Analysing Data assignment 3 part 2

Rachel Hartemink

Important note:
Unfortunately, I experienced a lot of problems with running ollama. The current notebook works on Google Colab, but I had to reduce the dataframe to just the first 15 rows. I tried running the whole dataframe and different subsets, but unfortunately, it kept crashing continuously with more than 15 rows. This, of course, also majorly affects the precision, recall, f1 score and this analysis. Only 4 genres were present in the subset of the dataset I used. Due to the extreme reduction to just 15 classifications, the results should not be seen as representative of the model's actual capabilities on a full dataset.

Discussion:

Overall metrics

The zero-shot approach achieves slightly higher precision (0.35 vs. 0.29) and recall (0.28 vs. 0.17) than the few-shot method. However, both strategies have relatively low F1 scores (0.23 and 0.19, respectively), indicating that these models struggle with balancing precision and recall. The few-shot approach does not significantly improve overall performance, suggesting that limited labeled examples do not provide substantial help. It can also be the case that either the dataset is too small or the model struggles to generalize from the few-shot examples.
The struggle with classifying lyrics is probably also due to the ambiguous nature of song lyrics. Lyrics can be poetic, metaphorical or vague and can often fit multiple genres. This makes it hard to categorise them, even for humans. The model might perform better if the prompt was even more detailed, with multiple examples per genre.

Per-class performance

Zero-shot: Electronic and hip-hop are predicted well, with relatively high precision and recall. Pop has perfect recall but low precision (many false positives). Rock is predicted with high precision (1.00) but very low recall (0.11), meaning the model predicts it rarely but with confidence.

Few-shot: Hip-Hop is perfectly classified (1.00 precision, 1.00 recall), but this could indicate overfitting to a small sample. Rock maintains high precision but struggles with recall, similar to zero-shot. Electronic and Pop receive no correct predictions, likely due to class imbalance or difficulty distinguishing their features.