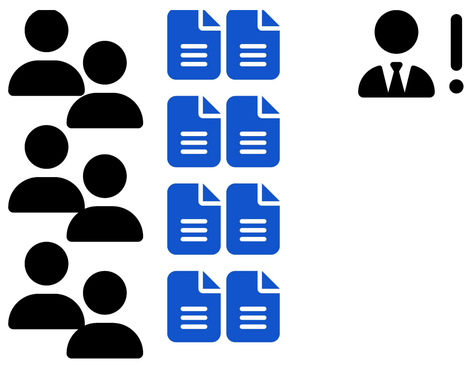
Data Set Size Analysis for Detecting the Urgency of Discussion Forum Posts

Valdemar Švábenský[1,2], François Bouchet[1], Francine Tarrazona[3], Michael Lopez II[3], Ryan S. Baker[4]
[1]Sorbonne University, LIP6 [2]Kyushu University [3]Ateneo de Manila University [4]University of Pennsylvania

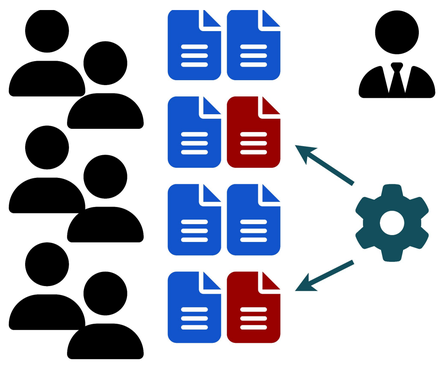
Full poster text

Contact info

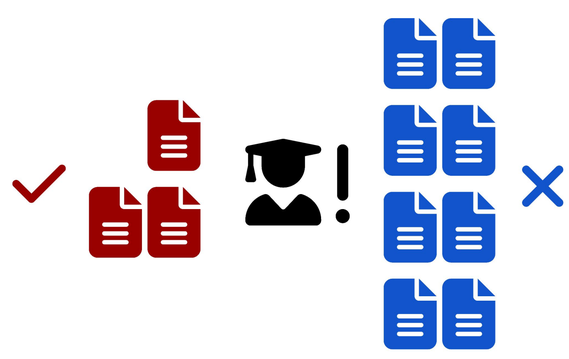
Which student posts in a course's online forum need an urgent instructor response?



Instructors face **many queries** in MOOCs and private courses.

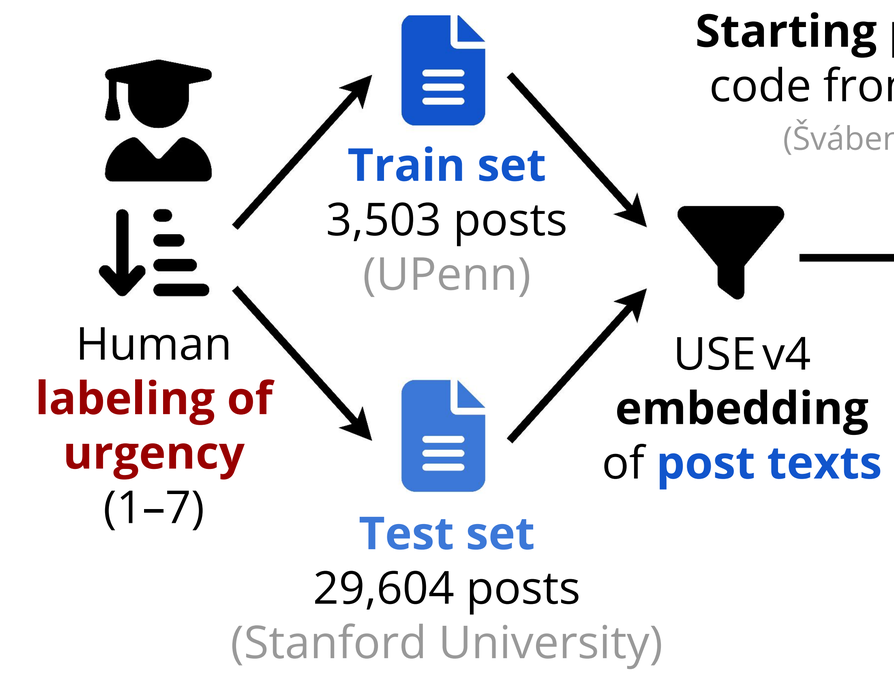


ML models can predict a post's **urgency** to **focus instructor's attention**.



Prior work limitation: **labeling** of **training data** is time-consuming.

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

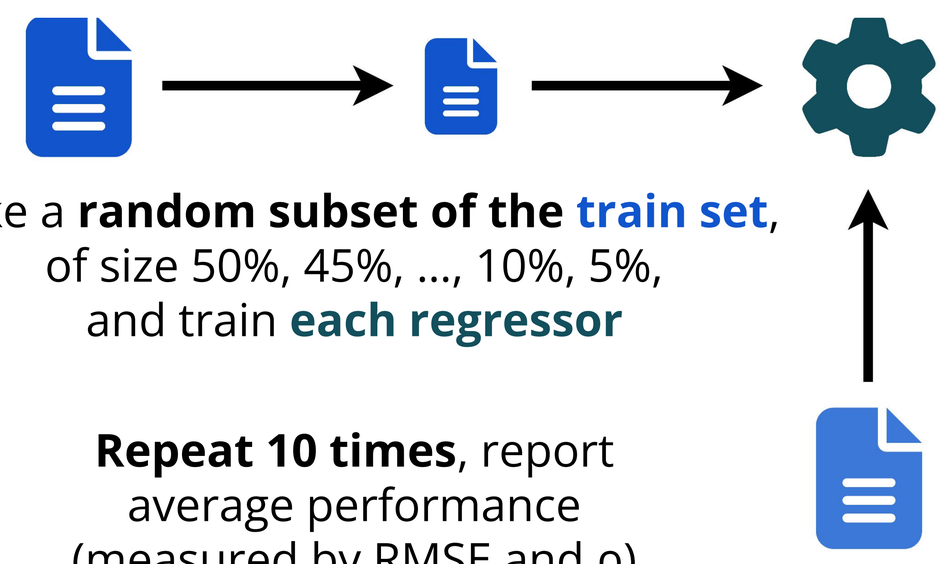


Starting point: data and code from a prior study (Švábenský et al., 2023)

USE v4 **embedding** of **post texts**

6 regressors

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

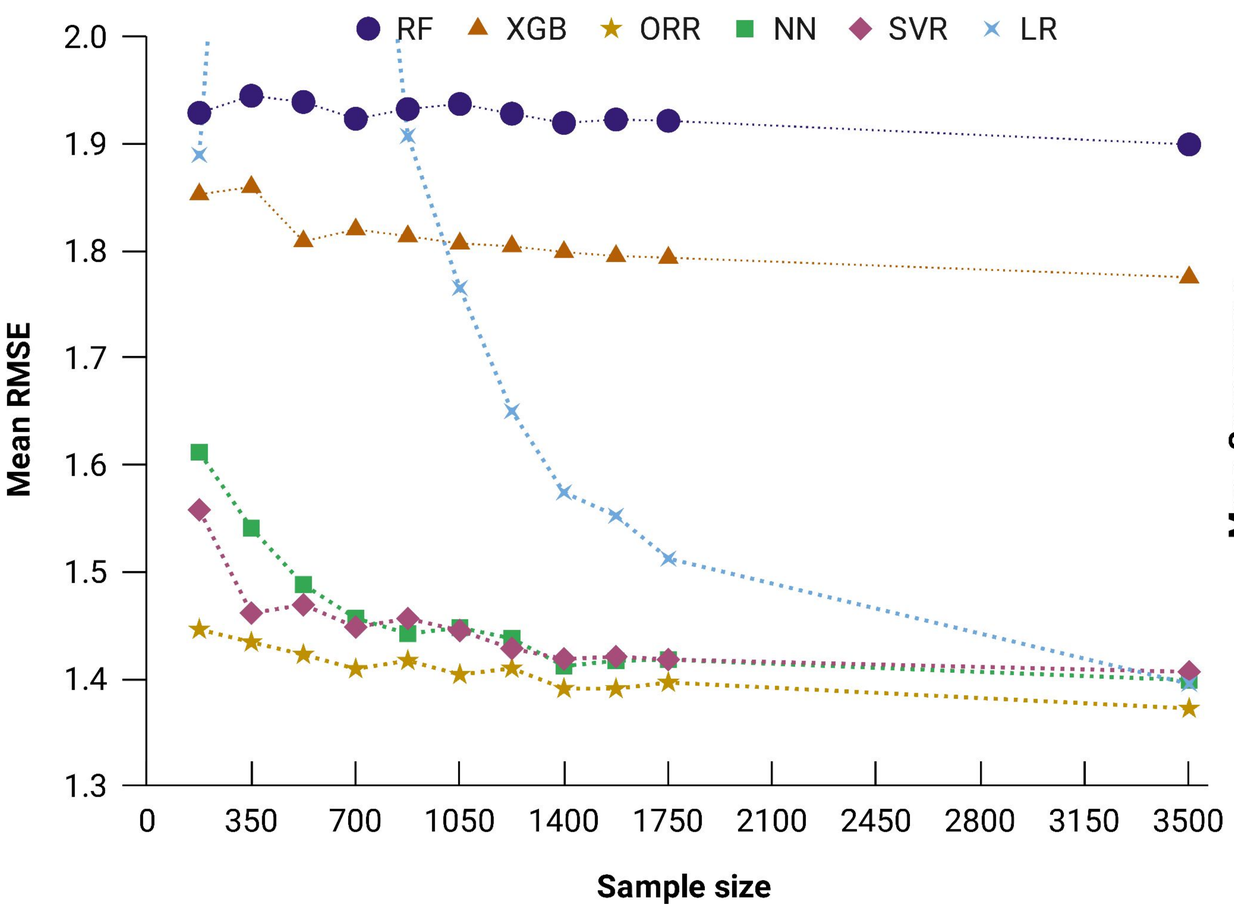


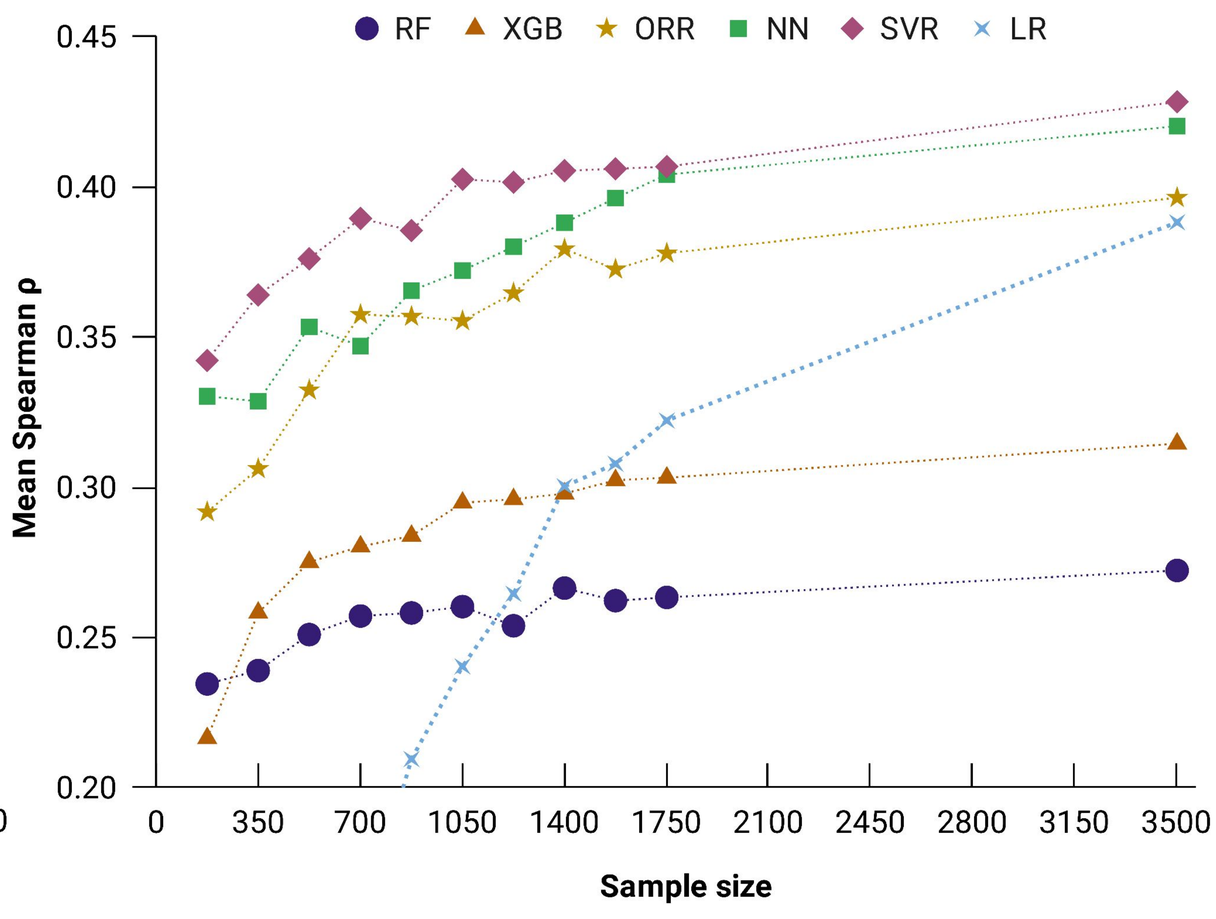
Take a **random subset of the train set**, of size 50%, 45%, ..., 10%, 5%, and train **each regressor**

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

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.

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

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

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