EduData: Rubric-Guided Topic Classification and Visualization of Student Course Evaluations

By: Sergio Zavala | Advised By: Dr. Cao Thang Bui

School of Computing and Design

Abstract

Large course evaluations often contain hundreds of open-ended comments, making it difficult for instructors to identify major themes efficiently. Prior NLP and LLM-based reporting methods can produce repetitive summaries, over-classify generic praise, and lack a structured rubric aligned with faculty needs.

We present a rubric-guided, multi-label NLP framework and interactive dashboard for analyzing course evaluation comments. Our approach organizes feedback into interpretable instructional categories and supports transparent exploration of model predictions, helping instructors review trends, inspect individual comments, and better understand student feedback at scale.

Objective: Evaluate whether rubric-guided NLP can classify and score course-evaluation comments in a way that is accurate, interpretable, and useful for instructors.

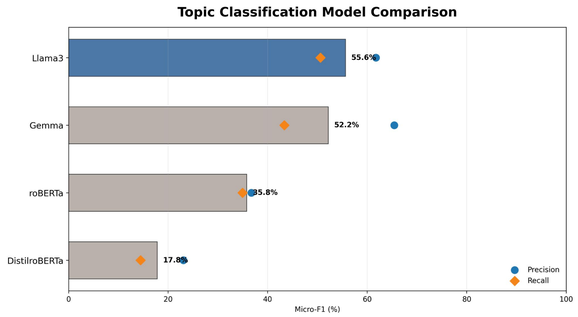
Methods

We analyzed 102 open-ended student course-evaluation comments using a two-stage NLP pipeline.

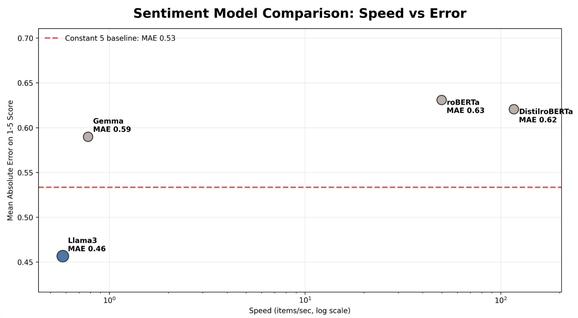
1. **Classification:** Each comment was classified into one or more instructional feedback categories, including course organization, pace, workload, clarity, assessment, communication, inclusivity, and learning resources. Generic praise without actionable detail was assigned to an “Other” category.
2. **Sentiment:** Each topic-comment pair was scored using a topic-specific 1-5 rubric. Unlike generic sentiment analysis, the rubric captures instructional meaning within each category; for example, pace and workload are scored based on whether they support learning rather than simply whether the language is positive or negative.

We compared local Llama3 and Gemma LLM models along with RoBERTa and DistilRoBERTa transformer models. Topic classification was evaluated against manually annotated reference labels using micro-F1, precision, recall, and exact set match. Rubric scoring was evaluated against manually scored reference scores using score accuracy, within-one accuracy, mean absolute error, macro-F1, and processing speed.

Model Comparison



Topic Classification: Llama3 had the strongest micro-F1, suggesting LLMs handle multi-label instructional themes better than zero-shot transformer baselines.



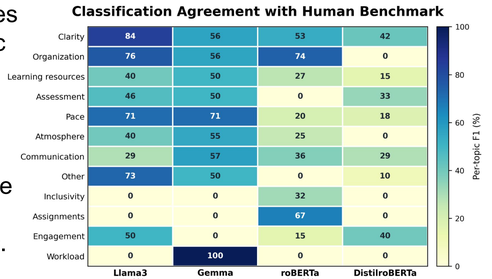
Rubric Scoring: Llama3 had the lowest MAE, while DistilRoBERTa processed comments fastest but with higher scoring error.

Dashboard Prototype

Dashboard prototype will display course-level summaries, topic scores, model confidence, and drill-down access to individual comments. The goal is to help instructors move from aggregate trends to inspectable evidence.

Classification Examples

The heatmap compares each model’s per-topic F1 score against the human benchmark. Llama3 showed the strongest overall agreement, while some categories remained difficult across models.



Conclusion

Rubric-guided NLP can turn open-ended course evaluations into structured topic labels, scores, and summaries that instructors can inspect. Llama3 achieved the strongest results, while DistilRoBERTa was fastest but less accurate.

Future work will require a larger human-labeled dataset across more courses. This would make it possible to fine-tune faster transformer models, improving accuracy while keeping the system practical for large-scale evaluation analysis.

Acknowledgement

I am extremely grateful for my advisor Dr. Cao Thang Bui for their endless support on the project and beyond the project. I would also like to thank to Dr. Roshini Ramachandran and Harsh Manohar for assisting me on this project.

References

Medina, M. S., Smith, W. T., Kolluru, S., Sheaffer, E. A., & DiVall, M. (2019). A Review of Strategies for Designing, Administering, and Using Student Ratings of Instruction. American Journal of Pharmaceutical Education, 83(5), 7177. https://doi.org/10.5688/ajpe7177

Poucke, M. V. (2025). Appraising Feedback Stance in Higher Education: A Corpus-Assisted Discourse Study of Student and Academic Perceptions, Perspectives and Preferences. Corpus Pragmatics, 9(3), 337–366. https://doi.org/10.1007/s41701-025-00196-3

Sun, J., & Yan, L. (2023). Using topic modeling to understand comments in student evaluations of teaching. Discover Education, 2(1), 25. https://doi.org/10.1007/s44217-023-00051-0

Zhang, M., Lindsay, E. D., Thorbensen, F. B., Poulsen, D. B., & Bjerva, J. (2024). Leveraging Large Language Models for Actionable Course Evaluation Student Feedback to Lecturers (arXiv:2407.01274). arXiv. https://doi.org/10.48550/arXiv.2407.01274