Analysis of Emotion Prediction in Human and LLM Annotated Datasets

soundbendor lab

Isaac Brickner & Patrick Donnelly

Oregon State University-Cascades

Oregon State University

Abstract

This study investigates ChatGPT's ability to recognize emotions in music. Using the list of Billboard Top 100 songs (1958–present) and four human-annotated datasets, songs were analyzed via three different ChatGPT prompts to extract emotional valence and arousal. Outputs were lemmatized and mapped to standardized emotional ratings. Results were compared against human annotations to evaluate ChatGPT's accuracy and consistency. Findings suggest that ChatGPT may serve as a scalable tool for emotion annotation in music, with implications for music information retrieval and automatic affective computing.

Background

Music is a deeply emotional form of expression, and researchers have long used human raters and surveys to measure its affective qualities using psychological models like the Valence-Arousal circumplex. With large language models, like ChatGPT, we have the opportunity to explore automated emotion annotation at large scale. This project examines whether ChatGPT can reliably estimate the emotional valence and arousal of popular songs and see how well its predictions align with human annotations—while also revealing long-term trends of emotion responses to American popular music.

Methods

- Developed structured prompts to elicit valence, arousal, and descriptive emotional content from `GPT-4o-mini` via the OpenAI API using three questioning paradigms:
  1. direct (numeric) estimate
  2. five descriptive words
  3. descriptive paragraph
- Utilized Python with `pandas` and `spaCy` to lemmatize and aggregate responses, mapping individual terms using the `XANEW` wordlist [1] for standardized emotion scoring.

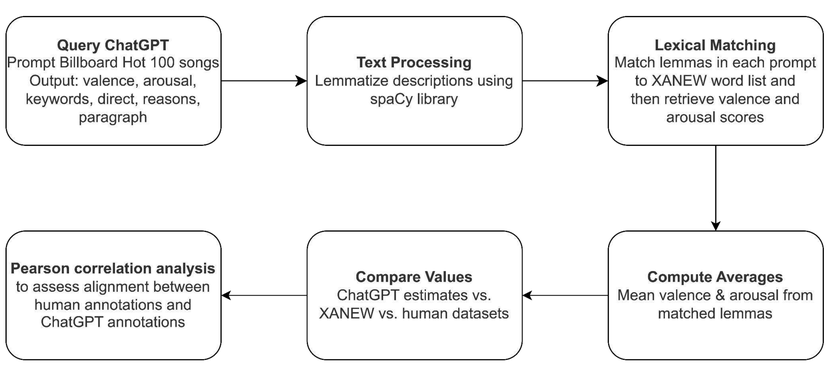

Figure 1: System diagram describing the steps of the process.

Datasets

- **AMG1608** (*n=1608*) gives high quality human-annotated valence and arousal ratings based on audio and lyrics.
- **Deezer Mood Dataset** (*n=18,644*) provides metadata, audio features, and algorithmically generated mood tags.
- **PMEmo** (*n=794*) contains static and dynamic annotations derived from noisy crowd-sourced listener ratings.
- **DEAM** (*n=1744*) includes valence/arousal annotations across an obscure set of older, copyright free music excerpts, along with comprehensive set of audio features.

Table 1: Pearson's correlation with humans annotations.

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

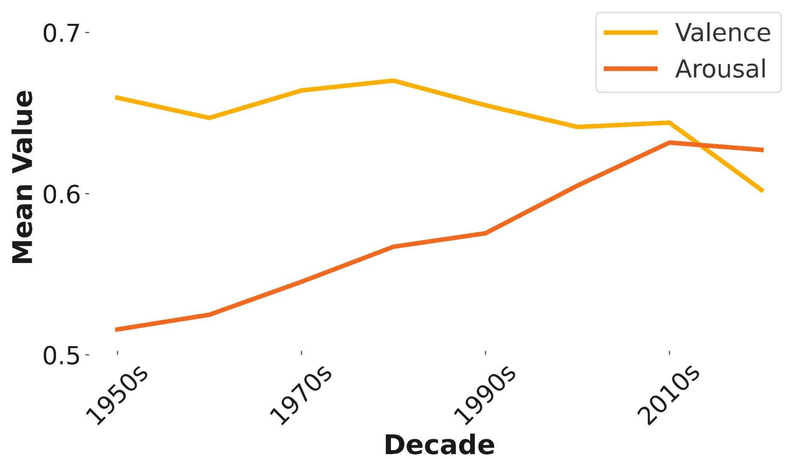

Figure 2: Trends of valence and arousal by decade.

Comparison Analysis

- AMG1608 showed the best results, compared to other datasets, but still only achieved moderate correlations.
- Deezer had the highest correlation with the lemmatized paragraph, however it is a synthetically annotated dataset.
- PMEmo's low results may stem from language differences from Chinese researchers interpreting Western pop music.
- Human correlations between shared songs in each of the datasets is also low; this disagreement further emphasizes the subjectivity of music emotion and difficulty of this task.

Table 2: Spearman's correlation with human annotations.

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

Discussion

- Our approach achieved moderate correlations between ChatGPT responses and human annotators for large and well-annotated datasets (AMG1608 and Deezer*).
- However, our approach found weak correlations with smaller datasets and older songs (PMEmo and DEAM).
- As future work, we will conduct a user annotation study to better validate the results of our automatic process.
- Overall, this experiment demonstrates the potential of LLM models to automatically annotate datasets with emotion.

References

[1] Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods, 45*(4), 1191–1207.