

Oregon State University

Analysis of Emotion Prediction in Human and LLM Annotated Datasets

Isaac Brickner

Project Partner: Dr. Patrick Donnelly

Oregon State University
College of Engineering

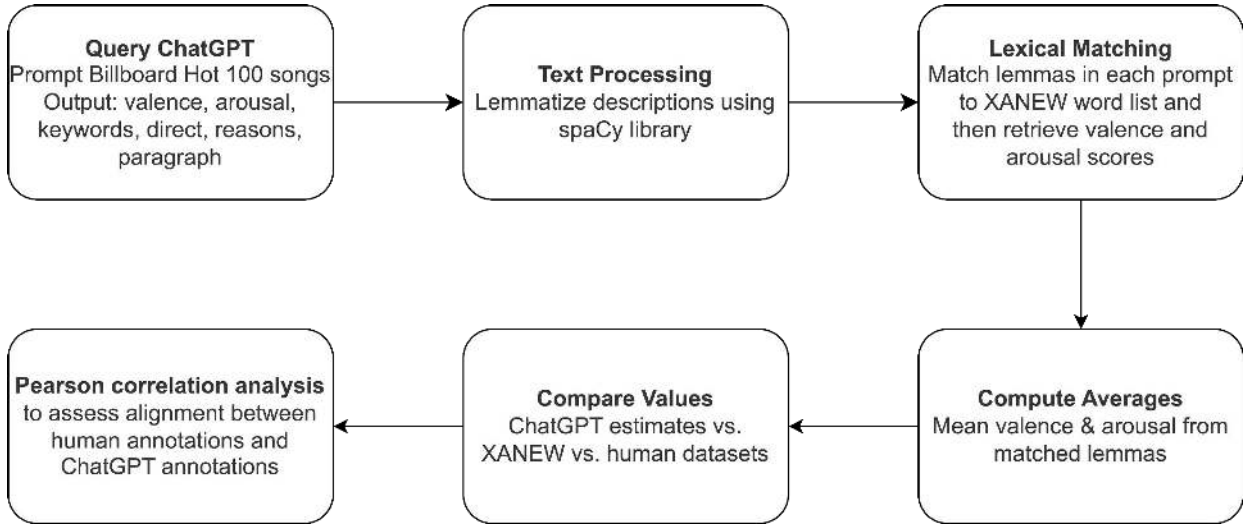
Background and Goals

- Music emotion recognition is subjective and challenging.
- Traditional datasets rely on human annotators, but agreement is often low.
- Exploring if LLMs (like GPT-4o-mini) can reliably annotate emotion in music datasets.

Oregon State University
College of Engineering

Methods

- Structured prompts to GPT-4o-mini via OpenAI API:
- Direct numeric (valence/arousal) Five emotion words
- Descriptive paragraph
- Processed responses with Python, pandas, spaCy.
  - Mapped responses using XANEW wordlist for standardized scoring.





Datasets Compared

AMG1608: Large, well-annotated, human-labeled.

Deezer: Synthetically annotated dataset.

PMEmo: Annotated by Chinese researchers, Western pop.

DEAM: Older dataset, smaller sample.

Oregon State University
College of Engineering

Results

AMG1608: Highest (but moderate) correlations with LLM annotations.

Deezer: Highest correlation for paragraph-based method.

PMEmo: Low correlation, likely due to language/cultural mismatch.

DEAM: Low correlations

Human-to-human correlations also low, showing subjectivity.

Oregon State University
College of Engineering

Figure: Pearson Correlation

Pearson's r	AMG1608	Deezer	PMEmo	DEAM
Valence *direct vs human*	0.6299	0.4333	-0.0213	0.0881
Arousal *direct vs human*	0.6746	0.3002	0.0114	0.2060
Valence *5 words vs human*	0.5397	0.3401	-0.0202	0.0373
Arousal *5 words vs human*	0.4973	0.2437	0.0096	0.2188
Valence *paragraph vs human*	0.5469	0.3625	-0.0616	0.0603
Arousal *paragraph vs human*	0.4428	0.2424	-0.0561	0.1704

Oregon State University
College of Engineering

Figure: Spearman Correlation

Spearman's ρ	AMG1608	Deezer	PMEmo	DEAM
Valence *direct vs human*	0.6366	0.4341	-0.0090	0.0848
Arousal *direct vs human*	0.6872	0.2755	0.0230	0.1948
Valence *5 words vs human*	0.5569	0.3465	-0.0168	0.0637
Arousal *5 words vs human*	0.5125	0.2287	0.0163	0.2125
Valence *paragraph vs human*	0.5617	0.3634	-0.0621	0.0651
Arousal *paragraph vs human*	0.4852	0.2381	-0.0387	0.1617

Oregon State University
College of Engineering

Discussion

- LLM emotion predictions moderately align with humans for large, high-quality datasets.
- Correlations are weaker on smaller or older datasets.
- Future work will include a user study to further validate the results.
- Overall, LLMs show promising potential for automatic emotion annotation of music datasets.