

Predicting Professor Review Ratings Using DistilBERT

A Natural Language Processing Approach Using PlanetTerp Reviews

The Problem

Goal:

Automatically predict a professor's **1–5 star rating** from the written text of their student reviews.

Why it matters:

- Helps students quickly evaluate courses
- Helps departments monitor teaching performance trends
- Demonstrates how transformer models interpret sentiment in long text

Challenge:

Ratings are subjective, noisy, and vary widely across reviewers.

Data Collection (PlanetTerp API)

Pulled hundreds of reviews for five UMD professors:

- Nelson Padua-Perez
- Timothy Pilachowski
- Pedram Sadeghian
- Stefan Doboszczak
- Jonathan Fernandes

For each review we collected

- Review text
- Numerical rating (1–5)
- Course and timestamp

Cleaned the dataset by:

- Removing empty reviews
- Standardizing rating format
- Filtering out missing values

Model Approach

I fine-tuned **DistilBERT**, a lightweight transformer model, to classify reviews into 5 rating levels.

Steps:

1. Tokenized review text using DistilBERT tokenizer
2. Fine-tuned a sequence classification head for 5 labels
3. Split data into training/testing sets
4. Evaluated performance using:
 - Accuracy
 - Mean Absolute Error
 - Confusion Matrix

Why DistilBERT?

- Fast to train
- Strong accuracy
- Lower GPU cost

Training Setup

- **Model:** DistilBERT-base-uncased
- **Classification head:** 5-way softmax
- **Loss function:** Cross-Entropy
- **Epochs:** 3
- **Batch size:** 8
- **Optimizer:** AdamW
- **Platform:** Google Colab

Training took only a few minutes due to the smaller transformer size.

Results

Accuracy: 0.73

Mean Absolute Error: 0.50

General observations:

- The model captures strong sentiment very well (1-star and 5-star reviews are the easiest).
- Most confusion occurs between **4 vs. 5** and **2 vs. 3**, which usually have similar language.
- The model performs well despite subjective, noisy review text.
- The fine-tuned DistilBERT classifier achieves solid performance with limited training data.

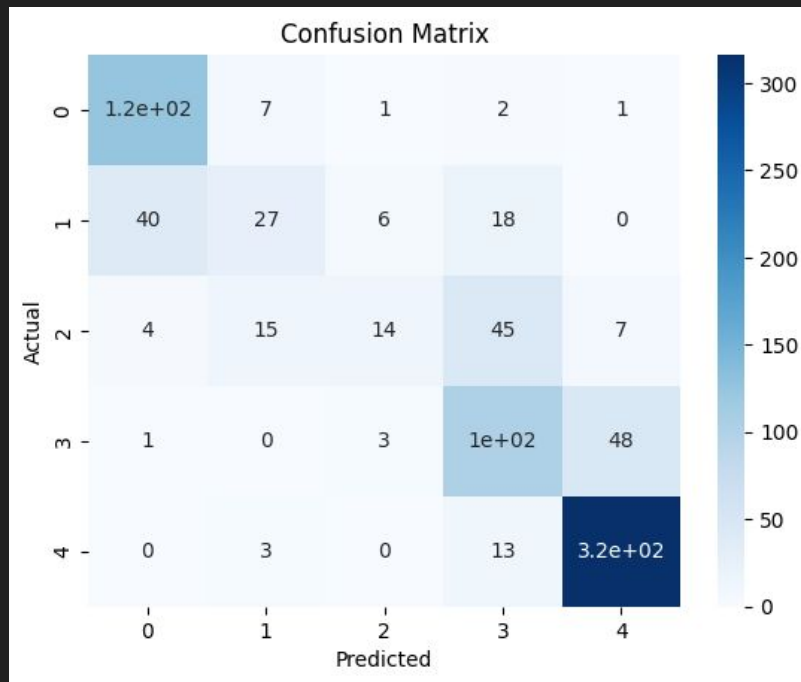
Confusion Matrix

Confusion Matrix (Model Predictions vs. True Ratings)

The confusion matrix on the right shows how often the model correctly predicted each rating from 1 to 5 stars. Darker squares represent higher counts of predictions.

Key insights:

- The model is strongest on **extreme ratings** (1-star and 5-star reviews).
- Most mistakes happen between **neighboring ratings** (2 vs. 3, and 4 vs. 5).
- Middle-range ratings are harder because they use similar wording in reviews.
- Overall, the diagonal pattern confirms the model is learning meaningful sentiment structure.



Example Prediction

Review Text	True	Predicted
“Great professor, very helpful and clear.”	5	5
“Lectures were boring and confusing	2	2
“Challenging but fair”	4	3

Conclusion

What worked well:

- DistilBERT effectively models sentiment from long text
- Training was fast and stable
- Predictions closely match human-written ratings

Limitations:

- Reviews are highly subjective
- Some mid-range ratings (2–4) look similar linguistically
- More data would improve model performance