Data Set Size Analysis for Detecting the Urgency of Discussion Forum Posts

Valdemar Švábenský[1], François Bouchet[1], Francine Tarrazona[2], Michael Lopez II[2], Ryan S. Baker[3]
[1]Sorbonne University, LIP6 [2]Ateneo de Manila University [3]University of Pennsylvania
valdemar@mail.muni.cz, francois.bouchet@lip6.fr

ABSTRACT: In both Massive Open Online Courses (MOOCs) and private courses, instructors face a large amount of queries in discussion forum posts that may merit a response. There has been ongoing research on how to employ machine learning to predict a post's urgency in order to focus instructors' attention. However, it is unclear how large a course is needed to develop these models. We took a publicly available data set of 3,503 labeled forum posts and code from one such prior study. We re-trained the six models described in the study, but with progressively smaller sample sizes, to determine if the models' performance would be preserved. Likewise, we demonstrate that using random subsets even as small as 10% of the original data set achieves comparable performance to full data sets in five out of six models.

Keywords: Learning analytics, educational data mining, urgency detection, replication

1 INTRODUCTION

When instructors reply to critical forum posts in MOOCs, it may decrease learners' inactivity and dropout rates (Almatrafi et al., 2018, Švábenský et al., 2023). However, it is difficult for instructors to identify which among the sheer volume of discussion forum posts require an urgent response. Thus, it is useful to determine priority posts by utilizing machine learning techniques.

Learning analytics researchers usually attempt to collect as large data sets as possible, but models for predicting post urgency may be useful in smaller courses as well. This paper explores whether models trained on small data sets can achieve comparable performance to using larger data sets.

In the past, few studies have assessed the generalizability of models trained on small data sets of forum posts. Yee et al. (2023) suggest that this approach has potential to be applied across courses in different academic disciplines. Training models on small textual data sets has been explored in other domains, such as urgency detection models in brief crisis messages (Kejriwal & Zhou, 2020). E.g., "Roof collapse in building on Main Street; multiple people trapped inside" is deemed urgent.

2 RESEARCH METHODS

We build upon a paper by Švábenský et al. (2023), which evaluated six models trained on a set of 3,503 posts from MOOCs at University of Pennsylvania. The models were validated on a separate test set of 29,604 posts from Stanford University. Post urgency was expressed on a 1 to 7 scale, with 7 as the most urgent. Each forum post text was encoded using *Universal Sentence Encoder v4* numerical feature embeddings. The six best-performing models were: Random Forest (RF), eXtreme Gradient Boosting (XGB), Linear Regressor (LR), Ordinal Ridge Regressor (ORR), Support Vector Regressor (SVR) with a Radial Basis Function kernel, and Neural Network (NN) regressor.

We took the original data and code, which are publicly available (see Švábenský et al., 2023), and implemented a slight modification. We progressively attempted to use data set sizes from 5% to 50% of the original, in increments of 5%. For each data set size, each of the six models was trained ten times, every time on a randomly chosen subset of the given size obtained from the original training set of 3,503 posts. Then, each model was evaluated on the original held out test set of 29,604 posts. Finally, as in the original study, model performance on the test set was assessed using Root Mean Squared Error (RMSE) and Spearman ρ correlation between the predicted and actual values of urgency. The final results were averaged across the ten training runs.

3 Results & discussion

As expected, the performance of all models degraded, but to a surprisingly limited extent. Figure 1 shows that across training subsets of different sizes, SVR, NN, and then ORR performed the best, and that a small sample size could be sufficient to train the models. The only exception is the LR model, whose performance degraded substantially until at least 25% of the data set was used. For 5–20%, the average RMSE was as high as 3.60, and the average Spearman correlation only 0.13. From 25% onward, the performance gradually improved with every step of adding more data, and at 50% it reached a satisfactory result of the RMSE of 1.51 and Spearman rho of 0.32.

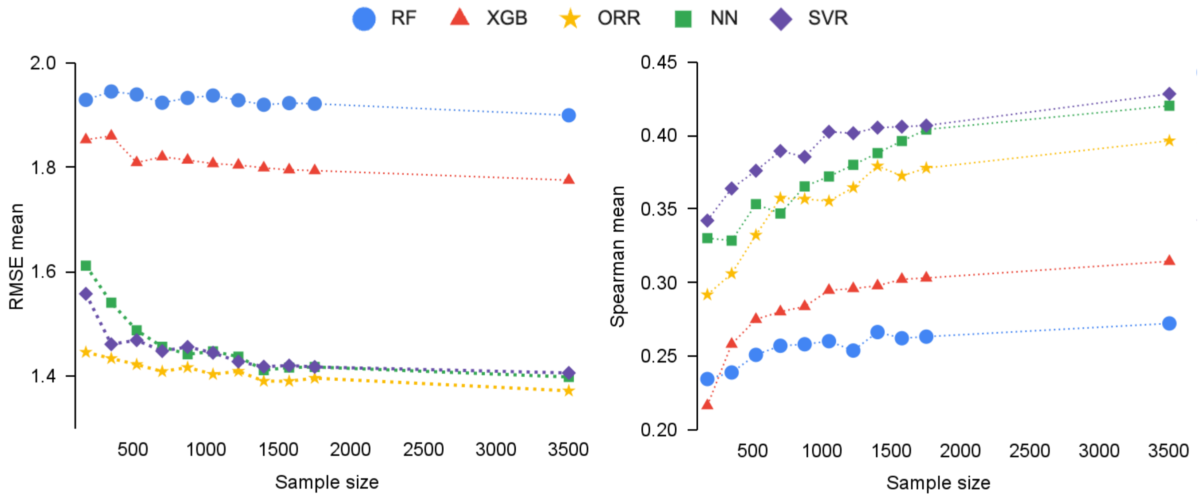

Figure 1: (left) RMSE and (right) Spearman coefficient with respect to the training data set size. LR is excluded from the figures due to its poor performance on very small samples.

There is no clear “elbow” in Figure 1 globally across all five models, but for further investigation we arbitrarily selected 10% (350 posts) to illustrate the difference in performance. Table 1 compares the models trained on the (a) original versus (b) partial data set of as little as 350 posts.

Table 1: Comparison of (a) the results reported by Švábenský et al. (2023), and (b) the same models trained on subsets of 350 posts, evaluated on the original test set, averaged across 10 runs.

	(a) Original models		(b) Models trained on the data subset	
Model	RMSE	ρ	RMSE avg, SD	ρ avg, SD
RF	1.8995	0.2723	1.9499, 0.0530	0.2354, 0.0168

XGB	1.7753	0.3145	1.8246, 0.0374	0.2633, 0.0165
LR	1.3953	0.3882	2.6063, 0.1997	0.1575, 0.0343
ORR	1.3723	0.3964	1.4349, 0.0306	0.3158, 0.0328
NN	1.3988	0.4202	1.5376, 0.0513	0.3220, 0.0355
SVR	1.4065	0.4283	1.4969, 0.0524	0.3752, 0.0215

4 Conclusion and recommendations

The limitation of the past work in this area (Almatrafi et al., 2018, Švábenský et al., 2023) is that it requires time-consuming human labeling of the training data, which not everyone can afford. Identifying the minimal amount of data needed for training prediction models is valuable for replicating this type of detection in other courses and contexts.

Although further research is needed to determine a precise cut-off and demonstrate generalizability, this paper suggests that hundreds, not thousands, of forum posts in the training data set can be sufficient for this problem. Recognizing a point where adding more labeled data does not have a substantial impact anymore can save time for future researchers. Doing so also lowers the barrier of entry to make this approach usable in different contexts, including smaller courses.

Alternatively, future work may utilize weak supervised learning (Zhou, 2018) for detecting urgent posts. Incomplete supervision uses data sets where only a small portion of training data is labeled.

REFERENCES

Almatrafi, O., Johri, A., & Rangwala, H. (2018). Needle in a haystack: Identifying learner posts that require urgent response in MOOC discussion forums. *Computers & Education*, *118*, 1–9. https://doi.org/10.1016/j.compedu.2017.11.002

Kejriwal, M., & Zhou, P. (2020). On detecting urgency in short crisis messages using minimal supervision and transfer learning. *Social Network Analysis and Mining*, 10(1), 58. https://doi.org/10.1007/s13278-020-00670-7

Švábenský, V., Baker, R. S., Zambrano, A., Zou, Y., and Slater, S. (2023). Towards Generalizable Detection of Urgency of Discussion Forum Posts. In Mingyu Feng, Tanja Käser, and Partha Talukdar (Eds.), *Proceedings of the 16th International Conference on Educational Data Mining* (pp. 302–309). https://doi.org/10.5281/zenodo.8115790

Yee, M., Roy, A., Perdue, M., Cuevas, C., Quigley, K., Bell, A., Ahaan Rungta, & Miyagawa, S. (2023). AI-assisted analysis of content, structure, and sentiment in MOOC discussion forums. *Frontiers in Education*, 8. https://doi.org/10.3389/feduc.2023.1250846

Zhou, Z.-H. (2018). A brief introduction to weakly supervised learning. *National Science Review*, *5*(1), 44–53. https://doi.org/10.1093/nsr/nwx106