The Journal of Open Source Education

AI and Teacher Performance Evaluations: A Tutorial with the Gemini API and Python

Eric Silberberg[1]

1 Assistant Professor, Librarian for Instructional Design and Education, Queens College, City University of New York

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Submitted: 24 January 2025
Published: unpublished

Summary

With a growing number of AI applications claiming to improve education, this article presents a hands-on tutorial designed to equip pre- and in-service teachers with the experience of building an AI tool. Participants develop a Python script that uses the Google Gemini 1.5 Flash language model to conduct sentiment analysis on a set of student feedback data. After building and interacting with the AI tool, the tutorial prompts participants to draw from their experience in the tutorial to evaluate the potential benefits and weaknesses in the application of AI to Teacher Performance Evaluations.

Statement of Need

Teacher performance evaluations (TPEs) in U.S. public schools are used by school administrators to coach teachers in refining their practice as well as to identify high performing teachers for promotions, salary increases, or tenure (National Center for Education Statistics, 2020). With the advent of enterprise AI systems, some schools have begun investing in AI tools to assist with TPEs. One proposed application is AI-powered sentiment analysis of teacher instruction. Concerns arose due to the lack of transparency surrounding the underlying AI models, which could introduce bias in sentiment analysis. This is particularly concerning with respect to aspects of communication that may be challenging for language models to accurately assess, such as accent, dialect, register, and humor (Elsen-Rooney, 2024; Langreo, 2023). However, outside of education, some research suggests that, in fact, many employees perceive AI as less biased than human evaluators during performance reviews (Brown et al., 2024).

The tutorial "Evaluating AI in Teacher Performance Reviews: Benefits, Biases, and Best Practices" aims to pull back the curtain on the impact of AI on TPEs. It gives pre- and in-service teachers the opportunity to develop their own AI-powered sentiment analysis tool, thus fostering a deeper understanding of these technologies and their potential impact on education. Teachers who are exposed to professional development with information and communication technologies demonstrate higher levels of empowerment and innovation (Yipeng, 2021). Thus, this hands-on tutorial seeks to equip them to critically evaluate changes to TPEs and enable them to use AI tools for their own creative solutions to problems they face in the classroom.

Learning Objectives

In the tutorial, students work through three exercises leading them to create a Python script using the Google Gemini 1.5 Flash language model that analyzes a collection of student feedback. By the end of the tutorial, students will be able to:

- Call the Gemini API in a Python function that conducts sentiment analysis on narrative student feedback as it's input.
- Refine this function's output by way of prompt engineering.

- Critique the use of AI in TPEs by pointing to their experiences in the classroom and completing the tutorial.

Course Content

The tutorial guides participants through the development of a Python script for sentiment analysis of TPEs using Gemini 1.5 Flash, a powerful (but more importantly free) language model accessible via an API and Python library. Designed for those with basic Python knowledge, the tutorial is structured around three hands-on exercises in which participants build a script that conducts sentiment analysis of student evaluations of teaching. It concludes with participants critically examining AI-assisted TPEs. The tutorial unfolds within `index.html`, which provides data files, links to relevant Python documentation, and code snippets stylized with Highlight.js.

The tutorial begins by framing the issues around AI-assisted TPEs in a similar fashion to the statement of need in this article. The participants are then prompted to obtain a Gemini API and install the Google AI Python SDK.

Practical exercises

The first exercise walks participants through writing a function that uses Gemini to create an anagram of a person's name. While not related to TPEs, the purpose is to familiarize participants with how to configure Gemini and integrate it into a function. This first exercise also asks participants to explore the response object that Gemini returns. This object contains both the text generated (anagram) and metadata about the generation. Unfortunately, the Gemini team continues to change the structure of this object's metadata, but the most interesting part is to explore the safety filter ratings (e.g. probability of harassment and hate speech generation).

The second exercise imagines a statistics course where students have submitted end of year feedback and introduces the idea that Gemini can rate the sentiment of this student feedback. The participants write a function that applies a Likert scale to the feedback, which is entirely narrative and not structured. The Likert scale ranges from 1, representing Very Negative, to 5, representing Very Positive sentiment. When the participants run a first iteration of the function, Gemini returns a numeric rating plus several lines of text to justify its response. This is useful to see whether Gemini understands the task. However, with the goal being to create structured data, the participants see how prompt engineering can be used to prompt Gemini to return only a single digit reflective of its Likert rating.

The third exercise builds on the previous one and asks participants to apply the sentiment analysis function to a set of 10 student reviews, which the tutorial provides in a downloadable spreadsheet. Participants use rudimentary aspects of pandas to apply the function to the set of student feedback and scrub the data: ensure that all outputs from Gemini are integers and there is no leading or trailing white spaces. This is necessary for the last part of the exercise where participants use pandas to find a simple average and interquartile range of student sentiment with respect to the statistics course.

Evaluation

The last section of the tutorial asks participants to reflect on the implications of bringing AI to bear on TPEs. The questions are:

- What are the potential benefits, limitations, and/or drawbacks of using AI sentiment analysis in teacher evaluations?
- Considering the limitations you may have encountered with the AI tool, propose one best practice for ensuring fairer and more reliable results when using AI for teacher performance evaluation.
- Based on your experience, do you think AI sentiment analysis could be a valuable tool in teacher evaluation, even with its limitations? Why or why not?

Experience of Use in Teaching and Learning Situations

I developed the tutorial while a fellow in the Building Bridges of Knowledge project, sponsored by the Lumina Foundation and the City University of New York. This project supported faculty in integrating ethical, responsible, and creative uses of AI into course materials.

I facilitated the workshop twice as part of Queens College Library's fall 2024 Data Services Workshop Series. Offered as a hybrid (online and in person) workshop, registration for these workshops was open to the entire college community. Outreach is a major part of librarianship, and this workshop was special because it enabled the Library to reach students that we do not traditionally see in our library instruction program. Students and faculty from the business, computer science, economics, education, and sociology departments attended the workshops.

Participant feedback (collected in an end-of-workshop survey) indicated that they appreciated the pacing of the modules and that the tutorial inspired them to pursue personal projects building on what they had learned. No sentiment analysis was conducted on participant feedback. Based on my experience, it is recommended to allot 2 hours to facilitate the workshop.

The workshops also proved to be an opportunity for students interested in building applications with an AI component to connect with each other. One student was looking to develop an app that generates cooking recipes, and she asked me a question about database management (something outside of my wheelhouse). However, a fellow student who had worked on a similar problem in the past, offered to connect and provide advice after the workshop.

Although I had sent several reminders before the workshop, urging registrants to complete module 2.1 (which provides instructions for installing the Gemini Python library and generating their API keys), I underestimated the number of students who would disregard these directions. This unfortunately led to delays during the first workshop. I anticipated this issue for the second workshop and instructed attendees to open module 2.1 upon arrival to proactively prevent delay.

The tutorials were also designed to be completed as self-directed learning. In fact, several workshop registrants who were ultimately unable to attend contacted me in the days following the workshop to see if I could share a recording or workshop notes. It was beneficial to have the full tutorial prepared and designed for independent study. Some students subsequently reached out while working on the tutorial with questions, to which I responded.

Acknowledgements

The tutorial "Evaluating AI in Teacher Performance Reviews: Benefits, Biases, and Best Practices" was made possible with support from the CUNY Office of Faculty Affairs' Building Bridges of Knowledge project and the Lumina Foundation.

References

Brown, J., Burke, J., & Sauciuc, A. (2024). Using artificial intelligence to evaluate employees: The effects on recruitment, effort, and retention. *Kelley School of Business Research Paper*, *2021-25*. https://doi.org/10.2139/ssrn.3861906

Elsen-Rooney, M. (2024, January 2). Can artificial intelligence help teachers improve? A network of NYC schools wants to find out. *Chalkbeat*. https://www.chalkbeat.org/newyork/2024/01/02/schools-to-use-artificial-intelligence-to-help-coach-teachers/

Langreo, L. (2023, May 10). Can AI do teacher observations and deliver PD? In some schools, it already does. *Education Week*. https://www.edweek.org/technology/can-ai-do-teacher-observations-and-deliver-pd-in-some-schools-it-already-does/2023/05

National Center for Education Statistics. (2020). *Teacher performance evaluations in U.S. public schools*. https://nces.ed.gov/pubs2020/2020133.pdf

Yipeng, T. (2021). Does information and communication technology (ICT) empower teacher innovativeness: A multilevel, multisite analysis. *Educational Technology Research and Development*, *69*. https://doi.org/10.1007/s11423-021-10052-1