Post-print version of the paper by Dalipi et al. in EDUCON 2018 - htttp://10.1109/EDUCON.2018.8363340

MOOC Dropout Prediction Using Machine Learning Techniques: Review and Research Challenges

Fisnik Dalipi[1,2], Ali Shariq Imran[3], Zenun Kastrati[3]
[1]Linnaeus University, Department of Computer Science, Sweden
[2]University College of Southeast Norway, Norway
[3]Faculty of IT and Electrical Engineering, Norwegian University of Science and Technology (NTNU), Norway
fisnik.dalipi@{lnu.se, usn.no}, {ali.imran, zenun.kastrati}@ntnu.no

***Abstract*— MOOC represents an ultimate way to deliver educational content in higher education settings by providing high-quality educational material to the students throughout the world. Considering the differences between traditional learning paradigm and MOOCs, a new research agenda focusing on predicting and explaining dropout of students and low completion rates in MOOCs has emerged. However, due to different problem specifications and evaluation metrics, performing a comparative analysis of state-of-the-art machine learning architectures is a challenging task. In this paper, we provide an overview of the MOOC student dropout prediction phenomenon where machine learning techniques have been utilized. Furthermore, we highlight some solutions being used to tackle with dropout problem, provide an analysis about the challenges of prediction models, and propose some valuable insights and recommendations that might lead to developing useful and effective machine learning solutions to solve the MOOC dropout problem.**

Keywords—MOOC; review; dropout prediction; machine learning; artificial intelligence.

I. INTRODUCTION

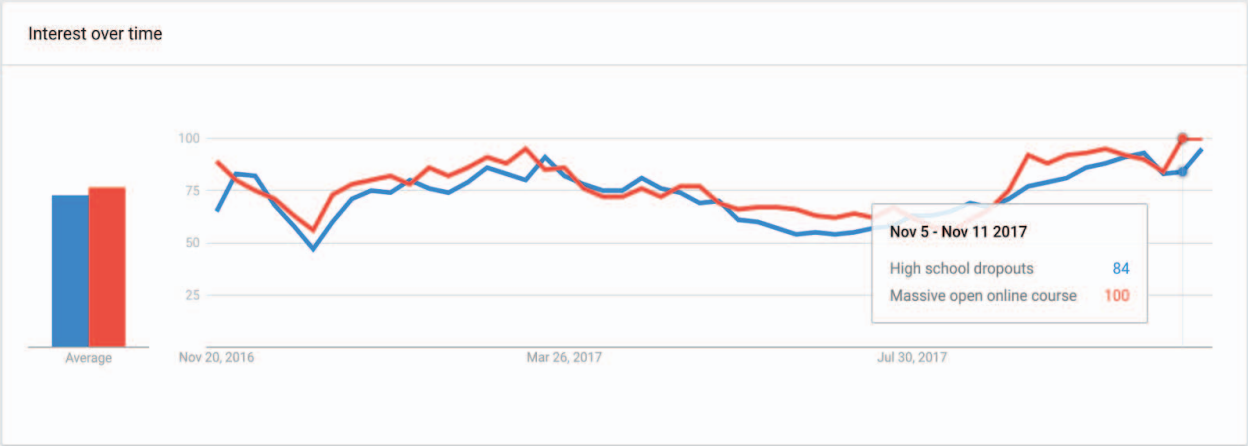
In order to respond to the rapidly changing job and study requirements, nowadays many students worldwide are increasingly considering MOOCs as learning content delivery platforms, which very often help them to bridge the gap between the skills required by the industry and the skills that they obtain at the universities. These modern trends make MOOCs an ideal choice because of many reasons: they provide open, online and low-cost high-quality education and access to the state-of-the-art courses from the elite universities and industry experts. These also make MOOCs an ideal tool for professional self-development and bringing courses to various student audiences across the world. The course materials are available to students at any given time so students can decide not only when to study, but also how, in which order, and at which pace. Furthermore, the MOOC phenomenon will represent the fourth stage in the evolution of online education, following the third stage where the Learning Management System was a central element [1]. As we indicated in our previous work [2], the MOOC approach when complemented with a flipped classroom pedagogy would also bring advantages toward enhancing the learning experience of students substantially. However, MOOCs are surrounded by different challenges such as a large number of dropouts, certification and graduation, verification of student's identity, and being unsuitable for complex and engineering education.

Though MOOCs are progressively becoming an integral part of a learning process in higher education settings, still there are many issues to be addressed when it comes to one of the leading currently unsolved problem revolving around MOOCs, and that is the student dropout problem. Student dropout is a common problem in brick and mortar educational settings especially in higher education with well-studied reasons and challenges [3], [4]. Despite the known factors for dropouts, there's still a vast interest in queries related to high school dropouts. A quick trend search on the Internet will reveal that in the past year there's a steady interest in people looking for high school dropout topic over time. On average 75 queries per week on the same subject as shown in Figure 1. The interest in MOOCs is no different. However, university students' dropping out of a MOOC often have different reasons and challenges in contrast to onsite students.

Some of the reasons for the low completion rates are course methodology, lack of social interaction and creativity. As the number of different MOOCs continues to grow, researchers along with educational technologists are proposing innovative dropout prediction solutions to heal the MOOC dropout headaches. Therefore, various machine learning (ML) techniques have been successfully applied to obtain statistically high dropout prediction accuracy, among which are [5], [6], [7] and [8].

The researchers' focus has been predominantly placed either on training and testing on student engagement data sampled from the same MOOC or the same course [9], [10], [11], or they have offered proposals and solutions that work only on clickstream data [12].

In this paper, we aim to provide a comprehensive overview of the ML application in solving the MOOC dropout problems. In addition, we investigate several different MOOC dropout prediction architectures that are widely used in literature. Moreover, we identify and analyze some of the remaining challenges, and discuss further the factors that might lead to statistically higher prediction accuracy.



The rest of paper is structured as follows. Section 2 gives an overview of the MOOC dropout phenomenon while Section 3 reviews the state-of-the-art on MOOC dropout prediction. In Section 4, we discuss some current challenges of dropout prediction models using ML techniques. Section 5 provides insights towards building useful and effective predictive solutions and Section 6 concludes the paper.

II. THE MOOC DROPOUT HEADACHE

In recent years, MOOCs are growing rapidly, providing the potential to open up the access to education for students all around the world. However, despite the potential benefits of MOOCs, the rate of students who did not complete or withdraw a course has been typically very high. The high students' drop out or withdraw rate of MOOCs has been attributed to many factors that generally can be classified as either student-related factors or MOOC-related factors. This is also illustrated in Table 1.

TABLE I. FACTORS INFLUENCING MOOC DROPOUTS

Two types of factors	
Student related	Lack of motivation
	Lack of time
	Insufficient background knowledge and skills
MOOC related	Course design
	Isolation and lack of interactivity
	Hidden costs

A. Student-related factors

Lack of student's motivations - is one of the most influential factors in preventing students from completing a MOOC. The motivation of students is, by itself, influenced by many factors which among others include the future economic benefit, development of personal and professional identity, challenge and achievement, enjoyment and fun [13]. It is therefore of great interest to explore about motives that drive students to enroll a MOOC. In this regard, [14] surveyed to examine student's motivations. They found that 95% of students see entertainment and enjoyment as the most important reason for enrolling a MOOC followed by a general interest in a topic as another reason selected by 87% of students. A small number of students (15%) chose to register the MOOC to help them decide if they want to take a higher education class, while a very small number of students (10%) decided to enroll the MOOC because they could not afford to attend a formal education.

Lack of time - the amount of time required to finish a course is shown to be another factor which has a significant impact on preventing students from completing the MOOCs requirements. A survey conducted by [14] figured out that watching online lectures and completing homework assignments and quizzes require too much time of student's schedule, and this is reported to be one of the factors that cause students to drop out of MOOCs.

Insufficient background knowledge and skills – another reason which may cause students to drop out from a registered MOOC is inadequate background knowledge and lack of required skills. Particularly, difficulties with math requirements are seen to be an issue that students were not able to complete a course [14]. As much of the interaction in MOOCs rely on text, it is required by students to have, beyond technical skills, strong skills in reading, writing, and typing. Lack of these skills is seen very often to be a cause for not completing or withdrawing a course [15].

B. MOOC-related factors

Course design – is one the key factors that cause students to dropout of MOOCs. Course design constitutes of three components, namely, course content, course structure, and information delivery technology. Among these three components, course content is the most significant predictor of MOOC retention [16]. Students who completed the MOOCs reported that these courses provided the content they were interested in learning accompanied with real cases and examples and practical implementation. Tips about soft skills and the learning material also were provided by theses course. On the contrary, issues related to the content of the MOOCs such as the courses were too complex or technical, the language used was

too complicated, the course had too many modules, etc., were some of the complaints reported by students who did not complete the MOOCs [17].

Feeling of isolation and lack of interactivity in MOOCs - is another factor which is shown to have a direct effect on students' dropout of MOOCs. Results of a survey conducted by [16] about MOOCs' dropout showed that there is no engagement of students in a discussion or brainstorming providing thus low interaction and poor feedback between the lecturer of the course and students. Students also mentioned in the survey that teamwork and communication between students were also not present and this creates the feeling of isolation.

Hidden costs – this is another reason that may cause high students withdraw rate of MOOCs. These costs represent an amount of money which sometimes is required to be paid by students to get their certificates or to purchase pricey textbooks recommended by lecturers of the courses [18].

Identification and exploration of factors that have a direct effect on student's dropout or withdrawal of MOOCs would enable researchers, lectures, and educational technologists to investigate and propose new strategies and techniques that will help students to persist longer and complete MOOCs successfully.

III. State-of-the-art Research On Mooc Dropout Prediction

Although published research works relating to MOOCs dropout was noticeably scarce at the time of their initial introduction, there is now a growing knowledge body of relevant literature. Comparing existing ML architectures for MOOC dropout prediction however is a challenging task due to a wide variety of features that are used, each of which is designed for a slightly different problem specification.

The current state-of-the-art approaches dealing with MOOC dropout prediction, as illustrated in Table 2, are mostly using clickstream features as engagement patterns. The clickstream features are directly computed from the clickstream log files, which contain the interaction events among students and the MOOC courseware including discussion forums, video lectures, answers to quiz questions, assignments and more.

Different from other works, [19] apply an approach that used K-means to make quantitative analysis by employing students clustering aiming at discovering inactive students automatically in MOOC environment. Qiu et al. [20] among other researchers focus on proposing dropout prediction models based on support vector machines. They conclude that students who exert higher effort and ask more questions are not necessarily more likely to get certificates, while on the other hand, the probability that a student obtains the course certificate increases significantly when she or he has one or more "certificate friends." In [12], the authors proposed a Support Vector Machine (SVM) framework focusing on clickstream data. They found that prediction is better at the end of course rather than at the beginning where they detected rather weak data signals. They also discovered significant interclass variations in the data which helped them achieve an increase in prediction accuracy up to 15% for some weeks of the course.

TABLE II. Overview of state-of-the-art research on MOOC dropout prediction using ML

Engagement Pattern - Clickstream Features	
Study	ML architecture
Liu & Li [19]	K-means
Wang et al. [30]	CNN+RNN
Al-Shabandar et al. [31]	DT
Al-Shabandar et al. [32]	DT
Whitehill et al. [9]	DNN
Nagrecha et al. [23]	DT+LR
Xing et al. [8]	Bayesian Net.+DT
Robinson et al. [34]	NLP
Qiu et al. [20]	SVM+LR
Liang et al. [24]	LR+SVM
Crossley et al. [5]	NLP
Whitehill et al. [27]	Multinomal LR
Boyer and Veeramachaneni [25]	LR
Chaplot et al. [33]	ANN
Coleman et al. [35]	NLP
Kizilcec et al. [26]	LR
He et al. [10]	LR
Taylor et al. [28]	LR
Stein & Allione [29]	Survival Analysis
Fei & Yeung [6]	RNN+HMM
Kloft et al. [12]	SVM
Amnueypornsakul et al. [21]	SVM
Balakrishnan &Coetzee [22]	HMM+SVM
Engagement Pattern - Other Features	
Study	ML architecture
Jiang et al. [36]	LR
Rose et al. [37]	Survival Analysis

Furthermore, SVM approaches were also used by [21] and [22]. In [21] it was observed that features related to the quiz submission and those that capture interaction with

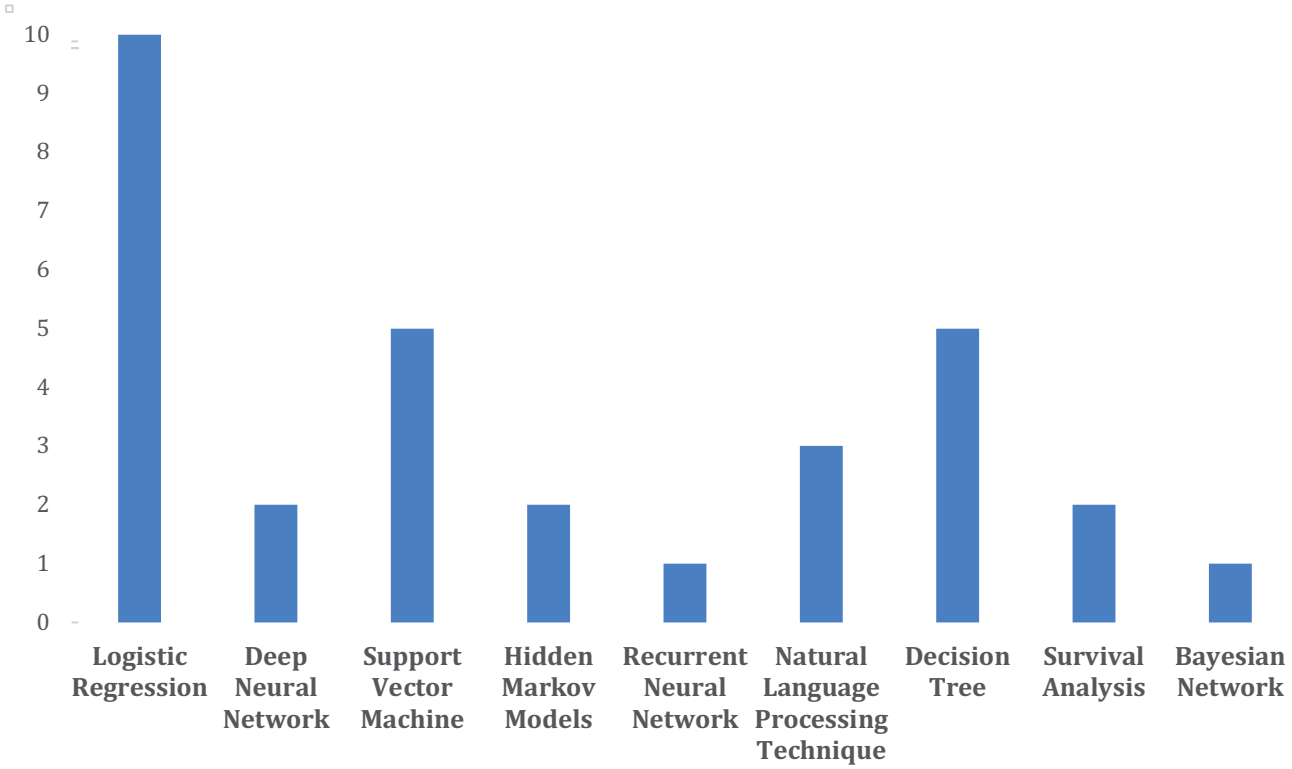


Fig. 2. Frequency of occurrences of ML architectures for MOOC dropout prediction in literature

various course components are found to be reasonable dropout predictors in a certain week. On the other hand, [22] used SVM and Hidden Markov Models (HMM) to design effective prediction models by producing a reasonable Receiver Operating Characteristic (ROC) curve with an Area Under Curve (AUC) value of 0.710. The study also observed that a student who never checks their course progress largely increases their probability of dropping out only after the fourth week. Furthermore, the HMM method has also been used by [6].

As shown in Figure 2, logistic regression (LR) has been the most frequently used technique. It has been implemented, among others, by [23], [24], [25], [26], [10], [27], and [28] to predict student dropout in MOOC environment based on clickstream data. In [23] authors incorporate interpretability in MOOC dropout prediction and demonstrate how existing dropout pipelines can be interpretable and enable analysis of both models and predictions longitudinally. Also, in [24], authors use LR for dropout prediction task, achieving 89% prediction accuracy. In another work [25], LR is applied for dropout prediction addressing the need of transferring knowledge across courses. A systematic investigation of MOOC dropout using LR based on data from 100,000 students in 21 courses was conducted by [26]. This study yielded that the likelihood that the student drops out increases substantially if the student disengages for 14 days or more, whereas the probability of such a re-engaging increase with the fraction of released videos the student has viewed prior to the absence.

The first steps towards automatic intervention in MOOC student dropout, and early and accurately identifying at-risk students were taken by [27] and [10]. Here, authors use and compare LR with many other prediction models in terms of performance on Coursera data and propose two transfer learning algorithms to trade-off smoothness and accuracy. In [27] an automatic classifier of MOOC dropout and it was shown that it generalizes to new MOOCs with high accuracy. For classification, the authors used multinomial logistic regression (MLR).

Another prediction methodology based on LR was documented in [28], which involves a meticulous engineering of over 25 predictive features. The study revealed that dropout prediction is a tractable problem, achieving an AUC as high as 0.95.

There are also works that considered Recurrent Neural Networks (RNN) to predict the MOOC dropouts [6], [30]. In [6], authors approach the dropout prediction from a sequence labeling perspective, and by using RNN model with long short-term memory cells (LSTM), they obtain significantly better dropout prediction results than HMM. Most recently, another successful attempt using RNN in combination with Convolutional Neural Network (CNN) was also made to automatically extract features from the raw MOOC data [30]. In this study, it was shown that the model could achieve comparable results to feature engineering-based methods, saving a lot of time and human efforts, and eliminate the potential inconsistency introduced by the manual process.

Among various ML algorithms, some researchers also focus on decision tree (DT) learning [8], [31], [32], deep neural network (DNN) [9], sentiment based artificial neural network (ANN) [33], and natural language processing statistical models (NLP) [5], [34], [35].

In contrast to the abovementioned clickstream feature-based prediction models, there are only a few works that apply ML techniques that use engagement pattern features other than clickstream data, such as grades or social networks [36], [37].

IV. Challenges Of Dropout Prediction Using Machine Learning

One prominent challenge facing dropout prediction for MOOC is the lack of enough sample data that not only corresponds to dropouts but also to graduates for unbiased classification results. The effectiveness of the ML relies on the availability of enormous amount of both positive and negative samples. Many students register on various MOOCs just to experience self-paced online learning, even with the intention to obtain a certification, dropout quite frequently during the course duration. The lack of negative samples is evident from the Harvard MOOC dataset [38], where out of 641138 registered students only 17687 obtained the certification, thus leaving many positive samples (623451) for dropouts but a handful of certified students. A significant difference in observable sample data affects the generalization of the deep neural network model during the training step but also comprises the classification accuracy.

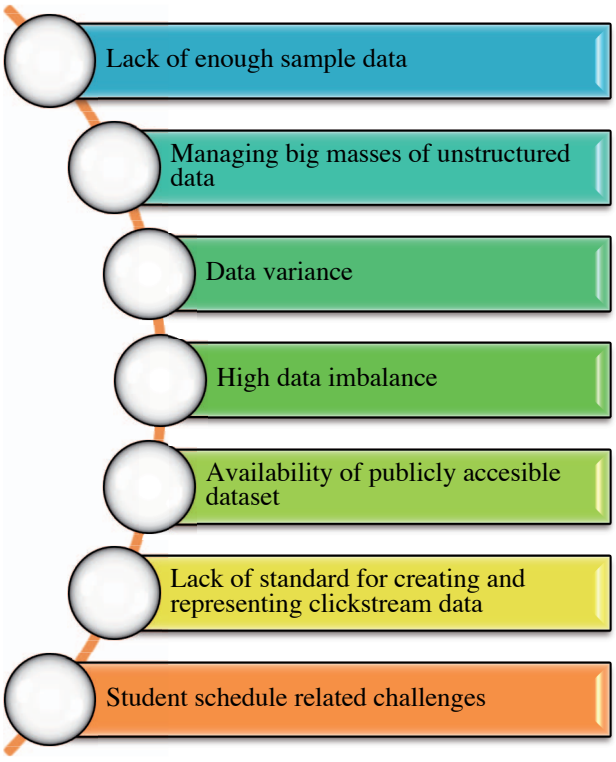


Fig. 3 Challenges of ML student dropout prediction in MOOCs

Moreover, due to the fact that MOOC contains big masses of unstructured data, within the paradigm of big data, management is perhaps the most challenging problem since missing data often occurs and is harder to validate. Consequently, a multitude of ML techniques, e.g., HMM cannot be applicable since this approach relies on a finite set of data and so cannot be applied if observations are missing. Researchers often tend to resolve this issue by replacing missing observations with mean values which may not be a realistic choice for many of the observed features.

Data variance is another factor that affects ML reliability. Namely, in MOOC's self-paced learning pedagogy, students have the freedom to decide what, when, and how to study. This might lead to considerable data variance, which may produce less accurate and reliable ML models, particularly for SVM and Naive Bayes, since their performance can quickly turn poor when dealing with imbalanced classes. In other words, the high data imbalance may result in biasing of the classifier towards the majority of the class.

Another challenge is the availability of publicly accessible dataset. MOOC platforms are often reluctant to publish the data due to confidentiality and privacy concerns. The de-identified information which is made public is also restricted to system generated events (clickstream data) only while most of the user-provided data is omitted. The lack of user-provided data can be a hindrance in successfully estimating the reasons behind the dropouts.

eLearning platforms on the other hand design and adapt MOOCs to their needs. The user data they gather and the clickstream data they generate doesn't necessarily conform to a standard format, due to lack of available standards - such as IEEE LOM for learning objects. The lack of a standard for creating and representing clickstream data for MOOC implies that a network trained and validated for one MOOC on a given platform might not be interoperable across different MOOC on another platform, or even on the same platform if a MOOC gathers and collects data differently. Lack of standards gives rise to another challenge, i.e., how to create compelling ML models and predictive solutions that can be generalized across different MOOCs.

Furthermore, it is noteworthy to mention that there are still many unsolved student schedule related challenges that MOOCs are currently facing. Students want to construct timetable based on their preferences, and this is especially problematic because of the wide range of such preferences. Some prefer to move through the course week by week, and some prefer to explore material during the entire run of the course, while others prefer to have all the lecture videos and assignments prior to the start of the course [39]. Satisfying such preferences of students to be accommodated on different timetables is a complicated problem which can be solved using artificial intelligence (AI) optimization techniques. These AI techniques are specialized to be used for scheduling and thus students' time constraints can be solved using these techniques.

A schematic representation of the identified challenges of ML student dropout prediction in MOOCs is given in Figure 3.

V. Towards Useful And Effective Predictive Solutions

This section presents various recommendations and proposals towards useful and effective predictive solutions for dropout predictions, which may not only assist in developing generalizable solutions across different MOOCs but also help lecturers to timely intervene if they foresee dropouts during a course.

A. Clickstream data standardization

Similar to learning object metadata standards such as the IEEE LOM, CanCore, SCORM, there is a need for a unification of clickstream data for MOOCs. The standard should incorporate all possible traceable events and interactions from various multimedia elements/components within a MOOC including frequency of interactions, user presence time, number of videos watched, documents viewed, and links opened - among others.

B. Student-provided data

ML techniques can benefit from the student-provided data in predicting dropouts. Student's collected data that doesn't reveal the identity of the student and may not require de-identification should also be made part of the standardization process. The data may be gathered either explicitly, i.e., provided by the student: such as age, gender, demography, prerequisite courses, degree level, background knowledge, or implicitly by employing feature engineering techniques using data mining. It is important during this stage to employ a data collection method that complies with ethical standards, since personal data privacy concerns may arise. Therefore, it is necessary to protect privacy by restricting data access and by utilizing access control to the data entries and by incorporating security control mechanisms such that sensitive information cannot be compromised.

C. Feature engineering techniques

ML prediction model can even learn from unstructured text to predict the dropouts. To further improve the prediction model performance, different feature engineering techniques need to be explored and incorporate other features such as test grades, student's prior experience, and more.

D. Engagement System

In a MOOC data-intensive environment as it is now, the lecturer cannot track how the whole group of students interact with the course. Hence, the MOOC platform designers need to employ a system which will monitor student engagement with course material. In addition, that engagement system or module would profile an entire course. For instance, to timely identify students that are likely to drop, the module should alert the lecturer if a particular student has not been engaging with the course for a certain period of time. In this way, the lecturer can design educational interventions and communicate with students to encourage before they completely disengage. Accordingly, the low engagement and the educational intervention reflect the existent opportunity to adapt and refine the future MOOC interfaces. Moreover, from the prediction standpoint, researchers should concentrate more on interventions and to those who are likely to benefit most from interventions.

E. Evaluating models and predictors

When it comes to designing interventions that are related to MOOCs, it is of paramount importance to know their accuracy. To obtain high accuracy estimates, it is essential that the classifier's evaluation method complies with the manner it will be applied to the actual intervention. In this regard, there are few attempts [25], [9] to transfer and evaluate models from one MOOC course to another and to perform live or real-time prediction with ongoing courses. However, more significant evidence in literature is needed towards evaluating more methods systematically on multiple courses since it is not yet clear whether a prediction model trained on a certain course can be extensible or effective to other courses. On the other hand, as observed by [40], identification of students who are less likely to complete the course is more accurate when the predictions are tailored for each course separately.

F. Improving student's social engagements

The students' social interactions via social networks potentially might provide interesting additional evidence for engagement and persistence in MOOC. However, we have not yet witnessed serious research efforts on (i) reliable models that integrate both the predictors of academic assessment and exploring social network structures, and (ii) how to improve students' social integration and interaction in the MOOC environment.

Stimulating and encouraging student's interaction via discussion forum activities and peer-evaluations could also possibly represent an efficient way to mitigate the dropout. Although engaging a large number of students in such MOOC forums is not a trivial task, it may however bring some enhancements and more knowledge to all students. The most inactive students in the forum could benefit from actively participating students through reading to existing discussions. In this way, they could find answers to their assignments and engage further by posting and asking for their concerns. Having said that, relevant studies are still missing to investigate how student involvement in peer-evaluated writing assignments and discussion forums is associated with learning results and dropout.

VI. Conclusion

This paper provides a comprehensive review of the most recent and relevant research endeavors on machine learning application toward predicting, explaining and solving the problem of student dropout in MOOCs. It highlights both student-related factors and MOOC related factors that lead to a high number of dropouts. The paper also identifies some of the critical challenges associated with student dropout prediction and provides recommendations and proposal to assist researchers employing various machine learning techniques in solving it timely and efficiently. Additionally, this paper floats the idea of unification of a clickstream data and student-provided data as a standard, similar to various learning objects metadata standards.

The MOOC student dropout topic is only partially researched and there is a wide space of research aspects to be investigated further in order to build well-defined and more precise predictive solutions to identify, understand and explain

the reasons of dropout. A deeper understanding of the reasons why student dropout could help course developers and lecturers improve course content and undertake educational interventions.

Future work will follow on data collection for the development and analysis of deep learning algorithms towards solving the MOOC student dropout.

REFERENCES

[1] K. Masters, "A Brief Guide to Understanding MOOCs", The Internet Journal of Medical Education 1(2011), pp. 392-410.

[2] F. Dalipi, A. Kurti, K. Zdravkova, L. Ahmedi, "Rethinking the conventional learning paradigm towards MOOC based flipped classroom learning", 2017 IEEE International Conference on Information Technology Based Higher Education and Trainint (ITHET), July 10-12, 2017, Ohrid, Macedonia.

[3] R. Chen, " Institutional characteristics and college student dropout risks: A multilevel event history analysis". *Research in Higher Education*, *53*(5), 487-505.

[4] L. Ulriksen, L.M. Madsen, H.T. Holmegaard, "Why do students in stem higher education programmes drop/opt out?–explanations offered from research". In *Understanding student participation and choice in science and technology education* (pp. 203-217). Springer Netherlands.

[5] S. Crossley, L. Paquette, M. Dascalu, D.S. McNamara, R.S. Baker, "Combining click-stream data with nlp tools to beter understand MOOC completion", in 6th International Conference on Learning Analytics and Knwoledge, pp. 6-14, ACM 2016.

[6] M. Fei, D.Y. Yeung, "Temporal models for Predicting Student Dropout in Massive Open Online Courses", 2015 IEEE International Conference on Data Mining Workshop (ICDMW).

[7] K.R.Koedinger, J. Kim, J.Z. Jia, E.A. McLaughlin, N.L. Bier, "Learning is Not a Spectator Sport: Doing is Better than Watching for Learning from a MOOC", L@S 2015, March 14-18, 2015, Vancouver, Canada.

[8] W. Xing, X. Chen, J. Stein, M. Marcinkowski, "Temporal predication of dropouts in MOOCs: Reaching the low hanging fruit through stacking generalization", Computers in Human Behavior, Volume 66, January 2017.

[9] J. Whitehill, K. Mohan, D. Seaton, Y. Rosen, D Tingley, "Delving Deeper into MOOC Student Dropout Prediction", arXiv:1702.06404v1.

[10] J. He, J. Bailey, B.I.P. Rubinstein, R. Zhang, "*Identifying At-Risk Students in Massive Open Online Courses*", 29th AAAI Conference on Artificial Conference, Austin, January 25-29, 2015, TX, USA.

[11] D. Yang, T. Sinha, D. Adamson, C.P. Rose, "Turn on, Tune in, Drop out: Anticipating student dropouts in Massive Open Online Courses", Proceedings of the 2013 NIPS Data-driven education workshop, Nevada, USA.

[12] M. Kloft, F. Stiehler, Z. Zheng, and N. Pinkwart, "Predicting MOOC dropout over weeks using machine learning methods". In Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs, pages 60–65, 2014.

[13] L. Yan, S. Bowel, (2013) MOOCs and Open Education: Implications for Higher Education. http://publications.cetis.org.uk/2013/667. (consulted 10 November 2017).

[14] Y. Belanger, and J. Thornton, (2013). Bioelectricity: A quantitative approach. Duke University's First MOOC. https://dukespace.lib.duke.edu/dspace/bitstream/handle/10161/6216/Duke_Bioelectricity_MOOC_Fall2012.pdf?sequence=1. (consulted 3 November 2017).

[15] B. Murray, (2001), "What makes students stay". eLearn Magazine. http://elearnmag.acm.org/archive.cfm?aid=566901. (consulted 1 November 2017).

[16] K. Hone, and G. Said, "Exploring the factors affecting MOOC retention: A survey Study". *Computers & Education* 98, 157-168.

[17] P. Adamopoulos, "What makes a great MOOC? An interdisciplinary analysis of student retention in online courses". In Proceedings of 34th International Conference on Information Systems, *pp. 1-21.*

[18] H. Khalil, M. Ebner, "MOOCs completion rates and possible methods to improve retention: a literature review", in EdMedia: World Conference on Educational Multimedia, Hypermedia and Telecommunications, Tampere, Finland, June 2014.

[19] T. Liu, X. Li, "Finding out Reasons for Low Completion in MOOC Environment: An Explicable Approach Using Hybrod Data Mining Methods", 2017 International Conference on Modern Education and Information Technology (MEIT 2017), June 24-25, 2017 at Chongqing, China.

[20] J. Qiu, J. Tang, T.X. Liu, J. Gong, C. Zhang, Q. Zhang, Y. Xue, "Modeling and Predicting Learning Behavior in MOOCs", 9th ACM International Conf. on Web Search and Data Mining, February 22 - 25, 2016, San Francisco, California, USA.

[21] B. Amnueypornsakul, S. Bhat, P. Chinprutthiwong, "Predicting Attrition Along the Way: The UIUC Model", Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), October 25-29, 2014, Doha, Qatar.

[22] G. Balakrishnan, D. Coetzee, "Predicting Student Retention in Massive Open Online Courses using Hidden Markov Models". Technical report, UC Berkeley, 2013.

[23] S. Nagrecha, J. Z. Dillon, N.V. Chawla, "MOOC Dropout Prediction: Lessons Learned from Making Pipelines Interpretable", WWW'17 Companion, April 3–7, 2017, Perth, Australia.

[24] J. Liang, C. Li, L. Zheng, " Machine Learning Application in MOOCs: Dropout Prediction", The 11th International Conference on Computer Science & Education (ICCSE 2016) August 23-25, 2016. Nagoya University, Japan.

[25] S. Boyer, K. Veeramachaneni, "Transfer learning for predictive models in massive open online courses", In International Conference on Articial Intelligence in Education, Springer, 2015.

[26] R. Kizilcec, S. Halawa. Attrition and achievement gaps in online learning. In L@S 2015, March 14-18, 2015, Vancouver, Canada.

[27] J. Whitehill, J. Williams, G. Lopez, C. Coleman, and J. Reich, "Beyond prediction: Toward automatic intervention to reduce mooc student stopout", In Educational Data Mining , 2015.

[28] C. Taylor, K. Veeramachaneni, U.-M. O'Reilly, "Likely to stop? Predicting stopout in massive open online courses" arXiv , 2014. http://arxiv.org/abs/1408.3382 .

[29] R. Stein, G. Allione, "Mass attrition: An analysis of drop out from a principles of microeconomics MOOC". PIER Working Paper , 14(031), 2014.

[30] W. Wang, H. Yu, C. Miao, "Deep Model for Dropout Prediction in MOOCs", Proceedings of the 2nd ACM International Conference on Crowd Science and Engineering, July 06-09, 2017, Beijing, China.

[31] R. Al-Shabandar, A. Hussain, A. Laws, R. Keight, J. Lunn, N. Radi, "Machine Learning Approaches to Predict Learning Outcomes in Massive Open Online Courses", IEEE International Joint Conference On Neural Networks, May 14-19, 2017, Anchorage, USA.

[32] R. Al-Shabandar, A. Hussain, A. Laws, R. Keight, J. Lunn, "Towards the Differentiation of Initial and Final Retention in Massive Open Online Courses", International Conference on Intelligent Computing, August 7-10, 2010, Liverpool, UK.

[33] D.S. Chaplot, E. Rhim, J. Kim, "Predicting Student Attrition in MOOCs using Sentiment Analysis and Neural Networks", In Proceedings of AIED 2015 Fourth Workshop on Intelligent Support for Learning in Groups (2015).

[34] C. Robinson, M. Yeomans, J. Reich, C. Hulleman, and H. Gehlbach, " Forecasting student achievement in moocs with natural language processing". In Proceedings of the Sixth International ACM Conference on Learning Analytics & Knowledge, April 25-29, 2016, Edinburgh, UK.

[35] C. Coleman, D. Seaton, and I. Chuang, "Probabilistic use cases: Discovering behavioral patterns for predicting certication". L@S 2015, March 14-18, 2015, Vancouver, Canada.

[36] S. Jiang, A. Williams, K. Schenke, M. Warschauer, and D. O'Dowd, "Predicting MOOC performance with week 1 behavior". In 7th

International Conference on Educational Data Mining, July 4-7, 2014, London, UK.

[37] C. P. Rose, R. Carlson, D. Yang, M. Wen, L. Resnick, P. Goldman, and J. Sherer, "Social factors that contribute to attrition in moocs". In Proceedings of the 1st ACM conference on Learning @ Scale cConference, March 4-5, 2014, Atlanta, USA.

[38] "HarvardX-MITx Person-Course Academic Year 2013 De-Identified dataset, version 2.0", doi:10.7910/DVN/26147, Harvard Dataverse, V10, https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/26147 (Consulted 6 November 2017).

[39] D. Bruff (2013), "Lessons Learned from Vanderbilt's First MOOCs" - Center for Teaching. https://cft.vanderbilt.edu/2013/08/lessons-learned-from-vanderbilts-first-moocs/. (Consulted 5 November 2017)

[40] A. Wolff, Z. Zdrahal, D. Herrmannova, J. Kuzilek, and M. Hlosta, "Developing predictive models for early detection of at-risk students on distance learning modules". In Machine Learning and Learning Analytics workshop at LAK'14 ACM Conference, 24-28 March 2014, Indianapolis, Indiana, USA.