Comparing supervised machine learning approaches to automatically code learning designs in mobile learning

Gerti Pishtari[1], Luis P. Prieto[1] Maríaa Jesús Rodríguez-Triana[1] and Roberto Martinez-Maldonado[2]

[1]Tallinn University, Narva maantee 25, 10120 Tallinn, Estonia
[gpishtar, lprisan, mjrt]@tlu.ee

[2]Monash University, Wellington Rd, Clayton VIC 3800, Australia
Roberto.MartinezMaldonado@monash.edu

Abstract. To understand and support teachers' design practices, researchers in Learning Design manually analyse small sets of design artifacts produced by teachers. This demands substantial manual work and provides a narrow view of the community of teachers behind the designs. This paper compares the performance of different Supervised Machine Learning (SML) approaches to automatically code datasets of learning designs. For this purpose, we extracted a subset of learning designs (i.e., their textual content) from Avastusrada and Smartzoos, two mobile learning tools. Later, we manually coded it guided by rel-evant theoretical models to the context of mobile learning and used it to train and compare several combinations of SML models and feature extraction techniques. Results show that such models can reliably code learning design datasets and could be used to understand the learning design practices of large communities of teachers in mobile learning and beyond.

Keywords: Supervised Machine Learning, Learning Design, Learning Analytics, Mobile Learning, Contextual Learning

1 Mobile Learning from a Learning Design perspective

Mobile Learning (m-learning) activities promote authentic and contextualized learning [18, 12]. These activities usually take place across spaces (physical and digital) and settings (formal, informal, or non-formal) [9, 11]. To enable teachers to design for m-learning, the field of Learning Design (LD) has come up with several authoring tools [13]. For instance, Smartzoos support the design of geo-localised learning activities outdoor [14], while with GLUESP-AR teachers design activities that happen across multiple physical and digital spaces [9].

Designing learning activities is already a strenuous task for teachers. In m-learning they also have to deal with the complexity of designing across settings and spaces (previously discussed), together with the need to possess substantial technical and pedagogical competencies, relevant to this context. Mettis and Väljataga [8] after manually analysing designs that teachers created in an m-learning training, concluded that most of the designs were decontextualized (i.e., not related with the situated learning environment) and scored low on the cognitive level (i.e., that mainly required from students to remember basic concepts, instead of performing analysis or evaluations). Considering that teachers should have been trained to produce adequate technology-enhanced designs (including m-learning ones) since their pre-service education [8], more research is needed to first understand and then support teachers' practices when designing for m-learning. A first step could be analysing of databases of design artifacts from existing m-learning tools.

To address this gap, researchers would have to analyse large communities of teachers that design for m-learning. Existing studies have already automatically analysed learning designs practices, focusing on (teachers, or students) action logs, or the structure of the designs (e.g., [3]). Nevertheless, when researchers want to consider more high-level aspects (e.g., the pedagogical approaches followed by teachers), the typical approach has been to manually code the designs (see, for instance [16]). For large datasets, it would be necessary an automatic coding strategy, as it is time consuming to follow a manual approach. Therefore, in this paper we *compare different supervised machine learning (SML) models and features extraction techniques to automatically code datasets of learning designs for m-learning.*

We started by compiling a dataset with learning designs from two m-learning platforms, Avastusrada (avastusrada.ee) and Smartzoos (smartzoos.eu). As a first step, we considered as input features for the algorithms only the textual content (in Estonian) of the learning tasks included in the designs. Although, the design artifacts in these tools also include other metadata that could be potentially used as features for the SML algorithms (such as different types of learning tasks and learning resources), these usually are tool-dependent and would not be useful for platform-independent algorithms that can be later used to analyse learning designs from multiple tools.

We manually labelled the dataset guided by theoretical models and taxonomies, relevant to the context of m-learning and also used in previous studies that manually labeled m-learning deisngs [8, 18]. These include the Revised Bloom's Taxonomy [7], the Inquiry Based Learning (IBL) model [10], and the categorization of the role of the context in a learning activity [18]. This dataset (with the textual content as input and the corresponding codes as the output that had to be predicted) was later used to train and compare the different SML models and feature extraction techniques (see section 3).

2 Machine Learning as analytics for LD in m-learning

Research in Learning Analytics (LA) has largely used SML to predict learners' performance [1, 21]. Furthermore, Prieto et al. [15] attempted to use SML to support researchers, by automatically coding diaries of students' learning progress. Yet, the automated analysis of artifacts created by teachers remains an underexplored area. Therefore, this paper presents *a comparison of the performance of different SML approaches, when trained to code datasets of m-learning designs*, guided by theoretical models that are pertinent in the context of m-learning (see section 3).

Analytics can inform LD practices in different levels: as LA (i.e., informed based on students data), as design analytics (i.e., informed from traces of the LD process), or as community analytics (such as metrics about LD practices of a community of teachers behind a specific m-learning tools) [6]. Few studies reflect this alignment between LD and LA in m-learning [11]. Cases that explicitly addressed this alignment, focus mainly on LA for LD [9, 17], while design and community analytics for LD remain unexplored. This paper aims to *explore the potential of SML techniques to automatically code learning designs.* Successful algorithms could be later used to analyse large databases of designs from multiple tools, as well as to create systems that provide design and community analytics in m-learning.

3 Methodology

This study is guided by the following research question (RQ): *To what extent can SML techniques automatically code datasets of m-learning designs, in terms of IBL phases, context and cognitive level?* To respond this question, we conducted an exploratory study that consists of two parts. During the first part we compiled a dataset of learning tasks (i.e. their textual content), extracted from existing learning designs in Avastusrada and Smartzoos, which are used by two complementary communities of teachers. Avastusrada is used in formal settings (by K-12 schools in Estonia), while Smartzoos in informal, or non-formal ones (used by zoos in Estonia, Sweden and Finland). The dataset had 1,472 learning tasks in Estonian, originating from 168 different designs (114 from Avastusrada and 54 from Smartzoos).

To determine the cognitive level (required from learners) in each learning task, we coded this dataset using a binary version of the *Revised Bloom's taxonomy* [7], consisting of: *lower-order thinking*, representing the two lowest categories (Remember and Understand); and *higher-order thinking* representing the rest (Apply, Analyse, Evaluate, Create). This choice was made to identify tasks that require students to (at least) apply their knowledge in different learning situations from tasks that did not (a relevant aspect of Avastusrada and Smartzoos). Furthermore, to understand the role played by the situated environment in the learning designs, we coded each task based on the following categories (inspired by [18]):*learning in context*, i.e., learning happening in a specific situated learning environment; *learning about context*, when the situated environment itself is the object of learning. Finally, to understand the extent to which IBL pedagogies (relevant to the context of Avastusrada and Smartzoos) were present in the learning designs, we used the following phases of the *IBL model* proposed by [10]: *Conceptualization*, during which learners have to come up with a hypothesis, or problem; *Investigation* that include activities such as experimentation and data interpretation; and *Conclusion* during which learners reflect upon the results and their implications.

Guided by these theoretical models, we coded the dataset using 6 binary codes that signaled if a learning task included: *higher-order thinking*, *learning in context*, *learning about context*, *conceptualization*, *investigation*, and *conclusion.* As Bloom categories are hierarchical, they are represented by a single code. For the rest we use a separate code for each category (e.g., a task can have more than one phase of IBL). The dataset was coded by two master students from the school of Digital Technologies, Tallinn University. We first conducted a test where each coder worked with the same subset of 100 tasks and compared the results to establish a common coding approach. The same procedure was repeated until the end, during which cases doubtful cases were consulted with the first author of this paper (see the full coded dataset in bit.ly/ManuallyLabelledDatasetJLA2021). During the second part of this study we used the dataset to train, evaluate and compare several common SML models and feature extraction techniques for natural language processing (for each of binary code in the dataset). We first preprocessed the textual content (see Figure 1, in green).

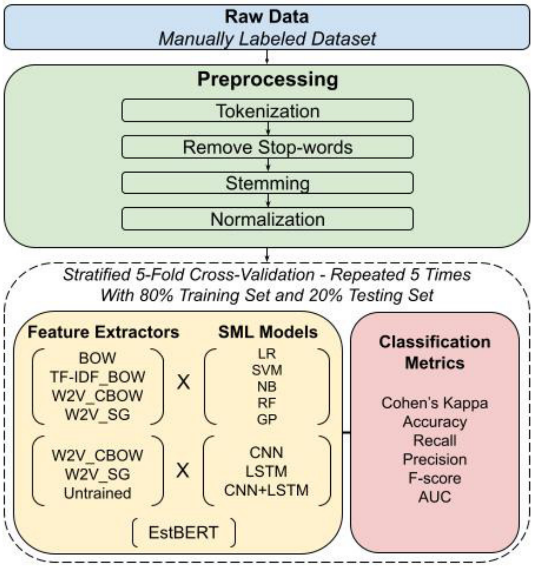

Fig. 1. The process of comparing different SML approaches.

Using 80% of the dataset as training and 20% as testing set, we tested a combination of classic SML models and neural networks with different feature extractors. The first group consisted of classic models, i.e., Logistic Regression (LR), Naive Bayes (NB), Random Forest (RF), Support Vector Machine (SVM) with a linear kernel and Gaussian Processes (GP), combined with feature extractors such as the pre-trained word2vec in Estonian with 100 embedding dimensions, both as a continuous bag of words (W2V_CBOW) and as a skip-gram (W2V_SG), bag of words (BOW), and bag of words with term-frequency inverse-document-frequency (TF-IDF_BOW). Neural networks included Long Short-Term Memory Recurrents (LSTM), Convolutionals (CNN) and a mixed model (CNN+LSTM). These were tested in combination with the word2vec mentioned above, and an untrained embedding layer. LSTM consisted of a single bidirectional layer, while CNN was a 1-dimensional layer, both with 64 hidden units. We used early stopping based on the validation loss to avoid overfitting. Finally, we also used the Estonian version of the Bidirectional Encoder Representations from Transformers (EstBERT) [19], with an AdamW optimizer with 2e-0.5 as the initial learning rate and a single layer. The process was a stratified 5-fold cross-validation, repeated 5 times, based on various classification metrics, used for the comparison (see Figure 1 below). Algorithms with kappa values (the inter-rater reliability between the manual and automatic process) lower than 0.65 were not considered as reliable [20]. Algorithms were written in Python, using Scikit-learn and Tensorflow packages.

4 Results

This section presents a comparison of the combination of SML models and feature extractors, guided by Cohen's kappa. The attached document includes results for all the metrics (bit.ly/ResultsStep2JLA2021). In Figure 2, we can see that classic models did not surpass the threshold value for kappa>0.65. Neural networks performed better, but only EstBERT significantly surpassed kappa>0.65. The prevalence, which considers the balance of the dataset for each code (see the horizontal line in 2, where the right side represents balanced datasets), had a direct influence over the performance of the classic models, but had no significant influence over the performance of the neural networks.

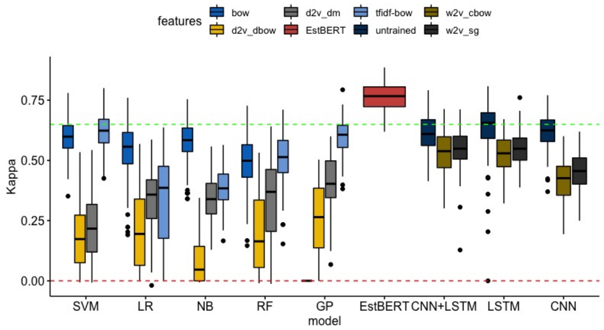

Fig. 2. Distribution of Cohen's kappa between human coders and the different combinations of SML models and feature extractors, for the six the codes.

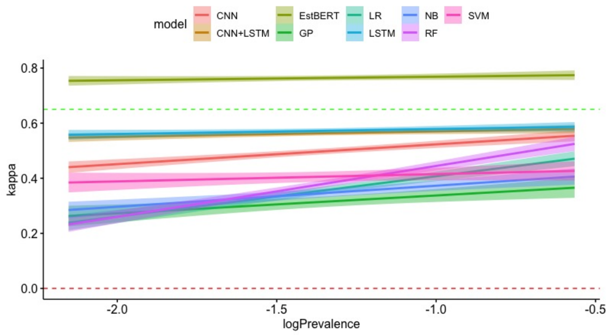

Fig. 3. Variation of the reliability for all the models with the logarithmic prevalence of each code.

5 Discussion

Regarding our RQ *(the performance of SML approaches when coding datasets of m-learning designs)*, we were able to train algorithms based on EstBERT that for our particular dataset, were reliable on all the six codes (with kappa>0.65). EstBERT algorithms also performed uniformly well on all the other classification metrics that we used. Thus, SML could be used in the future to support researchers in LD, when analysing large datasets of learning designs. In the context of m-learning, similar algorithms could be used to analyse the whole databases of Avastusrada and Smartzoos, providing a case of community analytics in m-learning [6], as well as enabling large-scale and in-the-wild studies about the open issue of teachers' design practices in m-learning [8]. Other m-learning tools could benefit from the same SML approach, such as GLUESP-AR [9], or QuestInSitu [17]. Beyond m-learning, our approach could be useful to analyse LD platforms used by big communities of teachers, such as ILDE [5].

Most of the codes in our dataset did not have a balanced distribution (see Figure 3), which is typical in qualitative coding tasks. However, EstBERT algorithms performed well with all the codes, despite their prevalence and constitute an example of dealing with unbalanced datasets (common in education). The dataset used to train and compare the SML approaches constitutes a limitation for this study as it is not a representative of all the kinds of designs in m-learning. Also, the manual coding process might have produced biases conditioning the performance of the algorithms. Nevertheless, while in this paper we present only preliminary results for our exploratory study, further optimizing the models could produce better performance results. We considered as a threshold value kappa>0.65. However, various researchers advocate for different threshold values, or for the inclusion of other metrics (e.g., *Shaffer's rho* [4]).

We used as input features the textual content of the learning tasks. In future work, features such as task type might improve the prediction for tool-specific analysis. The design artifacts were in Estonian, a contribution, as few SML algorithms exist in this language, but also a limitation, as English versions of word2vec, BERT, etc., are usually pre-trained based on larger amounts of data.

6 Conclusion

In this study, we provide an example of how SML approaches can mimic humans, in the context of coding datasets of m-learning designs. We compared different SML models and feature extraction techniques. Models based on EstBERT constantly provided values of kappa>0.65, thus could be used to conduct in-the-wild studies of how teachers design for m-learning.

Future work will include further steps of optimization for all the models that were considered in this study. In line with recent trends of providing models that are transparent to the related stakeholders [2], it is important to further tune-up the performance of classic models (such as LR) and compare it with black-box ones (such as the neural networks). Once further optimized, the best performing algorithms will be used to analyse the learning designs included in Avastusrada and Smartzoos. A similar approach might be useful to analyse other known LD tools in m-learning (e.g., QuestInSitu [17], or beyond (e.g., ILDE, [5]).

Acknowledgements

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). The authors would like to thank the coders for their contribution to this study. Roberto Martinez-Maldonado's research is partly funded by Jacobs Foundation.

References

1. Chen, F., Cui, Y.: Utilizing student time series behaviour in learning management systems for early prediction of course performance. Journal of Learning Analytics **7**(2) (2020) 1–17
2. Conati, C., Porayska-Pomsta, K., Mavrikis, M.: Ai in education needs interpretable machine learning: Lessons from open learner modelling. arXiv preprint arXiv:1807.00154 (2018)
3. de Jong, T., Gillet, D., Rodríguez-Triana, M.J., Hovardas, T., Dikke, D., Doran, R., Dziabenko, O., Koslowsky, J., Korventausta, M., Law, E., et al.: Understanding teacher design practices for digital inquiry–based science learning: the case of golab. Educational Technology Research and Development (2021) 1–28
4. Eagan, B.R., Rogers, B., Serlin, R., Ruis, A.R., Arastoopour Irgens, G., Shaffer, D.W.: Can we rely on irr? testing the assumptions of inter-rater reliability. In: International Conference on Computer Supported Collaborative Learning. (2017)
5. Hernández-Leo, D., Asensio-Pérez, J.I., Derntl, M., Prieto, L.P., Chacón, J.: Ilde: Community environment for conceptualizing, authoring and deploying learning activities. In: European conference on technology enhanced learning, Springer (2014) 490–493

6. Hernández-Leo, D., Martinez-Maldonado, R., Pardo, A., Muñoz-Cristóbal, J.A., Rodríguez-Triana, M.J.: Analytics for learning design: A layered framework and tools. British Journal of Educational Technology **50**(1) (2019) 139–152
7. Krathwohl, D.R.: A revision of bloom's taxonomy: An overview. Theory into practice **41**(4) (2002) 212–218
8. Mettis, K., Väljataga, T.: Designing learning experiences for outdoor hybrid learning spaces. British Journal of Educational Technology **52**(1) (2021) 498–513
9. Muñoz-Cristóbal, J.A., Rodríguez-Triana, M.J., Gallego-Lema, V., Arribas-Cubero, H.F., Asensio-Pérez, J.I., Martínez-Monés, A.: Monitoring for awareness and reflection in ubiquitous learning environments. International Journal of Human–Computer Interaction **34**(2) (2018) 146–165
10. 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.: Phases of inquiry-based learning: Definitions and the inquiry cycle. Educational research review **14** (2015) 47–61
11. 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.: Learning design and learning analytics in mobile and ubiquitous learning: A systematic review. British Journal of Educational Technology **51**(4) (2020) 1078–1100
12. Pishtari, G., Rodríguez-Triana, M.J., Väljataga, T.: A multi-stakeholder perspective of analytics for learning design in location-based learning. International Journal of Mobile and Blended Learning (IJMBL) **13**(1) (2021) 1–17
13. Pishtari, G., Rodríguez-Triana, M.J.: An analysis of mobile learning tools in terms of pedagogical affordances and support to the learning activity lifecycle. In Gil, E., Mor, Y., Dimitriadis, Y., Köppe, C., eds.: Hybrid Learning Spaces. Springer (2021)
14. Pishtari, G., Väljataga, T., Tammets, P., Savitski, P., Rodríguez-Triana, M.J., Ley, T.: Smartzoos: modular open educational resources for location-based games. In: European Conference on Technology Enhanced Learning, Springer (2017) 513–516
15. Prieto, L.P., Pishtari, G., Rodríguez-Triana, M.J., Eagan, B.: Comparing natural language processing approaches to scale up the automated coding of diaries in single-case learning analytics. In: Second International Conference on Quantitative Ethnography: Conference Proceedings Supplement. (2021) 39
16. Rodríguez-Triana, M.J., Prieto, L.P., Pishtari, G.: What do learning designs show aboutpedagogical adoption? an analysis approachand a case study on inquiry-based learning. (In Press)
17. Santos, P., Pérez-Sanagustín, M., Hernández-Leo, D., Blat, J.: Questinsitu: From tests to routes for assessment in situ activities. Computers & Education **57**(4) (2011) 2517–2534
18. Sharples, M.: Making sense of context for mobile learning. Mobile learning: The next generation (2016) 140–153
19. Tanvir, H., Kittask, C., Sirts, K.: Estbert: A pretrained language-specific bert for estonian. arXiv preprint arXiv:2011.04784 (2020)
20. Viera, A.J., Garrett, J.M., et al.: Understanding interobserver agreement: the kappa statistic. Fam med **37**(5) (2005) 360–363
21. Xu, X., Wang, J., Peng, H., Wu, R.: Prediction of academic performance associated with internet usage behaviors using machine learning algorithms. Computers in Human Behavior **98** (2019) 166–173