

# Shared Task on Prediction of Dropout Over Time in Massively Open Online Courses

**Carolyn P. Rosé**

Language Technologies Institute  
and Human-Computer Interaction Institute  
Carnegie Mellon University  
5000 Forbes Avenue, Pittsburgh, PA 15213  
[cprose@cs.cmu.edu](mailto:cprose@cs.cmu.edu)

**George Siemens**

Center for Distributed Education  
University of Texas at Arlington  
701 South Nedderman Drive, Arlington, TX  
76019  
[gsiemens@uta.edu](mailto:gsiemens@uta.edu)

## Abstract

The shared task on Prediction of Dropout Over Time in MOOCs involves analysis of data from 6 MOOCs offered through Coursera. Data from one MOOC with approximately 30K students was distributed as training data and consisted of discussion forum data (in SQL) and clickstream data (in JSON format). The prediction task was Predicting Attrition Over Time. Based on behavioral data from a week's worth of activity in a MOOC for a student, predict whether the student will cease to actively participate after that week. This paper describes the task. A full write up of the results is published separately (Rosé & Siemens, 2014).

## 1 Overview

Research on Massively Open Online Courses (MOOCs)<sup>1</sup> is an emerging area for real world impact of technology for analysis of social media at a large scale (Breslow et al., 2013). Modeling user experience in MOOCs supports research towards understanding user needs better so that experiences that are more conducive to learning can be offered. Beyond that, automated analyses enable adaptive technology to tailor the experience of users in real time (Rosé et al., 2014a). This paper describes a shared task designed to enlist the involvement of the language technologies community in this endeavor and to identify what value expertise within the field might bring.

One area for impact of natural language processing in the MOOC space is in modeling behavior within the threaded discussion forums. In a typical MOOC, between 5% and 10% of students actively participate in the threaded discussion forums. Previously published research demonstrates that characteristics of posting behavior are predictive of dropout along the way (Rosé et al., 2014b; Wen et al., 2014a; Wen et al., 2014b; Yang et al., 2013; Yang et al., 2014). However, ideally, we would like to make predictions for the other 90% to 95% of students who do not post. Thus, in this shared task, we challenge participants to use models of social interaction as displayed through the text-based interaction between students in the threaded discussions (from the minority of students who participate in them) to make meaning from the clickstream data we have from all students. If the discussion data can be thus leveraged to make more effective models of the clickstream data, then a meaningful prediction about drop out along the way can also be made about the students who do not post to the discussion forums.

One of the biggest challenges in the shared task is that the participants were only given data from one Coursera MOOC as training and development data. Their task was to produce a predictive model that could be applied to data from other MOOCs they did not have access to. A separate report describes a detailed analysis of the results applying submitted models to each of 5 test MOOCs (Rosé & Siemens, 2014).

12 research teams signed up for the shared task, including an international assortment of academic and industrial teams. Out of these 12 teams, only 4 submitted final models (Sinha et al., 2014; Sharkey & Sanders, 2014; Amnueypornsakul et al., 2014; Kloft et al., 2014).

In the remainder of this paper we describe the shared task in greater detail and discuss plans for future related research.

---

<sup>1</sup> <http://www.moocresearch.com/reports>

## 2 Shared Task

Participants in the shared task were given a complete SQL dump and clickstream dump from one Coursera MOOC as training data. The student-week was the unit of analysis. In other words, a prediction was made for each student for each week of their active participation to predict whether that week was the last week of their active participation. Scripts were provided to parse the data into a form that could be used for the task, e.g., aggregating entries per user per week. Scripts were also provided for running a test of the trained model on test data. The purpose of the scripts was to standardize the way in which each team's work would later be evaluated on the test MOOCs that participants did not have access to.

A major part of the work in doing the task is in determining what an effective representation would be of the behavior trace associated with each student-week that would enable making an accurate prediction. In other words, the question is what are the danger signs that a student is especially vulnerable to drop out? The rules of the task were such that the information the model was allowed to use for making the prediction could be extracted from the whole participation history of all training students (including both the SQL data and the clickstream data) up to and including the week a prediction was being made for.

Each of the four finalist teams submitted a final model trained on the training MOOC and a write up including result trained on a designated subset of students from the training MOOC and tested on the remaining students. Results were presented in terms of precision, recall, and fmeasure for the held out users.

We recommend that participants make use of the text data to bootstrap effective models that use only clickstream data. However, participants were welcome to leverage either type of data in the models they submitted. In our evaluation presented separately (Rosé & Siemens, 2014), we evaluated the models on the test MOOCs in three different ways: First, an evaluation was conducted on data from students who actively participated in the discussion forums. Second, an evaluation was conducted on data from students who never participated in the discussion forums. And finally, an evaluation was conducted on the set of students that includes both types of students.

Each submission consisted of a write up describing the technical approach and a link to a downloadable zip file containing the trained model and code and/or a script for using the trained model to make predictions about the test sets. The code was required to be runnable by launching a single script in Ubuntu 12.04. A code stub for streamlining the preparation of the submission was distributed with the data. The following programming languages were acceptable: R 3.1,

C++ 4.7, Java 1.6, or Python 2.7. The script was required to be able to run within 24 hours on a 2400 MHz machine with 6 cores.

## 3 Looking Forward

Computational modeling of massive scale social interaction (as in MOOCs and other environments for learning at scale) has the potential to yield new knowledge about the inner-workings of interaction in such environments so that support for healthy community formation can be designed and built. However, the state-of-the-art in graphical models applied to large scale social data provides representations of the data that are challenging to interpret in light of specific questions that may be asked from a learning sciences or social psychological perspective. What is needed are new methodologies for development and interpretation of models that bridge expertise from machine learning and language technologies on one side and learning sciences, sociolinguistics, and social psychology on the other side. The field of language technologies has the human capital to take leadership in making these breakthroughs.

The shared task described in this paper is the first one like it where a data set from a Coursera MOOC has been made publically available so that a wide range of computational modeling techniques can be evaluated side by side (Rosé & Siemens, 2014). However, there is recognition that such shared tasks may play an important role in shaping the future of the field of Learning Analytics going forward (Pea, 2014).

One of the major challenges in running a shared task like this is ensuring the protection of privacy of the MOOC participants. Such concerns have been the focus of much discussion in the area of learning at scale (Asilomar Convention, 2014).

Data sharing ethics were carefully considered in the design of this shared task. In particular, all of the students who participated in the MOOC that produced the training data were told that their data would be used for research purposes. The data was carefully preprocessed to remove personal identifiers about the students and the university that hosted the course. All of the workshop participants who got access to the data were required to participate in human subjects training and to agree to use the data only for this workshop, and not to share it beyond their team. Data was shared through a secure web connection. Approval for use of the data in this fashion was approved by the Institutional Review Board of the hosting university as well as the university that ran the MOOC.

It was a goal in development of this shared task to serve as a forerunner in what we hope will become a more general practice of community wide collaboration on large scale learning analytics (Suthers et al., 2013).

## Acknowledgements

The authors would like to thank Norman Bier for assistance in working through the data sharing logistics. This work was funded in part by NSF Grant OMA-0836012.

## References

- Amnueypornsakul, B., Bhat, S., & Chinprutthiwong, P. (2014). Predicting Attrition Along the Way: The UIUC Model, in *Proceedings of the 2014 Empirical Methods in Natural Language Processing Workshop on Modeling Large Scale Social Interaction in Massively Open Online Courses, Qatar, October 2014*.
- Asilomar Convention (2014). *The Asilomar Convention for Learning Research in Higher Education*, June 13, 2014.
- Breslow, L., Pritchard, D., De Boer, J., Stump, G., Ho, A., & Seaton, D. (2013). Studying Learning in the Worldwide Classroom: Research into edX's First MOOC, *Research & Practice in Assessment* (8).
- Kloft, M., Stiehler, F., Zheng, Z., & Pinkward, N. (2014). Predicting MOOC Dropout over Weeks Using Machine Learning Methods, in *Proceedings of the 2014 Empirical Methods in Natural Language Processing Workshop on Modeling Large Scale Social Interaction in Massively Open Online Courses, Qatar, October 2014*.
- Pea, R. (2014). *The Learning Analytics Workgroup: A Report on Building the Field of Learning Analytics for Personalized Learning at Scale*, Stanford University.
- Rosé, C. P. & Siemens, G. (2014). *Shared Task Report: Results of the EMNLP 2014 Shared Task on Predictions of Dropout Over Time in MOOCs*, Language Technologies Institute Technical Report.
- Rosé, C. P., Goldman, P., Sherer, J. Z., Resnick, L. (2014a). Supportive Technologies for Group Discussion in MOOCs, *Current Issues in Emerging eLearning*, Special issue on MOOCs, December 2014.
- Rosé, C. P., Carlson, R., Yang, D., Wen, M., Resnick, L., Goldman, P. & Sherer, J. (2014b). Social Factors that Contribute to Attrition in MOOCs, in *Proceedings of the First ACM Conference on Learning @ Scale*.
- Sinha, T., Li, N., Jermann, P., & Dillenbourg, P. (2014). Capturing 'attrition intensifying' structural traits from didactic interaction sequences of MOOC learners, in *Proceedings of the 2014 Empirical Methods in Natural Language Processing Workshop on Modeling Large Scale Social Interaction in Massively Open Online Courses, Qatar, October 2014*.
- Sharkey, M. & Sanders, R. (2014). A Process for Predicting MOOC Attrition, in *Proceedings of the 2014 Empirical Methods in Natural Language Processing Workshop on Modeling Large Scale Social Interaction in Massively Open Online Courses, Qatar, October 2014*.
- Suthers, D., Lund, K., Rosé, C. P., Teplovs, C., Law, N. (2013). *Productive Multivocality in the Analysis of Group Interactions*, edited volume, Springer.
- Wen, M., Yang, D., & Rosé, C. P. (2014b). Linguistic Reflections of Student Engagement in Massive Open Online Courses, in *Proceedings of the International Conference on Weblogs and Social Media*
- Wen, M., Yang, D., & Rosé, C. P. (2014a). Sentiment Analysis in MOOC Discussion Forums: What does it tell us? in *Proceedings of Educational Data Mining*.
- Yang, D., Sinha, T., Adamson, D., & Rosé, C. P. (2013). Turn on, Tune in, Drop out: Anticipating student dropouts in Massive Open Online Courses, in *NIPS Data-Driven Education Workshop*.
- Yang, D., Wen, M., & Rosé, C. P. (2014). Peer Influence on Attrition in Massively Open Online Courses, in *Proceedings of Educational Data Mining*.