JOURNAL OF LEARNING ANALYTICS

SOLAR SOCIETY for LEARNING ANALYTICS RESEARCH

Volume 9(2), 236–252. https://doi.org/10.18608/jla.2022.7551

Design Analytics for Mobile Learning: Scaling up the Classification of Learning Designs Based on Cognitive and Contextual Elements

Gerti Pishtari[1], Luis P. Prieto[2], María Jesús Rodríguez-Triana[3], Roberto Martinez-Maldonado[4]

Abstract

This research was triggered by the identified need in literature for large-scale studies about the kinds of designs that teachers create for mobile learning (m-learning). These studies require analyses of large datasets of learning designs. The common approach followed by researchers when analyzing designs has been to manually classify them following high-level pedagogically guided coding strategies, which demands extensive work. Therefore, the first goal of this paper is to explore the use of supervised machine learning (SML) to automatically classify the textual content of m-learning designs using pedagogically relevant classifications, such as the cognitive level demanded by students to carry out specific designed tasks, the phases of inquiry learning represented in the designs, or the role that the situated environment has in the designs. Because not all SML models are transparent, but researchers often need to understand their behaviour, the second goal of this paper is to consider the trade-off between models' performance and interpretability in the context of design analytics for m-learning. To achieve these goals, we compiled a dataset of designs deployed using two tools, Avastusrada and Smartzoos. With this dataset, we trained and compared different models and feature extraction techniques. We further optimized and compared the best performing and most interpretable algorithms (EstBERT and Logistic Regression) to consider the second goal with an illustrative case. We found that SML can reliably classify designs with accuracy > 0.86 and Cohen's kappa > 0.69.

Notes for Practice

- Design analytics looks at overall trends and gaps in online communities of teachers and other designers of learning experiences.
- We show how supervised machine learning can automatically categorize learning designs from such online communities at scale, according to high-level pedagogical features (e.g., Bloom's taxonomy of learning activities).
- The machine learning algorithm to use (e.g., interpretable, black box, or a combination of both) should be chosen based on evaluations of its respective performance and the needs of the target users (e.g., teachers need to understand the reasoning behind design recommendations suggested automatically by the system).

Keywords

Supervised machine learning, learning design, learning analytics, analytics for learning design, mobile learning, contextual learning

Submitted: 22/07/2021 — **Accepted:** 05/03/2022 — **Published:** 31/08/2022

Corresponding author [1] *Email: gerti.pishtari@donau-uni.ac.at Address: University for Continuing Education Krems (Danube University Krems), Dr.-Karl-Dorrek-Straße 30, 3500 Krems, Austria; Tallinn University, School of Digital Technologies, Narva Maantee 25, 10120 Tallinn, Estonia. ORCID ID: https://orcid.org/0000-0001-9451-6881*

[2] *Email: lprisan@tlu.ee Address: Tallinn University, School of Educational Sciences, Narva Maantee 25, 10120 Tallinn, Estonia. ORCID ID: https://orcid.org/0000-0002-0057-0682*

[3] *Email: mjrt@tlu.ee Address: Tallinn University, School of Digital Technologies, Narva Maantee 25, 10120 Tallinn, Estonia. ORCID ID: https://orcid.org/0000-0001-8639-1257*

[4] *Email: roberto.martinezmaldonado@monash.edu Address: Monash Univerersity, Faculty of Information Technology, 20 Exhibition Walk, Clayton, VIC, 3800, Australia. ORCID ID: https://orcid.org/0000-0002-8375-1816*

ISSN 1929-7750 (online). The Journal of Learning Analytics works under a Creative Commons License, Attribution - NonCommercial-NoDerivs 3.0 Unported (CC BY-NC-ND 3.0)

1. Introduction

Mobile learning (m-learning) environments promote learning across physical and digital spaces, with merged characteristics from formal and informal settings (Muñoz-Cristóbal, Rodríguez-Triana, et al., 2018; Pishtari et al., 2020). In m-learning, teachers often design learning activities that encourage students to learn in creative and contextualized settings (Sharples, 2015). While designing learning activities is already a demanding task, m-learning entails added complexity because teachers have to consider multiple spaces and settings (Muñoz-Cristóbal, Rodríguez-Triana, et al., 2018). Indeed, understanding how teachers design in m-learning is still an open issue (Mettis & Väljataga, 2021).

Understanding and supporting teachers' practices and needs when designing for learning are the main goals of the field of learning design (LD) (Persico & Pozzi, 2015). To achieve these goals, research in LD has proposed various theoretical and practical solutions that consider the different phases of the LD process (Pozzi et al., 2016) or that promote communities of teachers to co-create and share good design practices (Hernández-Leo et al., 2014; Dalziel et al., 2016). Yet, when trying to understand teachers' design practices, one of the common strategies in LD has been to analyze the design artifacts created by teachers (i.e., the end product of the LD process). The design artifacts may consist of, among others, the objectives and structure of a learning activity, as well as the set of tasks that students have to perform during it.

While research in analytics for LD has often applied automatic analysis of the LD artifacts, usually by focusing on (teachers' or students') logged actions (such as in Toetenel & Rienties, 2016, which analyzes teachers' pedagogical decisions), or the structural aspects of the designs (e.g., de Jong et al., 2021), when researchers want to investigate more abstract aspects (e.g., the cognitive level required by students to perform tasks that teachers have included in the designs), the common practice has been to manually code the artifacts based on a set of (high-level pedagogically grounded) codes. See, for instance, Mettis and Väljataga (2021) for an example that classifies the learning tasks that teachers have included in their designs based on the cognitive level required by students, or Rodriguez and colleagues (2021), which labels the designs based on the phases of inquiry learning that they include. This manual coding approach not only demands extensive and repetitive work by researchers (thus limiting the amount of data that can be analyzed) but also provides only a partial view of the elements under assessment (e.g., a limited picture of the kinds of learning designs created by a community of teachers behind a specific m-learning tool).

Therefore, educational researchers would benefit from systems that would scale up these manual approaches by automatically and reliably classifying/coding LD artifacts created by teachers. In the context of m-learning, such automatic classifications would enable researchers to better understand the design practices of large communities of LD (which, as we previously mentioned, is still an open issue) by analyzing big datasets of learning designs, based on a set of high-level pedagogically grounded codes.

Supervised machine learning (SML) approaches that first train algorithms that can learn from a small set of labelled training data and that, once trained, can be deployed to predict outcomes for large-scale unforeseen data could help to tackle this problem. While the communities of learning analytics (LA) and educational data mining (EDM) have already used SML techniques informed by learners' traces, mainly to assess learners' performance (Chen & Cui, 2020; Bulathwela et al., 2020; Xu et al., 2019), a less explored area of application in SML is to support the automated analysis of LD artifacts created by teachers. Moreover, recent research emphasizes that, when their performance allows it, interpretable and explainable machine learning models not only better inform teaching and learning practices (e.g., by informing teachers and students why the system is suggesting a specific action) but also foster trust in their users (Conati et al., 2018). Indeed, educational researchers might benefit from systems that are interpretable and can explain the reason behind decisions made or classifications done by the algorithm to better understand teachers' design practices (e.g., by providing more detailed feedback about the design practices of specific teachers).

The contribution of this paper is twofold. The main goal, as a first step toward using analytics to understand teachers' design practices in m-learning, is to *explore the potential use of SML approaches to automatically classify datasets of learning designs (more specifically, the textual content included in the designs) that are created by teachers in the context of m-learning*. The second goal is to provide *a comparative example of the trade-off between performance and interpretability in the context of analytics for LD in m-learning*.

To achieve these goals, we first extracted a dataset that included a representative sample of the learning designs (i.e., their textual content in Estonian) from two m-learning tools, Avastusrada and Smartzoos (further discussed in the following sections). We then manually labelled it following similar qualitative coding strategies as in Mettis and Väljataga (2021) and Rodriguez and colleagues (2021). Namely, we were guided by (1) the revised Bloom's taxonomy (Krathwohl, 2002) to understand the cognitive level of each learning task that teachers included in their designs and that learners had to perform and (2) the categorizations of the role that the context can play during a learning activity (Sharples, 2015) to inquire about the role played by the situated environment in each task. Furthermore, we also considered (3) the inquiry-based learning (IBL) model proposed by Pedaste and colleagues (2015) to understand the extent to which IBL pedagogies (relevant in the context of Avastusrada and Smartzoos) are present in the designs. We chose these coding strategies both for their relevance in the context of m-learning and

ISSN 1929-7750 (online). The Journal of Learning Analytics works under a Creative Commons License, Attribution - NonCommercial-NoDerivs 3.0 Unported (CC BY-NC-ND 3.0)

to compare them with other studies that used them to manually evaluate learning designs in m-learning (Mettis & Väljataga, 2021; Sharples, 2015; Rodríguez-Triana et al., 2021). Using this manually labelled dataset (the textual content as input features and the codes as the output that should be predicted), we trained, evaluated, and compared the performance of several SML models and feature extraction techniques (further discussed in the Methodology section). To tackle the second goal, we further optimized and compared the best performing algorithm (in our case, a black-box model based on EstBERT) with the most interpretable one (i.e., logistic regression) using an illustrative case.

To the best of our knowledge, this is the first paper exploring the potential use of SML to scale up the automatic classification of learning designs, using high-level pedagogically guided coding strategies, usually conducted manually by researchers (such as the strategy discussed above). Moreover, this paper also contributes to the research attempts at providing meaningful analytics for LD, in m-learning and beyond.

The rest of this paper is structured as follows. Section 2 describes the related work, while Section 3 explains the methodology followed. Section 4 presents the results. Section 5 discusses the implications and limitations. Finally, Section 6 concludes the paper with an overview of future work.

2. Related Work

2.1 M-Learning from an LD Perspective

M-learning is often connected in literature with learning environments augmented with mobile and sensor technologies that promote connectivity, interactivity, and learners' authenticity (Sharples, 2015; Pishtari et al., 2020). As mentioned above, in m-learning environments, learning may happen across multiple contexts (such as in a blended learning scenario happening partially online and partially physically in person) and spaces (e.g., across a digital space online and a physical space or outdoors) (Muñoz-Cristóbal, Rodríguez-Triana, et al., 2018).

Teachers are best placed to design m-learning activities that connect learning in the classroom, at home, and outdoors (Sharples, 2015). In this manuscript, we use *teachers* to refer to both educators of formal institutions and learning designers that design for m-learning in informal/non-formal settings, because we do not distinguish between them. Indeed, the field of LD has proposed various authoring tools that allow teachers to design m-learning activities (Pishtari & Rodríguez-Triana, 2022). For instance, GLUESP-AR allows teachers to design m-learning activities that happen across multiple digital and physical spaces (Muñoz-Cristóbal, Rodríguez-Triana, et al., 2018), QuesTInSitu supports the design of location-based learning activities outdoors (Santos et al., 2011), and EtiquetAR does the same in indoor environments (Pérez–Sanagustín et al., 2013).

The content and structure of the LD artifacts that teachers create depend on the authoring tool. For example, in QuesTInSitu, learning designs are composed of a number of questions (connected to specific geographical points in the map) that teachers create by using predefined templates (such as correct answers and free-form answers) and that include a combination of textual and audiovisual content (e.g., a question that also includes a video that students have to watch) (Santos et al., 2011).

Apart from the added complexity of designing across spaces, settings, and contexts, designing in m-learning also demands that teachers possess extensive technical and pedagogical competencies, due to both the novelty of the field and its related pedagogical and technical aspects. Indeed, Mettis and Väljataga (2021) analyzed 25 designs from teachers of a training program in m-learning and stated that teachers included mostly decontextualized learning tasks in their designs (i.e., tasks or questions that were not connected to the situated environment where learning was supposed to happen) and that were classified as lower-order thinking according to the revised Bloom's taxonomy (i.e., requiring students only to remember and understand concepts, instead of also being able to apply the learned concepts in new situations and perform related analysis, evaluations, etc.) (Krathwohl, 2002). The authors state that this happened even though the teachers that participated were supposed to have learned how to produce more elaborate designs since the early years of their pre-service education (Mettis & Väljataga, 2021). Thus, this suggests that further research is required to investigate the real-life pedagogical value of existing authoring tools used by teachers to design learning activities in m-learning, as well as teachers' needs and practices when designing for m-learning.

To address this gap, researchers would need to conduct large-scale studies on *how teachers design for m-learning*. According to Muñoz-Cristóbal, Hernández-Leo, and colleagues (2018), to respond to this question it is necessary to consider four dimensions: the temporal dimension, which includes the different stages of the design workflow followed by teachers; the physical dimension, which corresponds to the spaces in which the design work is carried out; the social dimension, which specifies the division of design labour among members; and the tasks dimension, which includes the creation of the learning tasks. In the current study, we focus mainly on the task dimension. More specifically, we focus on analyzing the design artifacts created by teachers, which could help to respond to the question, *what kind of design do teachers create for m-learning?* Responding to this question requires the analysis of large datasets of design artifacts that have been created by communities of teachers in m-learning, in real-life settings (e.g., communities behind specific authoring tools). Moreover, to conduct these large-scale studies, it would be necessary to apply an automatic coding strategy (because it is not feasible to follow a manual coding approach as in Mettis & Väljataga, 2021), which can also mimic abstract, or high-level, pedagogically guided

ISSN 1929-7750 (online). The Journal of Learning Analytics works under a Creative Commons License, Attribution - NonCommercial-NoDerivs 3.0 Unported (CC BY-NC-ND 3.0)

classifications that are relevant to the context of m-learning. For instance, Mettis and Väljataga (2021) manually classified m-learning designs by focusing on the the role played by the situated learning environment in each design and the cognitive level that is required by students to perform the tasks included in the designs, while Rodríguez-Triana and colleagues (2021) labelled the designs based on the different phases of IBL that they include.

Therefore, the current paper explores *the feasibility of using SML to automatically classify datasets of m-learning design artifacts, based on a high-level pedagogically guided coding strategy*. As a first step, we consider only the textual content of the learning tasks included in the design artifacts as input features for the SML algorithms. Although design artifacts also include other attributes (for instance, we previously mentioned the case of QuesTInSitu and the different types of questions and learning resources that it allows), these are usually case-dependent, differing based on the tool. Therefore, using only the textual content of the learning designs as an input feature for the SML algorithms allows for the development of platform-independent algorithms that can be used to analyze design artifacts from multiple m-learning tools.

2.2 SML in Technology-Enhanced Learning

Various researchers in the fields of LA and EDM have applied SML techniques mainly to predict students' performance. For instance, Chen and Cui (2020) use deep learning to analyze students' temporal behaviour in a learning management system in order to predict their course performance, Xu and colleagues (2019) use learners' internet usage behaviour to predict their success, while Bulathwela and colleagues (2020) use SML to predict student engagement in video lectures. There have also been attempts to use SML to support researchers, such as in Prieto and colleagues (2021), which inquires about the potential use of SML to automatically code diaries of students' learning progress.

A commonality of the aforementioned studies is that they all use SML techniques informed by learners' traces. Yet, a less explored area of application in SML is to support the automated analysis of artifacts created by teachers in *design time*. These analyses can help us understand and optimize not only teaching and learning practices but also the learning environment where they occur (Ferguson, 2012). Therefore, this paper presents *an example of SML techniques used to automatically classify datasets of m-learning design artifacts*, guided by relevant theoretical models (see Section 3.1.2), and using the textual content of the learning tasks that they include as input features for the algorithms.

Despite the potential of SML, the trade-off between accuracy and interpretability is an open question in the field of m-learning (Ribeiro et al., 2016; Rudin & Radin, 2019). Ribeiro and colleagues (2016) state that one should not blindly trust the decisions of SML models that are not interpretable and ask about the purpose of using black-box models instead of interpretable ones when they offer comparable performance. Depending on the context, interpretable models might also be essential to support different stakeholders in educational settings, for example, in intelligent tutoring systems that apply models that are open to learners (i.e., that not only provide suggestions about actions that learners should take but also inform them on why the system is suggesting a specific action) (Conati et al., 2018). Thus, in this paper we also *compare black-box and interpretable models in our context of automatically coding m-learning designs* using an example.

2.3 Aligning LD and LA in m-Learning

LD and LA solutions can complement each other (Macfadyen et al., 2020). LD can provide a context, guiding pedagogically grounded LA that can be more meaningful for teachers and students (Persico & Pozzi, 2015). LA can inform LD practices, such as design decisions or learning design evaluation. Hernández-Leo and colleagues (2019) describe three levels where analytics can inform LD practices: during the deployment of learning designs (i.e., through analytics that can inform LD practices), during the evaluation of the designs (i.e., through design analytics), and in the broader context of a community of teachers behind specific LD platforms (i.e., through analytics at the community level, or community analytics).

In the context of m-learning, Pishtari and colleagues (2020) found only three studies that explicitly address this alignment. For example, Munõz-Cristóbal, Rodríguez-Triana, and colleagues (2018) inquire about the use of LA to inform teachers' LD practices when designing learning activities that happen across multiple physical and digital spaces (e.g., online environments), while Melero and colleagues (2015) investigate the potential use of analytics to inform design decisions in outdoor situated learning environments. Nevertheless, these studies have mainly focused on LA for LD and have not considered the other levels discussed previously, such as design analytics and community analytics for LD. The current paper contributes to the research attempts of aligning LD and LA in m-learning by *providing evidence of the potential use of SML techniques that can automatically classify the textual content of design artifacts (based on a set of predefined codes) and that can later be used to analyze entire databases of LD artifacts from existing m-learning tools (i.e., connected to design and community analytics for LD)*.

3. Dataset and Methodology

This study is guided by the following research questions (RQs):

ISSN 1929-7750 (online). The Journal of Learning Analytics works under a Creative Commons License, Attribution - NonCommercial-NoDerivs 3.0 Unported (CC BY-NC-ND 3.0)

(RQ1) To what extent can different SML approaches automatically code datasets of m-learning design artifacts in terms of cognitive level, context, and inquiry phases?

(RQ2) What is the trade-off between performance and interpretability from the perspective of the end-users when choosing between the SML approaches in the context of LD in m-learning?

To respond to these RQs, we conducted an exploratory study consisting of four steps (S1–4), further depicted in Figure 1.

S1: We developed/constructed a dataset with the textual content of 1,472 learning tasks (i.e., questions), included in the m-learning design artifacts, and manually labelled them based on a predefined set of codes (discussed in Section 3.1.2).

S2: We trained, evaluated, and compared algorithms based on the most common SML models and feature extraction techniques for natural language processing, using the dataset described in S1.

S3: We further optimized the parameters of the best performing and most interpretable algorithms that resulted from S2.

S4: We used different techniques (e.g., coefficient interpretation) to provide a rather general interpretation for the best models of S3 (when possible, for interpretable ones), or only for individual classifications (for black-box models).

The rest of this section describes each of these steps in detail.

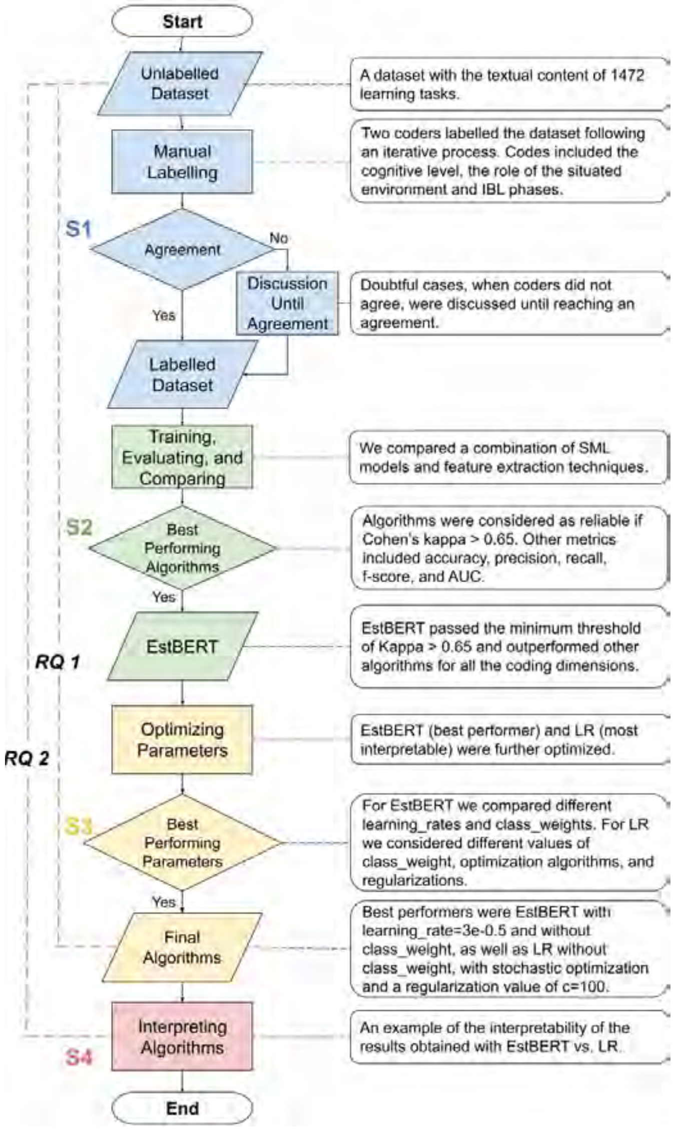
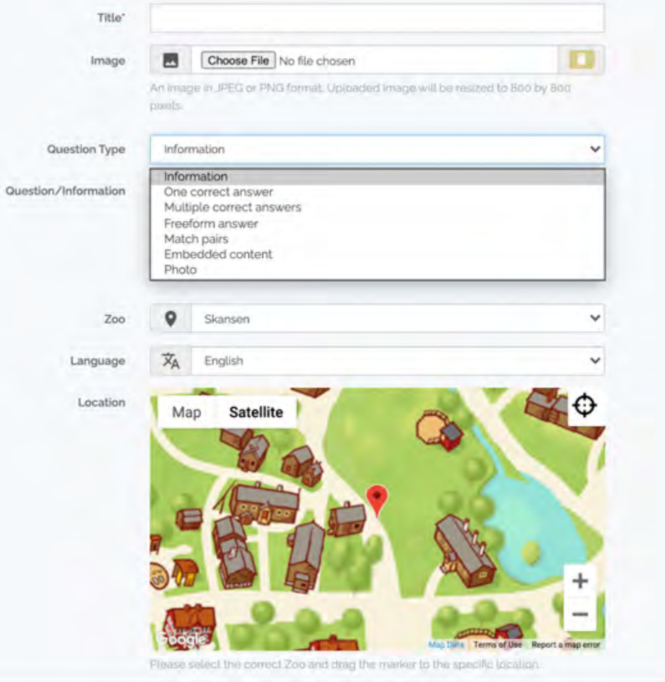

Start
Unlabelled Dataset
A dataset with the textual content of 1472 learning tasks.
Manual Labelling
Two coders labelled the dataset following an iterative process. Codes included the cognitive level, the role of the situated environment and IBL phases.
S1
Agreement
No
Discussion Until Agreement
Doubtful cases, when coders did not agree, were discussed until reaching an agreement.
Yes
Labelled Dataset
Training, Evaluating, and Comparing
We compared a combination of SML models and feature extraction techniques.
S2
Best Performing Algorithms
Algorithms were considered as reliable if Cohen's kappa > 0.65. Other metrics included accuracy, precision, recall, f-score, and AUC.
Yes
EstBERT
EstBERT passed the minimum threshold of Kappa > 0.65 and outperformed other algorithms for all the coding dimensions.
RQ 1
Optimizing Parameters
EstBERT (best performer) and LR (most interpretable) were further optimized.
RQ 2
S3
Best Performing Parameters
For EstBERT we compared different learning_rates and class_weights. For LR we considered different values of class_weight, optimization algorithms, and regularizations.
Yes
Final Algorithms
Best performers were EstBERT with learning_rate=3e-0.5 and without class_weight, as well as LR without class_weight, with stochastic optimization and a regularization value of c=100.
Interpreting Algorithms
S4
An example of the interpretability of the results obtained with EstBERT vs. LR.
End

Figure 1. Workflow Diagram Describing the Four Steps (in Colours) of the Study and Their Relation to the RQs

3.1 Step S1: Constructing and Manually Labelling the Dataset

3.1.1 Sources of Learning Designs

We constructed a dataset with the textual content of the design artifacts created with two m-learning tools, Avastusrada[1] and Smartzoos[2]. Both tools allow teachers to design exploratory and/or inquiry learning activities outdoors as geo-localized learning tracks (for more details on Smartzoos, see Pishtari et al., 2017). The learning designs in these two tools (which in this context correspond to the learning tracks) consist of a number of learning tasks (e.g., questions) that teachers create, using a set of predefined templates (which include one correct answer, match pairs, free answer, etc.). Each learning task may consist of a combination of textual and audiovisual content that teachers freely generate, or embed in it as links, and is connected to a specific geographical point in the map (see Figure 2).



Title*

Image: Choose File No file chosen

An image in JPEG or PNG format. Uploaded image will be resized to 800 by 800 pixels.

Question Type: Information

Question/Information

Zoo: Skansen

Language: English

Location

Please select the correct Zoo and drag the marker to the specific location.

Figure 2. Screenshot of the Interface Used by Teachers to Design Learning Activities in Smartzoos Showing the Different Templates (under Question Type) and the Location That They Can Choose for Each Learning Task

In this study, we use as input features for the SML models only the textual content included in the learning tasks that teachers created with Avastusrada and Smartzoos (excluding metadata such as task type or location coordinates). This allowed us to develop system-agnostic algorithms that can be used to classify the textual content of learning tasks (thus also the corresponding LD artifacts) from multiple m-learning tools. For instance, Avastusrada and Smartzoos have different types of templates for the learning tasks, so including *template type* as a feature would not have allowed training system-agnostic algorithms to be able to classify learning designs from both tools.

Avastusrada and Smartzoos also provide LD artifacts from two complementary communities of teachers in m-learning. Avastusrada is used by K–12 schools in Estonia, thus providing learning designs from formal settings. Smartzoos is used by several zoos in the Nordic region, therefore mainly including designs from informal and non-formal settings. The full database of Avastusrada had designs from 77 teachers, while that of Smartzoos from 24. We randomly selected the designs, stratifying the selection based on the teachers (i.e., on the condition that teachers would have at least one design included and, when possible, be equally represented). This resulted in 168 designs from Avastusrada and 54 from Smartzoos, or a total of 1,472 learning tasks in Estonian, which were included in the final unlaballed dataset.

3.1.2 Coding Strategy

We manually labelled the dataset following a similar approach as in Mettis and Väljataga (2021). Namely, we used the *revised Bloom's taxonomy* (Krathwohl, 2002), which has been widely used by researchers in the field of LD when analyzing design artifacts, in order to determine the cognitive level that specific tasks or questions that teachers have included in their designs

[1] https://www.avastusrada.ee

[2] https://www.smartzoos.eu

ISSN 1929-7750 (online). The Journal of Learning Analytics works under a Creative Commons License, Attribution - NonCommercial-NoDerivs 3.0 Unported (CC BY-NC-ND 3.0)

require from the students that are intended to respond (see, for instance, Rodríguez-Triana et al., 2021; Mettis & Väljataga, 2021). The taxonomy includes six hierarchical levels. In this study we use a binary version of the taxonomy consisting of

- *lower-order thinking*, including the first two categories of the original taxonomy, *remember* and *understand*;
- *higher-order thinking*, including the *apply*, *analyze*, *evaluate*, and *create* categories in the taxonomy.

This choice was made to differentiate between tasks that required students to (at least) apply their knowledge in a different real-life learning scenario (which is one of the main goals of Avastusrada and Smartzoos) from tasks that did not.

Furthermore, we used the *categories of the role of the context in m-learning* proposed by Sharples (2015), which include three (non-exclusive) categories of how the contextual and situated learning environment affects the learning process:

- *Learning in context* is learning that happens in a specific situated environment, such as an archaeological park.
- *Learning about context* is when the context is the target of learning (i.e., students have to learn about it).
- *Learning through context* is the physical or digital environment, as well as the social interactions that allow or facilitate learning (e.g., a specific learning tool).

In our case, we excluded *learning through context* because we considered Avastusrada and Smartzoos to be the digital context, while learning designs created with them by default describe the physical environment where the learning activity takes place (i.e., the physical context enabling learning). Tasks were labelled as *in context* if the students had to be in the situated learning environment to respond and as *about context* if students were learning something related to the situated environment itself.

Finally, since inquiry learning plays an important role in Avastusrada and Smartzoos, to understand the extent to which IBL pedagogies are present in the learning designs, we used the *IBL model* from Pedaste and colleagues (2015). This model includes the phases of *orientation*, *conceptualization*, *investigation*, *conclusion*, and *discussion*. We chose it because it represents a meta-model that synthesizes previous ones (Pedaste et al., 2015). We did not consider the *orientation* phase, which consists of an exploration done by the learners about the related topic, or of an introduction done by teachers, because this part is compulsory in Avastusrada and Smartzoos. We also did not consider the *discussion* phase, since it is a transversal process that happens during the entire IBL activity (Pedaste et al., 2015). Regarding the considered phases:

- *Conceptualization* requires students to come up with a research question, hypothesis, or problem.
- *Investigation* requires exploration, experimentation, or data interpretation.
- *Conclusion* requires students to reflect on the results of the experiment.

Following the suggestions of Shaffer (2017), the dataset was labelled using binary classification codes based on the models discussed so far. Specifically, we used six binary codes to mark the presence or absence of higher-order thinking, learning in context, and learning about context, as well as the IBL phases of conceptualization, investigation, and conclusion. Note that for the categories based on Bloom's taxonomy we used only one code, since the classification is hierarchical (i.e., a task can be classified as either higher-order or lower-order thinking), while for the rest of the theoretical models we used separate codes for each of the categories that we considered because they are not exclusive (e.g., a task can include more than one IBL phase).

3.1.3 Manually Labelling the Dataset

Two master's students from the School of Digital Technologies at Tallinn University labelled the dataset following an iterative process (based on the coding scheme explained in Section 3.1.2). The process started with a small test during which each coder separately labelled a subset of 100 tasks from the dataset. The results were later compared in order to identify border and doubtful cases, as well as to establish a common coding strategy. Thereafter, the same procedure was repeated for the rest of the dataset. Doubtful cases, when coders did not agree, were further discussed until a consensus was reached. While there is no single correct way to deal with the reliability of the coding process (McDonald et al., 2019), we opted for the iterative process described above instead of focusing on the inter-rater reliability (IRR) of the coders (e.g., where coders first label a sample, then their IRR is calculated and when it is high enough the rest of the dataset is divided among them). This approach allowed us to construct a consistent dataset, where each single doubtful case was discussed in detail. The resulting dataset is available online[3]. Table 1 presents the distribution of the tasks per code.

3.2 Step S2: Comparing Different SML Models and Feature Extraction Techniques

The textual content from the dataset of learning tasks was first preprocessed, including tokenization, removing stop-words in Estonian[4], and stemming and normalization (see Figure 3, in green). The dataset was split into 80% training set and 20% testing set.

[3] https://zenodo.org/record/6320368

[4] https://datadoi.ee/handle/33/78

ISSN 1929-7750 (online). The Journal of Learning Analytics works under a Creative Commons License, Attribution - NonCommercial-NoDerivs 3.0 Unported (CC BY-NC-ND 3.0)

Table 1. Distribution of Learning Tasks per Code

Code	Yes	No
Bloom (higher-order)	836	636
in_context	373	1,099
about_context	347	1,125
conceptualization	171	1,301
investigation	511	961
conclusion	330	1,142

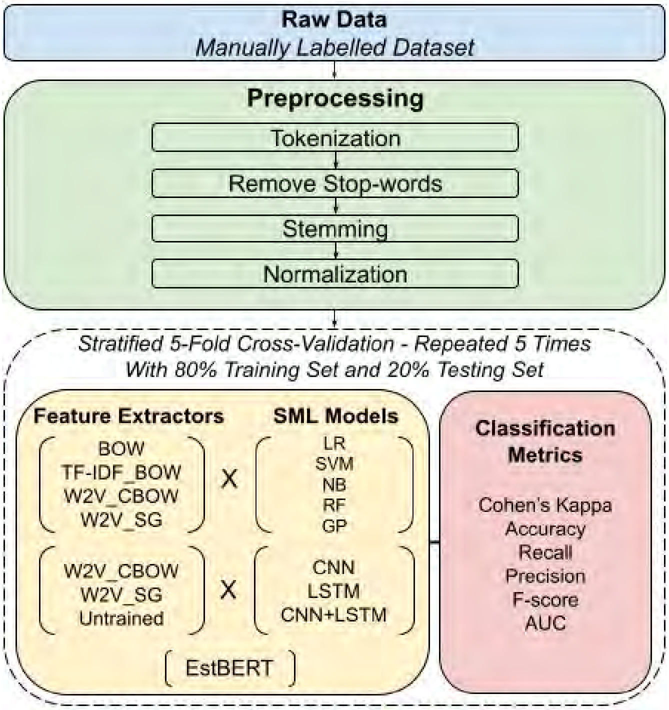


Figure 3. The Process of Training and Comparing Different Feature Extraction and SML Models, as well as the Classification Metrics Used

For each binary classification code discussed above, we trained and compared a combination of different feature extraction techniques and some of the most common classic and neural network models in SML, which have been reported in research considered by two systematic literature reviews, one on automatic question classifiers (Silva et al., 2019) and the other on deep learning (Minaee et al., 2021). Namely, classic SML models included logistic regression (LR), support vector machine (SVM) with a linear kernel, naive Bayes (NB), random forest (RF), and Gaussian processes (GP). These were used in combination with feature extractors such as bag of words (BOW); bag of words normalized with term-frequency inverse-document-frequency (TF-IDF_BOW); and the Estonian version of the pre-trained word2vec model[5] with 100 embedding dimensions, with both its skip-gram (W2V_SG) and its continuous bag of words (W2V_CBOW) variants.

Neural network models included convolutional networks (CNNs) and long short-term memory recurrent (LSTM) networks and a combination of the two (CNN+LSTM). We used them in combination with the same word2vec features mentioned above, as well as with an untrained embedding layer. In this phase, the CNN consisted of a one-dimensional layer, while the LSTM network was a bidirectional layer, both of which had 64 hidden units. To avoid overfitting, we implemented an early stopping rule guided by the validation loss, with a patience value equal to five epochs. Furthermore, we also considered bidirectional encoder representations from transformers (BERT), which have recently performed well in textual classification tasks (Devlin et al., 2019). Namely, we used the version pre-trained in Estonian (EstBERT) (Tanvir et al., 2021). During this phase we opted for a model with one layer and an AdamW optimizer with an initial learning rate of 2e-5, as suggested in Devlin and colleagues (2019).

[5] https://datadoi.ee/handle/33/91

Classification metrics that we considered to evaluate and compare the performance of the algorithms include Cohen's kappa, accuracy, f-score, precision, recall, and area under the curve (AUC). Only algorithms with kappa (IRR calculated comparing the manually coded dataset and the output of the algorithms) larger than 0.65 were considered as reliable, as suggested in Viera and Garrett (2005). The entire process included a stratified five-fold cross-validation that was repeated five times. Figure 3 represents the process and the classification metrics that were considered. We used the Python[6] programming language to build and evaluate the algorithms. Concretely, we used the Scikit-learn[7] package for the classic models and Tensorflow[8] for the neural networks.

3.3 Step S3: Optimizing the Parameters of EstBERT and LR

After the wide-spectrum exploration of SML methods in the previous step, we further trained the best performing algorithm for each of the six coding tasks (to get an understanding of how far the performance could be optimized). In our case, the best performing algorithms were always based on EstBERT (see Section 4.1). We also further optimized the most easily interpretable models (based on LR with a BOW feature extractor) and compared their performance versus the ones based on EstBERT to answer our second RQ (related to performance versus interpretability). For EstBERT and LR, we tried a combination of different balances of the dataset (through the class_weight parameter), solvers, and C values. Class_weight is used to adjust the weights of the classes (in our case the binary classes of each code). Concretely, we used the values None and Balanced, as well as ratios of 1 to 2, 1 to 4, and 3 to 4. The solver parameter specifies the algorithm used in the optimization problem. We considered newton-cg, lbfgs, liblinear, and sag. For the C value (smaller values of which specify stronger regularization), we used 100, 10, 1, 0.1, and 0.01.

Regarding EstBERT algorithms, we considered the same class_weight values as for LR, as well as three different learning rates (2e-5, 3e-5, 5e-5), as suggested by the original BERT paper (Devlin et al., 2019). The algorithms used during the different phases of this study are available online[9].

3.4 Step S4: An Example of Interpreting Results Obtained from EstBERT and LR

To respond to RQ2, we compared the interpretation of the predictions obtained using the LR model (as the most interpretable model) and EstBERT (best performer), based on the Bloom's taxonomy code (which was the only one where LR surpassed the threshold of kappa greater than 0.65; see Section 4.1). Concretely, to illustrate the interpretation of the LR models in the paper, we filtered the statistically significant BOW features (i.e., with p value < 0.05) and plotted the 10 most predictive per category (i.e., higher- versus lower-order thinking), together with their standard errors. To interpret particular predictions by the EstBERT models, we used LIME, a model-agnostic interpreter that helps to explain specific predictions done by any classifier (Ribeiro et al., 2016).

4. Results

This section presents the results organized alongside the aforementioned steps S2, S3, and S4.

4.1 Results from Comparing the Models and Feature Extraction Techniques (S2)

In this subsection, we compare the combination of models and feature extraction based on the IRR (Cohen's kappa). See the extended version of the results for step S2[10] for a full view of the results, including all the considered classification metrics.

Neural network models, especially EstBERT, outperformed the classic SML models. In general, we can see that the classic models are far away from the desired value of kappa > 0.65 (see Figure 4, left). Although neural network models performed better, we can observe that only for EstBERT is the value of kappa significantly above 0.65 (see Figure 4, right). The prevalence of the codes (i.e., how common or rare a code is in the dataset) influenced the classic models' performance, degrading as codes became rarer, but did not change significantly for neural networks (see Figure 5). In step S3, we further optimize the parameters of the best performing model (EstBERT) and the most interpretable one (LR with BOW).

4.2 Results from the Optimization of EstbERT and LR (S3)

The best performing LR algorithms included a combination of an unmodified class_weight (equal to None), the sag solver, and a c_value of 100 (see Figure 6). In Table 2, we can see that the value of kappa was greater than 0.65 only for the Bloom (higher-

[6] https://www.python.org/
[7] www.scikit-learn.org
[8] www.tensorflow.org
[9] https://zenodo.org/record/6320619
[10] https://bit.ly/ResultsStep2JLA2021

ISSN 1929-7750 (online). The Journal of Learning Analytics works under a Creative Commons License, Attribution - NonCommercial-NoDerivs 3.0 Unported (CC BY-NC-ND 3.0)

versus lower-order thinking) code. For the full list of results for LR, including all the classification metrics, see the extended version[11].

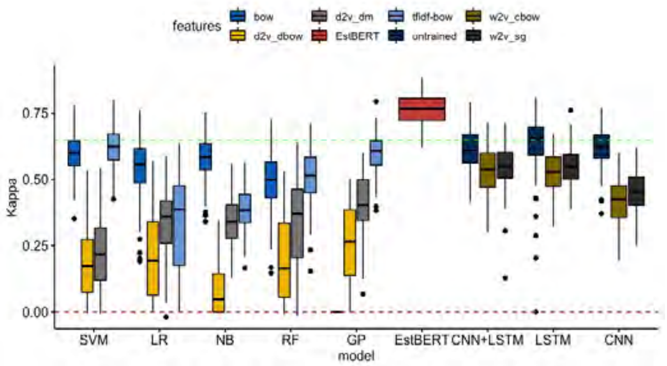


Figure 4. Distribution of IRR (Cohen's Kappa) between the Human Coder and the Combination of SML Models with the Feature Extraction Techniques, for All the Codes

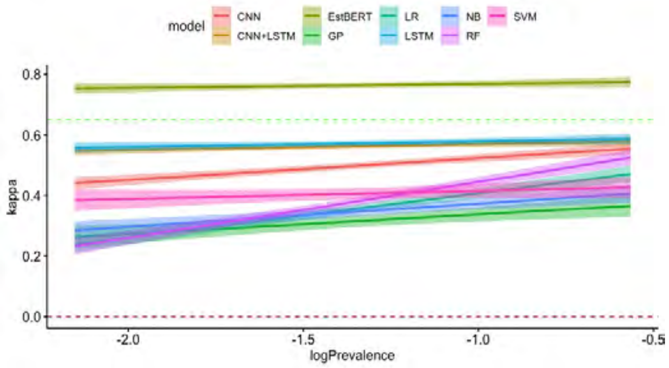


Figure 5. Variation of the Models' Reliability with Each Code's Prevalence (Logarithmic) in the Dataset

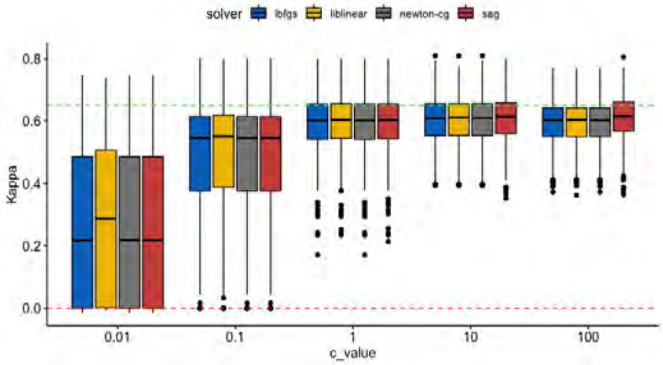


Figure 6. Distribution of Cohen's Kappa for LR Algorithms with Different Solver and C Values (class_weight = None), for All the Codes

EstBERT, on the other hand, performed well throughout the different combinations of the learning_rate and class_weight parameters (see Figure 7). As we can see in Table 3, EstBERT performed well throughout all codes. For the full list of results for EstBERT, including all the classification metrics, see the extended version[12].

[11] https://bit.ly/LRStep3JLA2021

[12] https://bit.ly/EstBERTStep3JLA2021

Table 2. Performance of the Best LR Algorithms (in Terms of Cohen's Kappa, Accuracy, AUC, and Recall) per Code

Code	Kappa	Accuracy	AUC	Recall
Bloom	0.69	0.85	0.85	0.84
in_context	0.57	0.85	0.75	0.54
about_context	0.54	0.86	0.73	0.51
conceptualization	0.48	0.92	0.68	0.37
investigation	0.62	0.83	0.79	0.63
conclusion	0.59	0.88	0.76	0.53

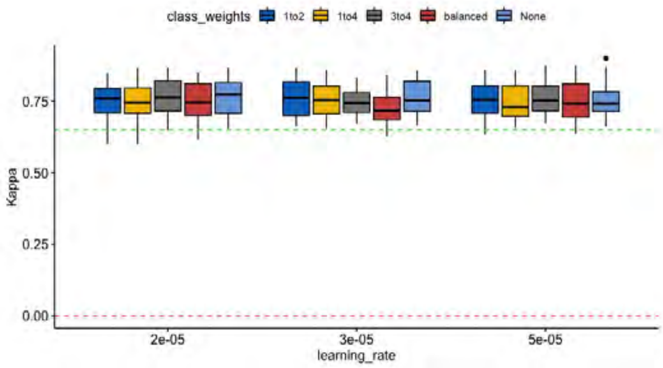


Figure 7. Distribution of Cohen's Kappa for EstBERT with Different learning_weight and class_weight Values, for All the Codes

4.3 Interpreting the Results of LR and EstBERT (S4)

This subsection offers two examples of the interpretation of the results with EstBERT and LR. As we can see in Figure 8, with LR it is possible to extract the statistically significant and most important features per category, as well as their standard errors. For instance, according to the Bloom code, words like *create* and *result* are strongly correlated with higher-order thinking, and *correct*, *list*, and *state* with lower-order thinking. Furthermore, words like *work* and *choose* are on the side of higher-order thinking in Figure 8, but their standard error is big enough to make them unreliable predictive features.

Using LIME (Ribeiro et al., 2016), we can interpret single predictions done with EstBERT, as in Figure 9, where the algorithm correctly classified the learning task as lower order based on the Bloom code, with a 99% probability. The most significant words were *correct* (30%) and *water* (19%).

5. Discussion

Regarding RQ1 (*the performance of different existing SML approaches to code datasets of m-learning designs artifacts*), we were able to train algorithms that reliably mimicked human coding. For this particular dataset, neural network models outperformed classic models. In particular, EstBERT performed uniformly well on all our classification metrics, providing values of Cohen's kappa greater than 0.65 (since step S2) for all six coding tasks. Therefore, similar SML approaches could be used in the future to support the LD practices of different stakeholders in m-learning. For instance, in the short term, algorithms based on EstBERT (or other BERT variants for other languages) could be used by researchers to conduct large-scale and in-the-wild case studies on the kinds of designs that teachers create for m-learning. Connected to this point, the next step would be to use the algorithms based on EstBERT to analyze the full databases of Avastusrada and Smartzoos, which would provide an overview of teachers' design practices at the community level (i.e., a case of community analytics for LD in m-learning)

Table 3. Results for EstBERT with class_weight = None and learning_rate = 3e-5

Code	Kappa	Accuracy	AUC	Recall
Bloom	0.82	0.91	0.90	0.96
in_context	0.84	0.94	0.92	0.89
about_context	0.75	0.91	0.86	0.76
conceptualization	0.76	0.95	0.90	0.84
investigation	0.69	0.86	0.84	0.79
conclusion	0.71	0.90	0.84	0.74

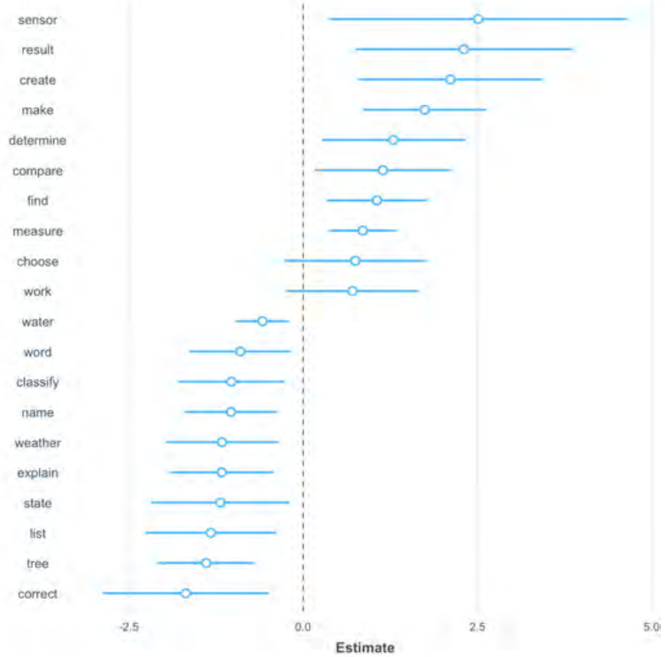
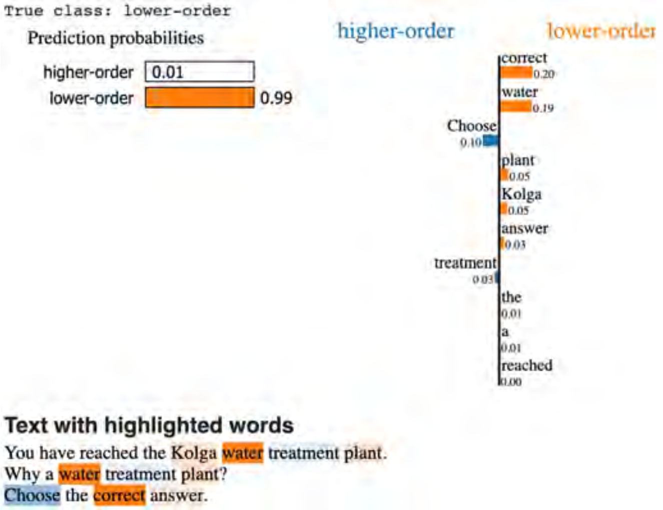


Figure 8. Ten Most Significant Predictive BOW Features (Translated into English) for the Lower-Order (Left) and Higher-Order (Right) Thinking Categories of the Bloom Code



Text with highlighted words

You have reached the Kolga water treatment plant.
Why a water treatment plant?
Choose the correct answer.

Figure 9. Interpretation of an EstBERT Prediction for the Bloom Code with LIME (Ribeiro et al., 2016), Showing the True Class and Prediction Probabilities (Upper Left), Feature Importance per Category (Right), and the Text of the Learning Task Translated into English, with the Relevant Words Highlighted (Lower Left)

(Hernández-Leo et al., 2019). The same approach could be applied to analyze the learning designs of other m-learning tools, such as QuesTInSitu (Santos et al., 2011), etiquetAR (Pérez–Sanagustín et al., 2013), and GLUESP-AR (Muñoz-Cristóbal, Rodríguez-Triana, et al., 2018).

In other cases beyond m-learning, similar classification models could provide a broader perspective on teachers' design practices. Platforms such as Graasp (which also follows an IBL approach; de Jong et al., 2021) and ILDE (which includes a wide community of teachers that use it to conceptualize, author, and deploy learning designs across platforms; Hernández-Leo et al., 2014) could be fertile starting points for this kind of inquiry. Conducting similar studies in other contexts would be a necessary step to provide further evidence on the feasibility of using SML to classify the textual content of design artifacts, in general. These studies would also require considering different sets of high-level pedagogically guided codes, each relevant to a specific context.

In the long term, similar SML approaches could also be used to support teachers' design practices and contribute to addressing the lack of m-learning design solutions that provide design analytics (Pishtari et al., 2020). For example, future research could explore the feasibility of integrating SML algorithms in digital assistants that can guide teachers during the LD process, in similar ways as they have been applied to guide learners in intelligent tutoring systems (Conati et al., 2018). For instance, during the conceptualization phase of a learning design, teachers usually define a set of goals that they want to achieve through the learning activity that they are planning. To illustrate this case, we might consider the example of a teacher that sets as goals the creation of an m-learning activity with Avastusrada that would enable students to learn in and about a specific endangered natural park, while as a second goal the activity should also prompt students to conduct higher-order thinking tasks (such as analyzing specific parts of the natural environment that are damaged, drawing conclusions from the collected results, and proposing creative solutions on how the situation of the park could be improved). In this scenario, during the authoring phase of the learning design, the SML algorithms could be used to analyze in real time the tasks created by the teacher and compare them with the goals. If the results do not match the goals (e.g., the learning activity planned by the teacher does not include higher-order thinking tasks), the digital assistant could inform the teacher about it or even go a step further and provide templates of designs that would help the teacher create higher-order thinking tasks. In the context of Avastusrada and Smartzoos, these SML algorithms could be embedded in the existing dashboards that provide design analytics (see, for instance, Pishtari et al., 2021).

This paper also provides an example of the potential use of SML approaches informed by features other than learners' traces, namely, the learning design artifacts. The field of LA could benefit from a wider variety of SML models, trained on a variety of traces from learners, teachers, and the learning environment, to support their research goals on understanding teaching and learning practices, as well as the environment where they occur (Ferguson, 2012).

It is worth noting that not all the codes had a balanced distribution in the dataset, which is a common occurrence in many qualitative coding tasks, where prevalence is much lower than 50% (Shaffer, 2017). Nevertheless, algorithms based on EstBERT performed well despite the varying prevalence of the codes, providing an example of how state-of-the-art SML techniques could be used to deal with unbalanced datasets, which are common in education, such as in the context of single-case LA (Prieto et al., 2021).

Regarding RQ2 (*the trade-off between performance and interpretability*), it is evident that highly interpretable models, such as LR, offer more insights into the overall features that influence the decisions. We showed that model-agnostic techniques such as LIME (Ribeiro et al., 2016) can also produce limited interpretations, such as the features that influence individual classifications of black-box models.

Although more interpretable models can foster trust and better decision making from the involved stakeholders (Conati et al., 2018; Minaee et al., 2021), in the end, the decision of which model should be used will depend on its performance in specific cases, as well as the requirements of the stakeholders. For example, when transparency is crucial, models that are more interpretable could be prioritized if they surpass the necessary classification metrics thresholds (such as kappa greater than 0.65, in our case). For instance, in the example discussed previously of offering guidance to teachers during the design, information similar to that shown in Figure 8 (i.e., BOW features) and Figure 9 (i.e., LIME output) would help the teacher better understand the elements that could classify a specific learning task as higher-order thinking. It is worth mentioning that visualizations similar to Figures 8 and 9 need to be adapted in a way that makes them easily interpretable by relevant stakeholders.

On the other hand, if transparency is not the priority and the stakeholders have a high confidence in the algorithm (e.g., an algorithm that performs well and a training dataset that is appropriate for the specific case of the deployment), the best performing one can be used, and techniques like LIME (Ribeiro et al., 2016) will still offer a limited degree of interpretability. Such is the case of deploying the algorithms based on EstBERT to analyze the full database of Avastusrada and Smartzoos that we discussed before.

In specific cases, it might be possible to use ensemble models that combine both interpretable and black-box models. For instance, black-box models such as our model based on EstBERT could be used to make accurate predictions about a particular design (and techniques like LIME are used to increase trust in the prediction of the case at hand), while an interpretable model could be used to provide general advice on improving a learning design (e.g., suggesting examples of keywords that are related to higher-order thinking or contextualized tasks).

Limitations of the current study derive mostly from the methods and techniques that were used. The dataset used to train and evaluate the SML models and feature extraction techniques might not represent all the relevant kinds of designs in m-learning. Biases of the human coders (e.g., considering as higher-order thinking according to Bloom's taxonomy only the tasks that would be so in their case, rather than depending on the level of the targeted students) might have been passed to the classifying algorithms. Nevertheless, we used an iterative labelling process (explained in the methodology section) that tries to minimize these kinds of biases. Moreover, we did not further optimize the architecture of the models that we trained (in steps S2 and S3). Further optimization, such as using several layers in the neural network models, could have produced different results. Other parameters could have been considered, such as trying optimizers other than AdamW with EstBERT. The threshold of kappa

ISSN 1929-7750 (online). The Journal of Learning Analytics works under a Creative Commons License, Attribution - NonCommercial-NoDerivs 3.0 Unported (CC BY-NC-ND 3.0)

greater than 0.65 that we considered to evaluate and compare the algorithms is still a debated issue, with various researchers advocating for different values, as well as for the adoption of complementary metrics such as *Shaffer's rho* (Eagan et al., 2017).

This study uses only the textual content of the learning tasks as features. On the one hand, this allowed us to develop platform-agnostic algorithms that can be later deployed to analyze the designs from different platforms (as discussed before). On the other hand, features such as the type of task (i.e., Avastusrada and Smartzoos specify different types of tasks, such as one correct answer, match pairs, free answer, etc.) might be relevant for the prediction and should be considered in future work. Also, the performance of the algorithm might differ on datasets of designs that include different features. For instance, Gervet and colleagues (2020) report that LR might outperform deep learning models when the dataset includes a very large number of user interactions with the system (which are missing in our dataset). The language of the textual content of the designs (Estonian) is both a contribution (because few SML algorithms exist in Estonian) and a limitation (since, for instance, other languages might offer language processing techniques, such as word2vec, that are pre-trained based on larger amounts of data). We have also made our dataset available to the LA community, thus contributing to addressing the need for more open educational datasets (Mihaescu & Popescu, 2021).

Last but not least, this study focuses on the automatic analysis of the design artifacts (i.e., focusing on what kinds of designs teachers create for m-learning questions). Future studies could also explore the other dimensions proposed by Muñoz-Cristóbal, Hernández-Leo, and colleagues (2018) (as previously discussed in Section 2) and address the question of *how teachers design for m-learning*.

6. Conclusion

This study provides an example of the potential use of SML approaches to mimic human coding (of pedagogically relevant features) in datasets of m-learning designs. Namely, we trained, evaluated, and compared a combination of several SML models and feature extraction techniques that use the textual content of the learning tasks (included in the designs) as input and can predict the cognitive level required by students to perform each task, the role played by the situated environment in them, and the phases of IBL that are represented by the tasks.

We showed that SML approaches, informed by features other than learners' traces, can reliably mimic human coders and thus be used by the LA and EDM communities. In our case, the algorithms based on EstBERT outperformed the other models and passed the threshold of kappa greater than 0.65 (i.e., necessary to deploy the algorithm in real settings). The same techniques could be used to complement existing work in the field of LD that have used a manual approach until now, such as Mettis and Väljataga (2021) in the context of m-learning, or that have focused on automatic analysis of design artifacts based on teachers' logs or the structure of the design artifacts (e.g., Toetenel & Rienties, 2016; de Jong et al., 2021).

We provided a comparative example on the interpretability of the results from EstBERT (a black-box model) and LR (a highly interpretable model). In real settings, the choice between such models, or the use of both kinds in support system implementation, will depend on the performance of the algorithms and on the specific needs of the stakeholders. For instance, systems that aim to support teachers during the authoring phase of the learning design (e.g., by analyzing the designs in real time, comparing them with the goals defined by teachers, and providing guidance) might prefer to prioritize performance, since agnostic models like LIME (Ribeiro et al., 2016) would still provide the necessary interpretation for single predictions. However, researchers could benefit more from an interpretable model, even if their performance is inferior to black-box models but still within the acceptable boundaries of reliability.

Future work will include the deployment of the algorithms based on EstBERT to analyze the full databases of m-learning tools, such as Avastusrada and Smartzoos. In the short term, the research community could benefit by using similar algorithms to explore other LD communities of practice in m-learning (e.g., QuesTInSitu; Santos et al., 2011), or beyond (Graasp; de Jong et al., 2021), (ILDE; Hernández-Leo et al., 2014); or other learning management systems. In the long term, if embedded in existing LD systems, similar SML approaches could support teachers during their LD practices (e.g., analyzing the design that they are creating and offering suggestions to improve upon them).

Acknowledgements

The authors would like to thank Sara Sinha and Vishnuraj Maniyil for manually coding the dataset.

Declaration of Conflicting Interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

This research has been partially funded by the European Union in the context of CEITER (Horizon 2020 Research and Innovation Programme, grant agreement no. 669074) and the Estonian Research Council (Personal Research Grant no. PRG1634). Roberto Martinez-Maldonado's research is partly funded by Jacobs Foundation.

References

Bulathwela, S., Pérez-Ortiz, M., Lipani, A., Yilmaz, E., & Shawe-Taylor, J. (2020). Predicting engagement in video lectures. In A. N. Rafferty, J. Whitehill, C. Romero, & V. Cavalli-Sforza (Eds.), *Proceedings of the 13th International Conference on Educational Data Mining* (EDM 2020), 10–13 July 2020, online (pp. 50–60). https://educationaldatamining.org/files/conferences/EDM2020/papers/paper_62.pdf

Chen, F., & Cui, Y. (2020). Utilizing student time series behaviour in learning management systems for early prediction of course performance. *Journal of Learning Analytics*, *7*(2), 1–17. https://doi.org/10.18608/jla.2020.72.1

Conati, C., Porayska-Pomsta, K., & Mavrikis, M. (2018). AI in education needs interpretable machine learning: Lessons from open learner modelling. In B. Kim, K. R. Varshney, & A. Weller (Eds.), *Proceedings of the 2018 ICML Workshop on Human Interpretability in Machine Learning* (WHI 2018), 14 July 2018, Stockholm, Sweden (pp. 21–27). https://doi.org/10.48550/arXiv.1807.00154

Dalziel, J., Conole, G., Wills, S., Walker, S., Bennett, S., Dobozy, E., Cameron, L., Badilescu-Buga, E., & Bower, M. (2016). The Larnaca Declaration on Learning Design. *Journal of Interactive Media in Education*, (1). https://doi.org/10.5334/jime.407

de Jong, T., Gillet, D., Rodríguez-Triana, M. J., Hovardas, T., Dikke, D., Doran, R., Dziabenko, O., Koslowsky, J., Korventausta, M., Law, E., Pedaste, M., Tasiopoulou, E., Vidal, G., & Zacharia, Z. C. (2021). Understanding teacher design practices for digital inquiry-based science learning: The case of Go-Lab. *Educational Technology Research and Development*, 1–28. https://doi.org/10.1007/s11423-020-09904-z

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*. https://doi.org/10.48550/arXiv.1810.04805

Eagan, B. R., Rogers, B., Serlin, R., Ruis, A. R., Arastoopour Irgens, G., & Shaffer, D. W. (2017). Can we rely on IRR? Testing the assumptions of inter-rater reliability. *Proceedings of the 12th International Conference on Computer Supported Collaborative Learning* (CSCL 2017), 18–22 June 2017, Philadelphia, PA, USA. https://repository.isls.org/handle/1/275

Ferguson, R. (2012). Learning analytics: Drivers, developments and challenges. *International Journal of Technology Enhanced Learning*, *4*(5-6), 304–317. https://doi.org/10.1504/IJTEL.2012.051816

Gervet, T., Koedinger, K., Schneider, J., & Mitchell, T. (2020). When is deep learning the best approach to knowledge tracing? *Journal of Educational Data Mining*, *12*(3), 31–54. https://doi.org/10.5281/zenodo.4143614

Hernández-Leo, D., Asensio-Pérez, J. I., Derntl, M., Prieto, L. P., & Chacón, J. (2014). ILDE: Community environment for conceptualizing, authoring and deploying learning activities. In C. Rensing, S. de Freitas, T. Ley, & P. J. Muñoz-Merino (Eds.), *European Conference on Technology Enhanced Learning* (EC-TEL 2014), 16–19 September 2014, Graz, Austria (pp. 490–493). Springer. https://doi.org/10.1007/978-3-319-11200-8_48

Hernández-Leo, D., Martinez-Maldonado, R., Pardo, A., Muñoz-Cristóbal, J. A., & Rodríguez-Triana, M. J. (2019). Analytics for learning design: A layered framework and tools. *British Journal of Educational Technology*, *50*(1), 139–152. https://doi.org/10.1111/bjet.12645

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory into Practice*, *41*(4), 212–218. https://doi.org/10.1207/s15430421tip4104_2

Macfadyen, L. P., Lockyer, L., & Rienties, B. (2020). Learning design and learning analytics: Snapshot 2020. *Journal of Learning Analytics*, *7*(3), 6–12. https://doi.org/10.18608/jla.2020.73.2

McDonald, N., Schoenebeck, S., & Forte, A. (2019). Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–23. https://doi.org/10.1145/3359174

Melero, J., Hernández-Leo, D., Sun, J., Santos, P., & Blat, J. (2015). How was the activity? A visualization support for a case of location-based learning design. *British Journal of Educational Technology*, *46*(2), 317–329. https://doi.org/10.1111/bjet.12238

Mettis, K., & Väljataga, T. (2021). Designing learning experiences for outdoor hybrid learning spaces. *British Journal of Educational Technology*, *52*(1), 498–513. https://doi.org/10.1111/bjet.13034

Mihaescu, M. C., & Popescu, P. S. (2021). Review on publicly available datasets for educational data mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *11*(3), e1403. https://doi.org/10.1002/widm.1403

ISSN 1929-7750 (online). The Journal of Learning Analytics works under a Creative Commons License, Attribution - NonCommercial-NoDerivs 3.0 Unported (CC BY-NC-ND 3.0)

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2021). Deep learning–based text classification: A comprehensive review. *ACM Computing Surveys (CSUR)*, *54*(3), 1–40. https://doi.org/10.1145/3439726

Muñoz-Cristóbal, J. A., Hernández-Leo, D., Carvalho, L., Martinez-Maldonado, R., Thompson, K., Wardak, D., & Goodyear, P. (2018). 4FAD: A framework for mapping the evolution of artefacts in the learning design process. *Australasian Journal of Educational Technology*, *34*(2). https://doi.org/10.14742/ajet.3706

Muñoz-Cristóbal, J. A., Rodríguez-Triana, M. J., Gallego-Lema, V., Arribas-Cubero, H. F., Asensio-Pérez, J. I., & Martínez-Monés, A. (2018). Monitoring for awareness and reflection in ubiquitous learning environments. *International Journal of Human–Computer Interaction*, *34*(2), 146–165. https://doi.org/10.1080/10447318.2017.1331536

Pedaste, M., Mäeots, M., Siiman, L. A., De Jong, T., Van Riesen, S. A., Kamp, E. T., Manoli, C. C., Zacharia, Z. C., & Tsourlidaki, E. (2015). Phases of inquiry-based learning: Definitions and the inquiry cycle. *Educational Research Review*, *14*, 47–61. https://doi.org/10.1016/j.edurev.2015.02.003

Pérez–Sanagustín, M., Martínez, A., & Delgado-Kloos, C. (2013). Etiquetar: Tagging learning experiences. In D. Hernández-Leo, T. Ley, R. Klamma, & A. Harrer (Eds.), *European Conference on Technology Enhanced Learning* (EC-TEL 2013), 17–21 September 2013, Paphos, Cyprus (pp. 573–576). Springer. https://doi.org/10.1007/978-3-642-40814-4_61

Persico, D., & Pozzi, F. (2015). Informing learning design with learning analytics to improve teacher inquiry. *British Journal of Educational Technology*, *46*(2), 230–248. https://doi.org/10.1111/bjet.12207

Pishtari, G., Rodríguez-Triana, M. J., Sarmiento-Márquez, E. M., Pérez-Sanagustín, M., Ruiz-Calleja, A., Santos, P., P. Prieto, L., Serrano-Iglesias, S., & Väljataga, T. (2020). Learning design and learning analytics in mobile and ubiquitous learning: A systematic review. *British Journal of Educational Technology*, *51*(4), 1078–1100. https://doi.org/10.1111/bjet.12944

Pishtari, G., & Rodríguez-Triana, M. J. (2022). An analysis of mobile learning tools in terms of pedagogical affordances and support to the learning activity life cycle. In E. Gil, Y. Mor, Y. Dimitriadis, & C. Köppe (Eds.), *Hybrid learning spaces* (pp. 167–183). Springer. https://doi.org/10.1007/978-3-030-88520-5_10

Pishtari, G., Rodríguez-Triana, M. J., & Väljataga, T. (2021). A multi-stakeholder perspective of analytics for learning design in location-based learning. *International Journal of Mobile and Blended Learning (IJMBL)*, *13*(1), 1–17. https://doi.org/10.4018/IJMBL.2021010101

Pishtari, G., Väljataga, T., Tammets, P., Savitski, P., Rodríguez-Triana, M. J., & Ley, T. (2017). Smartzoos: Modular open educational resources for location-based games. In É. Lavoué, H. Drachsler, K. Verbert, J. Broisin, & M. Pérez-Sanagustín (Eds.), *European Conference on Technology Enhanced Learning* (EC-TEL 2017), 12–15 September 2017, Tallinn, Estonia (pp. 513–516). Springer. https://doi.org/10.1007/978-3-319-66610-5_52

Pozzi, F., Asensio-Pérezc, J. I., & Persico, D. (2016). The case for multiple representations in the learning design life cycle. In B. Gros, Kinshuk, & M. Maina (Eds.), *The future of ubiquitous learning* (pp. 171–196). Springer. https://doi.org/10.1007/978-3-662-47724-3_10

Prieto, L. P., Pishtari, G., Rodríguez-Triana, M. J., & Eagan, B. (2021). Comparing natural language processing approaches to scale up the automated coding of diaries in single-case learning analytics. In A. R. Ruis & S. B. Lee (Eds.), *Second International Conference on Quantitative Ethnography: Conference Proceedings Supplement* (ICQE 2020), 1–3 February 2021, online (pp. 39–42). https://www.qesoc.org/images/pdf/ICQE20_Proceedings_Supplement_Final_web.pdf

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD 2016), 13–17 August 2016, San Francisco, CA, USA (pp. 1135–1144). ACM. https://doi.org/10.1145/2939672.2939778

Rodríguez-Triana, M. J., Prieto, L. P., & Pishtari, G. (2021). What do learning designs show about pedagogical adoption? An analysis approach and a case study on inquiry-based learning. In T. D. Laet, R. Klemke, C. Alario-Hoyos, I. Hilliger, & A. Ortega-Arranz (Eds.), *European Conference on Technology Enhanced Learning* (EC-TEL 2021), 20–24 September 2021, Bolzano, Italy (pp. 275–288). Springer. https://doi.org/10.1007/978-3-030-86436-1_21

Rudin, C., & Radin, J. (2019). Why are we using black box models in AI when we don't need to? A lesson from an explainable AI competition. *Harvard Data Science Review*, *1*(2). https://doi.org/10.1162/99608f92.5a8a3a3d

Santos, P., Pérez-Sanagustín, M., Hernández-Leo, D., & Blat, J. (2011). QuesTInSitu: From tests to routes for assessment *in situ* activities. *Computers & Education*, *57*(4), 2517–2534. https://doi.org/10.1016/j.compedu.2011.06.020

Shaffer, D. W. (2017). *Quantitative ethnography*. Cathcart Press. https://www.quantitativeethnography.org/

Sharples, M. (2015). Making sense of context for mobile learning. In J. Traxler & A. Kukulska-Hulme (Eds.), *Mobile learning: The next generation* (pp. 140–153). Taylor and Francis. https://doi.org/10.4324/9780203076095-9

Silva, V. A., Bittencourt, I. I., & Maldonado, J. C. (2019). Automatic question classifiers: A systematic review. *IEEE Transactions on Learning Technologies*, *12*(4), 485–502. https://doi.org/10.1109/TLT.2018.2878447

Tanvir, H., Kittask, C., & Sirts, K. (2021). EstBERT: A pretrained language-specific BERT for Estonian. *arXiv preprint arXiv:2011.04784*. https://doi.org/10.48550/arXiv.2011.04784

Toetenel, L., & Rienties, B. (2016). Analysing 157 learning designs using learning analytic approaches as a means to evaluate the impact of pedagogical decision making. *British Journal of Educational Technology*, *47*(5), 981–992. https://doi.org/10.1111/bjet.12423

Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: The kappa statistic. *Family Medicine*, *37*(5), 360–363. https://www.stfm.org/familymedicine/vol37issue5/Viera360

Xu, X., Wang, J., Peng, H., & Wu, R. (2019). Prediction of academic performance associated with internet usage behaviors using machine learning algorithms. *Computers in Human Behavior*, *98*, 166–173. https://doi.org/10.1016/j.chb.2019.04.015