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

Patrick J. Donnelly and **Aidan Beery**
Electrical Engineering and Computer Science
Oregon State University - Cascades
Bend, OR, USA
{patrick.donnelly, beerya}@oregonstate.edu

Abstract

The automatic prediction of emotional responses to music is a task of inherent interest to the field of music information retrieval. These efforts are often hindered by the absence of large datasets available for this task. In this work, we investigate the use of sentiment analysis on online social media conversations as an alternate data source to train computational models to predict the emotive responses to a piece of music. Using two datasets annotated with valence and arousal values, we create a corpus of social media commentary for these songs extracted from YouTube, Twitter, and Reddit. We evaluate our approach with transformer models to predict the affective values of the 2402 songs in our dataset. We achieve a moderate Pearson's correlation of 0.62 and 0.72 for valence and arousal, respectively, for discourse from YouTube. These promising results demonstrate that discourse about music may carry semantic information useful to making determinations about the music itself. Such an approach could potentially supplement music information retrieval systems to estimate emotion for pieces of music for which the audio is restricted by copyright or otherwise unavailable.

1 Introduction

The task of music emotion recognition employs computational methods to attempt to predict the emotions elicited by a listener while listening to a piece of music. Estimating the cultural average response of an audience to a song is of interest to the music information retrieval community. A system for automatic music emotion recognition would enable large music libraries to be rated for estimated emotive responses. Online music streaming platforms could then better tune their music recommendation algorithms by filtering by mood or emotion.

Although researchers have investigated many different approaches for music emotion recognition, such efforts have been hindered by the paucity of large datasets suited for this task. These datasets are expensive to create as they require multiple human listeners to manually annotate musical excerpts. Complicating matters, there is no standard definition of this task. Some studies consider four, six, or eight discrete labels used to approximate human emotion. More recent datasets favor the valence-arousal model which treats emotions as a set of continuous values in a multidimensional space (Russell, 1980).

Furthermore, most musical recordings are copyrighted and this usually precludes the release of audio data as part of the dataset. In an attempt to bypass this limitation, the Million Song Dataset[1] released pre-computed acoustic features instead of raw audio. However, this approach limited the ability of researchers to explore certain algorithms or discover innovative features. In the area of music emotion recognition, researchers have struggled with an apparent upper bound in the ability of low-level acoustic features to predict human affective responses to music (Panda et al., 2020). Some researchers have turned to multimodal approaches, such as incorporating natural language analysis of song lyrics to make predictions about the song itself (Laurier et al., 2008).

We hypothesize that social media discussions surrounding a song contain semantic information which can be used to help predict a song's affective qualities. In this work, we present an approach for estimating affective responses to music based solely on commentary from social media. We create datasets of social media discourse for the songs contained in two music emotion datasets. We then compare the efficacy of two popular transformer

[1] http://millionsongdataset.com/

models in the task of predicting affective responses to music trained solely on natural language without the use of signal processing or acoustic features. To our knowledge, this is the first attempt to estimate affective responses to music indirectly based on social media conversations.

2 Background and Related Work

In this section we briefly review some of the approaches for music emotion recognition. We also describe the transformer architectures that we employ in our experiments.

2.1 Acoustic Features

Traditionally, approaches for automatic music emotion recognition have relied on learning information from acoustic features derived from the raw audio of a song. One early approach tasked domain experts to annotate 250 pieces from the Classical repertoire with four broad categories: *contentment*, *depression*, *exuberance*, and *anxiety*. The authors achieved a classification accuracy of 86.3% against expert ratings by training a Gaussian mixture model on acoustic and temporal features (Lu et al., 2006).

Another such study crowdsourced online annotations for 30-second excerpts from film scores. Importantly, each of the 200 songs was tagged by multiple listeners, 28.2 annotators on average, with one of eight mood categories: *sublime*, *sad*, *touching*, *easy*, *light*, *happy*, *exciting*, and *grand*. The authors empirically selected 29 acoustic features and trained a Support Vector Machine, reporting a cosine similarity of 0.73 between predicted and user-annotated labels (Wu and Jeng, 2008).

More recently, there have been efforts to develop mid-level features that may be better understood by a knowledgeable listener. These explainable features include perceptual concepts such as tonal stability, articulation, and rhythm. The authors trained a convolutional neural network using such features extracted from a set of 110 movie soundtracks to achieve a correlation of 0.71 compared to expert emotion annotations (Chowdhury et al., 2019). This approach encourages feature-importance analysis on these mid-level features, perhaps enabling future recommendation systems to provide context for its mood-based suggestions.

2.2 Incorporating Song Lyrics

Approaches using acoustic features alone have not yet proven entirely effective in predicating affective responses of music. The semantic gap between low-level audio features and human affective responses potentially limits the ability of systems using only raw acoustic information to predict emotional response to music (Panda et al., 2020). It is likely that music emotion prediction systems must be augmented with additional data in order to improve emotion recognition performance in any meaningful capacity (Yang and Chen, 2012).

Subsequently, researchers turned to a song's lyrics as a potential source of data to aid in the prediction of a song's emotional qualities. For 1000 pop songs, one study generated synthetic labels by comparing the similarity of Last.FM[2] tags to one of four mood category descriptors (*angry*, *happy*, *sad*, *relaxed*) using the WordNet[3] database. The authors then were able to predict mood categories with 62.5% accuracy using only lyrics, compared to their baseline accuracy of 89.8% using acoustic features. When the authors combined acoustic and lyric features, they improved classification accuracy to 92.4% (Laurier et al., 2008).

Another study reported that their lyric-only model (63.7%) outperformed its audio-based counterpart (57.9%) (Hu and Downie, 2010). A multimodal model using both feature sets marginally increased performance to 63.7% using a dataset covering 18 mood categories.

2.3 Direct Prediction from Lyrics

Recently, investigators have explored emotion recognition models based only on the text in the lyrics. One such study determined the valence and arousal values of individual words in the lyrics using established word lists. These values were then aggregated to create a song-level prediction of valence and arousal. The authors achieved a 74.25% classification accuracy relative to the All Music Guide[4] mood tags (Cano and Morisio, 2017).

Agrawal et al. applied a transformer approach based solely on song lyrics to achieve a 94.78% classification accuracy on a large dataset of lyrics and four emotion categories (Agrawal et al., 2021). This promising result demonstrates the ability to extract meaningful semantic information from music lyrics without the need for acoustic analysis.

[2] https://www.last.fm/
[3] https://wordnet.princeton.edu/
[4] https://www.allmusic.com/

2.4 Transformers

Transformers are deep learning models based on the principle of self-attention. This mechanism allows each token in an input to be weighted based on the context provided by surrounding tokens in order to capture an internal representation of the dependencies between elements. First introduced in 2017 (Vaswani et al., 2017), this architecture has proved especially successful with natural language processing tasks. More recently, transformer models have been adapted for emotion recognition of natural language (Chiorrini et al., 2021). Transformer models have also been recently applied in the area of music mood categorization, using lyrics as model input (Agrawal et al., 2021).

Bi-directional Encoder Representations from Transformers, or BERT (Devlin et al., 2019), is a popular transformer model for natural language understanding. This model comes pre-trained on a large dataset of English literature and Wikipedia articles. Although pre-training allows BERT-like models to be fine-tuned relatively quickly, training can still require immense compute resources, especially in the case of many NLP tasks where datasets can be quite large.

The RoBERTa model improves upon the original BERT model, adding additional model parameters and increasing the size of the training dataset by an order of magnitude (Liu et al., 2019). Although RoBERTa is able to exceed BERT's performance on many benchmark NLP tasks, this performance comes at the cost of significantly greater resources required for model training.

In an alternate approach, the DistilBERT model aims to lower the computational cost of training transformer models on large datasets by reducing the size of the model, and thereby significantly improving training and inference times (Sanh et al., 2019). The DistilBERT model reduces the number of model parameters by almost a factor of two while retaining competitive performance when compared to BERT.

In this work, we compare RoBERTa and DistilBERT models in the task of predicting emotion of songs directly from social media discussions.

3 Datasets

In this section we describe our selection of songs to consider in our experiment. We then detail our procedure for collecting musical discourse from social media platforms.

3.1 Music Emotion Datasets

Although music emotion recognition has been of particular interest in recent years in the field of music information retrieval, research is limited by a lack of available datasets suited for this task. These datasets require listeners to annotate musical excerpts. These experiments must be carefully controlled and usually occur in a lab environment or using an online crowdsourcing platform. Furthermore, these studies must employ large sample sizes, since the interpretation of music is highly subjective. We identified only four such datasets that provide continuous affective measurements in the valence-arousal space.

The DEAM dataset consists of 1,803 songs selected from royalty-free platforms and are songs likely unknown to the participant (Aljanaki and Soleymani, 2018). Listeners provided continuous annotations over the duration of the 45-second excerpt. Unlike other datasets, DEAM is able to provide the accompanying audio for each song, since these are not restricted by copyright. However, this also meant that these songs are relatively unknown. We were unable to consider this dataset because of an insufficient presence of social media commentary.

The Deezer dataset is a large set of 18,644 songs with synthetically generated affective annotations (Delbouys et al., 2018). These annotations were created with affective modeling based on the song's set of user tags on the website Last.FM. Although the large size of the Deezer dataset makes it a potentially valuable tool in this area of research, we exclude its consideration here as its emotion labels were estimated from natural language rather than human annotation.

In this work, we consider two datasets of songs with valence-arousal annotations. The first is the AMG1608 dataset, 1608 songs selected from the All Music Guide (Chen et al., 2015). In a crowdsourced task, listeners annotated 30-second excerpts. The study employed 665 annotators and achieved between 15 and 32 annotations per song. The second, the PmEmo dataset, contains annotations for 794 songs selected for their popularity on record industry charts (Zhang et al., 2018). The study recruited 457 undergraduate students to annotate 30-second excerpts. Because this study took place in a controlled lab setting, the authors also collected measurements of electrodermal activity.

Each of these datasets provides an artist name,

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

Table 1: Comparison of the music emotion datasets

song title, and accompanying valence and arousal labels for each song. However, the scale used in each approach varies, as shown in Table 1. These differences reflect the differences in the methodology used in the data collection. We scale these values to $[-1, 1]$ for use in this study, but we concede that these differences limit the utility of cross-dataset comparisons. We show the distribution of valence and arousal labels as Figure 1.

From these two datasets, we extract the artist names and song titles to be used in our queries. In total, we consider 2402 songs, however duplicates were not removed, so that each dataset can be evaluated independently of one another.

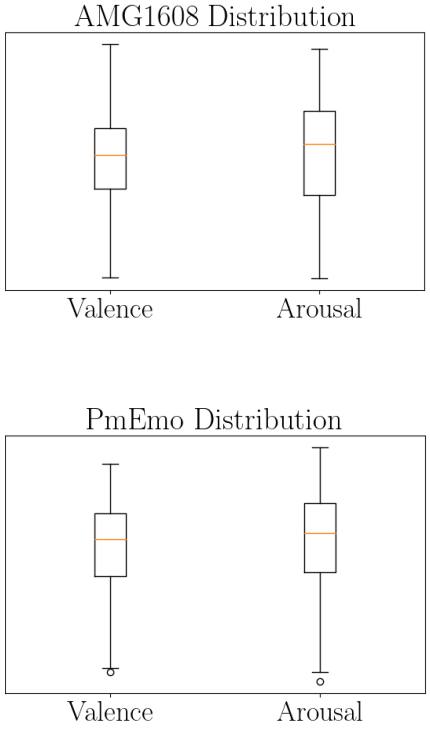

Figure 1: Box-and-whisker plots for the distribution of music emotion dataset labels in each dataset.

3.2 Social Media Data Collection

To explore the use of social media conversations as a feature space for music emotion prediction, we first must create a dataset of online discourse. We collect comment threads from social media for the songs featured in our two music emotion datasets: AMG1608 (Chen et al., 2015) and PmEmo (Zhang et al., 2018). Reddit, Twitter, and YouTube are large, popular social media platforms with active music subcultures, where individuals often converse about artists, songs, and concerts. In the case of YouTube, many use it as a platform to share and listen to music as well.

We collected our data over the course of two months in late 2021, harvesting all relevant commentary posted to date. For each song in our dataset, we query the platform for the artist name and track title. For YouTube and Reddit we extract the 10 highest rated submissions, based on likes and upvotes, respectively. We include all nested comments that appear as a response to a top-level comment. As a platform focused on short text posts, Twitter posts differ from the other two sites. Instead of a traditional reply chain, users retweet a post while potentially adding optional commentary. To avoid duplicating comments, we instead retrieve the top 100 tweets referencing the given song, excluding retweets. If a query for a song yields no submissions, we exclude that song from our dataset. In Table 2 we summarize our dataset of retrieved comments for each social media source.

Across both datasets, YouTube achieved the highest retrieval rates, finding more than 95% of the songs. This shows that YouTube supports a robust community for music-related discourse. The song retrieval rate for Reddit was also high, succeeding in finding at least 86% of the songs in either dataset. The retrieval rate for Twitter is notably lower, finding only 43% and 51% of songs in the two datasets. Because we could not find discussion of many songs, we conclude that Twitter is a less active medium to discuss opinions about music.

We show the distribution of retrieved comments as Figure 2. Overall we found more comments per song for songs in the PmEmo dataset across these three platforms than songs from AMG1608. This likely reflects the popularity of songs on record industry charts. We also observe that comments on Reddit tend to contain more words than those on YouTube, indicating that these discussions are frequently longer and perhaps more detailed than similar conversations on YouTube. As expected, given the 280-character limit, Twitter conversations are much shorter.

		Songs		Comments			Words	
		n	yield	n	μ	σ	μ	σ
AMG1608	Reddit	1431	89%	129,722	80.7	154.3	2400.8	69.1
	YouTube	1592	99%	217,093	135.0	57.7	2128.7	33.66
	Twitter	822	51%	5726	3.6	7.9	51.1	14.5
PmEmo	Reddit	627	86%	103,398	136.6	218.7	3810.5	56.8
	YouTube	730	95%	121,546	160.6	63.9	2172.1	44.1
	Twitter	331	43%	2699	3.6	7.3	46.0	15.2

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

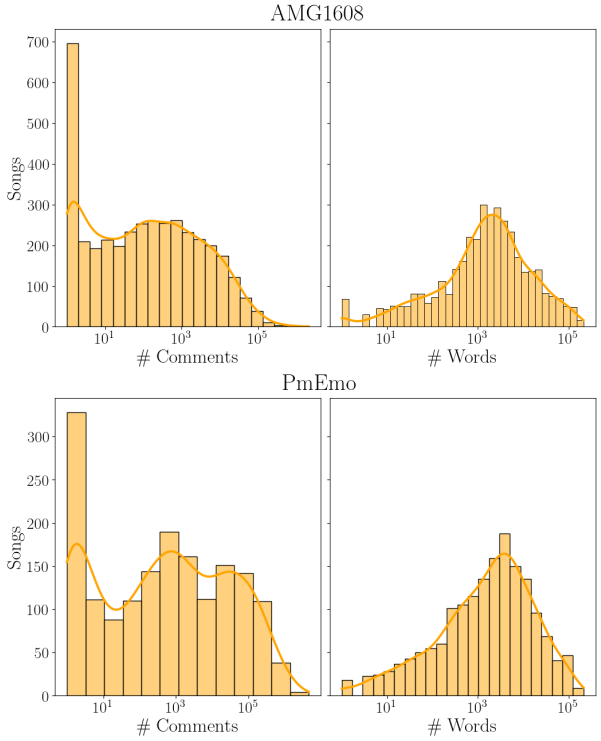

Figure 2: Comment and word distributions of our discourse datasets.

4 Experiment and Results

To test the utility of social media discourse towards the prediction of emotion in music, we conduct a deep learning experiment. We compare two powerful pre-trained transformer models for natural language understanding – DistilBERT (Sanh et al., 2019) and RoBERTa (Liu et al., 2019). In this section, we describe our model architecture, explain our experimental design, and present our results.

4.1 Model Architecture

Our model architecture consists of a pre-trained transformer model augmented with a densely connected neural network to predict regression targets from the last hidden state of the transformer, referred to as the regression head. We use the `TFDistilBertBase` and `TFRobertaBase` model implementations provided by the Huggingface deep natural language processing library[5].

Each input to the model consists of one text comment, with corresponding music valence and arousal labels. Sentences are tokenized using Huggingface's `TokenizerFast` library. We use the default input size of 128. Comments longer than 128 words will be truncated, and comments shorter than this sequence length will be right-padded with 0-tokens. For each model, we use the default language model architecture: six layers and twelve self-attention heads in the case of DistilBERT, and twelve layers with twelve self-attention heads for RoBERTa.

Deep learning models, such as transformers, naturally support multi-target regression through the use of a set of output nodes. This approach allows our model to predict valence and arousal as co-dependent values instead of independent labels as has often been done in prior approaches. As a regression head for each pre-trained transformer model, we append two fully connected layers and an output layer of two nodes, representing valence and arousal. We use mean-squared error as our loss function and a learning rate of 1×10^{-5}.

4.2 Experimental Design

We randomly partition our dataset into training (0.70), validation (0.15), and test (0.15) sets. We split our dataset at the song-level, rather than the comment-level to prevent potential information leakage from our test set. Valence and arousal labels are normalized and scaled to [0, 1].

We filtered the raw social media comments to remove URLs and HTML tags. Since transformer

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

models are pre-trained on large corpora of English text, they expect the input to adhere to standard grammatical structure. Therefore we do not filter stop-words or words with neutral sentiment.

The model outputs consist of a valence and arousal prediction for each comment. To aggregate a prediction at the song-level, we take the mean of the predictions across all comments for a particular song. We evaluate our models' performance using the Pearson's correlation between our predictions and ground truth values for valence and arousal.

4.3 Results

We begin by comparing the RoBERTa and DistilBERT models. We train each model using the combined social media commentary from Reddit, Twitter, and YouTube. We considered any song in the AMG1608 and PmEmo datasets as long as they are included in at least one of the three social media sources.

4.3.1 Model Comparison

BERT-like models are known to require minimal additional task-specific training due to their pre-trained nature (Devlin et al., 2019). In preliminary experiments we trained each model for 10 epochs and observed that our models converged between one and three epochs, depending on the dataset. In order to compare the performance of these two models while controlling for overfitting, we train each model for two epochs. We compare the results for each model and dataset in Table 3.

		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 3: Comparison of DistilBERT and RoBERTa performance after two epochs of training.

The performance between the two models is comparable and the differences are not statistically significant. However, the difference between computational cost for these models is considerable. In our experiments, the DistilBERT model completed training in less than half the runtime as RoBERTa. Because the models are comparable in predictive performance, we use DistilBERT in our subsequent experiments.

4.3.2 Source Comparison

Next, we train individual models for each social media source, Reddit, Twitter, and YouTube. Because not every song was found on each platform, the number of songs used to train each model varies (see Table 1). We compare these three source-specific models with another model that combines comments from all three sources. We show the results of this experiment for the AMG1608 song list using DistilBERT as Table 4.

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 4: Results of DistilBERT trained for two epochs for each social media source and song list.

We observe the highest overall performance on the YouTube subset achieving valence and arousal correlations of 0.62 and 0.72, respectively, on the AMG1608 dataset. Across both datasets the YouTube model tends to outperform the Reddit model, even though both data sources contain a comparably large number of comments. Although we matched fewer songs and collected fewer comments from Twitter, we still observe weak correlations between Twitter discourse and the song's valence and arousal annotations.

In addition to the source-specific models, we also examined the performance of the combined model. For the AMG1608 song list, the combined model demonstrated improvement over either the Reddit or Twitter model alone. However, aggregating all sources together reduced performance compared to the YouTube model alone by 21% for valence and 11% for arousal. Conversely, we observe the combined model outperformed any of the three source-specific models for the PMEmo songs.

Ultimately, these conflicting trends likely reflect the differences between the specific songs present in the two datasets, rather than differences in utility between the social media sources. This again underscores the field's need for much larger and

diverse datasets of music annotated with human affective responses.

5 Discussion

In this work, we present a novel approach to estimate the emotional qualities of a song solely through analysis of discussion of that song on social media. We create large datasets of conversations discussing music from three social media platforms. We train natural language transformer models to predict affective measurements of the song, directly from this discourse alone. Overall, we observe moderate correlations between our predictions across the three social media sources. These results indicate that the semantic information embedded in these comments can potentially be used to help predict affective responses to music.

5.1 Limitations and Future Work

We found that the distributions of our model's predictions tend to cluster closely to the center of the valence-arousal space. We show the distribution of our predictions and the actual value as Table 3.

We hypothesize that this occurs for two reasons. First, our unfiltered data may be too noisy for meaningful sentiment analysis at scale without some initial filtering of the comments. As future work, we will investigate approaches to clean the dataset while managing selection bias. We will consider dropping comments that may have adverse effects on our model performance, such as those of an insufficient length, those containing a low or negative score, or those generated by bots. We will also investigate dropping any comments which do not contain an affective word, using established affective word lists.

Secondly, we predict values for each comment and aggregate these comment-level predictions to estimate a value for the entire song. This approach was convenient to facilitate our exploration of existing pre-trained transformer architectures for this task. However, this aggregation risks losing valuable semantic information. For example, the sentiment contained within one comment may be cancelled by another with opposing sentiment, reducing them to an average neutral sentiment. As future work, we will investigate model architectures which may allow us to better retain inter-comment dependencies, such as Relation-Aware Transformers (Wang et al., 2021). We will also explore new architectures that support longer input sequences, such as the `xl-net` model (Yang et al., 2020) that has been recently applied to music mood classification from lyric analysis (Agrawal et al., 2021).

Our model is bounded by the requirement of a sufficient corpus of social media conversation pertaining to a song. This restricts this approach's efficacy in cases of newly released music or niche genres. An acoustic or lyrical approach, in contrast, would handle these scenarios equally to more popular song examples. In future experiments we will compare the performance of multimodal music emotion recognition systems when augmented with a social media input.

All our models trained exclusively on Twitter data performed poorly compared to the other source-specific models. However, our data collection method differed between social media platforms. We intend to repeat our Twitter data collection process in order to retrieve far more comments than available in this work. Also, we will revise our data-mining approach to include responses while explicitly filtering out reduplicated text. Despite these improvements, the combination of comment length restrictions and low yield rates for mentions of a song on the platform lead us to anticipate finding less available data on Twitter compared to YouTube or Reddit.

Additionally, we will explore other potential sources for social media commentary. For example, the community annotated tags on the site Last.FM have been used to generate features for music emotion recognition (Bischoff et al., 2009; Delbouys et al., 2018). Last.FM has recently added "Shouts", which allows users to post free-form comments in response to a song. To our knowledge no existing work has attempted to use sentiment analysis on Last.FM conversations for music emotion prediction. We will investigate commentary on the website SoundCloud[6] as well. SoundCloud is unique in that it associates posts with specific timestamps in the recording. This temporal information could be useful in determining changes of sentiment over the course of a piece of music. As we continue to collect additional data, we plan to make our dataset publicly accessible to facilitate further research into the use of social media commentary for music emotion recognition.

[6] https://soundcloud.com/

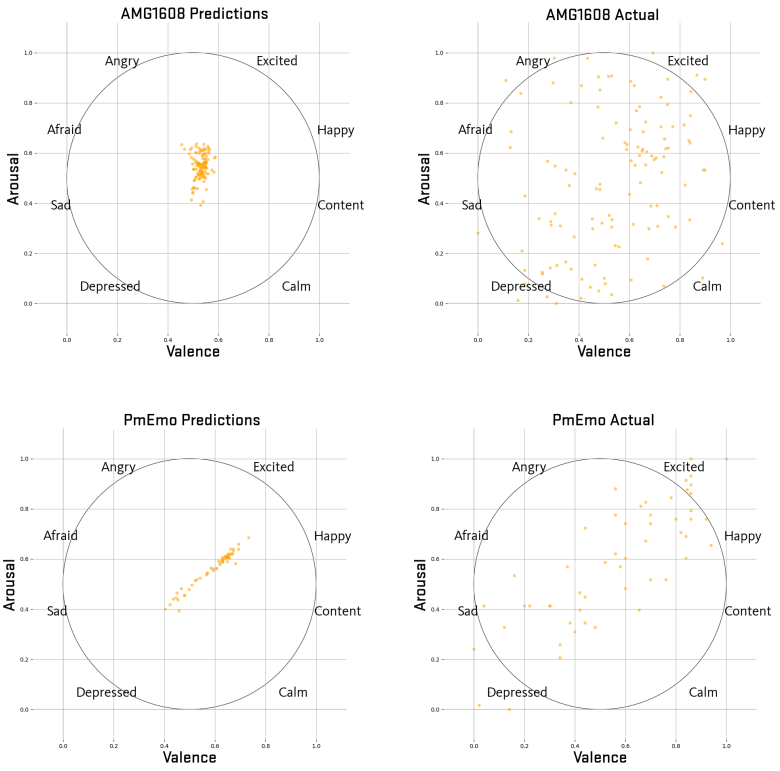

Figure 3: Distribution of DistilBERT predictions on songs in our test set for AMG1608 and PmEmo.

5.2 Contributions

In this work we assess the feasibility of predicting music emotion indirectly from social media discourse. By leveraging freely available social media commentary, we explore alternate modalities to make determinations about a musical affect when the raw waveform may not be available.

We create a novel dataset of conversations related to music from Reddit, YouTube, and Twitter. These comments correspond to the songs annotated in two music emotion datasets frequently cited in the literature. This correspondence allows comparison of the results of our supervised deep learning approach with human annotated labels of musical affect as well as to existing audio-based methods for estimation of musical affect.

Our results demonstrate that the conversations from social media platforms like Reddit, YouTube, and Twitter do contain semantic information which may be relevant to the task of music affect prediction. Although these correlations are moderate at best, they show the potential utility of this approach in music emotion recognition, especially if employed in a multimodal system that also compares audio and lyric-derived features. To our knowledge, this is the first approach to predict valence and arousal of musical songs using only conversational information from social media platforms.

5.3 Conclusion

Research investigating the automatic detection of the emotional qualities of music is often hindered by the absence of large-scale datasets annotated with affective responses to music. Such datasets are difficult to create and copyright concerns often limit the release of the raw audio needed by many machine learning approaches. Motivated by the use of song lyrics to predict emotion in music, in this work we explore the novel task of leveraging social media discourse to predict affective responses to music. We trained natural language transformer models using only discourse from three social media sites to predict the valence and arousal of pieces of music. We found moderate correlations between discourse about songs on Reddit, Twitter, and YouTube and the human annotated values of affective responses to those songs. Therefore, it is possible to predict the affective qualities of some songs directly from online conversations.

References

Yudhik Agrawal, Ramaguru Guru Ravi Shanker, and Vinoo Alluri. 2021. Transformer-based approach towards music emotion recognition from lyrics. *Advances in Information Retrieval. ECIR 2021*, 12657:167–175.

Anna Aljanaki and Mohammad Soleymani. 2018. A data-driven approach to mid-level perceptual musical feature modeling. In *Proceedings of the 19th International Society for Music Information Retrieval Conference*, ISMIR, pages 615–621.

Kerstin Bischoff, Claudiu S Firan, Raluca Paiu, Wolfgang Nejdl, Cyril Laurier, and Mohamed Sordo. 2009. Music mood and theme classification - a hybrid approach. In *Proceedings of the 10th International Society for Music Information Retrieval Conference*, ISMIR, pages 657–662.

Erion Cano and Maurizio Morisio. 2017. Moodylyrics: A sentiment annotated lyrics dataset. In *Proceedings of the 2017 International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence*, page 118–124. ACM.

Yu-An Chen, Yi-Hsuan Yang, Ju-Chiang Wang, and Homer Chen. 2015. The AMG1608 dataset for music emotion recognition. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing*, page 693–697. IEEE.

Andrea Chiorrini, Alex Mircoli, Claudia Diamantini, and Domenico Potena. 2021. Emotion and sentiment analysis of tweets using BERT. In *Proceedings of the EDBT/ICDT 2021 Joint Conference*.

Shreyan Chowdhury, Andreu Vall, Verena Haunschmid, and Gerhard Widmer. 2019. Towards explainable music emotion recognition: The route via mid-level features. In *Proceedings of the 20th International Society for Music Information Retrieval Conference*, ISMIR, pages 237–243.

Remi Delbouys, Romain Hennequin, Francesco Piccoli, Jimena Royo-Letelier, and Manuel Moussallam. 2018. 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, pages 370–375.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, pages 4171 – 4186.

Xiao Hu and J. Stephen Downie. 2010. Improving mood classification in music digital libraries by combining lyrics and audio. In *Proceedings of the 10th annual joint conference on Digital libraries*, page 159–168. ACM.

Cyril Laurier, Jens Grivolla, and Perfecto Herrera. 2008. Multimodal music mood classification using audio and lyrics. In *2008 Seventh International Conference on Machine Learning and Applications*, page 688–693.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *arXiv:1907.11692*, 1907(1907.11692).

Lie Lu, D. Liu, and Hong-Jiang Zhang. 2006. Automatic mood detection and tracking of music audio signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1):5–18.

Renato Panda, Ricardo Manuel Malheiro, and Rui Pedro Paiva. 2020. Audio features for music emotion recognition: a survey. *IEEE Transactions on Affective Computing*, pages 1–20.

James A. Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv:1910.01108*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010.

Bailin Wang, Richard Shin, Xiaodong Liu, Oleksandr Polozov, and Matthew Richardson. 2021. Rat-sql: Relation-aware schema encoding and linking for text-to-sql parsers. *arXiv:1911.04942*.

Tien-Lin Wu and Shyh-Kang Jeng. 2008. Probabilistic estimation of a novel music emotion model. In *Proceedings of the 14th international conference on Advances in multimedia modeling*, MMM'08, page 487–497. Springer-Verlag.

Yi-Hsuan Yang and Homer H. Chen. 2012. Machine recognition of music emotion: A review. *ACM Transactions on Intelligent Systems and Technology*, 3(3):1–30.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2020. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Kejun Zhang, Hui Zhang, Simeng Li, Changyuan Yang, and Lingyun Sun. 2018. The PMEmo dataset for music emotion recognition. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, page 135–142. ACM.