneau et al., 2020).

Research hypotheses – statistical evaluation The best-performing classifier was then used to evaluate the exploratory hypotheses based on a reflectivity in an original set of 1,070 reflective journals.

[4] https://github.com/EduMUNI/reflection-classification/tree/master/data (in camera-ready will be replaced by dataset DOI)

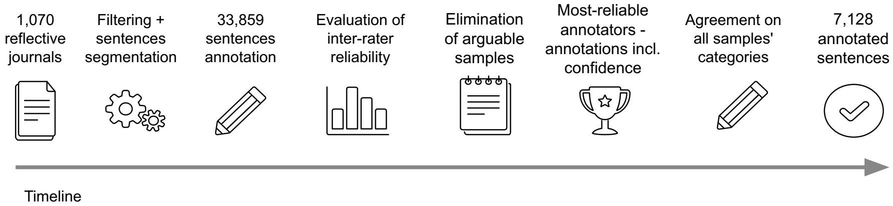

1,070 reflective journals
Filtering + sentences segmentation
33,859 sentences annotation
Evaluation of inter-rater reliability
Elimination of arguable samples
Most-reliable annotators - annotations incl. confidence
Agreement on all samples' categories
7,128 annotated sentences
Timeline

Fig. 1 A flowchart of the data collection and annotation process outlines the two iterations of annotation collection that we performed to reach an inter-annotator agreement sufficient for training machine learning algorithms

3.3 Dataset collection and annotation

Data collection process follows steps outlined in Fig. 1. We obtain 1,070 reflective journals (33,859 sentences, the length of reflective journals: M = 2,817 characters, SD = 1,287) from Czech student teachers (4th year of 5-year master degree programme, n = 220). Students created reflective journals five times over one year of study during their teaching practice. In these journals, students are asked to provide free writing text about their experience from the teaching practice. Distribution of reflective journals per subject given by student teachers is following: Czech (n = 112), English (122), Geography education (88), History (54), Biology education (67), Civics education (147), Physics and chemistry education (66), Music education (39), Art education (115), Russian and French (31), Special education (56), German (74), Vocational subjects (44), Mathematics education (55). This dataset CEReD - Czech-English Reflective Dataset Authors (Štefánik & Nehyba, 2021) is available on Github, see footnote in Section 3.2.

In the following section, we describe the process of collecting reflective annotations for this journal set, in the scope of the preceding results and issues connected to the annotation collection methodology that we have addressed in the second iteration of data collection.

In the first phase, the annotators were asked to annotate arbitrary units of text, whose meaning determines its belonging to one of the selected reflective categories. If applicable, annotators could assign more than one category for a piece of journal content and annotated pieces of text could have arbitrary mutual overlays.

Each of the annotated journals was then anonymised, segmented to sentences, based on the heuristics combining a punctuation and upper casing of the subsequent word. The sentences were then assigned to one of nine predefined categories, if a majority of the annotators assigned at least 1/2 of the words of a given sentence to the same category. To avoid ties, we assured that each journal is annotated by exactly three distinct annotators. Once the initial experiments of automatic classification of such annotated sentences were performed, we observed very low accuracy of the classifiers on the held-out set of sentences. The best performing Random Forest classifier reaches only 48.85% of Accuracy.

Further investigation of this situation uncovered that inter-annotators' κ suggest annotations' perplexity that is over our expectations. Inter-annotators' Cohen's κ for pairs of all five annotators ranges between 10.45% and 50.76%.

To evaluate the effect of aforementioned voting strategy on mitigating annotator's divergence, we have randomly picked 100 sentences with the assigned category using this strategy, and re-evaluated the assigned labels using a more experienced, expert annotator[5]. We have found that 36.6% of assigned annotations were not agreed upon by the expert annotator.

We suspected that the measured inter-annotator mismatch is a cause of variable understanding of the reflective categories among annotators. To confirm this, we have performed a classification with held-out evaluation on annotations of each of the five annotators separately. The Accuracy varied significantly, between 18% and 44%. This, however, was still below our expectations.

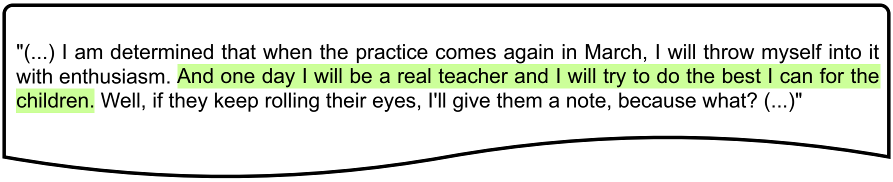
We further hypothesize that a poor quality of classification can be attributed to a high inner inconsistency of annotators, rather than a mutual inconsistency of their annotating strategy.

Using the obtained measures of the situation, we have decided to re-collect all the annotations. We have adjusted the methodology as follows.

We pick only the sentences, where the majority of annotators of the first round agreed on the true category, aiming to collect data set being as separable and decisive as possible. We have requested to collect annotations for all of these sentences duplically from the two best-performing annotators, based on the classification performance of a random forrest model fit on their personal annotations. Additionally, we requested the annotators to evaluate the typicality of given sample within their assigned category on a scale of non-decimal values between 7 and 1; 7 presents the most typical samples of a given category, while samples of typicality, i.e. confidence 1 are the least typical samples of given category. Note that we further refer to this value as *confidence*; We also use this confidence for segmentation of samples in evaluation.

Once completed, we have measured a number of cases where annotators did not match the assigned category. We have found that this was the case for 418 out of a total of 7,128 sentences, i.e. 5.86% of sentences. We asked the annotators to discuss the assigned category and agree on a single, most relevant one for each of the mismatched sentences. The annotators were able to agree on each of the sentences. As a result of this process, the dataset that we have collected and make public for the community, contains also an estimation of "typicality" of the assigned category by two annotators with highest inner consistency. We have observed that some of the sentences in the database are repetitive, or only minor mutual modifications. To provide a standard held-out evaluation set, that eliminates the theoretical possibility of training data occurrence in evaluation (in other literature referred to as *data leakage*), we have sorted the sentences alphabetically, and use only the last 5% for held-out evaluation. The results reported in Figs. 3, 4 and 6 report the accuracy on this

[5] Expert annotator has the most profound knowledge of category labelling and is responsible for the final decision in case of ambiguity (Fort, 2016). A similar design is used, for instance, by Hu (2017).



"(...) I am determined that when the practice comes again in March, I will throw myself into it with enthusiasm. And one day I will be a real teacher and I will try to do the best I can for the children. Well, if they keep rolling their eyes, I'll give them a note, because what? (...)"

Fig. 2 The samples of the CEReD contain a classified sentence (highlighted in the figure) and its surrounding context. A sample from CEReD data set in the figure was annotated as *intention* with mean confidence of 4 on a scale from 1 (least typical -) to 7 (most typical for the category). CEReD data set is anonymised and we make it publicly available for future research (see Section 3.3)

held-out set of sentences, or it's subset of minimal specified confidence. For reproducibility, we give our test split publicly available to reuse and compare the results on, in the project repository (Section 3.5).

3.4 Reflectivity classification

In this section, we start with locating the problem of reflectivity identification in the field of NLP. We proceed with a description of classification algorithms and respective textual representations that we experiment with.

We identify that reflection identification can be approached either as Text classification, or as Named entity recognition (NER) problem. In the classification approach, a system assigns a sequence of tokens, in our case a sentence, into a single category.

As an expression of reflection might not be necessarily aligned with a single sentence, the traditional classification approach might include less relevant parts into a given category. This problem can be eliminated by the latter-mentioned NER approach, which allows for a classification of arbitrary units of texts, usually words, where the same category can span over a sequence of words.

In this work, we only approach the problem as sentence classification due to the more straightforward interpretability of the results, better comparability with previous studies, lower demands for volumes of annotated data, and its ability to constrain the annotators towards more consistent annotations (Section 3.3). However, we acknowledge the mentioned benefits that the NER approach could bring.

Together with a representation of the classified sentence itself, a system might consider the context of the sentence, which we find to be a significant determiner of the reflective category of the sentence itself. As illustrated on a specific training sample in Fig. 2, we experiment with including a context of two preceding and one succeeding sentence. This additional segment of text is separated from the classified sentence in a method-specific manner, that we describe separately for each type of classifiers.

We experiment with two distinct categories of classification algorithms that we refer to as "shallow" based on non-neural machine learning algorithms commonly used for text classification and "deep" classifiers that compose a final prediction using a stack of linear classifiers, also referred to as deep neural networks.

3.4.1 Shallow classifiers

Following an example of Ullmann (2019), we first experiment with simple shallow classifiers. Based on the overview of the related work, we pick the classifiers from the following categories: (i) based on linear discriminators: Support Vector Machine, Logistic Regression and Logistic regression, (ii) conditional probabilistic classifier of Naïve Bayes, (iii) Tree-based: Random Forest classifier. Following an example of Ullmann (2019), these classifiers use one-hot single-word representation, i.e. unigram bag-of-words representation of sentences of size that is hyperparameter-tuned for their optimal performance in a range of 100–2000 predictor words, i.e. tokens. Noticeably, the best-performing Random Forest Classifier uses a bag-of-words representation pruned to top 800 most-frequent tokens of both sentence and context. We experiment with both including and not including the same representation for the context into the classification process. When the context is included, it is represented in the same way as the classified sentence and the two unigram representations are then concatenated. We find a slight improvement of results, when the context is included: in case of the best-performing Random Forest classifier, including same unigram representation of context improves the accuracy from 70.8% to 73.6%. We perform hyperparameter tuning of specific parameters of each of the classifiers and report the accuracy of the best-performing system in Fig. 3. An accuracy overview of all evaluated shallow classifiers can be found in Appendix 1.

3.4.2 Deep classifiers

Second, we experiment with selected state-of-the-art neural sequence classification models of the transformer family (Vaswani et al., 2017). Such classifiers are pre-trained on a large corpora of texts for objectives related to general language understanding, such as masked language modeling (MLM) (Devlin et al., 2019), or denoising (Lewis et al., 2020) that allows them to fit the specific objective with lower demand for an amount of training samples. For the sentence classification problem, we fit the deep, neural classifiers on cross-entropy loss objective commonly used for classification problems (Cox, 1958). We follow well-established representation conventions of transformer models: We represent input sentences as a bag-of-words of SentencePiece subwords (Kudo & Richardson, 2018) that differ from the previous bag-of-words representation in that a minority of words is internally split to subwords frequently occurring prefixes and suffixes, for instance, in word inflexions. We make sure that the subword vocabulary that we use covers the vocabulary used for pretraining of each particular model. In addition to shallow classifiers' representation, transformers also use the encoding of mutual ordering of input words. We find it crucial to provide the neural classification models with a context of the classified sentence. We add the contextual segment of text to the given representation, preceded with the special token "<s>" on a position separating the classified sentence from its context, similarly to Devlin et al. (2019). We find that if the context window is not consistently provided with the classified sentence, neural classifiers struggle to converge and

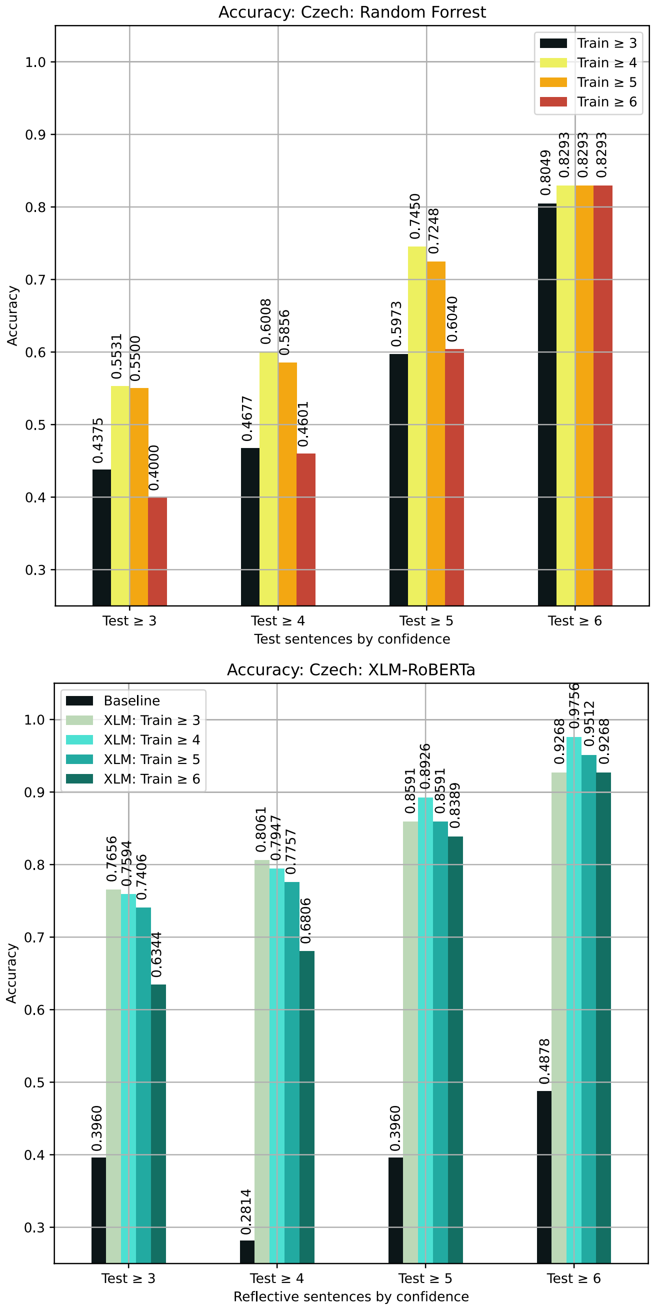


Fig. 3 Per-sentence accuracy of reflection classification of the best-performing shallow (left) and deep model (right): Random Forrest on BoW and XLM-RoBERTa transformer, respectively, trained on Czech reflective sentences over selected thresholds of mean confidence of annotators. Each of the trained classifiers is evaluated on Czech test sets with distinct confidence thresholds. Refer to Sections 3.4.1 and 3.4.2 for methodology description and Section 4.1.1 for a discussion. *Baseline* results denote a performance acquired by classifying every sentence into the most common category, i.e. *Experience*

significantly underperform even compared to the shallow classifiers. We managed to partially eliminate the need for context using the UDA (Xie et al., 2020) data augmentation technique but such classifiers still underperform the same architectures utilizing context by 5–10% of accuracy points.

We perform a systematic tuning of training batch size, learning rate and a number of epochs for each of the models, but we find the models to perform almost consistently among the trials. We report the accuracy of classifiers using the optimal parameters based on a performance on a validation set in Fig. 4 for English, and 3 for Czech classification. See Appendix 1 for an extended technical description of the neural networks and training procedure.

3.5 Methodology of exploratory hypotheses

We use the trained classifier to identify reflectivity in a set of novel, unannotated reflective journals. We perform analyses based on a distribution or temporal dynamics of the characteristics of such identified reflection. The analyses were performed using XLM-RoBERTa classifier with samples of training confidence of 4 and higher.

We make sure that the whole procedure of each of our exploratory analyses is fully reproducible. Together with the anonymised data sources, a neural classification model and a python library allowing one-line extraction of reflective segments from a text, each analysis can be run using a single click in the corresponding notebook. Notebooks corresponding to each of the hypotheses can be found in our Github repository[6].

3.5.1 H1: There exists a "critical reflection" threshold such that, we can see a statistically-significant difference in the performance of the students.

We start by associating a set of the reflective journals from our dataset with a corresponding evaluation of the mentor, i.e. supervising teacher, reported at the same time as the journal submission. Student performance is reported via a questionnaire collected from the mentors. The questionnaire is composed of 7 questions with 6-point Likert scale of answers, where 6 presents the *best* performance, and 1 presents the *worst* one. Number of participants who fill the questionnaire, $n = 223$. The full dataset and items of the questionnaire are also available in project repository (see footnote in Section 3.2).

We performed qualitative assessment of the collected results. Its parallel analysis suggests one factor and one component of questionnaire, Cronbach α is .91, and item-total correlation is between .66 to .77. The recommended size for Cronbach α is more than .7 (Nunnally & Bernstein, 1994), and the item-total correlation is more than .30 (Cristobal et al., 2007). Thus, we can use the questionnaire's overall score as an index of student teachers' quality performance on their practice. Despite

[6] https://github.com/EduMUNI/reflection-classification/tree/master/analyses

the best efforts of the mentors for objective evaluation, we acknowledge that mentor's assessment of student performance may be subject to some confounding factors (such as the length of the mentor's internship, level of education, relationship with the student, and the like). We sum the ranks of each evaluation and normalize the values to $\langle 0, 1\rangle$; we further refer to this value as a *performance* of the student, i.e. the performance of the student at a time of creating the corresponding journal.

Having both the journals and their respective performance, we formulate the problem to systematically identify the threshold that could outline the level of "critical reflection" (Cohen-Sayag & Fischl, 2012; Lepp et al., 2020) as follows: We look for a value of relative reflection that can best distinguish the performance of the students.

Specifically, we consider the relative proportion (number of *reflective* sentences divided by a number of *all* sentences) within the journal as a predictive variable, and the reported performance associated with the given journal as a target variable. In this framework, we fit a regression decision tree on the single predictive variable, in order to best-predict the target, or dependent variable. Presuming that the groups of higher-performing students can be distinguished by an amount of their reflection in the diaries, this way, we systematically search for a "critical reflection" threshold as a proportion of reflection best-splitting values of students' performance. For illustration, the decision tree, with the optimal splitting value of relative reflection ($X[0]$) is also visualized in Fig. 5.

After fitting the decision tree and obtaining the found reflection threshold, we test whether the performance of the two groups separated by such identified level of relative reflection are significantly different. Specifically, we perform a T-Test to uncover if there is a statistically-significant difference between a mean in performance of the two groups.

Additionally, we repeat the process separately for each of the categories. The results for a separate analysis based on a relative ratio of each of the categories can be found in the notebook of this analysis in project repository.

3.5.2 H2: The number of categories in reflective journals depends on the ordering of submission of reflective journals.

We construct a set of causal models for each category that use a temporal ordering and a length of the journals as predictors and a number of occurrences of each predicted category as the outcome variable. Journal length is used for a normalization, as each journal has a different length and a total number of classified items, i.e. sentences.

As the causal inference models (cf. Gelman & Hill, 2006), we build a set of Generalized linear mixed models (GLMMs) (Jiang, 2017; Stroup, 2012; Faraway, 2016). GLMMs provides some advantages compared to simpler models, such as ANCOVA: it does not require listwise deletion for a treatment of the missing values, and multivariate models provide a higher level of expressivity for testing on fixed effects (cf. Hoffman & Rovine, 2007).

To ease the mapping of our approach to the standard notation, having *j* samples for every *i*-th category, the following GLMMs were used:

$$\begin{aligned}\text{category}_i &\in \{\,\text{Experience, Reflection, Feeling, Difficulty}\\ &\qquad \text{Perspective, Belief, Learning, Intention, Other}\,\}\\ \text{category}_i\text{count} &= \gamma_{00} + \gamma_{10}\text{ordering}_{ij} + \gamma_{20}\,\text{journal length} + u_{0j} + e_{ij}\end{aligned}$$

We test the distribution of each category and proceed with an assumption of the distribution that best matches our observations; models for categories *Other* and *Experience* are GLMMs with Conway–Maxwell–Poisson distribution and models from *Reflection* to *Intention* are best fit by Negative binomial distribution. Conway–Maxwell–Poisson distribution and Negative binomial distribution is the special distribution used for non-normal distribution in GLMMs. The Conway–Maxwell–Poisson and Negative binomial distribution are commonly used for under and over-dispersed count data, that is when the conditional variance is below or above the conditional mean.

The analyses were performed using Python (Van Rossum & Drake, 2009) and R scripts (R Core Team, 2020), in particular using the lme4 package (Bates et al., 2012), glmmTMB package (Magnusson et al., 2017) and DHARMa package for residual diagnostic (Hartig, 2019).

All details of the procedure of analysis (creating matrix for analysis, choosing of distributions, hypotheses testing and residuals diagnostics) are available to reproduce on the reference in Section 3.5.

4 Results

4.1 Reflectivity classification

In this section, we respond to our primary research question, introduced in Section 3.1, quantifying an accuracy of classification that we can reach using automated system.

Figure 3 compares accuracy of two best-performing traditional classifier among the surveyed ones, referred to as “shallow”, with the best-performing “deep” classifier, on a classification of held-out Czech sentence set: a Random Forest classifier (left) and XLM-RoBERTa transformer (right). Both classification models were iteratively trained and evaluated on a subset of sentences with confidence over the labeled confidence threshold.

Analogically, Fig. 4 shows the classification accuracy of these models trained and evaluated on English reflective sentences (left) and the model trained on English and evaluated on Czech reflective sentences (right).

Table 1 Per-category performance of the best-performing shallow classifier (Random Forrest) together with the best performing neural classifier (XLM-RoBERTa) on English data set. Categories are ordered by the training sample size (N). Both classifiers are trained on *confidence threshold* ≥ 4 and evaluated on *confidence threshold* ≥ 5. No samples of *Intention* category in test set exceed given test threshold, hence it does not appear here

	Random Forrest			XLM-RoBERTa			
	Precision	Recall	F1-score	Precision	Recall	F1-score	N
Learning	.0	.0	.0	1.0	1.0	1.0	45
Perspective	.0	.0	.0	1.0	.5	.67	52
Belief	.0	.0	.0	.25	.67	.36	232
Difficulty	.5	.11	.18	.7	.78	.74	288
Reflection	.29	.4	.33	1.0	1.0	1.0	764
Feeling	.58	.83	.68	.96	.96	.96	861
Experience	.44	.29	.35	.74	.96	.84	1, 011
Other	.73	.80	.76	.97	.85	.91	1, 547

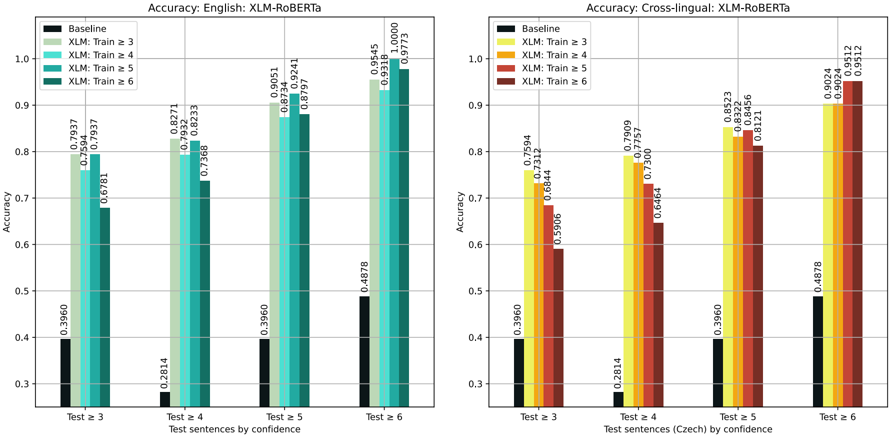


Fig. 4 Per-sentence accuracy of reflection classification of XLM-RoBERTa transformer trained on English reflective sentences over selected thresholds of mean confidence of annotators. Each of the trained classifiers is evaluated on Czech test sets with distinct confidence thresholds. Refer to Sections 3.4.1 and 3.4.2 for methodology description and Section 4.1.1 for a discussion. *Baseline* results denote a performance acquired by classifying every sentence into the most common category, i.e. *Experience*

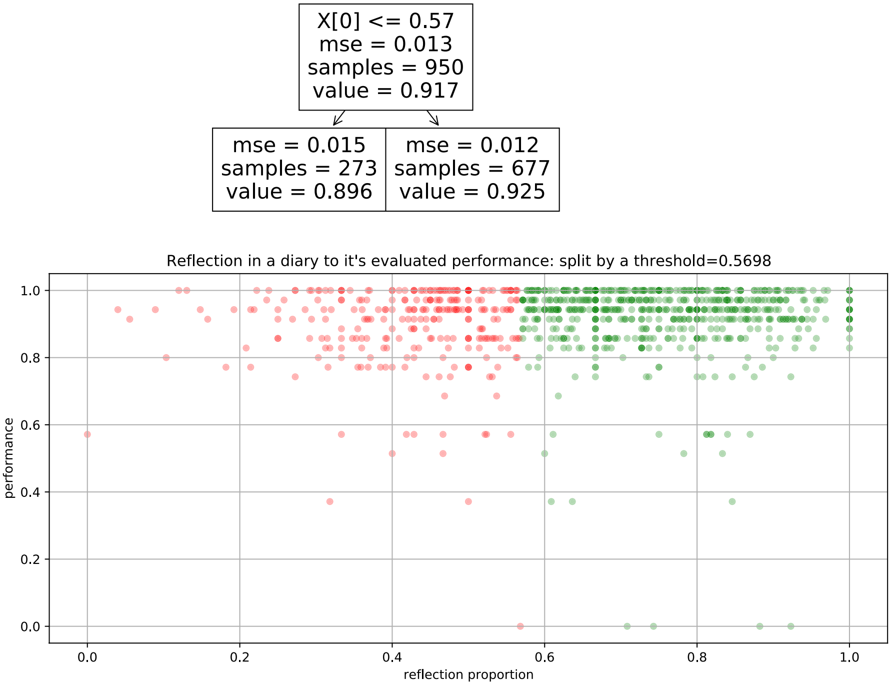


Fig. 5 Top: Decision Tree Regressor with optimal mean squared error with respect to predicted performance. A value of *X[0]* denote a found value of *critical reflection*. Items denoted as *mse*, *samples* and *value* denote a mean squared error, number of samples and a mean performance value, respectively, in each group occurred by the tree split. Bottom: reflective proportion of journals from our samples associated with the evaluated performance of the students from the time of the writing. The reported performance is assigned as a mean of the graded evaluation of the supervisor teacher. Color indicates two groups separated by a *Critical Threshold*, picked by the displayed Tree Regressor

4.1.1 Monolingual settings

In the Fig. 3, we observe that shallow classifiers are a strong baseline, especially for lower-confidence settings, reaching 55.31% of accuracy on test samples of the confidence score over 3. The low-confidence samples are also the ones where we see the highest gains of the deep neural model (XLM-RoBERTa), which on the same confidence threshold reaches accuracy ranging from 63.44% to 76.56%. On the highly-confident samples, the accuracy of the deep model ranges from 92.68% to 97.56%, while the best non-neural classifier (Random Forrest) ranges between 80.49% and 82.93% of accuracy.

Slight overperformance in English settings, reported in Fig. 4 outlines the fact that XLM-Roberta has been pre-trained on a larger amount of English texts, hence having a better base understanding of the language. Consequently, the accuracy of the neural classifier on the English test set ranges from 93.18% to full

100%. The difference in accuracy between English and Czech suggests that further pre-training of the multilingual base model especially on language-specific data, or for applications in lower-resource languages might further improve the eventual classification accuracy. The difference might also be mitigated by further increasing the amount of annotated reflective sentences. Note that the amount of training reflective sentences in our experiments was equal for both languages.

Table 1 exposes strong correlation between a sample size for given category (*N*) and selected performance measures (Precision, Recall, F1-score), suggesting that an extension of the training dataset for underrepresented categories might significantly improve the performance on these. Despite the fact that for the top four least-common categories a number of test samples with confidence over the given threshold is lower than 10, results suggest that deep classifier perform significantly better in low-resource settings. We also observe that a category of *Experience* is particularly difficult to distinguish by a simple, rule-based system of a decision tree, but the situation improve significantly with the more expressive neural models.

4.1.2 Cross-lingual settings

As XLM neural models are pre-trained on multilingual texts, we also experiment with cross-lingual classification, i.e. how well a classifier trained on English performs in another language. Figure 4 (right) shows that in such settings, the English model performs comparably to the model fine-tuned exclusively for the language on which it is applied: comparing the English model evaluated on Czech sentences, we observe the decay of accuracy only between 0 and 4.2%, as compared to the Czech model evaluated on Czech. This suggests that the models trained on languages where the data might be easier to obtain can be successfully used for other languages that XLM has been pre-trained on (Conneau & Lample, 2019). We also believe that further classification quality could be obtained by concurrently fine-tuning on classification in different languages if resources in multiple languages are available.

Importantly, this result shows that the multilingual model that we train on our data set can be applied to any of the 100 pre-trained languages of RoBERTa Conneau and Lample (2019) with only a small loss of accuracy. We provide the evaluated English model to free use as a part of this work to foster the research of reflectivity in other languages, where a large-scale collection of annotated data might not be feasible.

4.2 Exploratory hypotheses

In this section, we summarize the results of the exploratory analyses introduced in Section 3.5.

Table 2 Odds ratios for all categories denote a change in a number of occurrences of the given category on a given ordering of student submissions, relative to the first submission. These values suggest that for top six categories of reflection, a change is consistently positive or negative, while for the bottom three, the relative change to a first submission is not consistent

Journal ordering	2nd	3rd	4th	5th
Experience	**1.21**	**1.46**	**1.17**	**1.42**
Difficulty	**1.46**	**1.68**	1.06	**1.72**
Perspective	**1.39**	**1.51**	1.02	1.33
Learning	1.00	**1.46**	1.29	**1.69**
Other	−.92	**−.76**	−.98	**−.85**
Feeling	**−.89**	−.91	−.96	**−.78**
Belief	−.78	−.91	−.98	1.07
Reflection	1.02	**−.84**	**−.74**	−.89
Intention	−.86	1.13	−.91	−.85

Values in **bold** indicate values that are significant on the 95% confidence level

4.2.1 H1: There exists a "critical reflection" threshold such that, we can see a statistically-significant difference in the performance of the students.

As described in Section 3.5.1, we fit the decision tree of the depth=1 using the MSE objective (shown in Fig. 5 – top), getting the candidate *critical reflection threshold* with a relative reflection ratio of .57.

We find that the mean performance of the groups separated by this threshold, as shown on Fig. 5 are significantly different, on a significance level $\alpha = .95$. Specifically, a mean performance of the students associated with the journals below this threshold is by .031 relative performance points lower than the ones above the threshold.

We repeat the approach for each reflection category separately, i.e. we observe whether there exists a critical reflection threshold on such that the difference in performance of the two groups below and above this threshold is statistically significant. We find that the group means are also significantly different for each of the categories on a significance level $\alpha = .95$. However, in cases of categories *Experience* and *Belief*, we find that the group with higher proportion of *Experience* and *Beliefs* occurrences has significantly lower performance than the other.

The results suggest that having a high proportion of *Belief* category might not relate to a better performance in practice. Such interpretation is consistent with a view of contemporary epistemology on a relationship between beliefs and behavior (Schwitzgebel, 2010, cf.). Similarly, our data show that a mere capture of *Experience* has no connection to a higher performance in students' practice as perceived by mentors. On the contrary, our data suggest that the relationship between higher perceived student performance in practice is associated with a higher incidence of the categories: *Feeling*, *Reflection*, *Difficulty*, *Perspective*, *Learning*, *Intention*.

This finding is applicable in pedagogical applications of reflection writings. Among others, it raises an assumption that it is recommended not to solemnly pursue a clean description of *Experience* and *Beliefs* in students writing, but also to analyze the *Experience* and motivate students to reflect on other mentioned categories, relating to a higher perceived performance of students in practice. This assumption

Table 3 Mean and standard deviation of the occurrence of the category per journal

Category	Mean	Standard deviation
Other	9.83	9.12
Experience	6.70	5.31
Feeling	4.47	3.64
Reflection	3.98	3.43
Difficulty	1.44	1.77
Perspective	.69	1.08
Belief	.91	1.44
Learning	.36	.72
Intention	.27	.59

is consistent with, for example, the previous assertion that in order for reflection to affect performance in practice, it must exceed a critical threshold (Cohen-Sayag & Fischl, 2012). We could thus assume that critical reflection is associated with the depth of Reflection category and a proportion of categories such as *Feeling*, *Reflection*, *Difficulty*, *Perspective*, *Learning*, *Intention*.

4.2.2 H2: The number of categories in reflective journals depends on the ordering of submission of reflective journals.

To address the second research exploratory hypothesis, we compute the odds of occurrence of each of the reflective categories in the journals, sorted by the time of the handover. Then, we evaluate the odds values against the hypothesis that the number of categories does not evolve at all, i.e. that it remains the same as in time of the first handover. For each category, depending on the order of submission of the reflective journal, we calculated the odds ratio - exp (β) (Table 2) from the relevant model (see Appendix Table 5).

Table 2 shows that only the category of *Experience* is significantly affected over four submissions following the first submission in time. This means that the incidence of the *Experience* category in the second, third, fourth, and fifth submissions of the journal is higher by approximately one occurrence, as compared to the first submission. The subsequent most significant increase is in the *Difficulty* category, where a significant increase is in average by 1.5 occurrence in the second, third and fifth compared to the first submission. In the *Perspective* category, there is a significant increase by 1 occurrence in the second and third submissions of the journal compared to the first. The *Learning* category experiences a significant increase in the third and fifth submissions of the journal.

On the other hand, for the *Other* category, which is a non - reflective category, there is a significant reduction by less than one occurrence, namely for the third and fifth submissions of the journal. There is also a decrease compared to the first submission in the *Feeling* category, namely in the second and fifth submissions, and by less than one category. In the category of *Beliefs*, *Intention* and *Reflection*, there is an increase or decrease in the submission of journals as part of individual submissions. However, there is a significant decrease only in the category of *Reflection*, namely in the third and fourth submissions of the journals, as compared to the first submission.

It is important to notice the reference number of occurrences for each of the categories predicted in the dataset of 1,070 reflective journals (see Table 3). Clearly, the most extensive category is *Other* (9.8 occurrences per journal), followed by categories such as *Experience*, *Feeling*, *Reflection*, and *Difficulty*. The categories *Perspective*, *Belief*, *Learning*, and *Intention* do not reach a single occurrence in an average journal. The average length of a journal is about 29 sentences (M = 28.65, SD = 14.45). The *Experience* category represents a descriptive explanation of the experienced situation, which is the lowest level of the depth of reflection, which is rather a prerequisite for reflection. From this point of view, more than half of the sentences (16.5 sentences) are non-reflective (*Other*) or just describing student experience.

With respect to the reference counts, the increase or decrease in counts of the categories is not large. These results are in line with Spalding et al. (2002); Hanafi (2019); Pasternak & Rigoni (2015). These state that without a personalized and specific feedback, a quality of reflective writing of student teachers does not consistently improve. In the development of reflective writing for student teachers, we would not expect to observe a decrease in the reflective categories and we would also assume a more considerable decline in the *Other* category for a more developed student in reflective writing. However, we note that a possible decrease of *Other* certainly has its limitations because a certain number of sentences with the *Other* category plays an essential role as a ”binder” among other categories in the reflective journal text. We refer to this aspect as a “relative reflection” of the text.

5 Discussion & Conclusions

Our work applied the selected machine learning approaches to automatic identification of reflective writing in reflective journals and used the resulting toolset to answer some of the relevant hypotheses identified in the literature. In contrast to the previous work, we investigate the quality of the most recent deep language models based on transformers architecture (Vaswani et al., 2017) and critically compare their results to the previously best-performing methods instantiating traditional ensemble, statistical or probabilistic models. Following the taxonomy of categories proposed by Ullmann (2019), we reproduce the reported ranking of simple models, dominated by Random Forrest ensemble model, but find significant qualitative improvements of the novel models over the best models of previous work (Section 4.1.1). Additionally, we identify that novel neural models provide out-of-box applicability in other languages that multilingual models were trained on (Section 4.1.2). Further, we highlight the following findings of our work.

- Data collection process is the most significant determinant for the quality of the subsequent machine learning application. In applications where the definitions and differences of categories can be ambiguous, it is crucial to work on a data

collection framework that maximises data consistency. If the discriminative features are continuous, a diversification of samples by confidence allows distinguishing samples which are ’clear’ from the ones that might be arguably correct.
- As shown in Section 4.1, more complex neural language models bring significant qualitative benefits, as compared to the traditional machine learning methods. Neural models are also able to further benefit from cleaned samples.
- Multilingual language models might democratize the research in machine learning applications in social sciences, as the research based on languages other than English can further build upon work such as ours and contribute globally.
- Our analyses based on reflection identification found that an amount of reflection does not consistently grow over time in all the observed dimensions, i.e. categories during the practice. However, we have shown that there exists a relation between the student reported performance and an amount of reflection in their writing, which motivates future efforts to enhance students to incorporate more reflection in their practice.

In addition to these findings, our work provides the community with the anonymized, annotated corpus of reflective journals with transparent annotation guidelines, the pre-trained neural model applicable in more than 100 pre-trained languages and the library for reflection identification, with a reproducible demonstration of its use in our exploratory hypotheses. We believe that the future research will easily build upon our results to further enhance the reflection in education on all levels, for instance, for the use-case outlined, but not limited to the ones proposed in the following section.

5.1 Implication

We briefly summarize the implications of our work that we can identify in the related future research (1, 2) and pedagogical practice (3, 4).

1. We show how to conduct a data collection leading to a data set, which can be well-utilised for creating automated machine learning tools in the context of the inexactly-defined problem. Such a situation is characteristic of many other pedagogical and social applications. We believe that following our framework of multi-step data collection and refinement based on sample confidence and annotators agreement can circumvent one of the main blocks of widespread automated tools in other social applications, possibly reducing monotonic or trivial work in many social applications.
2. Making our CEReD data set available makes it significantly easier to further enhance the quality of the reflection identification tools. CEReD also eases the mutual comparability of the future methods by standardising the test benchmark.
3. Our reflection identification toolset makes it easy and accurate to provide students with personalised, automated, or semi-automated feedback. This is particularly

useful in situations where it can be difficult to deliver other qualitative assessments due to capacity or personal reasons.

4. High accuracy of our freely-available multilingual model allows us to quantify the methodical adjustments in the education process and evaluate their impact. The adjustment measure can have a form of change in the reflective ratio of student writings or a change in counts of specific categories. We argue that such quantification can also reflect on more latent properties of the education, for instance, an engagement of the students in the learning process.

In addition to the implications for automated or semi-automated feedback delivery, we acknowledge that a form of such feedback is a crucial covariate of its efficiency. In our future research, we will elaborate further into the possible forms of such feedback, for instance, in the form of automated scoring, automatically-selected textual feedback based on a student's personal trends, or visual analysis of the student's diary. We aim to compare these approaches with respect to their impact on students' perceived performance and its development in time and deliver specific suggestions for the form of such feedback for specific situations.

5.2 Limitations

We acknowledge several limitations that should be considered when using our results in future work.

As mentioned in Section 3.4, a problem of reflection identification in its full complexity does not map to a sentence classification: reflection determinant can span multiple sentences, or a complex, single sentence can contain multiple classes of reflection. Further, our categories of reflection are picked based on a review of the literature in an attempt to maximise operationalisation (Section 3.3), but we acknowledge that a more advanced scheme of reflective categories can exist.

Although we have shown that the model trained on one language can be used in other languages with the additional error ranging from 0 to 4.2%, our cross-lingual experiment does not reflect methodological and cultural discrepancies in different educational systems and should to be thoroughly considered before applying our models to other languages. Further, the accuracy of cross-lingual applications might vary among languages and is conditioned by the size of the language-specific corpus that RoBERTa model was pre-trained on – refer to the corresponding article (Liu et al., 2019). We have considered this aspect and measured the listed cross-lingual error bounds using the target language (Czech) of close-to mean size, hence providing a close-to mean estimate of additional cross-lingual error. However, the application quality in other languages will most likely vary. A qualitative difference caused by all these covariates can be quantified by collecting a small set of annotations compliant with our methodology from one's own domain and evaluating the prediction quality of these new annotations.

In the analyses using our toolset in future research, one should consider that the classifier exposes an error that varies among the categories. In particular, the underrepresented categories expose a higher level of error than the more common ones. This behaviour suggests that such a flaw can be further eliminated by collecting more samples of the underrepresented categories. Nevertheless, the error should be considered, especially in cases where the significance of the experiments lay close to the borderline.

We also acknowledge that the introduced data set exposes some errors related to a naive segmentation to sentences, automatic translation and possibly other methodological technicalities described in Section 3.3.

From the pedagogical perspective, note that we base our work on one specific concept from the definition of reflective writing, defining reflection in the scope of a set of categories. However, literature on reflection contains other concepts, such as the one that emphasises reflection as a holistic and embodied process (Bass et al., 2020; Kinsella, 2007).

The perceived performance of students by mentors in practice that we base our first analysis on may be biased, for instance, by the relationship between the mentor and the student. We note that the evaluation of performance by an independent observer in combination with the student's self-evaluation might minimise the risk of bias introduction.

A: Classifiers: technical details

In this section we describe technical specifics of training of the selected shallow and deep classifiers.

A.1: Data splits

To make sure that no sentences are present multiple times in our dataset, we start by removing the duplicate sentences regardless of the context. To minimize the chance of multiple occurrences of sentences that are close-to the same among the splits, we order the sentences alphabetically. We split the ordered list of sentences with their associated contexts in a ratio of *90:5:5* to a *train*, *validation* and *test split*, resulting in a total of 6,097 of training, 340 of validation and 340 of test sentences. We pick a validation set based on a confidence threshold matching the training confidence threshold, i.e. the classifier trained on sentence of minimal confidence of 4 is tuned on a validation set with a minimal threshold of 4 as well. Confidence filtering on all splits is applied only to the selected split so that it is reproducible and no samples can be exchanged between splits when testing on different confidence threshold.

A.2: Shallow classifiers

We report the tuned hyperparameters of the shallow classifiers, and report their comparative results of accuracy with their respective optimal parameters.

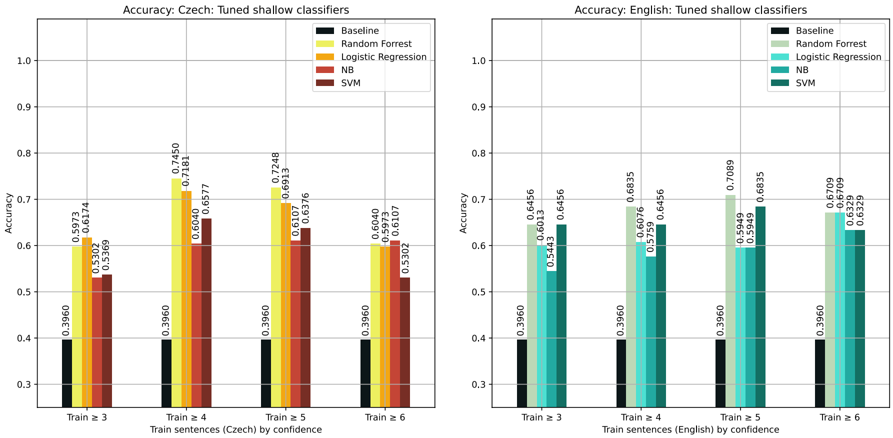


Fig. 6 Per-sentence accuracy of reflection classification of the evaluated shallow classifiers, with their respective hyperparameters tuned on a validation set. Results for both Czech instance (left) and English instance (right) of the introduced CEReD dataset

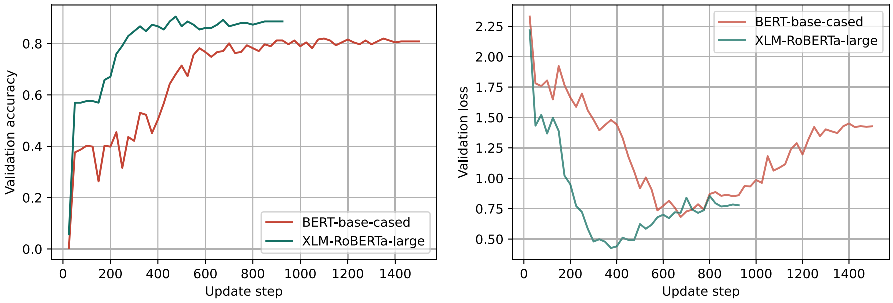


Fig. 7 Values of validation accuracy (left) and loss (right) during the training process, for XLM-RoBERTa-large and BERT-base-cased trained and validated on sentences with confidence over 4 propose a clear dominance of accuracy by XLM-RoBERTa model

A.2.1: Preprocessing

To minimize a dimensionality of bag-of-words representation, we lowercase and remove the stemming all the words of sentence and its context. We limit the vocabulary of bag-of-words to top-*n* most-common words, except the stop-words occurring in more than 50% of sentences. We consider *n* as a hyperparameter of every shallow classifier.

The final bag-of-words representation with context, used for all the shallow classifiers, is a concatenation of a bag-of-words vector of a sentence and its context.

Hyperparameters

We have performed an exhaustive hyperparameter search on a validation data set of the training confidence, on the following parameters of shallow classifiers:

- (Shared) use of context $\in \{\text{True}, \text{False}\}$
- (Shared) preceding and succeeding context window size $\in \{1, 2, \dots, 5\}$,
- (Shared) BoW size: limitations to top-n tokens: number of tokens $\in \{100, 150, \dots, 1000\}$
- (Shared) Tokenization for a specific language $\in \{True, False\}$
- Random Forrest: depth $\in \{2, 3, 4, 5\}$,
- Random Forrest: number of estimators $\in \{100, 200, \dots, 500\}$
- SVM: kernel function $\in \{linear, polynomial, radial\}$

Results

Figure 6 shows results for all the evaluated classifiers with the hyperparameters listed in Section 1 tuned on a validation split.

B: Deep classifiers

We extend the description of the training process from Section 3.4.2 with technical specifics relevant for reproduction or easier further customization.

We experiment with two multilingual representatives of transformer family: Multilingual BERT-base-cased (Devlin et al., 2019), and XLM-RoBERTa-large (Conneau et al., 2020). We split the samples using the same methodology as described in 1.

We segment the units of text based on Wordpiece (Turc et al., 2019) or Sentencepiece (Kudo & Richardson, 2018) model, respectively, built for the supported languages of the particular model. We utilize a context of the classified sentences in a concatenation, separated by a special symbol "<s>", similarly to other sequence-pair problems, such as answer extraction, or entailment classification.

By default, we train the neural classifiers using an effective batch size of 32, i.e. the weights of the models are adjusted using the gradients aggregated over the 32 training samples. We set the *warmup* to 10% of the total training steps and by default, we schedule the training for 20 epochs. We use early-stopping on evaluation accuracy with a patience over 500 training steps. A training of a single classifier takes approximately 8 hours on two GPUs of Nvidia Tesla T4.

For final evaluation, we pick the model with the highest validation accuracy, as measured in every 50 training steps. Figure 7 shows a comparison of validation loss and accuracy of Multilingual-BERT-base-cased and XLM-RoBERTa-large trained on samples with minimal confidence threshold of 4.

Due to the computational complexity and the fact that we find our results quite consistent to adjustments of aforementioned parameters, we do not perform a systematic hyperparameter search of parameters of deep classifiers and set these only by our best knowledge and intuition.

C: Literature review of applying machine learning to reflective writing

Table 4 Overview of a literature applying machine learning to reflective writing

Authors	Size of dataset for annotation	Students	Background of participants	Categories of annotations	Inter-rater reliability
Carpenter et al. (2020)	728 responses	153	Middle school students	Five-point scale	ICC = .67
Cheng (2017)	10,002 sentences	398	University students (multidisc.)	Analysis, Strategy, Report, Reformulation, Influences	Cohen's κ=$.60 - .73$
Hu (2017)	584 sentences	29 from 120 reports	Secondary school in Hong Kong	Levels of thinking orders: High, Medium, Low	-
Jung and Wise (2020)	1,500 statements	369	Dental students	Description, Analysis, Feeling, Perspective, Evaluation, Outcome	Krippendorff's $\alpha > .70$
Jung and Wise (2020)	1,500 statements	369	Dental students	None, Shallow, Deep	Krippendorff's $\alpha > .67$
Kovanović et al. (2018)	4,430 sentences	77	Undergraduate courses in performing arts	Observation, Goal, Motive, Other	Cohen's $\kappa > .75$
Liu et al. (2017)	random sample 2,000 posts	6,650	In-service K12 teachers	Focus: Technical, Personalistic Level: Description, Analysis, Critique	Cohen's $\kappa = .747$
Liu et al. (2019)	301 statements	43	Pharmacy students	Experience, Feelings, Knowledge, Integration, Validation, Appropriation	ICC =$.55 - .69$
Ullmann (2019)	5,080 setences	76	University students (multidisc.)	Reflection, Experience, Feelings, Belief, Difficulties, Perspective, Outcome (learned, intention)	Cohen's κ=$.48 - .98$
Wulff et al. (2021)	1,966 segments	17	Student teachers of physics	Circumstances, Description, Evaluation, Alternatives, Consequences	Cohen's κ= .74

Table 4 (continued)

Authors	Algorithms	Language representations	Language features + best algorithm	Split (train / test)	Performance
Carpenter et al. (2020)	RF, SVM, NN	BiGram, TF-IDF, GloVe, Finetuned GloVe, ELMo	average ELMo + SVM	10-fold cross-validation	R-squared = .40–.64 *MSE* = .26–.43 MAE = .40–.51
Cheng (2017)	NB Binary Relevance	LSA	LSA + NB (ove-vs-all)	5-fold cross-validation	Accuracy =.72−.82
Hu (2017)	-	n-grams, Part-of-Speech, Semantic Relations (all as Bag-of-Words), recursive feature elimination	all + SVM	5-fold cross-validation	Accuracy =.68−.81 F1-score = .52 − .70
Jung and Wise (2020)	RF, SVM, NB PART (rule-based)	unigram BoW, \linguistic features" (from LIWC)	all + RF	80% / 20%	Accuracy = .75 − .96 Precision = .74 − .96
Jung and Wise (2020)	RF, SVM, NB PART (rule-based)	unigram BoW, \linguistic features" (from LIWC)	all + RF	80% / 20%	Accuracy = .77 Precision = .76
Kovanovic et al. (2018)	-	300 uni-, bi-, tri-grams + 93 LIWC linguistic features + 109 \cohesion features" (noun, verb, negation density, LSA, word overlap), SMOTE oversampling (pairwise linear approximation)	all + RF	75% / 25%10-foldcross-validation	Accuracy = .81 − .89
Liu et al. (2017)	LR, NB, SVM Decision Tree Boosting	n-gram BoW (of chinese marks), top 50−−250 features ranked by Document Frequency, weighted by TF-IDF	? + NB	70%/30%	F1-score =.79−.86 Precision = .78 − .91
Liu et al. (2019)	RF, PART (rule-based) SVM, NB	LIWC + AWA): features as feedback for analytical/reflective writing	best 22 + MetaCost (cost cat. balancing) + RF	10-fold cross-validation	F1-score =.41−.87 Precision = .59 − .87
Ullmann (2019)	RF, NB, SVM, NN	unigram Bag-of-Words	all + random oversampling + RF	80%/20%	Accuracy =.71−.96
Wulff et al. (2021)	Decision Tree MultiNB MultiLR SGDClass	-	BoW + Similar results for: MultiNB, MultiLR, SGD	81 reflections split into training and test datasets	F1-score =.62−.84 Precision = .60 − .83 (LR) F1-score = .58 − .86 Precision = .52 − −1 (MultiNB) F1-score = .59 − .77 Precision = .56 − .79 (SGD)

D: Parameters of Generalized linear mixed models

Table 5 Estimated regression parameters, standard errors, z-values and P-values for GLMMs presented

	Estimate	Std. Error	z value	Pr(> $\|z\|$)	
Other*					
(Intercept)	1.53	.05	28.12	$< 2e^{-16}$	***
2nd	−.08	.05	−1.79	.07	.
3rd	−.27	.05	−5.59	.00	***
4th	−.02	.05	−.32	.75	
5th	−.17	.05	−3.31	.00	***
journal_length	.03	.00	24.36	$< 2e^{-16}$	***
Fit statistics					
Deviance	5,543.10				
AIC	5,561.10				
Experience*					
(Intercept)	.94	.07	14.05	$< 2e^{-16}$	***
2nd	.19	.06	3.19	.001	**
3rd	.38	.06	6.59	$4.55e^{-11}$	***
4th	.16	.06	2.41	.016	*
5th	.35	.06	5.87	$4.31e^{-09}$	***
journal_length	.02	.00	17.20	$< 2e^{-16}$	***
Fit statistics					
Deviance	5,072.9				
AIC	5,090.9				
Reflection **					
(Intercept)	.70	.08	8.56	$< 2e^{-16}$	***
2nd	.02	.06	.31	.75	
3rd	−.17	.06	−2.68	.007	**
4th	−.31	.07	−4.36	$1.32e^{-5}$	***
5th	−.12	.07	−1.81	.07	.
journal_length	.02	.00	11.60	$< 2e^{-16}$	***
Fit statistics					
Deviance	4,310.3				
AIC	4,326.3				
Feelings **					
(Intercept)	.85	.08	11.32	$< 2e^{-16}$	***
2nd	−.12	.06	−2.00	.045	*
3rd	−.09	.06	−1.57	.12	
4th	−.04	.06	−.65	.52	
5th	−.25	.06	−3.96	$7.42e^{-05}$	***
journal_length	.02	.00	12.36	$< 2e^{-16}$	***
Fit statistics					
Deviance	4,470.5				
AIC	4,486.5				

Table 5 (continued)

	Estimate	Std. Error	z value	Pr(> $\|z\|$)	
Difficulty **					
(Intercept)	−.82	.14	−6.03	$1.62e^{-09}$	***
2nd	.38	.11	3.29	.001	**
3rd	.52	.11	4.61	$4.04e^{-06}$	***
4th	.06	.13	.45	.65	
5th	.54	.12	4.73	$2.28e^{-06}$	***
journal_length	.02	.00	7.99	$1.38e^{-15}$	***
Fit statistics					
Deviance	3001.7				
AIC	3017.7				
Perspective **					
(Intercept)	−1.52	.16	−9.27	$< 2e^{-16}$	***
2nd	.33	.15	2.24	.03	*
3rd	.41	.14	2.86	.004	**
4th	.02	.17	.14	.89	
5th	.29	.15	1.88	.06	.
journal_length	.03	.00	8.06	$7.58e^{-16}$	***
Fit statistics					
Deviance	2,090.5				
AIC	2,074.5				
Belief **					
(Intercept)	−.90	.16	−5.61	$2.02e^{-08}$	***
2nd	−.25	.14	−1.80	.07	.
3rd	−.09	.14	−.69	.49	
4th	−.03	.15	−.17	.86	
5th	.07	.14	.50	.62	
journal_length	.02	.00	5.54	$3.04e^{-08}$	***
Fit statistics					
Deviance	2,426.4				
AIC	2,410.4				
Learning **					
(Intercept)	−1.88	.22	−8.68	$< 2e^{-16}$	***
2nd	.00	.19	.02	.98	
3rd	.38	.18	2.13	.03	*
4th	.26	.20	1.30	.20	
5th	.53	.18	2.91	.004	**
journal_length	.01	.00	2.50	.013	*
Fit statistics					
Deviance	1,456.1				
AIC	1,472.1				

Table 5 (continued)

	Estimate	Std. Error	z value	Pr(> $\|z\|$)	
**Intention ** **					
(Intercept)	−2.09	.22	−9.44	$< 2e^{-16}$	***
2nd	−.15	.21	−.68	.49	
3rd	.12	.20	.59	.55	
4th	−.09	.23	−.41	.68	
5th	−.16	.22	−.74	.46	
journal_length	.02	.00	5.97	$2.31e^{-09}$	***
Fit statistics					
Deviance	1,209.8				
AIC	1,225.8				

Signif. codes: 0 ’***’ .001 ’**’ .01 ’*’ .05 ’.’ .1 ’ ’ 1

* Conway−Maxwell−Poisson distribution

** Negative binomial distribution

Abbreviations AWA: Academic Writing Analysis; CEReD: Czech-English Reflective Dataset; ELMo: Embeddings from Language Models; LIWC: Linguistic inquiry and word count; LR: Linear Regression; LSA: Latent Semantic Analysis; MultiLR: Multinominal Linear Regression; MultiNB: Multiniminal Naive Bayes; NB: Näive Bayes; NN: Fully-connected neural network; RF: Random Forest; SGD: Stochastic Gradient Descent; SVM: Support Vector Machines; TF-IDF: Term frequency–inverse document frequency

Acknowledgements We wish to give special thanks to Vít Novotný for their help with LateX.

Author Contributions **Both authors contributed equally in this work.** Jan Nehyba identified the need for identification of reflection in the educational applications and proposed the specific application of the classifier for the purpose of exploratory hypotheses. Based on the related work, Michal Štefánik identified the technologies applicable for a classification of reflection, implemented the reproducible sources for the experiments and collected the results presented in this article.

Funding This publication was written at Masaryk university as part of the project “Analysis of Reflective Journals of Teacher Students” number MUNI/A/1238/2019 with the support of the Specific University Research Grant, as provided by the Ministry of Education, Youth and Sports of the Czech Republic in the year 2020.

Data Availability All the resources and instructions for reproduction can be found in the repository of our library: https://github.com/EduMUNI/reflection-classification. Data set has been collected as part of this work and is available as well in anonymised version on the referenced repository with additional details in /data directory (this will be replaced by permanent, but not anonymised DOI in camera-ready version). The repository also contains reproducible sources for all our experiments:

- Training of the language models, which can be useful for adapting the model to significantly different domain of data.
- Qualitative evaluation of the trained language model on given set of data.
- Analyses: sources for a reproduction of the results of exploratory hypotheses introduced in Section 3.5 can be found in /analyses directory. Please refer to the main README of the referenced repository for specific steps to perform a reproduction.

Declarations

Conflicts of interest Authors are required to disclose financial or non-financial interests that are directly or indirectly related to the work submitted for publication. Please refer to "Competing Interests and Funding "below for more information on how to complete this section.

References

Alger, C. (2006). 'what went well, what didn't go so well': Growth of reflection in pre-service teachers. *Reflective practice, 7*(3), 287–301.

Arrastia, M. C., Rawls, E. S., Brinkerhoff, E. H., & Roehrig, A. D. (2014). The nature of elementary preservice teachers' reflection during an early field experience. *Reflective Practice, 15*(4), 427–444.

Bahdanau, D., Cho, K., Bengio, Y. (2014) Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

Bain, J. D., Mills, C., Ballantyne, R., & Packer, J. (2002). Developing reflection on practice through journal writing: Impacts of variations in the focus and level of feedback. *Teachers and Teaching, 8*(2), 171–196.

Bass, J., Sidebotham, M., Creedy, D., & Sweet, L. (2020). Midwifery students' experiences and expectations of using a model of holistic reflection. *Women and Birth, 33*(4), 383–392.

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., & Grothendieck, G. (2012). *Package 'lme4'*. R Foundation for Statistical Computing, Vienna, Austria: CRAN.

Bean, T. W., & Stevens, L. P. (2002). Scaffolding reflection for preservice and inservice teachers. *Reflective Practice, 3*(2), 205–218.

Beaumont, A., Al-Shaghdari, T. (2019) To what extent can text classification help with making inferences about students' understanding. International conference on machine learning, optimization, and data science (372–383)

Bolton, G. (2010) Reflective practice: Writing and professional development. Sage publications

Boud, D., Keogh, R., & Walker, D. (1985). *Reflection: Turning experience into learning*. London: Kogan Page.

Bruno, A., Galuppo, L., & Gilardi, S. (2011). Evaluating the reflexive practices in a learning experience. *European Journal of Psychology of Education, 26*(4), 527–543.

Cardenas, D. G. (2014). Learning networks to enhance reflectivity: key elements for the design of a reflective network. *International Journal of Educational Technology in Higher Education, 11*(1), 32–48.

Carpenter, D., Geden, M., Rowe, J., Azevedo, R., Lester, J. (2020) Automated analysis of middle school students' written reflections during game-based learning. International conference on artificial intelligence in education (67–78)

Cattaneo, A.A., Motta, E. (2020) "I reflect, therefore i am... a good professional". on the relationship between reflection-on-action, reflection-in-action and professional performance in vocational education. Vocations and Learning (1–20)

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.

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)

Chou, P. N., & Chang, C. C. (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 course. *Turkish Online Journal of Educational Technology-TOJET, 10*(3), 101–114.

Cochran-Smith, M. (2005). Studying Teacher Education: What we know and need to know. *Journal of Teacher Education, 54*(4), 301–306. https://doi.org/10.1177/0022487105280116

Cohen-Sayag, E., & Fischl, D. (2012). Reflective writing in pre-service teachers' teaching: What does it promote? *Australian Journal of Teacher Education, 37*(10), 2.

Colton, A. B., & Sparks-Langer, G. M. (1993). A conceptual framework to guide the development of teacher reflection and decision making. *Journal of teacher education, 44*(1), 45–54.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Stoyanov, V. (2020) Unsupervised cross-lingual representation learning at scale. Proceedings of the 58th annual meeting of the association for computational linguistics (8440–8451). Association for Computational Linguistics. https://aclanthology.org/2020.acl-main.747. https://doi.org/10.18653/v1/2020.acl-main.747. Accessed 16 June 2022

Conneau, A., Lample, G. (2019) Cross-lingual Language Model Pretraining. H.M. Wallach, H. Larochelle, A. Beygelzimer, F. D'Alché-Buc, E.B. Fox, R. Garnett. Advances in neural information processing systems 32: Annual conference on neural information processing systems 2019, neurips 2019, december 8-14, 2019, Vancouver, BC, Canada (7057–7067). Retrieved from https://papers.nips.cc/paper/2019/hash/c04c19c2c2474dbf5f7ac4372c5b9af1-Abstract.html. Accessed 16 June 2022

Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological), 20*(2), 215–232.

Cristobal, E., Flavian, C., & Guinaliu, M. (2007). Perceived e-service quality (PeSQ): Measurement validation and effects on consumer satisfaction and web site loyalty. *Managing Service Quality: An International Journal, 17*(3), 317–340. https://doi.org/10.1108/09604520710744326

Cui, Y., Wise, A. F., & Allen, K. L. (2019). Developing reflection analytics for health professions education: A multi-dimensional framework to align critical concepts with data features. *Computers in Human Behavior, 100*, 305–324.

Darling, L. F. (2001). Portfolio as practice: The narratives of emerging teachers. *Teaching and Teacher Education, 17*(1), 107–121.

Devlin, J., Chang, M.W., Lee, K., Toutanova, K. (2019) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers) (4171–4186). Minneapolis, Minnesota ACL. https://doi.org/10.18653/v1/N19-1423

Dyment, J. E., & O'Connell, T. S. (2011). Assessing the quality of reflection in student journals: A review of the research. *Teaching in Higher Education, 16*(1), 81–97. https://doi.org/10.1080/13562517.2010.507308

Fallon, M.A., Brown, S.C., Ackley, B.C. (2003) Reflection as a strategy for teaching performance-based assessment. Brock Education Journal 13(1)

Faraway, J. J. (2016). *Extending the linear model with r: generalized linear, mixed effects and nonparametric regression models*. CRC Press.

Finlay, L. (2008). *Reflecting on reflective practice. PBPL Paper, 52*, 1–27.

Fort, K. (2016) Collaborative annotation for reliable natural language processing: Technical and sociological aspects. Wiley

Fox, R. K., Dodman, S., & Holincheck, N. (2019). Moving beyond reflection in a hall of mirrors: developing critical reflective capacity in teachers and teacher educators. *Reflective Practice, 20*(3), 367–382.

Fox, R.K., White, C.S. (2010) Examining teachers' development through critical reflection in an advanced master's degree program. The purposes, practices, and professionalism of teacher reflectivity: Insights for twenty-first-century teachers and students (3–24)

García-Gorrostieta, J. M., López-López, A., & González-López, S. (2018). Automatic argument assessment of final project reports of computer engineering students. *Computer Applications in Engineering Education, 26*(5), 1217–1226.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press.

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

Hanafi, M. (2019) Perceptions of reflection on a pre-service primary teacher education programme in teaching english as a second language in an institute of teacher education in malaysia (Unpublished doctoral dissertation). Canterbury Christ Church University

Hartig, F. (2019) DHARMa: residual diagnostics for hierarchical (multi-level/mixed) regression models. R package version 0.2,4

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

Hedlund, D. E. (1989). A dialogue with self: The journal as an educational tool. *Journal of Humanistic Education and Development, 27*(3), 105–13.

Hoffman, L., & Rovine, M. J. (2007). Multilevel models for the experimental psychologist: Foundations and illustrative examples. *Behavior Research Methods, 39*(1), 101–117.

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.

Hu, X. (2017) Automated recognition of thinking orders in secondary school student writings. *Learning: Research and Practice, 3*(1) 30–41

Hume, A. (2009). A Personal Journey: Introducing Reflective Practice into Pre-Service Teacher Education to Improve Outcomes for Students. *Teachers and Curriculum, 11*, 21–28.

Jiang, J. (2017). *Asymptotic analysis of mixed effects models: theory, applications, and open problems*. CRC Press.

Jung, Y., Wise, A.F. (2020) How and how well do students reflect? multi-dimensional automated reflection assessment in health professions education. Proceedings of the Tenth International Conference on Learning Analytics & Knowledge (595–604)

King, P. M., & Kitchener, K. S. (2004). Reflective judgment: Theory and research on the development of epistemic assumptions through adulthood. *Educational psychologist, 39*(1), 5–18.

Kinsella, E. A. (2007). Embodied reflection and the epistemology of reflective practice. *Journal of Philosophy of Education, 41*(3), 395–409.

Klein, S. R. (2008). Holistic reflection in teacher education: Issues and strategies. *Reflective Practice, 9*(2), 111–121.

Knight, S., Shibani, A., Abel, S., Gibson, A., Ryan, P., Sutton, N., et al. (2020). Acawriter: A learning analytics tool for formative feedback on academic writing. *Journal of Writing Research, 12*(1), 141–186.

Kolb, D. (2014) Neprobádaný život nestojí za život. J. Nehyba, B. Lazarová (Eds.), Reflexe v procesu učení. desetkrát stejně a přece jinak. (23-30). Masaryk University Press

Kovanović, V., Joksimović, S., Mirriahi, N., Blaine, E., Gašević, D., Siemens, G., Dawson, S. (2018) Understand students' self-reflections through learning analytics. Proceedings of the 8th international conference on learning analytics and knowledge (389–398)

Krol, C.A. (1996) Preservice Teacher Education Students' Dialogue Journals: What Characterizes Students' Reflective Writing and a Teacher's Comments. ERIC. Retrieved from https://files.eric.ed.gov/fulltext/ED395911.pdf (Paper presented at the Annual Meeting of the Association of Teacher Educators (76th, St. Louis, MO). Accessed 16 June 2022

Kudo, T., Richardson, J. (2018) SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. Proceedings of the 2018 conference on empirical methods in natural language processing: System demonstrations (66–71). Brussels, Belgium, Association for Computational Linguistics. Retrieved from https://aclanthology.org/D18-2012https://doi.org/10.18653/v1/D18-2012. Accessed 16 June 2022

LaBoskey, V. K. (1994). *Development of reflective practice: A study of preservice teachers*. New York: Teachers College Press.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R. (2020) ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. Proceedings of International Conference on Learning Representations, ICLR. Retrieved from https://openreview.net/forum?id=H1eA7AEtvS. Accessed 16 June 2022

Larrivee, B., Cooper, J. (2006) An educator's guide to teacher reflection. Houghton Mifflin. Retrieved from https://books.google.cz/books?id=tVaYL-x67ekC. Accessed 16 June 2022

Lee, H. J. (2005). Understanding and assessing preservice teachers' reflective thinking. *Teaching and Teacher Education, 21*(6), 699–715.

Lee, I. (2008). Fostering preservice reflection through response journals. *Teacher Education Quarterly, 35*(1), 117–139.

Lepp, L., Kuusvek, A., Leijen, Ä., Pedaste, M., Kaziu, A. (2020) Written or video diary-which one to prefer in teacher education and why? 2020 ieee 20th international conference on advanced learning technologies (icalt) (276–278)

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Zettlemoyer, L. (2020) BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. Proceedings of the 58th annual meeting of the association for computational linguistics

(7871–7880). Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.acl-main.703. Accessed 16 June 2022

Lindroth, J.T. (2015) Reflective journals: A review of the literature. *Update: Applications of Research in Music Education, 34*(1)66–72. https://doi.org/10.1177/8755123314548046

Liu, Q., Zhang, S., Wang, Q., & Chen, W. (2017). Mining online discussion data for understanding teachers reflective thinking. *IEEE Transactions on Learning Technologies, 11*(2), 243–254.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Stoyanov, V. (2019) RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692

Loughran, J. (2007) Enacting a pedagogy of teacher education. Enacting a pedagogy of teacher education (11–25). Routledge

Loughran, J., & Corrigan, D. (1995). Teaching portfolios: A strategy for developing learning and teaching in preservice education. *Teaching and teacher Education, 11*(6), 565–577.

Magnusson, A., Skaug, H., Nielsen, A., Berg, C., Kristensen, K., Maechler, M., Brooks, M.M. (2017) Package 'glmmTMB' Package 'glmmTMB'. R Package Version 0.2.0

Maloney, C., & Campbell-Evans, G. (2002). Using interactive journal writing as a strategy for professional growth. *Asia-Pacific Journal of Teacher Education, 30*(1), 39–50.

Mena-Marcos, J., Garcia-Rodriguez, M. L., & Tillema, H. (2013). Student teacher reflective writing: what does it reveal? *European Journal of Teacher Education, 36*(2), 147–163.

Moon, J. A. (2006). Learning journals: A handbook for reflective practice and professional development. *London, Routledge.*https://doi.org/10.4324/9780429448836-8

Nakayama, H., Kubo, T., Kamura, J., Taniguchi, Y., Liang, X. (2018) doccano: Text annotation tool for human. Software available from https://github.com/doccano/doccano. Accessed 16 June 2022

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric Theory*. McGraw Hill.

Pasternak, D.L., Rigoni, K.K. (2015) Teaching reflective writing: thoughts on developing a reflective writing framework to support teacher candidates. Teaching/Writing: The Journal of Writing Teacher Education 4(1) 5

Pedro, J. Y. (2005). Reflection in teacher education: exploring pre-service teachers' meanings of reflective practice. *Reflective practice, 6*(1), 49–66.

R Core Team (2020) R: A language and environment for statistical computing [Computer Software Manual]. Vienna, Austria. Retrieved from https://www.R-project.org/. Accessed 16 June 2022

Ryken, A. E., & Hamel, F. L. (2016). Looking again at "surface-level'' reflections: Framing a competence view of early teacher thinking. *Teacher Education Quarterly, 43*(4), 31–53.

Schwitzgebel, E. (2010). Acting contrary to our professed beliefs or the gulf between occurrent judgment and dispositional belief. *Pacific Philosophical Quarterly, 91*(4), 531–553.

Shoffner, M. (2008). Informal reflection in pre-service teacher education. *Reflective Practice, 9*(2), 123–134.

Shum, S. B., 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.

Spalding, E., Wilson, A., & Mewborn, D. (2002). Demystifying reflection: A study of pedagogical strategies that encourage reflective journal writing. *Teachers college record, 104*(7), 1393–1421.

Štefánik, M. & Nehyba, J. (2021). Czech-English Reflective Dataset (CEReD). (Version V1) [Data set]. *GitHub*. 11372/LRT-3573

Stiler, G.M., Philleo, T. (2003) Blogging and blogspots: An alternative format for encouraging reflective practice among preservice teachers. Education 123(4)

Stroup, W. W. (2012). *Generalized linear mixed models: modern concepts, methods and applications*. CRC Press.

Tan, J. (2013). Dialoguing written reflections to promote self-efficacy in student teachers. *Reflective Practice, 14*(6), 814–824.

Turc, I., Chang, M.W., Lee, K., Toutanova, K. (2019) Well-Read Students Learn Better: On the Importance of Pre-training Compact Models. https://www.youtube.com/watch?v=LoyyKVJgHKoarXiv:1908.08962. Accessed 16 June 2022

Ukrop, M., Švábenský, V., Nehyba, J. (2019) Reflective diary for professional development of novice teachers. Proceedings of the 50th ACM technical symposium on computer science education (1088–1094)

Ullmann, T. D. (2019). Automated analysis of reflection in writing: Validating machine learning approaches. *International Journal of Artificial Intelligence in Education, 29*(2), 217–257.

Van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual*. CA CreateSpace: Scotts Valley.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Polosukhin, I. (2017) Attention is All you Need. I. Guyon et al. (Eds,), Advances in neural information processing systems (Vol 30). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a8 45aa-Paper.pdf. Accessed 16 June 2022

Wallin, P., & Adawi, T. (2018). The reflective diary as a method for the formative assessment of self-regulated learning. *European Journal of Engineering Education, 43*(4), 507–521.

Ward, J. R., & McCotter, S. S. (2004). Reflection as a visible outcome for preservice teachers. *Teaching and Teacher Education, 20*(3), 243–257.

Whipp, J., Wesson, C., & Wiley, T. (1997). Supporting collaborative reflections: Case writing inanurban pds. *Teaching Education, 9*(1), 127–134.

Wilcox, B. L. (1996). Smart portfolios for teachers in training. *Journal of Adolescent & Adult Literacy, 40*(3), 172–179.

Wulff, P., Buschhüter, D., Westphal, A., Nowak, A., Becker, L., Robalino, H., & Borowski, A. (2021). Computer-Based Classification of Preservice Physics Teachers' Written Reflections. *Journal of Science Education and Technology, 30*(1), 1–15.

Xie, Q., Dai, Z., Hovy, E., Luong, M.T., Le, Q.V. (2020) Unsupervised data augmentation for consistency training. arXiv: Learning. Retrieved from. arXiv: 1904.12848

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education-where are the educators? *International Journal of Educational Technology in Higher Education, 16*(1), 1–27.

Zeichner, K., & Wray, S. (2001). The teaching portfolio in US teacher education programs: What we know and what we need to know. *Teaching and Teacher Education, 17*(5), 613–621.