Learning Affective Responses to Music from Social Media Discourse

Aidan Beery and Patrick J. Donnelly

Abstract The ability to automatically estimate typical affective responses to music would enable the development of emotion-aware music recommendation systems. However, the lack of suitable datasets for this task has hindered attempts to design such systems. In this work, we introduce social media conversational data as a new feature space for music emotion recognition. We create a large dataset of social media musical discourse with over 11.8 million comments from Reddit and YouTube discussing 19,627 different songs. We fine tune large language models on this conversational data in a two-target regression task to predict music valence and arousal annotations. We demonstrate a modest ability to estimate human annotated music emotion targets directly from social media comments. Our highest performing model achieves Pearson's correlations of 0.80 and 0.79 for valence and arousal, respectively. These results imply that emotive qualities of a song may be inferred directly from social media conversations, without access to the audio or lyrics.

1 Introduction

The ability of music to elicit powerful emotional responses in listeners is widely-recognized across human cultures and societies. Although the connection between human emotions and music is not fully understood, the fact that listening to music engages the limbic system [32] hints at the importance of music throughout our biological and cultural evolutions [16]. In recent years, there has been growing interest in the music information retrieval (MIR) community to design computational approaches capable of estimating human emotional responses to music. If researchers studying music emotion recognition (MER) tasks could automatically estimate typical human responses to any piece of music, music recommendation systems would be able to make more emotionally-informed suggestions.

Oregon State University, e-mail: {beerya,donnellp}@oregonstate.edu

Methods for music mood evaluation typically rely on human annotators to each listen to a musical excerpt and provide a subjective rating based on their perceived emotive response. These subject studies are both expensive and time-consuming, requiring many annotators to rate each song to ensure statistically significant sampling. This dirth of annotated data has hindered the advancement of music emotion recognition systems. Furthermore, no standard for emotion modeling has emerged among researchers. Models of affective response often range from mood classification tasks [10] to prediction of continuous valence-arousal values [14].

Despite these challenges, there have been a variety of attempts to predict a song's emotive qualities automatically. Early methods attempted to learn associations with emotion from manually engineered acoustic features [42]. However, such approaches have been insufficient, and researchers have declared a "semantic gap" between low-level acoustic descriptors and the perceptual features observed by human listeners [48]. Furthermore, copyright concerns restrict MIR researchers from distributing audio recordings alongside music emotion datasets, hindering the exploration of audio information as a feature space. Lyrics have been used to augment acoustic feature models [34, 63] and by themselves [26, 2]. However not all music contains lyrics, limiting the generalizability of this approach. Researchers have also explored other modalities, including heart rate [28], electrodermal activity [67], and video of facial expressions [33], often with little success. A few studies have reported limited success with estimating music emotion values from the tags provided by users on online music metadata aggregators such as Last.fm[1] [17, 5, 6]. However, this feature space is relatively small, often only consisting of a few dozen single-word descriptors which members of the Last.fm community deemed relevant to a piece of music.

We hypothesize that the conversations users have about a piece of music might contain semantic clues about typical affective responses to that piece of music. We present a novel approach [20] for learning the continuous valence and arousal values of a song using only the social media conversations referencing that song. To achieve this, we compile a large dataset of social media music discourse using the songlists from four music emotion datasets: AMG1608 [14], PMEmo [67], DEAM [3], and Deezer [17]. We train several large language models to predict music emotion values from these social media comments alone without relying on audio signal analysis or lyrics information. We believe this to be the first approach to estimate the affective qualities of a song solely from social media conversations.

2 Related Work

Music emotion recognition is the task of training computational models to estimate a culturally-average emotive response for a piece of music. The ability to automatically understand the relationship between the audio signal of a piece of music and the anticipated human emotion is of great interest to the field of music information

[1] https://www.last.fm/

retrieval. This is a particularly difficult problem because affective responses about music are subjective and vary both within and between different culturally-entrained groups of listeners. Researchers studying this problem typically train models to estimate an emotional response based on the average of multiple different annotators.

Furthermore, there is a large semantic gap between high level music concepts and low level acoustic features extracted directly from an audio signal. To overcome these difficulties, researchers have explored many different modalities, including descriptive features from audio, music scores, song lyrics, music videos, and even physiological signals monitoring the listener.

Research studies in music emotion recognition typically seek to classify a categorical label for the entire song with classification [36, 27, 34, 65, 40] or to estimate dimensional values with regression [65, 13, 53, 44, 54]. Researchers have also explored probabilistic mappings between categorical and dimensional emotion semantics of music [58]. These predictions are most typically made at the song-level [27, 43, 39, 15, 19, 11] although there is also active research attempting to track dynamic changes in emotion over time [61, 53, 37, 11].

2.1 Acoustic Features

Traditionally, researchers exploring models to recognize music emotion have relied upon low-level features extracted from the audio signal of a song. Many studies rely upon features extracted from common audio toolkits and frameworks, such as PsySound [9], MARSYAS [56], jAudio [46], YAFFe [45], OpenSmile [22], or Essentia [7]. In other cases, the authors craft customized signal processing methods to attempt to capture information from the audio signal that might be useful in attempts to predict human emotional responses to music. Over the years, researchers have explored thousands of features measuring pitch, melody, harmony, rhythm, dynamics, timbre, and expression (see [48] for a review). Using these descriptive audio features to train machine learning models, researchers have explored many different algorithms, such as linear regression [13, 40, 53], support vector machines [36, 34, 65, 27, 23], support vector regressors [65, 53, 61], random forests [34], and Gaussian models [42, 53], and in recent years, deep learning approaches such as autoencoders [12], generative adversarial networks[29], convolutional neural networks [19, 15] and recurrent neural networks [61, 37, 43, 39, 47, 11].

In one early approach, researchers applied support vector machines in an attempt to classify 13 different emotions, using features extracted from 30-second excepts of 499 audio files across 128 different albums covering four genres of music [36]. The authors reported an F_1 score of only 0.41, highlighting the difficulty of the problem. One major limitation of this study is that the emotion labels were all labeled by a single expert listener. In another study, which considered only a single genre, the authors employed several domain experts to manually annotate a dataset of 250 pieces of classical music with one of four emotions: *contentment*, *depression*, *exuberance*, and *anxiety*. After extracting numerous rhythm and timbre features,

the authors applied a Gaussian mixture model to achieve 86.3% accuracy [42]. Acknowledging that annotator fatigue may lead to inconsistent emotion labels, one study designed a music emotion prediction tool to help reduce annotator fatigue in the hopes of yielding more robust datasets [63]. Because emotional reactions to music are highly subjective, another study sought to increase the number of annotations available for each of the 200 songs in their dataset. Crowd-sourcing the task online, the authors collected an average of 28.2 annotations across their dataset of 30-second excerpts of film score soundtracks [62], labeling eight different moods: *sublime*, *sad*, *touching*, *easy*, *light*, *happy*, *exciting*, and *grand*. For this task, the study reported a cosine similarity of 0.73 after training support vector machines with acoustic features. Although the authors extracted a total of 88 features, they reported that they achieved similar efficacy with only the best 29 features.

More recently, Chowdhury et al. investigated the development of mid-level features with the hope of helping to close the semantic gap between low-level audio features and human emotive responses to music [15]. These mid-level features describe perceptual concepts, such as tonal stability, articulation, and rhythm. The authors performed feature-importance analysis and trained convolutional neural networks to predict emotions from a dataset of 110 movie soundtracks, achieving a correlation of 0.71 relative to annotations by experts. In general, intelligent systems have struggled to predict human responses to music based on acoustic features alone. There remain disconnections between audio descriptors and high level music concepts. Because of this semantic gap between low-level audio features and human affective responses, researchers are limited in their ability to predict emotional response from acoustic information alone [48]. To improve the prediction of affective responses from audio, it seems necessary to supplement audio features with additional modalities [64].

2.2 Natural Language Processing Approaches

Given the predictive limitations of learning from audio alone, researchers considered the potential of song lyrics to aid in the prediction of the emotional qualities of a song. Investigators first began by examining statistical correlations between features extracted from the audio and the lyrics as well as the relationship between these features and the emotion annotations themselves [44]. To compensate for the lack of annotated data, one study synthetically generated emotion labels for a dataset of 100 pop-genre songs. They extracted popular tags from Last.FM and compared against the lexical database WordNet[2]. They applied latent semantic analysis, training self-organizing maps to annotate songs with four mood categories (*angry*, *happy*, *sad*, *relaxed*), manually verifying over two-thirds of labels [35]. The authors reported lower accuracy using lyrics alone (62.5%) compared to the models built on acoustic features (89.8%) [34]. The authors found that by combining acoustic and lyric features together, they were able to increase accuracy by three percent (92.4%) [5].

[2] https://wordnet.princeton.edu/

In a sequence of studies, Hu and Downie examined the relationship between emotion labels and text-based features extracted from the lyrics. To annotate their dataset with synthetic annotations of 18 moods, the authors used Last.FM tags and their WordNet distance to words in the ANEW word list [8] in order to estimate valence, arousal, and dominance values for 5,296 songs in their dataset [27]. The authors then compared various approaches to lyric sentiment analysis [25] in order to identify cases in which the performance of lyric-only models exceeded those of acoustic feature models [26]. Overall, the authors found their lyric-only model (63.7%) outperformed their audio-based model (57.9%). A fusion model combining text and audio features showed moderate improvement (67.5%) over lyrics alone. Similarly, another study reported that a late fusion of audio features and text-based features derived from the lyrics improved accuracy of their models from 46.6% to 57.1% [65]. More recently, researchers have investigated the performance of emotion recognition models based solely on the lyrics. In one such study, the authors estimated the valence and arousal values of the words in the lyrics using established word lists to create a song-level predictions of valence and arousal. The authors reported a 74.3% classification accuracy relative to the All Music Guide[3] mood tags [10].

2.3 Deep Learning Approaches

Following the many advances in deep neural algorithms and architectures over the last decade [52], researchers have begun exploring music emotion recognition tasks using both acoustic features and text-based lyrics using deep learning.

2.3.1 Deep Learning on Acoustic Features

Authors have investigated different deep neural architectures to attempt music emotion prediction using acoustic features. For this task, recurrent neural networks outperform feedforward neural networks [61]. Among recurrent architectures, bidirectional long short-term memory (BLSTM) models appear to improve prediction of musical affect over unidirectional long short-term memory (LSTM) models [37]. Additionally, researchers have reported that attentive-LSTM models improve prediction performance of arousal and valence estimations over baseline LSTM models without attention [43, 11]. By-passing preprocessing and feature extraction altogether, one research team trained bidirectional gated recurrent units directly on raw audio to attempt to classify discrete music emotions [47].

One study designed a custom experimental pipeline that makes use of both convolutional and recurrent neural networks. The authors employed a convolutional neural network to learn to select which acoustic features were subsequently used to train an LSTM model. On a custom dataset of 30 second excerpts of 124 pieces of

[3] https://www.allmusic.com/

Turkish traditional music, the authors achieved classification accuracy of 92.7% for three broad categories of emotion, which outperformed their baseline algorithms of support vector machines, random forest, and k-nearest neighbor [24]. Another recent study proposed adapting a generative adversarial network with double-channel attention mechanism (DCGAN) in order to learn the dependence between music features across channels [29]. To evaluate their architecture, the authors designed an experiment classifying five emotional characteristics (*happy*, *sad*, *quiet*, *lonely*, *longing*) on a custom dataset of 637 songs, reporting that the DCGAN (89.4%) outperformed both convolutional and recurrent architectures.

2.3.2 Deep Learning on Lyrics

In addition to the approaches to estimate emotional responses to music directly from the audio of a song, other researchers have studied the use of the lyrics of a song using deep learning to estimate human responses to music. Using only the text of the song lyrics, Agrawal et al. trained the `xl-net` transformer [66] and achieved around 95% classification accuracy on a large dataset of lyrics [2]. This encouraging result may imply large-language models have the ability to capture meaningful semantic relationships from music lyrics without additional acoustic descriptors.

More recently, investigators have begun adapting large language models (LLM) and algorithms to learn embeddings directly from the audio signal. One recent study combined a large-scale pretrained language model with an audio encoder to attempt to generate interpretations from cross-modal inputs of song lyrics and musical audio [68]. Another research team explored representations of music using a joint embedding of natural language and audio [30]. Using these embeddings of over 5000 music and text pairs, the team trained generative acoustic models able to produce music based on a text description given as input [1].

2.4 Large Language Models

Transformers are deep learning models based on the principle of self-attention [57]. This LLM architecture, first introduced in 2017, has quickly become popular in the areas of natural language processing (NLP) and computer vision (see review [38]) where large pretrained models have achieved state-of-the-art performances in a wide variety of tasks. In this section, we briefly review the four transformer-derived models that we compare in this study.

2.4.1 BERT

BERT, or Bi-directional Encoder Representations from Transformers, is a popular transformer model for learning representations of natural language [18]. BERT

leverages a large dataset of unstructured English text from Wikipedia and assorted literature. By taking unlabeled sequences of English text, corrupting parts of the input, and attempting to predict the missing tokens, the model encodes complex relationships between words, and demonstrates the ability to learn robust language representations. This self-supervised pretraining objective function is referred to as masked language modeling, and has become the foundation for many similar models. Because BERT and other LLMs are pretrained on very large amounts of data, these model can be fine-tuned to new tasks relatively quickly. However, model training still requires significant compute resources, especially when learning large datasets. BERT is widely used in many NLP tasks, including machine translation, dialogic generation, question answering, and sentiment analysis.

2.4.2 DistilBERT

DistilBERT seeks to address the immense computational requirements of BERT while retaining the same capability to learn effective language representations. To accomplish this, they leverage knowledge distillation to train a smaller model to emulate the behavior of BERT [51]. By optimizing DistilBERT to predict the same output probabilities as BERT during pretraining, the authors design a model which retains 97% of the performance on benchmark NLP tasks. DistilBERT bases its architecture on BERT, but reduces the number of hidden layers from 12 to 6. This lowers the number of model parameters by 40%, enabling faster training and fine-tuning.

2.4.3 RoBERTa

RoBERTa aims to surpass the performance of BERT by both leveraging a larger pre-training corpus and by modifying the pretraining objective task. [41]. The authors replicate the architecture of BERT while empirically studying how various factors in the pretraining of large language models impacts downstream task performance. They find that encoder transformer models respond positively to significantly larger pretraining datasets and longer pretraining schedules with larger batch sizes. Furthermore, they propose a dynamic masking method, randomly altering which tokens are masked during the masked language modeling task, which helps to improve the performance of RoBERTa on NLP benchmarks compared to BERT.

2.4.4 `xl-net`

`xl-net` is another transformer architecture which seeks to improve upon limitations of large transformer models with autoencoder-based pretraining objectives such as BERT and RoBERTa [66]. The `xl-net` architecture features an autoregressive objective function, doing away with the masked language modeling approach used

by BERT and its derivatives. This, the authors propose, should reduce the asymmetry between the distributions learned during the pre-training process and those of the inputs in downstream tasks. Unlike previous autoregressive approaches, `xl-net` models all possible permutations of the input sequence, enabling it to learn long-range and bi-directional dependencies of words in context. Like other LLMs, it is pretrained on a large corpus of text and subsequently fine-tuned with additional training for specific NLP tasks. `xl-net` has been shown to achieve competitive results on many NLP benchmarks.

3 Music Emotion Datasets

In this work, we study the task of music emotion recognition over four datasets of songs with song-level annotations of valence and arousal. These datasets were created using similar procedures: tasking human annotators to listen to a musical excerpt and provide a numeric description of their emotive response. The labor-intensive nature of collecting these annotations has hindered research in music emotion prediction. The advent of crowdsourcing platforms has enabled experimenters to reach a wider audience, however these annotations are still expensive and time-consuming [14]. Researchers have also created synthetic valence-arousal annotations [17] by mapping community-provided features, such as Last.fm tags or metadata from the All Music Guide, to existing word-affect datasets [60].

Dataset	Songs	Label Type	Scaling
AMG1608	1608	Crowdsourced	$[-1, 1]$
DEAM	1803	Crowdsourced	$[0, 10]$
PmEmo	767	Lab Survey	$[0, 1]$
Deezer	18,648	Synthetic [4]	$[-3, 3]$

Table 1: Details of the valence and arousal labels in selected MER datasets.

3.1 AMG1608

The AMG1608 dataset provides 1,608 songs selected from the All Music Guide (AMG) and rated for valence and arousal by 665 annotators [14]. The dataset's creators aimed to develop a large and state-of-the-art music emotion recognition dataset. To conduct an annotation experiment at this scale, the authors used Amazon Mechanical Turk−an online crowdsourced work platform−to reach a large subject pool. AMG users rated songs for 34 different mood categories, which are converted from mood labels to valence-arousal estimates using the `tag2VA` algorithm [59]. From this, a subset evenly distributed in the valence-arousal space was selected. 665

annotators participated in the experiment, and between 15 and 32 annotations were collected per song. 46 annotators provided ratings for over 150 songs, presenting a unique opportunity for a study of emotion-aware music recommender personalization as well as introducing a potential bias in the label distribution. Each annotator listened to a 30-second excerpt of the sample and provided a dimensional rating in the circumplex model of emotion. Each coordinate was treated as a valence-arousal label, and individual coordinate labels were averaged between annotators to produce an emotion label for a given song.

3.2 PMEmo

Music emotion recognition datasets typically fall into one of two categories: small datasets with few annotators rating samples in lab environments to yield high-quality individual annotations or larger datasets with many annotations of relatively lower quality gathered using online crowdsourcing platforms such as Amazon Mechanical Turk. The PMEmo dataset fills the need for music emotion datasets with both high-quality annotations and many samples by conducting a large-scale human subject study in a laboratory setting [67]. 457 annotators, including 366 Chinese university students, participated to annotate valence and arousal over a collection of 794 songs.

1000 songs were selected from record label industry charts between 2016 and 2017, such as Billboard Top 100, iTunes Top 100, and UK Top 40 Singles. After deduplication, 794 songs remained, primarily representing Western pop music. A 30-second sample representing the chorus of each song was manually excerpted by music students. Annotators were instructed to listen to each sample and provide dynamic valence-arousal ratings at a 2 Hz sample rate using an annotation interface derived from the Self Assessment Mannikin [8]. At the end of the sample, annotators were then asked to provide a single valence-arousal rating representing their overall emotive response to the song. These static annotations were averaged to provide valence-arousal labels for each song. Electrodermal activity was also recorded from participants during the listening and annotation experiment.

3.3 DEAM

Research in music information retrieval continues to be hindered by a lack of annotated datasets with accompanying audio data. Copyright restrictions on the majority of publicly released music prohibits researchers from distributing audio recordings in music datasets, posing significant challenges to approaches that learn from audio data. In response, Soleymani et. al. provide a dataset of royalty-free songs annotated for affective qualities using continuous emotion labels [3].

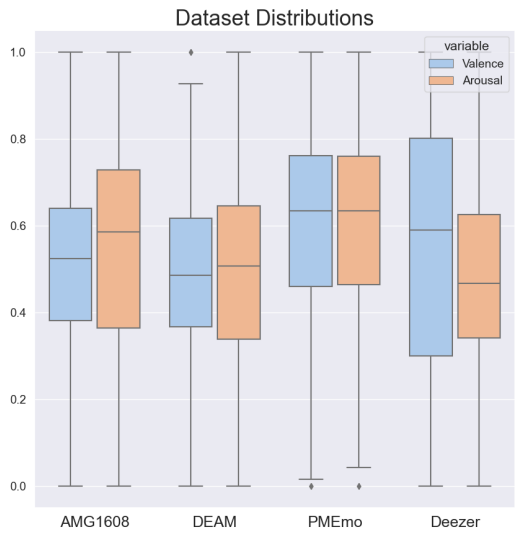


Fig. 1: Distributions of the valence-arousal labels for each of the datasets.

The DEAM dataset consists of 1,803 songs collected from `freemusicarchive`, `jamendo`, and medleyDB, online repositories of royalty-free music. 45-second segments were selected randomly from each sample, and annotators were asked to provide dynamic valence and arousal annotations at a 2 Hz sample rate while listening to this excerpt. These dynamic ratings were averaged over time to provide a single per-participant valence-arousal annotation. Annotators were recruited from Amazon Mechanical Turk, and each song received a minimum of five annotations. Along with averaged valence-arousal annotations, the authors provided both the excerpt and full-length audio for each song. Although the royalty-free licensing of these songs permits the distribution of the audio recordings, it seems that these copyright-free samples are more obscure than songs included in other datasets. For this reason, we expect to find less online discourse about these songs compared to popular songs used in other datasets.

3.4 Deezer

Despite these efforts towards the creation of large-scale annotated music emotion recognition datasets, even the largest manually annotated datasets only consist of a few thousand samples at most. To evaluate the utility of deep learning approaches for valence-arousal estimation, significantly more data is necessary than what is

currently available. The cost of manually annotating datasets at sufficient scale would be prohibitive. Researchers at Deezer developed a dataset consisting of synthetic valence-arousal labels, which we refer to as the Deezer dataset[17].

In this study, songs available in both the Million Song Dataset [4] and Deezer's music streaming library were selected. Each song's associated tags were aggregated from Last.fm, providing a list of community-provided key descriptors for each song. From these tags, a synthetic valence and arousal label was generated by comparing the tags against the Extended ANEW dataset [8], a collection of 14,000 English words annotated for valence and arousal [60]. The Deezer dataset operates on the fundamental assumption that the valence and arousal annotations of English words from Warriner et. al.'s experiments transfer to the music emotion space, and that the descriptors added by community users on Last.fm meaningfully relate to the song's emotive qualities. For these reasons, the authors concede that their dataset is not as robust of a ground truth as manually annotated music emotion datasets.

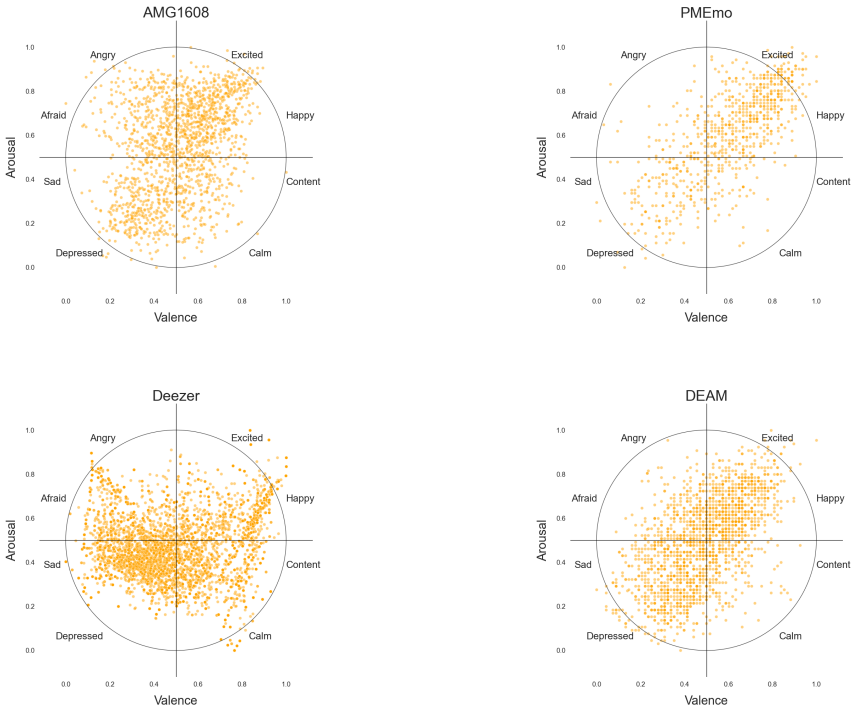
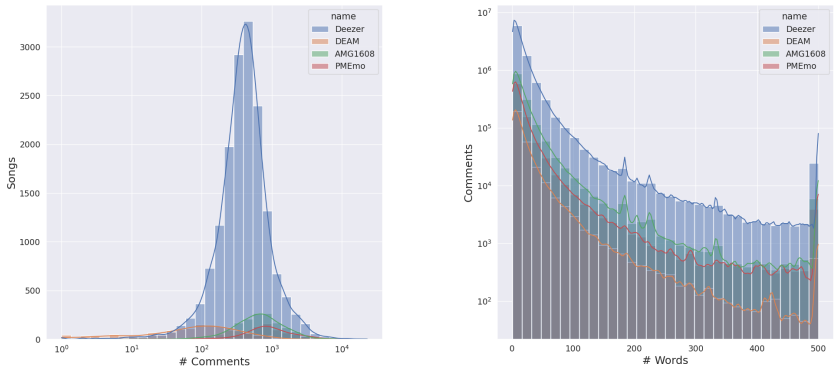


Fig. 2: Circumplex models, representing distributions of the labels from four music emotion datasets across Russell's emotional space [50].

4 Musical Discourse Model

We propose a system for the automatic prediction of static valence and arousal targets using the social media discourse related to a song to estimate users' average emotive response to that song. To accomplish this, we collect social media commentary from Reddit[5] and YouTube[6], both platforms with active music sub-cultures engaging in discussion about music. From these conversations alone, and without considering the song's audio or lyrics, we attempt to predict a song's valence and arousal by fine-tuning pretrained large language models. We focus our investigation on transformer-based encoder models, such as BERT [18] and its derivatives, on a two-target regression task of estimating song-level estimates of music emotion from online comments associated with a musical sample.

4.1 Collecting Social Media Commentary



(a) Number of comments per song

(b) Length of comments

Fig. 3: Social media discourse dataset distributions, stratified by dataset.

We collect social media discourse which references the songs from AMG1608 [14], PMEmo [67], DEAM [3], and Deezer [17]. In total, our dataset gathers social media comments about 19.627 songs, of which 4,179 are manually annotated. For each sample in the four datasets, we query the two social media platforms for posts which make direct reference to both the song title and the artist. From each platform,

[5] https://www.reddit.com/

[6] https://www.youtube.com/

we select the 50 highest rated submissions as ranked by each platform's search API. We then collect every comment and reply which responds to any of these top-level submissions. If a song has not been discussed on Reddit or YouTube, we omit that song from our discourse dataset. Table 1 describes our dataset of retrieved comments.

We collected data over a six month period between November 2022 and April 2023, scraping both past and recent comments. We achieved higher retrieval rates from YouTube across all four songlists; 84% to 97% of all queried songs had at least one matching post. Retrieval rates from Reddit were lower, although we found reference to over 80% of songs from AMG1608 and PMEmo. However, only 11% of songs from DEAM had corresponding comments on Reddit. When comparing DEAM against similarly sized datasets, we observed that our data collection totals 1,472,021 and 881,931 comments for the songs in AMG1608 and PMEmo datasets, respectively, but only 303,667 comments reference songs from DEAM. In total, our dataset of musical discourse contains more than 11 million comments. Figure 3 shows the distribution of the retrieved comments and their associated lengths. Unsurprisingly, the popular songs from the PMEmo dataset were associated with higher rates of discourse. Conversely, for the relatively obscure songs in the DEAM dataset, we found significantly fewer comments per song than other datasets.

		Songs		**Comments**			**Words**	
		n	Yield	n	μ	σ	μ	σ
AMG1608	Reddit	1,412	88%	578,283	409.5	1,180.5	15,796	43,847
	YouTube	1,563	97%	893,738	571.8	268.1	11,424	5,917
PmEmo	Reddit	624	81%	391,325	627.1	1,122.1	21,065	47,179
	YouTube	736	96%	490,606	666.6	267.1	11,333	5,806
DEAM	Reddit	205	11%	69,943	341.2	1,873.1	13,562	67,127
	YouTube	1,508	84%	233,724	155.0	194.8	3,153	5,030
Deezer	Reddit	11,122	60%	2,497,517	224.6	767.2	8,963	30,649
	YouTube	16,435	88%	6,685,202	406.8	229.2	7,524	4,792
Total		19,627	86%	11,840,338	603.3	911.5	14,904	32,125

Table 2: Summary statistics for our dataset of social media commentary by source.

4.2 Model Design

We evaluate our musical discourse dataset as a feature space in the task of music emotion recognition by applying large pretrained transformer models designed for natural language understanding tasks to our corpora. These models learn language representations from very large datasets of unstructured text using self-supervised language modeling tasks. From this pretraining, these models can be fine-tuned to learn downstream tasks in relatively few epochs [18, 41]. We fine-tune one such model, BERT, on our multi-target regression task by assigning each song's valence

and arousal label to the comments relating to that song, and attempting to predict a song's emotive annotations directly from this social media discourse. We then compare the performance of BERT on this task to a selection of other pretrained large language models: DistilBERT [51], RoBERTa [41], and `xl-net` [66].

We use the implementations of these large language models provided by the Hugggingface deep natural language processing library[7]. Each input consists of a single social media comment, labeled by the valence and arousal annotation for the song which the comment is associated with. We tokenize each input using the `TokenizerFast` library and use a maximum sequence length of 128 tokens, truncating and right-padding inputs to coalesce sequences to this dimension.

Our model consists of two components: a pretrained transformer to learn a representation for input comments, and a fully-connected neural network with one hidden layer to learn a regression target from this language representation. We output the last hidden state of the `[CLS]` token, as is standard for designing classifiers using BERT's language representations [18]. Our fully-connected layer, serving as the regression head, learns a valence and arousal output based on these `[CLS]` last-hidden-state vectors. We fine-tune BERT to adapt the learned representations to our downstream task. This does incur a risk of overfit, as BERT and its derivatives have significantly more parameters than our dataset has examples. We limit our fine tuning to two epochs to mitigate overfit as recommended in [18, 51]. We use a mean-squared-error loss and the Adam optimization algorithm [31] with a learning rate of 1×10^{-5}.

4.3 Experimental Design

We randomly partition each songlist into training, validation, and test subsets with $0.70 \times 0.15 \times 0.15$ split, respectively. All comments associated with a song are then placed in that song's corresponding subset. Valence and arousal labels are normalized and scaled to $[0, 1]$. Inputs are filtered to remove URLs and HTML tags. Further text pre-processing is unnecessary for fine-tuning of pretrained large language models, as BERT and similar models use unfiltered text from online sources for their pre-training tasks[18]. These models expect inputs to adhere to standard grammatical structure, and as such we do not lemmatize nor remove stopwords from our comments. Each comment is assigned a music valence and arousal prediction from our regression model. To produce song-level valence and arousal labels from these comment-level outputs, we take the average of all output labels for each comment associated with a song to produce the final valence-arousal estimation for that song.

[7] https://huggingface.co/models

5 Music Emotion Recognition with Large Language Models

We test the performance of a BERT-based regression model for predicting the emotive qualities of a piece of music. We begin by measuring the performance impact of language models trained to be case-sensitive versus their uncased counterparts. Next, we test comment-level filtering schemes, evaluating the impact of dropping short comments or those below a certain score threshold, where score is a measure of likes or upvotes from the comment's platform of origin. We compare several large language models to investigate the impact of different transformer architectures and pretraining schemas for this task. For our tuning experiments, we test each model configuration on both AMG1608 and PMEmo, chosen for their manually annotated labels and active discussion on the social media platforms we investigate. We select the filtering strategy and pretrained model with the best performance on these two datasets and investigate its performance on the DEAM and Deezer datasets.

5.1 Model Parameters and Dataset Preprocessing

We compare model implementation and dataset preprocessing methods for predicting music emotion targets from AMG1608 and PMEmo using our musical discourse dataset. First, we evaluate both cased and uncased versions of BERT for this task and measure the performance implications of case sensitivity in BERT pre-training on our task. Informed by this experiment, we then test dataset filtering methods to identify a preprocessing strategy to reduce potential noise in our social media data.

5.1.1 Case-Sensitivity of Language Model

We explore two versions of the BERT model: one case-sensitive (`bert-base-cased`) and the other, case-insensitive (`bert-base-uncased`). `bert-base-uncased` uses the same pre-training tasks as its cased counterpart. However, during pretraining, all text is transformed to lower-case and accent markers are removed.[8] We compare the performance of these two models on our task. For each model, we run experiments on AMG1608 and PMEmo. In both cases, the model is trained on a combination of comments from YouTube and Reddit.

In Table 3 we show the Pearson's correlation between our models' predictions and the datasets' annotated labels. When training on AMG1608, we observe a slight yet measurable improvement in model performance with the cased variant of BERT. We expect this improvement results from the use of capitalization as an important mechanism for conveying tone or intent, and therefore provides important semantic information for our emotion recognition tasks. We use `bert-base-cased` in following experiments.

[8] See https://huggingface.co/bert-base-uncased

	PMEmo		AMG1608	
	Valence	Arousal	Valence	Arousal
bert-cased	0.68	0.47	0.51	0.75
bert-uncased	0.68	0.49	0.45	0.74

Table 3: Pearson's correlation of cased and uncased variants of BERT.

5.1.2 Comment Filtering

Though we perform some preprocessing on the input text, this does not mean that all inputs are useful for our downstream task. To address the innate noisiness of social media data, we evaluate a selection of strategies for rejecting certain comments from our training set. We begin by filtering comments based on the number of likes or upvotes they receive from other users on their respective social media platform. We assume that highly rated comments are more likely to express sentiments shared by the community. By filtering out comments with lower scores, we hope to prune off-topic discussion and spam, as these types of responses are less likely to be informative. We initially begin with a score threshold of 3, based on the criteria used by other large language models which rely on Reddit data for their pre-training dataset [49]. This aggressive filtering method removes a total of 71% of all comments across AMG1608 and PMEmo. We also explore a weaker filtering threshold, requiring only that the score be positive (≥ 1), excluding a more modest 36% of our total comments.

When training on data filtered with a score threshold of 3, we observe marginal improvements in performance over the unfiltered BERT baseline across all dimensions except valence on PMEmo, for which we observe a decrease in correlation by 28%. The less aggressive score threshold of 1 does not appear to have as drastic impact on performance on the PMEmo dataset, and in fact outperforms baseline by 4% to valence and 10% to arousal. However, on the AMG1608 dataset, lowering the score threshold weakens performance overall and it does not exceed baseline performance. We suspect that dataset size is an important factor in this large difference in per-dataset performance. A filtering regimen which works well for AMG1608 may remove too many comments from PMEmo, and one which optimizes for PMEmo may leave too many noisy inputs for AMG1608. Additionally, we test filtering comments by length. Our model expects inputs of at most 128 words, adding zero-tokens to pad all inputs to this dimension. Though attention masking allows our model to handle zero-padded inputs without introducing excess noise, longer inputs contain more semantically meaningful tokens in each input tensor. We identify that the bottom quartile of comments in the combined AMG1608 and PMEmo musical discourse dataset has at most 30 characters. We drop this lower quartile of comments.

We combine our score threshold and length threshold filters to require all comments both be longer than 30 characters and have a positive score. Comparatively, this yields improved performance when predicting arousal labels for PMEmo and valence labels for AMG1608 than any other individual preprocessing method. How-

ever, this comes at the cost of a reduction in predictive performance in the other dimensions. When we filter short comments in the bottom quartile of character length, we achieve the best performance overall, with a 2.1% increase over baseline. We apply this technique in our comparison of language models in subsection 5.3.

	PMEmo			**AMG1608**		
	Valence	Arousal	Count	Valence	Arousal	Count
Baseline	0.68	0.46	688,712	0.51	0.75	1,281,473
Score ≥ 3	0.49	0.47	221,360	0.53	**0.79**	352,481
Score ≥ 1	0.71	0.51	448,631	0.45	0.74	822,623
Length ≥ 30	**0.80**	0.48	474,827	0.52	0.65	976,081
Joint Filtering	0.63	**0.57**	334,215	**0.62**	0.57	654,316

Table 4: Impact of comment filtering strategies on model performance.

5.2 Comparison of Social Media Sources

Next, we compare the utility of conversations from each social media platform for the purpose of music emotion recognition in Table 5. YouTube is widely used for sharing and listening to music, and we anticipate that users commenting most likely recently watched the music video. Reddit, on the other hand, is more suitable for longer conversations but lacks the YouTube's popularity as source for music. We believe both types of conversations contain semantic information relevant to the task of music emotion recognition. However, the difference in both user experience and intent warrants an investigation into models trained on individual sources.

	PMEmo		**AMG1608**	
	Valence	Arousal	Valence	Arousal
Youtube	0.54	0.47	0.60	0.65
Reddit	0.51	0.50	0.34	0.54
All	0.80	0.48	0.52	0.65

Table 5: Performance of source-specific models.

We find that a model trained on exclusively YouTube comment data exceeds our baseline model's performance on the AMG1608 dataset. However, neither source-specific model outperforms the best scores for AMG1608 presented in Table 4. A YouTube specific model underperforms relative to our multi-source baseline on PMEmo dataset, again demonstrating that additional filtering adversely impacts performance for datasets lacking large amounts of commentary. We find that models

trained on only YouTube conversations outperform those trained only on Reddit data, which we attribute to the fact that our dataset includes 135% more YouTube comments than Reddit submissions. Overall, our combined source model outperformed both models trained on individual social media sources.

5.3 Comparison of Language Models

As detailed in subsection 2.4, various BERT-derived models have sought to address limitations with the model and improve its performance on downstream tasks. We compare three of these models, fine-tuning each on our emotion prediction task. Specifically, we investigate DistilBERT [51], RoBERTa [41], and `xl-net` [66], each addressing different limitations of the original approach.

	PMEmo		**AMG1608**	
	Valence	Arousal	Valence	Arousal
BERT	0.80	0.48	0.52	0.65
DistilBERT	0.80	0.46	0.49	0.64
RoBERTa	0.79	0.53	0.55	0.67
`xl-net`	0.80	0.50	0.55	0.63

Table 6: Performance of selected pretrained large language models after fine-tuning.

We observe a small improvement in performance using language models pretrained on larger corpora of text. Predictions generated from models using DistilBERT achieve correlations within 97% of baseline, reinforcing claims made in [51]. `xl-net` and RoBERTa outperform BERT by 1.3% and 4.3%, respectively. Both of these models use significantly larger datasets for their respective pretraining approaches. We select RoBERTa for our final model evaluations in subsection 5.4.

5.4 Dataset Comparison

In our previous experiments, we focused on the PMEmo and AMG1608 datasets because the songs in these datasets were annotated by human subjects. In our final experiment, we compare our approach's performance across the available datasets, including both DEAM, which suffers from a lack of relevant social media commentary, and Deezer, whose annotations were synthetically generated. In this experiment, we first fine-tune RoBERTa using the combined commentary from both Reddit and YouTube. We then filter all comments shorter than 30 characters, dropping the shortest 25% of comments. We evaluate this model's performance on the DEAM and

Deezer datasets, whose distributions of labels differ from those of AMG1608 or PMEmo (see Figure 1).

	Valence	**Arousal**
AMG1608	0.55	0.67
PMEmo	0.79	0.53
DEAM	0.08	0.02
Deezer	0.47	0.43

Table 7: Pearson's correlations to the ground truth labels for the four datasets.

We find that our model's predictions achieve weak but measurable correlations to the synthetically annotated labels in the Deezer dataset. This is not surprising since the synthetically generated annotation are less likely to reflect the range and nuance of affective responses reported by human subjects when listening to music. We hypothesize that the social media discourse our model uses to predict a song's emotive properties may not correlate well with the word-level representations of text tags used to synthesize the Deezer dataset. Although we expected our model to struggle with the DEAM dataset, we observe that our model completely fails to predict emotional responses to the songs from DEAM. The labels in the DEAM dataset are provided by crowdsourced human annotators in a method similar to the labels provided in AMG1608. However, as shown in Table 2, our data collection of social media commentary yields significantly fewer comments associated with songs from DEAM than any other datasets. The average number of YouTube comments per song in DEAM is only 155, compared to between 400 and 650 comments per song for the other datasets. Given the insufficient quantity of discourse, our model was unable to learn from the DEAM dataset.

6 Discussion

We present a novel approach to predict dimensional music emotion labels through sentiment analysis of social media conversations discussing a piece of music. To assess the potential for a model to estimate a song's affective qualities solely from social media discourse, we create a large corpus of online conversations related to the songs in four published datasets for music emotion recognition. We construct a music emotion prediction system using pretrained large language models, leveraging the language representations learned by these transformer models to fine-tune each to this task. Overall, we observe modest correlations between the predictions made by these models and the dimensional emotion labels provided by human annotators.

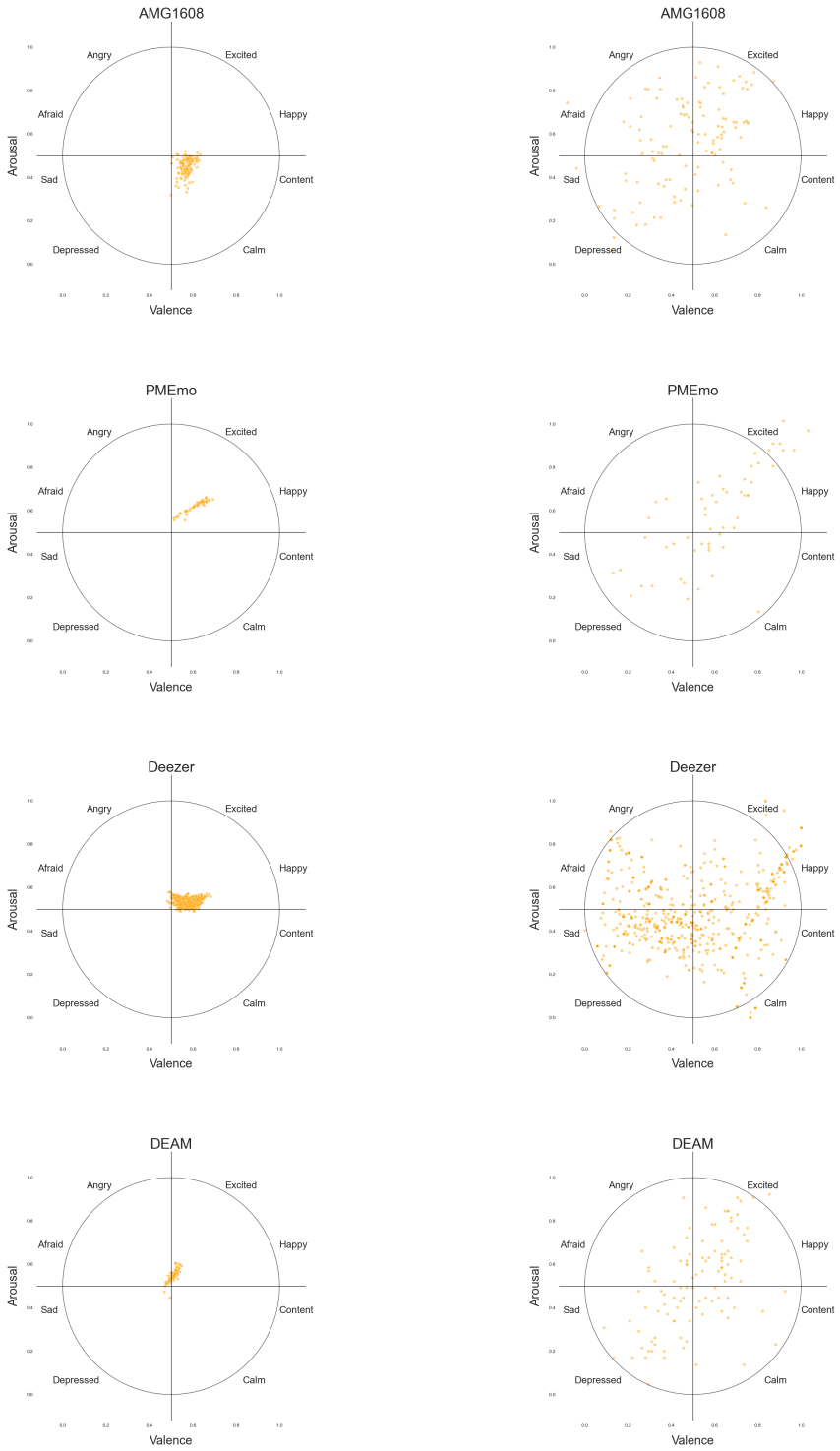


Fig. 4: Comparisons of the distributions of the ground truth labels (right) against our model's predictions (left) in the valence-arousal space.

6.1 Limitations

We visualize our output model predictions in the valence-arousal space in Figure 4. Our model's predictions tend to cluster closely to the center of this space. This indicates that, despite a moderate linear relationship between our estimates and true labels, these predictions are often collapsing to the average within the distribution. This issue persists despite our attempts to reduce the noise in our dataset. We anticipate this phenomenon results from our approach to aggregating comment-level predictions to generate a single overall valence and arousal estimate for a song. In our approach, we predict a valence and arousal value for each comment, then average the comment-level predictions to produce a song-level estimate. This process discards valuable semantic information. Comment-level predictions cannot capture the relationship between the comments, including those that serve as replies to another. Furthermore, this aggregation process may collapse comments belonging to the same song with conflicting sentiments to a neutral value.

Our model requires songs to correspond to a sufficient volume of comments available on the social media platforms. This requirement restricts our model's ability to make inferences about the emotive qualities of newly released songs or those that belong to a particularly niche genre. Models which use the audio information and lyrics of a song to make these predictions would not share such limitations. In our analysis of the DEAM dataset [3], we find that the copyright-free nature of these songs was correlated with our inability to find relevant conversational activity online. Without sufficient data, our model was unable to make meaningful predictions.

6.2 Future Work

We will explore methods to preserve the relationships between comments associated with the same song. In the current work, our model expects a single comment, paired with an accompanying music valence-arousal label, for each input. However, this does not reflect the task we seek to learn: a single valence-arousal prediction for a song given a set of comments. In future work, we will investigate new model architectures capable of receiving a single song, with all accompanying discourse packed into a single input tensor, and learning a music affect label without reliance on an aggregation of individual comment-level predictions. A potential approach to address this problem is to simply concatenate comments together to form a single input. BERT and its derivatives use `[SEP]` tokens to indicate the beginnings and ends of distinct sequences within a single input. However, the runtime complexity of transformer models scales quadratically with input size [18], and most large language models limit to a maximum of 1024 tokens per input. This restricts how many comments can be included as input. Further investigation is needed to design a custom architecture to learn across the different comments of a single thread.

We observe potential in our joint filtering approach in Table 4. Filtering comments in our musical discourse dataset marginally improves model performance.

Furthermore, it appears that filtering by both length and score improves performance along dimensions on which our model underperforms, namely PMEmo arousal and AMG1608 valence prediction. However, these filtering strategies resulted in inconsistent performance between datasets, which results from the differences in number of comments available by dataset. By applying dynamic filtering methods, which adjust score and length thresholds based on the number of samples which exist in a dataset, we may address the inconsistent effects of filtering techniques across datasets. Additional criteria for comment filtration should also be introduced and compared against the methods we demonstrate. For example, comments not associated with the expression of an affective response, as determined by existing dictionaries of affective terms[60], could be removed to filter out comments of neutral sentiment.

Additionally, we plan to explore additional sources for music-relevant online discourse. Last.fm and Soundcloud[9] are both online platforms that focus on music, and these communities are a potential source of information directly related to specific pieces of music. The community-provided tags on Last.fm have been used in prior music emotion recognition experiments [35, 17, 6, 14]. This platform also allows users to post comments in response to a specific track with comment mechanism known as "Shouts". Similarly, Soundcloud users can respond to a song with a public comment. Because these comments stem from communities on music-specific platforms, and are in direct response to a song as opposed to responding to a post about a song, we anticipate that the conversations on these platforms may be valuable to a social media-based music emotion recognition system.

To address the cases of songs with limited or no presence on these social media platforms, we intend to explore feature spaces beyond social media data to augment our existing approach. We expect the inclusion of lyrics, song metadata, and acoustic features in conjunction with our social media information to yield a more robust estimator. We hope that the exploitation of these feature spaces will improve a model's performance on songs for which there is comparatively little online conversation.

7 Conclusion

The development of an automatic system for estimating the emotive qualities of a piece of music has been impeded by a lack of large, high-quality, annotated music emotion recognition datasets. Such annotation experiments are expensive and time-consuming to perform. Furthermore, the distribution of the audio samples used in such datasets is prohibited by copyright law in many cases, restricting the use of an important feature modality for music emotion prediction tasks. We demonstrate the feasibility of predicting continuously-valued music emotion labels using only musical discourse from social media platforms. Such an approach does not rely on a song's audio nor its lyrics, enabling inference to be drawn about a song's affective qualities indirectly and without access to copyrighted information.

[9] https://soundcloud.com/

We create a large dataset of social media conversations about musical samples using the songlists provided by four music emotion recognition datasets. In total, we gather over 11 million comments discussing nearly 20 thousand songs. We use this dataset to design and evaluate a system for music emotion prediction using pretrained transformer models. We find that, with relatively few training epochs, these large language models can fine-tune to our music valence-arousal prediction task and provide emotion estimations with moderate correlation to human-provided annotations. To our knowledge, this is the first attempt to predict musical valence and arousal labels using exclusively conversational data from social media platforms.

The ability to predict how an average listener may respond to a piece of music could be used to improve existing music recommender systems. Without the need for costly annotation experiments nor licensing for song audio, large music libraries could be rated for dimensional emotional values. The granularity afforded by continuous valence-arousal annotations would allow music streaming services to categorize songs with greater respect to affective characteristics. As another potential broader impact of this research, the ability to quickly and autonomously annotate large libraries of music would enable intelligent systems to automatically generate affect-aware music playlists [21], with potential uses in music therapy [55].

We demonstrate that the conversations people have online about a piece of music can be used to train a model to predict the average affective response elicited in a listener by that song. Our model achieves moderate performance on the prediction of human annotated music emotion targets. Without access to song audio, pre-computed acoustic features, or song lyrics, we are able to fine-tune a large language model to estimate valence and arousal labels corresponding to affective response to a piece of music using only this online musical discourse.

References

1. Agostinelli, A., Denk, T.I., Borsos, Z., Engel, J., Verzetti, M., Caillon, A., Huang, Q., Jansen, A., Roberts, A., Tagliasacchi, M., Sharifi, M., Zeghidour, N., Frank, C.: MusicLM: Generating music from text (2023). DOI 10.48550/arXiv.2301.11325. ArXiv:2301.11325 [cs.SD]
2. Agrawal, Y., Shanker, R.G.R., Alluri, V.: Transformer-based approach towards music emotion recognition from lyrics. Advances in Information Retrieval (ECIR) **12657**, 167–175 (2021). DOI 10.1007/978-3-030-72240-1_12. ArXiv: 2101.02051
3. Aljanaki, A., Yang, Y.H., Soleymani, M.: Developing a benchmark for emotional analysis of music. PloS one **12**(3), 1–22 (2017). DOI 10.1371/journal.pone.0173392
4. Bertin-Mahieux, T., Ellis, D.P.W., Whitman, B., Lamere, P.: The million song dataset. In: Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR) (2011)
5. Bischoff, K., Firan, C.S., Paiu, R., Nejdl, W., Laurier, C., Sordo, M.: Music mood and theme classification - a hybrid approach. In: Proceedings of the 10th International Society for Music Information Retrieval Conference, ISMIR, pp. 657–662 (2009). DOI 10.5281/zenodo.1417317
6. Bischoff, K., Firan, C.S., Paiu, R., Nejdl, W., Laurier, C., Sordo, M.: Music mood and theme classification - a hybrid approach. Poster Session p. 6 (2009)
7. Bogdanov, D., Wack, N., Gómez Gutiérrez, E., Gulati, S., Boyer, H., Mayor, O., Roma Trepat, G., Salamon, J., Zapata González, J.R., Serra, X., et al.: Essentia: An audio analysis library

for music information retrieval. In: D.S. Britto A Gouyon F (ed.) Proceedings of the 14th of the International Society for Music Information Retrieval Conference (ISMIR), ISMIR, pp. 493–498. International Society for Music Information Retrieval (ISMIR) (2013)
8. Bradley, M.M., Lang, P.J.: Affective norms for english words (anew): Instruction manual and affective ratings (1999)
9. Cabrera, D., et al.: Psysound: A computer program for psychoacoustical analysis. In: Proceedings of the Australian Acoustical Society Conference, vol. 24, pp. 47–54. AASC Melbourne, Australia (1999)
10. Cano, E., Morisio, M.: Moodylyrics: A sentiment annotated lyrics dataset. In: Proceedings of the 2017 International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence, ISMSI '17, p. 118–124. Association for Computing Machinery, New York, NY, USA (2017). DOI 10.1145/3059336.3059340
11. Chaki, S., Doshi, P., Patnaik, P., Bhattacharya, S.: Attentive rnns for continuous-time emotion prediction in music clips. In: Proceedings of the 3rd Workshop on Affective Content Analysis, pp. 36–46. AAAI (2020)
12. Chang, W.H., Li, J.L., Lin, Y.S., Lee, C.C.: A genre-affect relationship network with task-specific uncertainty weighting for recognizing induced emotion in music. In: Proceedings of the 2018 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6. IEEE (2018). DOI 10.1109/ICME.2018.8486570
13. Chen, Y.A., Wang, J.C., Yang, Y.H., Chen, H.: Linear regression-based adaptation of music emotion recognition models for personalization. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2149–2153. IEEE (2014). DOI 10.1109/ICASSP.2014.6853979
14. Chen, Y.A., Yang, Y.H., Wang, J.C., Chen, H.: The amg1608 dataset for music emotion recognition. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), p. 693–697. IEEE, South Brisbane, Queensland, Australia (2015). DOI 10.1109/ICASSP.2015.7178058
15. Chowdhury, S., Vall, A., Haunschmid, V., Widmer, G.: Towards explainable music emotion recognition: The route via mid-level features. In: Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR, pp. 237–243 (2019). DOI arXiv:1907.03572
16. Cross, I.: Music, cognition, culture, and evolution. Annals of the New York Academy of sciences **930**(1), 28–42 (2001). DOI https://doi.org/10.1111/j.1749-6632.2001.tb05723.x
17. Delbouys, R., Hennequin, R., Piccoli, F., Royo-Letelier, J., Moussallam, M.: Music mood detection based on audio and lyrics with deep neural net. In: Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR, pp. 370–375 (2018)
18. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: "Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)", pp. 4171 – 4186 (2019). DOI 10.18653/v1/N19-1423
19. Dong, Y., Yang, X., Zhao, X., Li, J.: Bidirectional convolutional recurrent sparse network (bcrsn): an efficient model for music emotion recognition. IEEE Transactions on Multimedia **21**(12), 3150–3163 (2019). DOI 10.1109/TMM.2019.2918739
20. Donnelly, P.J., Beery, A.: Evaluating large-language models for dimensional music emotion prediction from social media discourse. In: M. Abbas, A.A. Freihat (eds.) Proceedings of the 5th International Conference on Natural Language and Speech Processing (ICNLSP 2022), pp. 242–250. Association for Computational Linguistics (2022)
21. Donnelly, P.J., Gaur, S.: Mood dynamic playlist: Interpolating a musical path between emotions using a KNN algorithm. In: T. Ahram, R. Taiar (eds.) Human Interaction & Emerging Technologies: Artificial Intelligence & Future Applications (IHIET-AI 2022), vol. 23. AHFE Open Access (2022). DOI 10.54941/ahfe100894
22. Eyben, F., Wöllmer, M., Schuller, B.: Opensmile: The Munich versatile and fast open-source audio feature extractor. In: Proceedings of the 18th ACM international conference on Multi-

media, MM '10, pp. 1459–1462. Association for Computing Machinery, New York, NY, USA (2010). DOI 10.1145/1873951.1874246
23. Fan, J., Tatar, K., Thorogood, M., Pasquier, P.: Ranking-based emotion recognition for experimental music. In: Proceedings of the 18th International Society for Music Information Retrieval Conference, *ISMIR*, vol. 2017, pp. 368–375 (2017). DOI 10.5281/zenodo.1416946
24. Hizlisoy, S., Yildirim, S., Tufekci, Z.: Music emotion recognition using convolutional long short term memory deep neural networks. International Journal of Engineering Science and Technology **24**(3), 760–767 (2021). DOI 10.1016/j.jestch.2020.10.009
25. Hu, X., Downie, J.S.: Improving mood classification in music digital libraries by combining lyrics and audio. In: Proceedings of the 10th Annual Joint Conference on Digital libraries, JCDL '10, p. 159–168. Association for Computing Machinery, New York, NY, USA (2010). DOI 10.1145/1816123.1816146
26. Hu, X., Downie, J.S.: When lyrics outperform audio for music mood classification: A feature analysis. In: Proceedings of the 11th International Society for Music Information Retrieval Conference, ISMIR, pp. 619–624 (2010)
27. Hu, X., Downie, J.S., Ehmann, A.F.: Lyric text mining in music mood classification. In: Proceedings of the 10th International Society for Music Information Retrieval Conference, *ISMIR*, vol. 183, pp. 2–209 (2009). DOI 10.5281/zenodo.1416790
28. Hu, X., Li, F., Ng, T.D.J.: On the relationships between music-induced emotion and physiological signals. In: Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR, pp. 362–369 (2018). DOI 10.5281/zenodo.1492425
29. Huang, I.S., Lu, Y.H., Shafiq, M., Ali Laghari, A., Yadav, R.: A generative adversarial network model based on intelligent data analytics for music emotion recognition under IoT. Mobile Information Systems **2021**, 1–8 (2021). DOI 10.1155/2021/3561829
30. Huang, Q., Jansen, A., Lee, J., Ganti, R., Li, J.Y., Ellis, D.P.W.: MuLan: A joint embedding of music audio and natural language. In: Proceedings of the 23rd International Society for Music Information Retrieval Conference, ISMIR, pp. 559–566 (2022). DOI 10.5281/zenodo.7316724
31. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. International Conference on Learning Representations (2017). DOI 10.48550/arXiv.1412.6980. ArXiv:1412.6980 [cs]
32. Koelsch, S.: Brain correlates of music-evoked emotions. Nature Reviews Neuroscience **15**(3), 170–180 (2014). DOI 10.1038/nrn3666
33. Koelstra, S., Muhl, C., Soleymani, M., Lee, J.S., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A., Patras, I.: Deap: A database for emotion analysis using physiological signals. IEEE Transactions on Affective Computing **3**(1), 18–31 (2012). DOI 10.1109/T-AFFC.2011.15
34. Laurier, C., Grivolla, J., Herrera, P.: Multimodal music mood classification using audio and lyrics. In: 2008 7th International Conference on Machine Learning and Applications, p. 688–693 (2008). DOI 10.1109/ICMLA.2008.96
35. Laurier, C., Sordo, M., Serra, J., Herrera, P.: Music mood representations from social tags. In: Proceedings of the 10th International Society for Music Information Retrieval Conference, ISMIR, pp. 381–386 (2009). DOI 10.5281/zenodo.1415600
36. Li, T., Ogihara, M.: Detecting emotion in music. In: Proceedings of the 4th International Society for Music Information Retrieval Conference, ISMIR, pp. 1–2 (2003). DOI 10.5281/zenodo.1417293
37. Li, X., Tian, J., Xu, M., Ning, Y., Cai, L.: Dblstm-based multi-scale fusion for dynamic emotion prediction in music. In: 2016 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6. IEEE (2016). DOI 10.1109/ICME.2016.7552956
38. Lin, T., Wang, Y., Liu, X., Qiu, X.: A survey of transformers. AI Open **3**, 111–132 (2022). DOI https://doi.org/10.1016/j.aiopen.2022.10.001
39. Liu, H., Fang, Y., Huang, Q.: Music emotion recognition using a variant of recurrent neural network. In: Proceedings of the 2018 International Conference on Mathematics, Modeling, Simulation and Statistics Application (MMSSA), pp. 15–18. Atlantis Press (2019). DOI 10.2991/mmssa-18.2019.4
40. Liu, Y., Liu, Y., Zhao, Y., Hua, K.A.: What strikes the strings of your heart?—feature mining for music emotion analysis. IEEE Transactions on Affective Computing **6**(3), 247–260 (2015). DOI 10.1109/TAFFC.2015.2396151

41. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach (2019). DOI 10.48550/arXiv.1907.11692. ArXiv:1907.11692 [cs]
42. Lu, L., Liu, D., Zhang, H.J.: Automatic mood detection and tracking of music audio signals. IEEE Transactions on Audio, Speech, and Language Processing **14**(1), 5–18 (2006). DOI 10.1109/TSA.2005.860344
43. Ma, Y., Li, X., Xu, M., Jia, J., Cai, L.: Multi-scale context based attention for dynamic music emotion prediction. In: Proceedings of the 25th ACM international conference on Multimedia, pp. 1443–1450. ACM (2017). DOI 10.1145/3123266.3123408
44. Malheiro, R., Panda, R., Gomes, P., Paiva, R.P.: Emotionally-relevant features for classification and regression of music lyrics. IEEE Transactions on Affective Computing **9**(2), 240–254 (2016). DOI 10.1109/TAFFC.2016.2598569
45. Mathieu, B., Essid, S., Fillon, T., Prado, J., Richard, G.: Yaafe, an easy to use and efficient audio feature extraction software. In: Proceedings of the 11th International Society for Music Information Retrieval Conference, *ISMIR*, vol. 2010, pp. 441–446 (2010). DOI 10.5281/zenodo.1418321
46. McKay, C., Fujinaga, I., Depalle, P.: jaudio: A feature extraction library. In: Proceedings of the 6th International Conference on Music Information Retrieval, ISMIR, pp. 600–3 (2005). DOI 10.5281/zenodo.1416648
47. Orjesek, R., Jarina, R., Chmulik, M., Kuba, M.: Dnn based music emotion recognition from raw audio signal. In: 2019 29th International Conference Radioelektronika (RADIOELEKTRONIKA), pp. 1–4. IEEE (2019). DOI 10.1109/RADIOELEK.2019.8733572
48. Panda, R., Malheiro, R.M., Paiva, R.P.: Audio features for music emotion recognition: a survey. IEEE Transactions on Affective Computing (2020). DOI 10.1109/TAFFC.2020.3032373
49. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners
50. Russell, J.A.: A circumplex model of affect. Journal of Personality and Social Psychology **39**(6), 1161–1178 (1980). DOI 10.1037/h0077714
51. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter (2019). DOI 10.48550/arXiv.1910.01108. ArXiv:1910.01108 [cs.CL]
52. Shrestha, A., Mahmood, A.: Review of deep learning algorithms and architectures. IEEE Access **7**, 53040–53065 (2019). DOI 10.1109/ACCESS.2019.2912200
53. Soleymani, M., Aljanaki, A., Yang, Y.H., Caro, M.N., Eyben, F., Markov, K., Schuller, B.W., Veltkamp, R., Weninger, F., Wiering, F.: Emotional analysis of music: A comparison of methods. In: Proceedings of the 22nd ACM International Conference on Multimedia, pp. 1161–1164 (2014). DOI 10.1145/2647868.2655019
54. Soleymani, M., Caro, M.N., Schmidt, E.M., Sha, C.Y., Yang, Y.H.: 1000 songs for emotional analysis of music. In: Proceedings of the 2nd ACM International Workshop on Crowdsourcing for Multimedia, CrowdMM '13, p. 1–6. Association for Computing Machinery, New York, NY, USA (2013). DOI 10.1145/2506364.2506365
55. Tang, Q., Huang, Z., Zhou, H., Ye, P.: Effects of music therapy on depression: A meta-analysis of randomized controlled trials. PLOS ONE **15**(11), 1–23 (2020). DOI 10.1371/journal.pone.0240862
56. Tzanetakis, G., Cook, P.: Marsyas: A framework for audio analysis. Organised sound **4**(3), 169–175 (2000). DOI https://doi.org/10.1017/S1355771800003071
57. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, p. 6000–6010. Curran Associates Inc., Red Hook, NY, USA (2017). DOI arxiv.org/abs/1706.03762v5
58. Wang, J.C., Yang, Y.H., Chang, K., Wang, H.M., Jeng, S.K.: Exploring the relationship between categorical and dimensional emotion semantics of music. In: Proceedings of the 2nd International ACM Workshop on Music Information Retrieval with User-centered and Multimodal Strategies (MIRUM), pp. 63–68. ACM Press, Nara, Japan (2012). DOI 10.1145/2390848.2390865

59. Wang, J.C., Yang, Y.H., Chang, K., Wang, H.M., Jeng, S.K.: Exploring the relationship between categorical and dimensional emotion semantics of music. p. 63–68. ACM, Nara Japan (2012). DOI 10.1145/2390848.2390865
60. Warriner, A.B., Kuperman, V., Brysbaert, M.: Norms of valence, arousal, and dominance for 13,915 english lemmas. Behavior Research Methods **45**(4), 1191–1207 (2013). DOI 10.3758/s13428-012-0314-x
61. Weninger, F., Eyben, F., Schuller, B.: On-line continuous-time music mood regression with deep recurrent neural networks. In: 2014 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 5412–5416. IEEE (2014). DOI 10.1109/ICASSP.2014.6854637
62. Wu, T.L., Jeng, S.K.: Probabilistic estimation of a novel music emotion model. In: Proceedings of the 14th international conference on Advances in Multimedia Modeling, MMM'08, p. 487–497. Springer-Verlag, Berlin, Heidelberg (2008). DOI 10.1007/978-3-540-77409-9_46
63. Yang, D., Lee, W.: Disambiguating music emotion using software agents. In: Proceedings of the 5th annual meeting of the International Society for Music Information Retrieval, p. 6 (2004). DOI 10.5281/zenodo.1415271
64. Yang, Y.H., Chen, H.H.: Machine recognition of music emotion: A review. ACM Transactions on Intelligent Systems and Technology **3**(3), 1–30 (2012). DOI 10.1145/2168752.2168754
65. Yang, Y.H., Lin, Y.C., Cheng, H.T., Liao, I.B., Ho, Y.C., Chen, H.H.: Toward multi-modal music emotion classification. In: Proceedings of the 9th Pacific Rim Conference on Multimedia, pp. 70–79. Springer (2008). DOI 10.1007/978-3-540-89796-5_8
66. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., Le, Q.V.: Xlnet: Generalized autoregressive pretraining for language understanding (2020). DOI 10.48550/arXiv.1906.08237. ArXiv:1906.08237 [cs]
67. Zhang, K., Zhang, H., Li, S., Yang, C., Sun, L.: The PMEmo dataset for music emotion recognition. In: Proceedings of the 2018 International Conference on Multimedia Retrieval, p. 135–142. ACM, Yokohama Japan (2018). DOI 10.1145/3206025.3206037
68. Zhang, Y., Jiang, J., Xia, G., Dixon, S.: Interpreting song lyrics with an audio-informed pre-trained language model. In: Proceedings of the 23rd International Society for Music Information Retrieval Conference, pp. 19–26. ISMIR, Bengaluru, India (2022). DOI 10.5281/zenodo.7316584