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

Valdemar Švábenský¹,², François Bouchet¹, Francine Tarrazona³, Michael Lopez II³, Ryan S. Baker⁴
¹Sorbonne University, LIP6 ²Kyushu University ³Ateneo de Manila University ⁴University of Pennsylvania

Which student posts in a course’s online forum need an urgent instructor response?

Instructors face many queries in MOOCs and private courses. ML models can predict a post’s urgency to focus instructor’s attention. Prior work limitation: labeling of training data is time-consuming.

Will the model performance change if trained on progressively smaller data subsets?

Starting point: data and code from a prior study (Švábenský et al., 2023)

Train set 3,503 posts (UPenn)

Human labeling of urgency (1–7)

Test set 29,604 posts (Stanford University)

Use v4 embedding of post texts

6 regressors
- Random Forest
- eXtreme Gradient Boosting
- Linear Regression
- Ordinal Ridge Regression
- Neural Network
- Support Vector Regression

Take a random subset of the train set, of size 50%, 45%, ..., 10%, 5%, and train each regressor

Repeat 10 times, report average performance (measured by RMSE and p)

Evaluate on the test set

Smaller data sets perform comparably to larger data sets for 5 out of 6 models.

Lower entry barrier and less overhead for smaller courses and other contexts.

This can make the data collection and analysis less time-consuming.

Future work can determine a precise cut-off and show generalizability.

No major change was observed after a several hundred data points.