Towards Generalizable Detection of Urgency of Discussion Forum Posts

V. Švábenský, R. S. Baker, A. Zambrano, Y. Zou, S. Slater

EDM 2023

Introduction: Context

- Computer-supported learning environments (e.g., MOOCs)
- Students ask **questions** (email, chat, forum, …)
- Instructor **responses** promote student learning + engagement

Introduction: Problem Statement

- MOOCs have many students → they post **many questions**
- **Urgency** = how quickly is the instructor's response needed?

Introduction: Problem Statement

- Goal: **categorize questions** based on their urgency
  - **Ordinal scale (1–7)** from the Stanford MOOCPosts dataset
- Help instructors decide **where to allocate their time**

- Previous work: only binary classification, only one data set
- Current work: fine-grained classification, two data sets

Methods: Data Collection

- **9 UPenn MOOCs** on various topics
- Remove posts that were only: website links, math formulas…
- Cleaned dataset: **3,503 posts** from **2,882 students**
- Each data point: **student ID**, **timestamp**, **post text**
- Students' personal information was redacted
  - Dataset available for download!
  - https://github.com/pcla-code/forum-posts-urgency

Methods: Data Labeling

- **3 human coders**
- Completed coder training and followed a pre-defined protocol
- **Manually labeled** the 3,503 post texts

Methods: Data Labeling

- 1: **No reason to read** the post
- 2: **Not actionable**, read if time
- 3: **Not actionable**, may be interesting
- 4: **Neutral**, respond if spare time
- 5: **Somewhat urgent**, good idea to reply, a TA might suffice
- 6: **Very urgent**: good idea for the instructor to reply
- 7: **Extremely urgent**: instructor definitely needs to reply

Methods: Data Labeling

Ex. label 1: “Hi my name is [REDACTED], looking forward to this course!”

Ex. label 5: “When will the next quiz be released? I'd like to get a head start on it since I've got some extra time these days.”

Ex. label 7: “The website is down, [link] seems down and I'm not able to submit the Midterm. Are the servers congested?”

Methods: Data Automated Pre-Processing

- Convert all text to **lowercase**
- Replace all characters, except a-z | 0-9, with spaces
- Remove duplicate whitespace
- Remove common **stopwords** in the English language
- Perform **stemming**

Methods: Model Training

Multi-class classification and regression

- To capture the ordering on the 1–7 scale
- **6 models**: Random Forest (RF), XGBoost (XGB), Linear Regression, Ordinal Regression, SVM Regression, Neural Network Regression
- **3 types of features**: Bag of Words (BoW), TF-IDF, USE
- **Cross-validation**: student-level 10-fold
- **Metrics**: RMSE, Spearman rho

Methods: Model Training

Binary classification

- To compare with related work
- 1–4 → not urgent, 4.5–7 → urgent
- **3 models**: RF, XGB, NN
- **3 types of features**: BoW, TF-IDF, USE
- **Cross-validation**: student-level 10-fold
- **Metrics**: AUC ROC, F1

Methods: Model Evaluation

- **Separate test set**: Stanford MOOCPosts
- Publicly available data set
- 29,603 posts

Results: Word Count Features

Multi-class classification and regression

- **TF-IDF** slightly better than BoW
- **SVReg** best on both data sets (NN next):

  Training/CV: **RMSE** 1.09, **Spearman** 0.55

  Test: **RMSE** 1.41, **Spearman** 0.40

Results: Word Count Features

Binary classification

- **NN** best on both data sets (XGB next):

  Training/CV: **AUC** 0.67, **F1** 0.91

  Test: **AUC** 0.57, **F1** 0.78

Results: USE Embeddings as Features

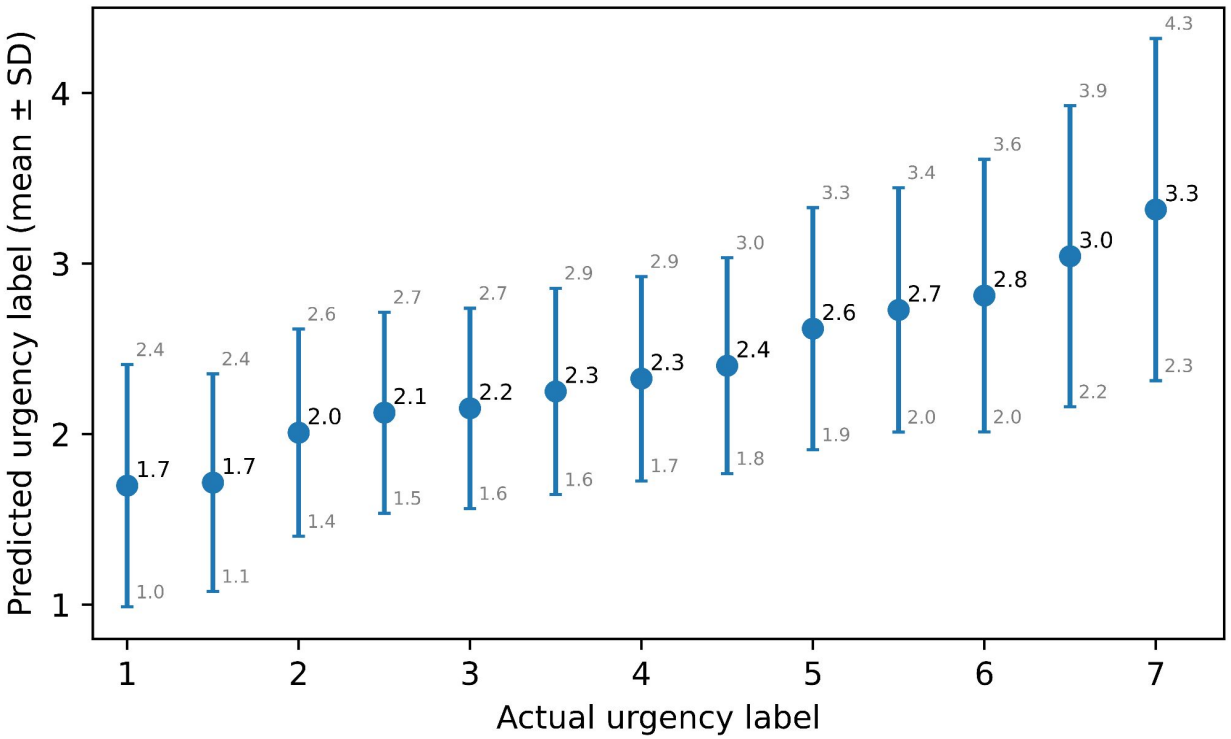
Multi-class classification and regression

- Very slightly better than word count models
- **SVReg** best on both data sets (NN next):

  Training/CV: **RMSE** 1.10, **Spearman** 0.57

  Test: **RMSE** 1.41, **Spearman** 0.43

Results: USE Embeddings as Features



Results: USE Embeddings as Features

Binary classification

- Similar to word count features
- **NN** best on both data sets (XGB next):

  Training/CV: **AUC** 0.64, **F1** 0.91

  Test: **AUC** 0.57, **F1** 0.78

Conclusion

- **Automatically determine the urgency** of forum posts
- Two data sets with different distributions – demonstrates **generalizability** of models
- **SVReg models with USE embeddings** best overall
- Comparable or slightly better performance than past work

Conclusion

- **Model quality**: unlikely that an urgent post will be labeled non-urgent and vice versa
- **Contribution**: support learners by providing feedback to instructors in large courses
- **Application**: integrate into MOOC platforms to provide automated notification on urgent posts

Learn More

Penn Center for Learning Analytics
University of Pennsylvania

Ryan Baker

twitter.com/BakerEDMLab

Baker EDM Lab

"Big Data and Education", running on edX now
Online Masters in Learning Analytics
Computer-Based Learning in Context
All lab publications available online – Google "Ryan Baker"