Evaluating Large Language Models for Music Emotion Prediction from Social Media Discourse

Aidan Beery Advisor: Dr. Patrick Donnelly

Abstract

The task of music emotion recognition has been of interest in the music information retrieval domain. We investigate the use of online social media discussions as a potential input for music emotion prediction models. Using this commentary, we evaulate the performance of two pre-trained large language models in predicting the valence and arousal of a piece of music. We query three social media platforms to build a corpus of conversations surrounding 2,402 songs. We achieve modest Pearson's correlations of 0.62 and 0.72 to valence and arousal targets respectively. These results demonstrate that there may be a connection between the sentiment expressed in the online discourse around a song and a listener's affective response to said song.

Introduction

Music emotion recognition (MER) is the application of computational methods to the understanding of the emotions a listener may experience when listening to a song. Historically, this task has relied on manual surveys using crowdsourced annotators [1]. The assumption is that a cultural average response can be derived from individual emotion annotations. However, manual annotation is both expensive and time consuming. The subjective nature of the perceived emotive qualities of a piece of music make it such that very large sample sizes are necessary to achieve statistically significant results. As a result, there has been a lack of available datasets for dimensional music emotion analysis, impeding research in the music information retrieval domain.

Figure 1. The 100 most frequently occurring words in our social media dataset

An automated system for estimating music emotion would enable large music libraries to be rated for affective response, aiding the development of music recommender algorithms as well as expediting future MER research. Our hypothesis is that the online conversations surrounding a song contain semantic information, which can be used to predict a song's emotive qualities.

We present a framework for music emotion recognition using pre-trained transformer models to estimate a song's valence and arousal directly from its social media commentary. We create a dataset of musical discourse from conversations aggregated from Reddit, YouTube, and Twitter, consisting of discussions in reference to songs included in two existing music emotion datasets. We then compare two large language models (LLMs), popular in natural language understanding tasks, for the prediction of a song's affective features from only social media discourse.

Datasets

For our experiment, we choose two music emotion datasets annotated for valence and arousal. The AMG1608 dataset consists of 1,608 songs selected from the All Music Guide [1]. A subset of the "Western Contemporary" genre was selected to create a dataset with an even distribution across the valence-arousal space. Annotations were gathered using the Amazon Mechanical Turk crowdsourcing platform, in all yielding 665 unique annotators. The PmEmo dataset consists of 794 songs identified from the Billboard Hot 100 and other record industry charts between 2016 and 2017 [5]. These samples were then rated by 457 undergraduate students from the authors' institution in a lab setting. In total, we collect data for 2,402 songs.

Dataset	Songs	Label Type	Scaling
AMG1608	1608	Crowdsourced	$[-1, 1]$
PmEmo	794	Lab Study	$[\ 0, 1\ ]$

Table 1. Comparison of the music emotion datasets

Methods

We query three social media platforms: Reddit, YouTube, and Twitter, for conversations explicitly mentioning both artist name and track title. For Reddit and YouTube, the ten highest rated posts matching each song were selected, and all comments in response to these posts were gathered. For Twitter, we scrape the top 100 matching tweets. We retrieve comment data for 95% of our dataset's songs from YouTube, 86% from Reddit, and 43% from Twitter.

We choose two popular pre-trained transformer models: DistilBERT [4] and RoBERTa [3]. Each input to the model consists of one text comment, labeled by its associated song valence and arousal values. Labels are scaled to $[0, 1]$ for inter-dataset consistency. We create word embeddings using HuggingFace's `TokenizerFast` library, with a vector length of 128 tokens. For each estimator, we fine tune the base language model on social media text, then take the final hidden layer and append a fully-connected neural network to perform multi-target regression. All resulting comment-level predictions for a song are then averaged to generate a song-level prediction.

Results

We begin with a comparison of prediction performance between our two models. Both LLMs were fine-tuned for two epochs on the downstream task. We find no statistically significant difference in Pearson's correlation between the target and either model's predictions. However, RoBERTa proves to be significantly more computationally expensive to train, as a result of having more than five times as many parameters as DistilBERT. As a result, we choose to use DistilBERT for future experiments.

		DistilBERT	RoBERTa
AMG1608	Valence	0.49	0.51
	Arousal	0.64	0.63
PmEmo	Valence	0.72	0.71
	Arousal	0.64	0.64

Table 2. Comparison of DistilBERT and RoBERTa performance, measured in Pearson's correlation

AMG1608

	Reddit	Twitter	YouTube	All
Valence	0.32	0.23	0.62	0.49
Arousal	0.56	0.34	0.72	0.64

PMEmo

	Reddit	Twitter	YouTube	All
Valence	0.56	0.26	0.68	0.72
Arousal	0.60	0.16	0.52	0.66

Table 3. Results of DistilBERT trained for two epochs for each social media source and song list

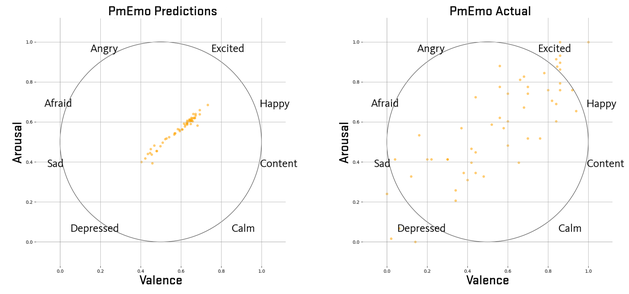

Figure 2. Distribution of DistilBERT predictions on songs in our test set for PmEmo

We split our dataset into three constituent subsets for each social media source to compare the importance of each source to the model's predictive ability. We find that models trained on YouTube comments exclusively tend to outperform any other single-source model. Overall, we achieve best correlations of 0.62 and 0.72 to valence and arousal on the AMG1608 dataset using a YouTube model, and 0.72 and 0.66 on PmEmo with a joint model.

Discussion

We demonstrate the feasibility of using social media conversations about a song to directly estimate the emotional qualities of that song using pre-trained large language models. As opposed to existing methods, our approach does not rely on access to audio samples, nor expensive annotation experiments. However, social media data can only be reliably gathered for currently released music. Data may be sparse in the case of niche genres, resulting in inadequate performance. Furthermore, by averaging together the estimated song valence and arousal at each comment, we risk losing the semantic information embedded in comments of conflicting sentiment.

In the future, we plan to explore alternative model architectures to better encode the relationship contained between comments of the same song. We also intend to augment our current dataset with additional data, both by scraping existing sources more thoroughly as well as including new platforms such as SoundCloud. A music emotion prediction model can accelerate research in music information retrieval, improve music recommender algorithms, and enable mood-dynamic playlist generation at scale [2]. To our knowledge, this is the first attempt to directly predict valence and arousal of musical samples using only social media discourse.

References

[1] Yu-An Chen et al. "The AMG1608 dataset for music emotion recognition". en. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. South Brisbane, Queensland, Australia: IEEE, Apr. 2015, pp. 693–697.

[2] Patrick Donnelly and Shaurya Gaur. "Mood Dynamic Playlist: Interpolating a musical path between emotions using a KNN algorithm". International Conference on Human Interaction and Emerging Technologies. Lausanne, Switzerland, 2022.

[3] Yinhan Liu et al. "RoBERTa: A Robustly Optimized BERT Pretraining Approach". (July 2019). arXiv:1907.11692 [cs].

[4] Victor Sanh et al. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter". (Oct. 2019).

[5] Kejun Zhang et al. "The PMEmo Dataset for Music Emotion Recognition". en. *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*. Yokohama Japan: ACM, June 2018, pp. 135–142.