The Contribution of Lyrics and Acoustics to Collaborative Understanding of Mood

Shahrzad Naseri[1*], Sravana Reddy[2*], Joana Correia[3], Jussi Karlgren[3], Rosie Jones[3]

[1] University of Massachusetts Amherst
[2] ASAPP
[3] Spotify Research
shnaseri@cs.umass.edu, sravana.reddy@gmail.com, {joanac, jkarlgren, rjones}@spotify.com

Abstract

In this work, we study the association between song lyrics and mood through a data-driven analysis. Our data set consists of nearly one million songs, with song-mood associations derived from user playlists on the Spotify streaming platform. We take advantage of state-of-the-art natural language processing models based on transformers to learn the association between the lyrics and moods. We find that a pre-trained transformer-based language model in a zero-shot setting – i.e., out of the box with no further training on our data – is powerful for capturing song-mood associations. Moreover, we illustrate that training on song-mood associations results in a highly accurate model that predicts these associations for unseen songs. Furthermore, by comparing the prediction of a model using lyrics with one using acoustic features, we observe that the relative importance of lyrics for mood prediction in comparison with acoustics depends on the specific mood. Finally, we verify if the models are capturing the same information about lyrics and acoustics as humans through an annotation task where we obtain human judgments of mood-song relevance based on lyrics and acoustics.

Introduction

Lyrics are important for the musical experience, providing us with the rich stories and messages that artists want to convey through their music. However, the perceived mood of a song may not stem from its lyrics alone. Take, for example, the song *Cardigan* by Taylor Swift:

> *'Cause I knew you*
> *Steppin' on the last train*
> *Marked me like a bloodstain, I*
> *I knew you*
> *Tried to change the ending*
> *Peter losing Wendy, I*
> *I knew you*

The lyrics are about the end of a relationship, and imply moods of sadness, longing, and heartbreak. However, the acoustic mood of the song with its familiar chord progressions and somewhat high tempo is calm and upbeat. This

example is not an exception; in fact, a recent analysis of the lyrics and acoustics of popular music (Interiano et al. 2018) identifies a trend where songs have been getting sadder lyrically in the last three decades, but also more 'danceable' and 'relaxed.'

In this paper, we investigate the association between song lyrics and *moods* – terms that describe affectual qualities of a song – and conduct a data driven analysis using state of the art natural language processing models to compare how lyrics contribute to the understanding of mood as defined collaboratively by the playlisting behavior of users of the Spotify music streaming platform.

Previous studies in the psychology of music have shown that acoustics and lyrics play different roles in listener perceptions, and that these roles depend on the specific mood. One study (Ali and Peynircioğlu 2006) showed that lyrics detract from emotion in happy and calm music in their participants, but enhance emotion in sad and angry music. Results from an fMRI study (Brattico et al. 2011) lend support to the hypothesis that lyrics are more important for the perception of sad emotions than acoustics, but that acoustics are of primary importance for the perception of happy emotions.

One motivation for our project is to tackle the above questions of music perception outside the laboratory, from the perspective of a large-scale data set of music originally tagged with moods by listeners on the Spotify music streaming platform. The second motivation is to explore how machine learning models might be designed to automatically associate moods with songs in order to enable listeners to search and discover music. While previous work (Zaanen and Kanters 2010; Laurier, Grivolla, and Herrera 2008; Hu and Downie 2010) has explored this modeling problem, our study uses a data set that is orders of magnitude larger. In addition, while previous work largely relies on bag-of-words models, we use state of the art natural language processing models based on the transformer architecture (Vaswani et al. 2017) to better capture the semantic nuances of lyrics in order to learn the associations between lyrics and moods.

We break down our overarching research question, *"How much do lyrics and acoustics of a song each contribute to understanding of the song's mood?"*, into the following subquestions:

RQ1 What can lyrics tell us about moods with *no* training on lyric-mood associations?

RQ2 Can training a lyrics-based model on listener-generated mood tags produce accurate mood associations?

RQ3 How much do lyrics contribute to moods compared to acoustics?

RQ4 Do models capture the same information about lyrics and acoustics as humans?

In order to answer these questions, we take advantage of a transformer model to represent lyrics as well as bag-of-words representations. The results show that using a zero-shot model (i.e., a model pretrained on web data with no training on our data set) is powerful in capturing the mood associations of songs. Training on listener-generated mood-song associations results in a model that captures these associations more accurately.

By comparing the prediction of models based on lyrics and acoustics, we find that the contribution of lyrics varies depending on the specific mood. We observe a similar result on conducting a manual annotation task where we elicit judgments on mood-song relevance based on lyrics and acoustics in isolation.

Mood in Music and Text: A Brief Survey

Music Psychology

The study of how humans connect moods and music has been explored through experiments in the fields of psychology and neuroscience. One study (Ali and Peynircioğlu 2006) finds that acoustics are more dominant than lyrics in eliciting emotions in study participants, and that lyrics play a bigger role in perceived sad or angry music compared to music with positive moods. This is similar to the finding of an fMRI experiment (Brattico et al. 2011) showing that lyrics define moods for sad music whereas acoustics are of primary importance for happy moods. The relative unimportance of lyrics on mood compared to acoustics is corroborated by other studies (Sousou 1997), but contradicted by others: for example, one paper (Stratton and Zalanowski 1994) finds that sad lyrics with upbeat acoustics elicits negative emotion, and yet another contradicting paper (Mori and Iwanaga 2014) demonstrates that sad lyrics seem to enhance pleasant feelings induced by happy-sounding music compared the acoustics alone. Finally, it has been shown (Besson et al. 1998) that humans appear to process song melodies and lyrics independently, raising questions about how conflicting moods in the acoustics and lyrics are processed.

One should keep in mind that some of the contradicting results from these laboratory studies may be due to variations in experimental design, the selection of participants and the music, small sample sizes, and the time period or cultural landscape in which the studies were conducted. We believe that a computational lens on these questions using web-scale data, while not perfect, is a complementary angle to such works.

Prediction of Mood with Lyrics and Acoustics

The annual Annual Music Information Retrieval Evaluation eXchange (MIREX) introduced a music mood classification task in 2007 (Downie, Laurier, and Ehmann 2008). This task explicitly disallowed consideration of lyrics in classification or evaluation. Submitted models were found to have overall better classification performance using acoustics for mood clusters like {'wistful', 'brooding'} and {'volatile', 'fiery'} compared to clusters like {'rousing', 'confident'} and {'fun', 'cheerful'}. Laurier, Grivolla, and Herrera (2008) collect a 1000-song and 4-mood data set of song-mood associations by eliciting judgments from annotators, and build classifiers using bag-of-words and latent semantic analysis features for lyrics, and acoustic features like timbre, tempo, and pitch. They find that acoustics and lyrics combined improve prediction of 'happy' and 'sad' moods, but that the performance is relatively saturated with acoustics alone for 'anger.' A follow-up study (Hu and Downie 2010) with a slightly larger data set of a few thousand songs and 18 moods finds that lyrics significantly outperform acoustics for prediction. A smaller scale and more detailed approach (Schmidt and Kim 2011) models the temporal dynamics of emotional approaches to songs.

Since annotations of mood for large collections of music are hard to obtain, McVicar, Freeman, and De Bie (2011) use unsupervised methods to correlate acoustic attributes and lyrics and find through canonical correlation analysis that the top correlation components correspond to mood, suggesting that lyrics and acoustic attributes in a song tend to have consistent moods overall.

Recently, there have been a body of works that applied deep neural network models to capture the association of mood/emotion and song by taking advantage of audio features (Saari et al. 2013; Panda 2019; Korzeniowski et al. 2020; Panda, Malheiro, and Paiva 2020; Medina, Beltrán, and Baldassarri 2020), lyrics features (Fell et al. 2019; Hrustanović, Kavšek, and Tkalčič 2021) as well as both lyrics and audio (Delbouys et al. 2018; Parisi et al. 2019; Wang, Syu, and Wongchaisuwat 2021; Bhattacharya and Kadambari 2018) features. Delbouys et al. classify mood of a song to either 'arousal' or 'valence' by utilizing a 100-dimensional word2vec embedding vector that is trained on 1.6 million lyrics in several different neural architectures such as GRU, LSTM, Convolutional Networks for their lyrics-based model. Further, they utilize audio mel-spectrogram as input to a convolutional neural network model. Parisi et al. show a comparison between text-based and audio-based deep learning classification models to classify the mood of a song to 5 discrete crowd-based adjectives: {'sad', 'joy', 'fear', 'anger', 'disgust'}.

Sentiment Analysis of Language

Some papers (Yang and Lee 2009; Zaanen and Kanters 2010; Fell and Sporleder 2014, etc) have used lyrics with bag-of-words based models to predict mood without comparison to acoustics. An analysis of proxies for linguistic creativity (Hu and Yu 2011) shows that sad or negative lyrics score higher in creativity than positive songs.

Outside of music, the detection of mood and affect in human language use more generally has been the object of systematic computational study since the first AAAI Symposium on the topic (Qu, Shanahan, and Wiebe 2004)

and traces its modern beginnings to the study of human emotional expression by Charles Darwin (1872) and others (James 1884, e.g.).

Sentiment analysis models have historically been based on bag of words classifiers and lexical look-ups, with some syntactical finesse to process e.g. negation or amplification (Pang and Lee 2008). This type of approach is deterministic and interpretable but at a cost for coverage. Deep learning approaches that can model contextual relationships have been showing improvements over traditional methods (Zhang, Wang, and Liu 2018). In particular, the use of pretrained transformer models such as BERT (Devlin et al. 2019) have seen success in recent years (Li et al. 2019; Yin, Meng, and Chang 2020). However, these models have not been applied to music or mood categories to the best of our knowledge.

Data

Our study analyzes the association of a song's lyrics with the set of terms describing its mood. Our palette of moods consists of 287 terms in English. This set of moods includes terms like "chill", "sad", "happy", "love', and "exciting". The moods are not limited to a specific part-of-speech, covering not only adjectives ("sad", "somber", etc), but also nouns ("motivation", "love", etc.) and verbs ("reminisce", "fantasize", etc.).

The association between a song and the mood is calculated using collaborative data (by "wisdom of the crowd"). More specifically, Spotify music streaming platform provides playlists of songs as well as enabling users to create their own playlists. The playlists have a name and an optional description. To calculate the association between song and mood, the Spotify music streaming platform starts from their collection of ($\approx$4 billion) playlists and filters down to those playlists that have words corresponding to the mood lexicon in their titles and/or descriptions. The co-occurrence between each mood and song is then computed. Finally, the association (Wang, Anderson, and McCurry 2021) between a mood and a song is calculated according to the Pointwise Mutual Information (PMI):

$$\text{PMI}(s, m) = \log \frac{p(s, m)}{p(s)p(m)} \quad (1)$$

where s and m are the songs and moods, respectively. $p(s, m)$ is the probability of co-occurrence of the song and mood. The PMI score shows how much more likely is the song s to co-occur with the mood m.

The final association score is a slight variation of the PMI score called Bayesian Normalized Pointwise Mutual Information (BNPMI) where instead of using empirical probability of the probability of a song given a mood, a conjugate beta prior (Schlaifer and Raiffa 1961) with parameters estimated by method of moments (Hall et al. 2005) is used. This change compensates for the rarer song and moods that *coincidentally* are co-occurring or not co-occurring.

BNPMI scores are between [-1,+1], where -1 represents negative association (the mood and song *never* co-occur), 0

	Total	Train	Test
song	955,109	716,331	238,778
(song, mood)	3,083,727	2,309,083	774,644

Table 1: Number of song and (song, mood) pairs.

of independence, and +1 of perfect co-occurrence. For more details about the BNPMI score, see the Appendix.

Associating songs and moods through terms in user playlists is not as precise as explicitly eliciting mood tags, but avoids biases in elicitation, and allows us to scale massively to a large number of songs, with the association collected from millions of diverse users. Our data set contains $\approx$ 955K songs. The lyrics of these songs are obtained from a commercial service[1] which precludes the sharing and distribution of these specific data as a separate collection. However, to give a sense of our dataset we provide 18 pairs of (song, mood) for 3 different songs in the Appendix. We break down the songs into train and test sets by reserving 75% and 25% of samples for train and test sets, respectively.

Moreover, in the process of creating the dataset we run into the duplicate (song, lyrics) tuples because of different spacing and new line insertions in the lyrics text. Since each song has a unique identifier, we only select one of the instances of duplicate tuples with the same song identifier.

For training and evaluation of our models, we label the association between a song and a mood by binning the BNPMI scores. More specifically, if the BNPMI score is greater than the threshold τ it represents a positive association and if it's lower than the $-\tau$ it represents a negative association. The values that fall between $-\tau$ and τ correspond to a neutral association. Figure 1 shows the histogram of BNPMI scores which is a normal distribution with mean equal to -0.0034 and standard deviation equal to 0.0861. In our experiments, we select the threshold $\tau = 0.1$, which the neutral association will fall within approximately one standard deviation of the mean.

Limiting the association only to positive and negative association results in $\approx$ 2 million (song, mood) pairs for training and $\approx$ 774K pairs for testing. The exact number of instances is shown in table 1. Moreover, we do not perform any negative sampling and we utilized all of the negative bnpmi score association in training our model.

Lastly, Figure 2 shows the frequency of top 20 mood descriptors with positive association to their corresponding song. "Chill" mood descriptors is the most frequent mood positively associated to songs.

Methodology and Experiments

In this section, we empirically investigate the answer to the overarching research question: "How much the lyrics and the acoustics of a song each contribute to understanding of the song's mood?".
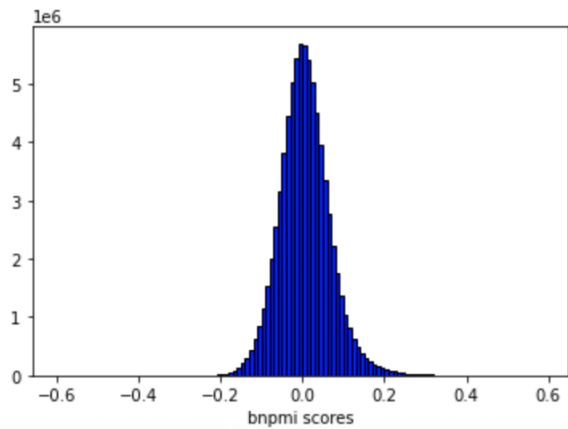
---

[1]https://www.musixmatch.com

Figure 1: BNPMI score distribution

RQ1. What Can Lyrics Tell us about Moods with *No* Training on Lyric-Mood Associations?

To answer this question, we formulate the problem as a zero-shot text classification task in which the model predicts the class of a text without having seen a single labeled example, instead using linguistic knowledge from a model pretrained on a different domain. Natural Language Inference (NLI) (Bowman et al. 2015), the task of determining whether or not a premise sentence semantically entails a hypothesis sentence, is used as a proxy to tackle the problem of zero-shot text classification of moods. This approach casts the target label into a sentence and judges whether or not the input text entails the target sentence.

We take advantage of the BART (Lewis et al. 2019) model that is trained for the NLI task on the Multi-genre Natural Language Inference (2017) corpus (Williams, Nangia, and Bowman 2018) as our model for the zero-shot text classification task, according to the approach proposed by Yin, Hay, and Roth (2019). The BART model fine-tuned on the MNLI corpus is a transformer-based model that has a classification head with 3 classes: Contradiction, Neutral, and Entailment.

Under the NLI formulation, lyrics are inputs and moods are targets. The BART model is used without any further training. We cast the mood terms into sentences using templates according to their parts of speech, with manual re-annotations for corner cases. Table 2 shows the examples of some moods and the corresponding sentences.

Following Yin, Hay, and Roth, the output for the Neutral class is ignored. To be more specific, the association between the lyrics and each mood is formulated as a binary classification task and the final probability of association is calculated according to the Contradiction and Entailment class probabilities.

We evaluate our zero-shot NLI approach against the binned association labels derived according to the BNPMI score (described in the Data section) in Table 3. We observe that the zero-shot NLI model has high precision, however it under-predicts the positive association between a song and mood, resulting in low recall.

We further illustrate the performance of the zero-shot classifier with the example song shown earlier: *Cardigan* by Taylor Swift. The listener-generated mood associations as characterized by high BNPMI scores include "heartbroken", "calm", "sad", "bittersweet", "vulnerable", and "obsessed". The classifier predicts all the above moods except "calm" to be associated with the song. Note that while our evaluation penalizes the classifier for missing "calm", it is in fact doing the right thing since that mood is not consistent with the lyrics, and is presumably derived from acoustic perceptions. RQ3 will address this aspect of the evaluations.

Ma et al. (2021) recently pointed out a few issues with the NLI approach for tackling the zero-shot text classification problem; in particular, they suggest that NLI is not a good proxy for text classification. They propose instead using models like BERT that are fine-tuned for the task of *next sentence prediction* (NSP), with the reasoning that NSP is a closer proxy of text classification than NLI.

Next sentence prediction is a binary classification task that given two sentences, predicts whether the second sentence follows the first input 'sentence'. As before, we use lyrics as the first sentence, and the moods cast as sentences as the second. We can see from the top row of Table 3 that the F1 score for the NSP-Zeroshot model is indeed higher than MNLI-Zeroshot. However, a closer examination of the confusion matrix shows that the model is biased toward predicting the positive class, resulting in lower precision.

Takeaways: Pretrained transformer models are much better than chance at predicting the association between song lyrics and moods, despite being trained on completely different tasks and domains. Next sentence prediction is more effective than natural language inference as a proxy for song-mood association prediction, but the latter shows higher precision. We hypothesize that fine-tuning the models on training data of song-mood associations will result in higher performance.

RQ2. Can Training a Lyrics-based Model on Listener-generated Mood Tags Produce Accurate Mood Associations?

To address this question, we take two modeling approaches. Following previous literature in mood predictions from lyrics, we represent the lyrics by a bag-of-words model. We also use a transformer-based model. We train both the bag-of-words and transformer-based models on the (lyrics, mood) pairs with the BNPMI-based binned association as the target labels.

Bag of Words (BoW) Lyrics are represented by the $tf.idf$ scores of their unigrams, estimated on the training set. For each mood, we train a binary logistic regression classifier, classifying whether the mood and the song are positively associated with each other.

Transformers We fine-tune the BART model pre-trained on the MNLI corpus ('fine-tuned NLI') as well as a BERT model trained on the next sentence prediction task (fine-tuned NSP). Similarly to the zero-shot learning setup in the previous section, the input is a pair of texts, with the model
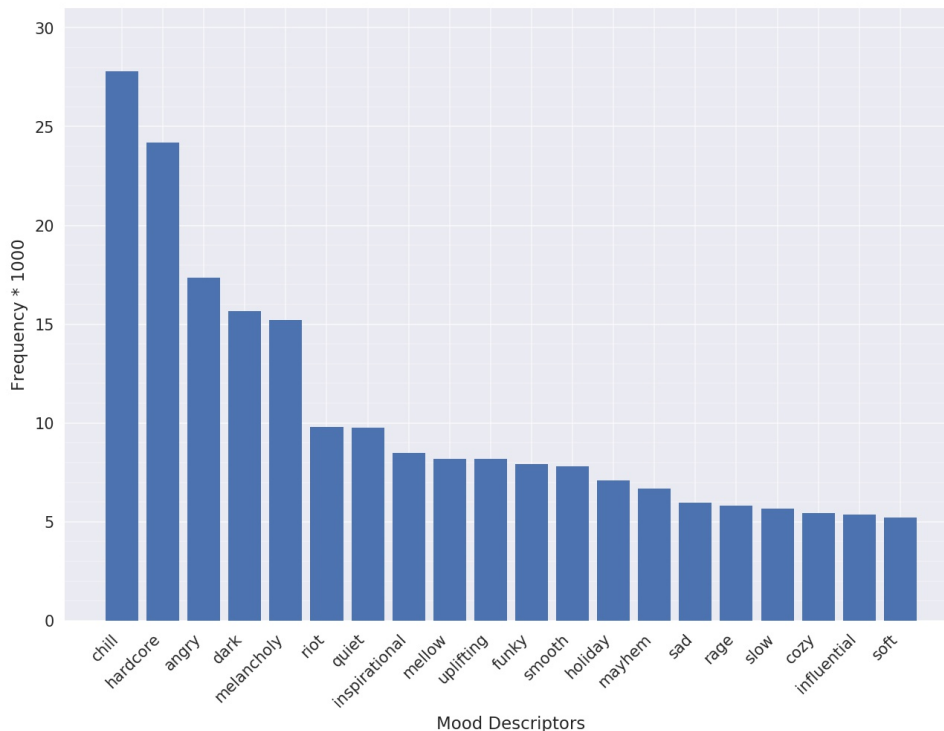
Figure 2: Frequency of top 20 mood descriptors with positive association to their corresponding song in the test split.

Mood	Part of Speech	Sentence
Chill	adjective	This is a **chill** song.
Heartbroken	adjective	This song is about being **heartbroken**.
Love	noun	This song is about **love**.
Fantasize	verb	This song makes you **fantasize**.

Table 2: Mood terms rewritten as sentences

is classifying whether the first text entails the second, while in the next sentence prediction the model predict if the second text would follow the first in a corpus. The models are trained on the BNPMI-based associations. Table 4 shows the label mappings used for each model.

We fine-tune the NLI model in two ways: excluding and including the neutral moods. Since the number of the neutral mood descriptors is high for each recording, we consider only one neutral mood descriptor for each recording to make the fine-tuning feasible (in the defined train set the total number of recordings with neutral descriptors is approximately 20M pairs).

The results for the BoW, fine-tuned NLI and fine-tuned NSP model are shown in the second row of table 3. We observe that fine-tuning shows large improvements in both precision and compared to the zero-shot approach. While the BoW model performs well with a F1 score of 91.74, the transformer models perform better, especially for recall. We also see that the difference between the fine-tuned NLI models (both w/o and w. neutral) and the fine-tuned NSP model

is negligible, illustrating that training on listener-generated song-mood associations provides both models with powerful signals that were missing in the zero-shot approach. Moreover, we observe fine-tuning with neutral descriptors decrease the performance across all measures, however this drop is negligible and therefore, throughout the rest of the paper NLI model refers to the NLI (w/o. Neutral) fine-tuning scheme.

Takeaways: Overall, fine-tuning on the training data of song-mood associations results in models with high precision and recall that can be valuable for predicting listener-generated moods of new songs, which could then be leveraged for conversational search and recommendation.

RQ3. How Much Do Lyrics Contribute to Moods Compared to Acoustics?

To understand the contribution of the acoustics features to a piece of music's mood, we use a logistic regression classifier with precomputed acoustics features of a song. We also build hybrid models which take advantage of both lyrics and

Modality	Approach	Lyrics Model	Precision	Recall	F1
Lyrics	Zero-shot Learning	NLI	83.14	43.95	57.50
		NSP	75.51	98.22	85.38
	Fine-tuned	BoW	93.85	90.35	91.74
		NLI (w/o. Neutral)	96.48	97.63	97.05
		NLI (w. Neutral)	95.90	96.79	96.34
		NSP	96.17	97.46	96.81
Acoustics	Fine-tuned	-	95.64	81.54	87.48
Hybrid	Fine-tuned	BoW	95.53	94.54	94.86
		NLI	96.57	97.93	97.24

Table 3: Lyrics-based, acoustics-based and hybrid models evaluated against the ground truth song-mood association.

Model	BNPMI Score	Association
NLI	[-1, -0.1]	Contradiction
	(-0.1, 0.1)	Neutral
	[0.1, 1]	Entailment
NSP	[-1, -0.1]	NotNextSentence
	[0.1, 1]	IsNextSentence

Table 4: Model targets derived from the BNPMI scores.

acoustics.

Acoustics The model represents each song by a set of numerical features corresponding to the acoustics of the song as provided by Spotify API. Table 5 shows the list of acoustic features. Similarly to the bag of words model, we train a binary logistic regression classifier for each mood to classify whether the mood and the song's acoustics are associated with each other.

The third row in Table 3 shows the results of this model. We observe that the acoustic model does better than the zero-shot approaches with respect to precision, and in the case of the NLI model with respect to recall as well; however, it is clearly outperformed by the fine-tuned lyrics models with respect to recall, where the lyrics are considerably better for coverage.

In the example of the *Cardigan* song, we find that the acoustics model predicts the mood "calm", which is associated with the song according to listener-generated playlist data, but is not predicted by the zero-shot lyrics classifier.

Hybrid We investigate two hybrid lyrics+acoustics models – representing a song based on both its lyrics and acoustics features. One uses the bag of words representation of the lyrics ('Hybrid-BoW') and the other uses the NLI transformer-based representation ('Hybrid-NLI'). It is worth mentioning since the difference between the performance of fine-tuned NSP and fine-tuned NLI models is negligible, we only select the NLI model to be used in our hybrid architecture for transformer-based representation of the lyrics.

Hybrid-BoW represents the lyrics and acoustics features by a concatenation of the $tf.idf$ vectors of lyrics and the acoustic features. It is trained the same way as the bag of words and acoustic models.

Hybrid-NLI uses the final layer before the classification layer of the BART transformer fine-tuned on the NLI task to represent lyrics. The acoustic features are fed into a multilayer perceptron (MLP) model to construct the acoustic representation. The lyrics and acoustic representations are concatenated, which result in the hybrid representation of the song. The hybrid representation is input to a classification head, similar to the classification head in the BART model fine-tuned on the sequence classification task[2]. Figure 3 shows the architecture of the Hybrid-NLI model.

We compare the results of the hybrid models to the lyrics-only and acoustics-only models in the final rows of Table 3. We observe that the Hybrid-BoW model has better recall than the acoustics model and is overall better than the fine-tuned BoW representation of the lyrics. The Hybrid-NLI model behaves similarly to the Hybrid-BoW model, outperforming acoustics-based and lyrics-based models in both recall and precision. The Hybrid-NLI model performance compared to the Hybrid-BoW model highlights the effectiveness of a transformer-based representation of the lyrics.

To better understand the behaviour of the lyrics-based, acoustics-based, and hybrid models for different moods, we compare the performance of these models for two selected moods, "chill" and "love", in Table 6. We observe that the fine-tuned NLI lyrics-based model outperforms the zero-shot and fine-tuned bag of words lyrics-based and acoustics-based models for both cases. Notably, lyrics-based models consistently perform worse for "chill" than they do for "love"; for "love", the lyrics alone do better than the acoustics alone. We can see the Hybrid-NLI model does best of all models for "chill" and has competitive performance with the fine-tuned NLI model for "love".

Takeaways: These results show that some listener-generated mood tags, e.g. "chill", are better predicted from acoustics than from lyrics, in contrast to other mood tags such as "love". Table 6 suggests that listeners may be paying differential attention to lyrics and acoustics when describing playlists with different moods, and that the differ-

---

[2]https://huggingface.co/transformers/v2.11.0/_modules/transformers/modeling_bart.html\#BartForSequenceClassification

Acoustic Features	acousticness, bounciness, beat strength, danceability, energy, flatness, instrumentalness, liveness, loudness, longest silence ratio, mechanism, organism, runnability, speechiness, tempo, valence, mean of dynamic range

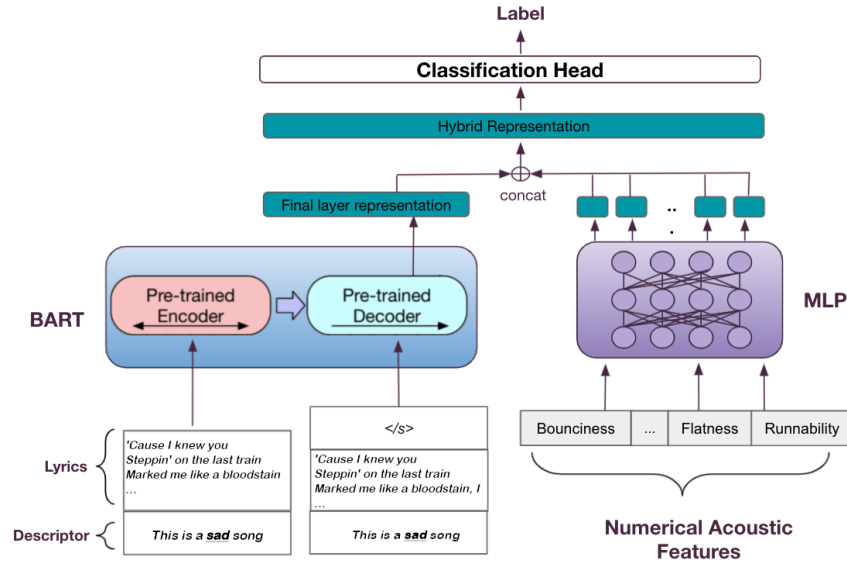Table 5: The list of numerical acoustic features provided by the Spotify API.

Figure 3: Architecture of the Hybrid NLI model.

ence between the semantics of "love" and "chill" plays a role here: the latter term is arguably less specific than the former and can thus be assumed to be realised with a larger variety of expression in lyrics, some of which may have to with temperature and meteorological considerations rather than mood and emotion. This polysemy is reduced by the introduction of acoustic features and through fine-tuning the model, which is evident in the scores given in the table.

Next, in RQ4, we will show that these results are consistent with human judgments in an annotation study.

RQ4. Do Models Capture the Same Information about Lyrics and Acoustics as Humans?

While the song-mood associations reflected by BNPMI scores are derived from listener playlisting data, there is no way to decompose these scores into associations provided by lyrics, by acoustics, and other factors.

We therefore conduct a human annotation task on 101 songs and 302 (song, mood) pairs, which are selected randomly, using 3 groups of annotators. The annotation task includes two subtasks:

- **Lyrics Annotation:** Judge whether a mood is relevant to a song only by reading the lyrics.
- **Acoustics Annotation:** Judge whether a mood is relevant to a song only by listening to the *instrumental* version.

To obtain 'instrumental' versions of the songs with the singing voice – and hence the lyrics – removed, we apply vocal source separation on the audio using a U-Net architecture (Jansson et al. 2017, 2019).

The annotators were given the options of {Yes, No, Uninformative} to annotate each (song, mood) pair. The final annotation is decided based on the majority vote, and in a case of disagreement between all three annotators, a fourth annotator resolved the disagreement. Table 7 shows the inter annotator agreement using Fleiss Kappa (Falotico and Quatto 2015) (κ). According to the Fleiss Kappa interpretation table (Landis and Koch 1977)[3], the annotators have 'fair' agreement; a qualitative examination of the annotations suggests that the subtasks are highly subjective.

Even so, we observe that the degree to which lyrics and acoustics contribute to the association between songs and mood depends on the specific mood. Consistent with the model predictions, "chill" is more inferable from acoustics, while "love" is more inferable from lyrics according to the annotations. Moreover, we find that a subset of the moods in our set are neither inferable from the lyrics nor the acoustics. This subset includes moods like "minimalist", "sunshine", and "obsessed". The association of these moods with songs may be determined by aspects other than acoustics or lyrics, such as cultural or personal associations. For example, some listeners may playlist a song under the term "obsessed" be-

---

[3]see Appendix "Fleiss Kappa Interpretation" for the interpretation table

Mood	Feature	Approach	Lyrics Model	Precision	Recall	F1
Love	**Lyrics**	Zero-shot Learning	NLI	77.62	86.82	81.97
		Fine-tuned on BNPMI	BoW	87.11	83.70	85.37
			NLI	92.73	90.87	91.79
	Acoustics	Fine-tuned		76.67	82.19	79.33
	Hybrid	Fine-tuned	BoW	89.13	87.13	88.12
			NLI	93.04	90.36	91.68
Chill	**Lyrics**	Zero-shot Learning	NLI	11.53	7.98	9.43
		Fine-tuned	BoW	27.92	78.19	41.15
			NLI	80.84	67.43	73.53
	Acoustics	Fine-tuned	-	39.84	88.09	54.86
	Hybrid	Fine-tuned	BoW	51.21	90.56	65.43
			NLI	83.48	71.48	77.01

Table 6: The results of the lyrics-based, acoustics-based and the hybrid models for the moods "love" and "chill".

Annotation	κ
Lyrics	0.2846
Acoustics	0.2910

Table 7: Inter annotator agreement based on the Fleiss κ.

cause it is recently popular and those listeners pay attention to the charts; others may do so because that song evokes personal memories and emotions.

The listener-generated BNPMI scores are composite reflections of both lyrics and acoustics (as well as other factors) from the listeners' point of view. In the previous research questions, we use the BNPMI-derived song-mood association as ground truth to train and evaluate our models. However, in this section, we investigate how much the BNPMI associations are aligned with the human evaluation; i.e., we use the human annotation as ground truth.

We define an annotation consensus as follows. If at least one source of the annotation, either lyrics or acoustics, annotate the association between the song and mood as positive, the consensus ground truth will be positive association. If one source of the annotation is negative and the other is uninformative, the consensus will be negative. Lastly if both of the sources are uninformative the consensus is uninformative.

Table 8 shows a comparison of BNPMI-derived song-mood associations against human judgments based on lyrics and acoustics annotations separately and in consensus. The scores demonstrate that lyrics correlate better with BNPMI association scores than acoustics, and that the the BNPMI association scores are most closely consistent with a consensus of lyrics and acoustics. We also see that there are more $(song, mood)$ pairs where the lyrics alone are insufficient for annotators to assign a mood to the track compared to music only.

Moreover, we investigate the BNPMI threshold set previ-

ously to assign the weak association labels by plotting the "Precision-Recall vs. Threshold" for BNPMI scores. Figure 4 shows the precision and recall of BNPMI score evaluation when the true labels are obtained from lyrics annotation as well as audio annotation. By looking at the figure we observe the selected threshold $\tau = 0.1$ is the appropriate threshold since the precision and recall are both maximized.
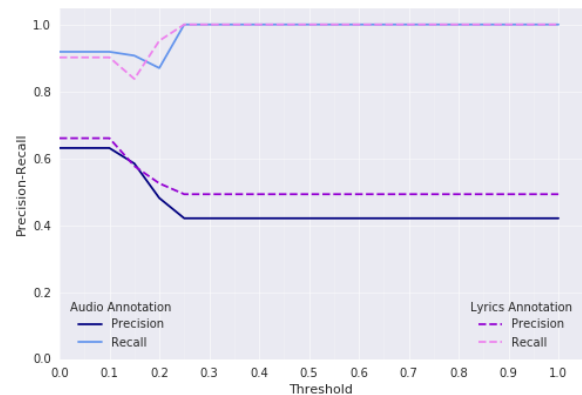

Figure 4: Precision Recall vs. Threshold for BNPMI score

We further evaluated our models with transformer-based representations (BART model fine-tuned on the MNLI corpus) for lyrics against the lyrics, acoustics and consensus annotations ground truths in Table 9. We observe that the zero-shot NLI model captures the association between lyrics and mood very well. Although the difference of the zero-shot NLI and the fine-tuned NLI models in F1 score is negligible, the fine-tuned NLI model over-predicts positive associations, resulting in a lower precision. The performance of the Hybrid model is slightly worse than the fine-tuned NLI model, which is the result of the model being more conser-

Source of Ground Truth	Precision	Recall	F1	TP	TN	FP	FN	uninformative
Lyrics Annotation	57.28	69.09	62.63	114	51	85	18	34
Acoustics Annotation	51.23	61.53	55.91	104	65	99	15	19
Lyrics & Acoustics Consensus	72.01	74.40	73.19	157	54	61	27	3

Table 8: BNPMI score evaluation against human judgments.

Ground Truth	Feature	Approach	Precision	Recall	F1	TP	TN	FP	FN
Lyrics Annotation	**Lyrics**	Zero-shot	72.58	68.18	70.31	90	102	34	42
		Fine-tuned	58.33	90.15	70.83	119	51	85	13
	Hybrid	Fine-tuned	57.00	87.02	68.88	114	51	85	18
Acoustics Annotation	**Lyrics**	Zero-shot	52.62	50.42	51.50	60	110	54	59
		Fine-tuned	50.00	88.23	63.82	105	59	105	14
	Hybrid	Fine-tuned	49.75	85.71	62.92	102	60	103	17
Lyrics & Acoustics Consensus	**Lyrics**	Zero-shot	79.06	55.43	65.17	102	88	27	82
		Fine-tuned	71.87	87.50	78.92	161	52	63	23
	Hybrid	Fine-tuned	70.90	85.24	77.41	156	51	64	27

Table 9: Evaluation of the NLI-based lyrics and hybird models against the lyrics annotation, acoustics annotation and the consensus of the lyrics and acoustics annotations.

vative in predicting positive associations.

As expected, we find the zero-shot NLI model does not capture the acoustics, and the fine-tuned NLI model improves the F1 score. The hybrid model performs slightly worse than the fine-tuned NLI model but the difference in F1 score is negligible.

Lastly, evaluating models against the consensus shows the fine-tuned NLI model captures both the lyrics as well as acoustics. As we expected the performance of the fine-tuned NLI model represented by F1 score improves.

Discussion

In this work, we study the association between song lyrics and moods, and compare how much the lyrics and acoustics of a song contribute to understanding the mood of the song. We investigate what lyrics tell us about the song's mood without any training data by formulating the problem as a zero-shot text classification task to classify whether the song's lyrics and the mood are associated with each other. We explore two transformer-based approaches, natural language inference and next sentence prediction for addressing the text-classification task. We find that the natural language inference model is good at capturing literal meaning of the song by having a high precision, however, it has low recall. On the other hand, the next sentence prediction model is biased towards predicting positive associations, resulting in high recall.

Furthermore, we investigate if training a lyrics-based model on the listener-generated mood associations results in a high performance model to predict the associations for unseen songs. We represent the lyrics with a bag-of-words

model as well as transformer-based models. We find that the fine-tuned transformer-based models outperform the bag of words model, and predict mood associations with high enough precision and recall that they can be valuable for predicting moods of unseen songs in music applications.

We also compare the relative contributions of lyrics and acoustics to the mood of a song by exploring models built on each modality, as well as by combining both into a hybrid model. We observe the contribution of lyrics varies depending on the mood. However, both lyrics and acoustics are a source of information for correctly classifying moods, and using a hybrid model trained on both lyrics as well as acoustic results in higher performance.

Lastly, through a human annotation task, we study whether the models capture the same information about lyrics and acoustics as humans. The annotations demonstrate that understanding the mood of a song by its lyrics or acoustics is a highly subjective task. However, similar to the lyrics-based and acoustics-based models, we observe that a subset of moods are more inferrable from the lyrics than acoustics and the other way around. Furthermore, we evaluate how much the collaborative listener-generated mood associations are aligned with human judgments, and find that they are mostly aligned with lyrics, suggesting that users seem to pay primary importance to lyrics with naming playlists.

One shortcoming of our combination of lyrics and acoustics in the models is that we use a rich transformer based representation for the former, but summary features with a linear classifier for the latter. As such, the acoustics-based models are best compared on the same footing with the bag-of-

words lyrics models rather than the transformer approaches. Future work should include deep models of acoustics for a fair comparison and hybrid combination.

It is worth keeping in mind that the mood expressed in a work of art may differ from the mood experienced by its audience—a picture, a text, or a song which expresses sadness or melancholia may well elicit enjoyment, exhilaration, or admiration in its viewer, reader, or listener (Sachs, Damasio, and Habibi 2015). An area of future work is to understand whether seemingly incorrect predictions by the models, or contradictions between listener-generated associations and predictions from models using lyrics, arise from these differences of songwriter intent and user perception.

Appendices

BNPMI Score Calculation

Pointwise mutual information (PMI) is a common measure of association between two possible events. In our context we want to measure the association between songs and words through their occurrence and cooccurrence in playlists of songs. A word is part of a playlist when it is present in the playlist's name or description. In this way the pointwise mutual information between a song, s, and a mood, m, is calculated as:

$$\text{PMI}(s, m) = \log \frac{p(s, m)}{p(s)p(m)} \quad (1)$$

A common normalization of PMI (NPMI) is given by:

$$\text{NPMI}(s, m) = \frac{\text{PMI}(s, m)}{\log p(s, m)} \quad (2)$$

which has values on [-1, 1], taking on the value -1 if they can never occur together and 1 one if the can never occur apart. If the events happen independently of one another, it takes on the value 0. It will be convenient to rewrite the formula for NPMI as:

$$\text{NPMI}(s, m) = \frac{\log p(s) - \log p(s|m)}{\log p(m) + \log p(s|m)} \quad (3)$$

In order to estimate NPMI, $p(s), p(m), p(s|m)$ is estimated. It is assumed that each mood and song occurs often enough in playlists that empirical estimates of $p(s)$ and $p(m)$ should be "good enough".

$$\hat{p}(s) = \frac{\#\{\text{playlists containing } s\}}{\#\text{playlists}} \quad (4)$$

$$\hat{p}(m) = \frac{\#\{\text{playlists containing } m\}}{\#\text{playlists}} \quad (5)$$

However, there is no assumption that there will be enough cooccurrences of s and m to get good enough estimates of $p(s|m)$. In particular, there might be spurious correlations from "lucky" cooccurences. To account for this there is this assumption that the $p(s|m)$ for a fixed mood m are drawn

from a conjugate prior distribution, $\text{Beta}(\alpha_m, \beta_m)$, for some α_m and β_m. $p_{s,m}$ is denoted to be the empirical estimate of $p(s|m)$ and estimate α_m and β_m for each mood m through method of moments as follows:

$$p_{s,m} = \frac{\#\{\text{playlists containing } s \text{ and } m\}}{\#\{\text{playlists containing } m\}} \quad (6)$$

$$\bar{p}_m := \frac{1}{\#\text{songs}} \sum_s p_{s,m} \quad (7)$$

$$\bar{v}_m := \frac{1}{\#\text{songs} - 1} \sum_s (p_{s,m} - \bar{p}_m)^2 \quad (8)$$

$$\hat{\alpha} = \bar{p}_d \left(\frac{\bar{p}_m(1 - \bar{p}_m)}{\bar{v}_m} - 1 \right) \quad (9)$$

$$\hat{\beta} = (1 - \bar{p}_m) \left(\frac{\bar{p}_m(1 - \bar{p}_m)}{\bar{v}_m} - 1 \right) \quad (10)$$

From this the posterior estimate for $p(s|m)$ is calculated as:

$$\hat{p}(s|m) = \frac{\#\{\text{playlist containing } s \text{ and } m\} + \hat{\alpha_m}}{\#\{\text{playlist containing } m\} + \hat{\alpha_m} + \hat{\beta_m}} \quad (11)$$

The final association score, BNPMI is defined to be the estimate of NPMI that comes from substituting our estimates for $p(s)$, $p(m)$, and $p(s|m)$ into equation 3:

$$\text{BNPMI}(s, m) = \frac{\log \hat{p}(s) - \log \hat{p}(s|m)}{\log \hat{p}(m) + \log \hat{p}(s|m)} \quad (12)$$

Examples of Mood and Song Pairs

Table 10 shows 18 pairs of mood and song for 3 different songs along with their lyrics and audio human annotation.

Fleiss Kappa Interpretation

Table 11 shows the interpretation of the κ value described by Landis and Koch (1977).

Descriptor	Song / Artist	Informativeness	
		Lyrics	Audio
calm	Stormbringer / deep purple	No	No
sad	Stormbringer / deep purple	Yes	No
relaxing	Stormbringer / deep purple	No	No
lit	Stormbringer / deep purple	No	Yes
slow	Stormbringer / deep purple	U	No
depression	Stormbringer / deep purple	Yes	No
smooth	Stormbringer / deep purple	No	No
chill	Stormbringer / deep purple	No	No
influential	Who's Gonna Take The Weight? / Gang Starr	Yes	U
sad	Who's Gonna Take The Weight? / Gang Starr	No	No
happy	Who's Gonna Take The Weight? / Gang Starr	No	No
soft	Who's Gonna Take The Weight? / Gang Starr	No	No
militant	Who's Gonna Take The Weight? / Gang Starr	No	No
motivation	What I'd Say / Earl Thomas Conley	No	No
upbeat	What I'd Say / Earl Thomas Conley	No	No
chill	What I'd Say / Earl Thomas Conley	No	Yes
good vibes	What I'd Say / Earl Thomas Conley	No	Yes
lit	What I'd Say / Earl Thomas Conley	No	No

Table 10: Examples of mood and song pairs along with lyrics and audio annotation. U indicates that the judgment source (lyrics or audio) is uninformative.

κ	interpretation
< 0	Poor agreement
(0.01, 0.20]	Slight agreement
(0.20, 0.40]	Fair agreement
(0.40, 0.60]	Moderate agreement
(0.60, 0.80]	Substantial agreement
(0.80, 1.00]	Almost perfect agreement

Table 11: Fleiss κ interpretation

References

Ali, S. O.; and Peynircioğlu, Z. F. 2006. Songs and emotions: are lyrics and melodies equal partners? *Psychology of Music*, 34(4): 511–534.

Besson, M.; Faïta, F.; Peretz, I.; Bonnel, A.-M.; and Requin, J. 1998. Singing in the Brain: Independence of Lyrics and Tunes. *Psychological Science*, 9(6): 494–498.

Bhattacharya, A.; and Kadambari, K. 2018. A multimodal approach towards emotion recognition of music using audio and lyrical content. *arXiv preprint arXiv:1811.05760*.

Bowman, S. R.; Angeli, G.; Potts, C.; and Manning, C. D. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Brattico, E.; Alluri, V.; Bogert, B.; Jacobsen, T.; Vartiainen, N.; Nieminen, S.; and Tervaniemi, M. 2011. A Functional MRI Study of Happy and Sad Emotions in Music with and without Lyrics. *Frontiers in Psychology*, 2: 308.

Darwin, C. 1872. *The Expression of the Emotions in Man and Animals*. London: John Murray.

Delbouys, R.; Hennequin, R.; Piccoli, F.; Royo-Letelier, J.; and Moussallam, M. 2018. Music mood detection based on audio and lyrics with deep neural net. *arXiv preprint arXiv:1809.07276*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 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 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Downie, X.; Laurier, C.; and Ehmann, M. 2008. The 2007 MIREX audio mood classification task: Lessons learned. In *Proc. 9th Int. Conf. Music Inf. Retrieval*, 462–467.

Falotico, R.; and Quatto, P. 2015. Fleiss' kappa statistic without paradoxes. *Quality & Quantity*, 49(2): 463–470.

Fell, M.; Cabrio, E.; Korfed, E.; Buffa, M.; and Gandon, F. 2019. Love me, love me, say (and write!) that you love me: Enriching the WASABI song corpus with lyrics annotations. *arXiv preprint arXiv:1912.02477*.

Fell, M.; and Sporleder, C. 2014. Lyrics-based analysis and classification of music. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers*, 620–631.

Hall, A. R.; et al. 2005. *Generalized method of moments*. Oxford university press.

Hrustanović, S.; Kavšek, B.; and Tkalčič, M. 2021. Recognition of Eudaimonic and Hedonic Qualities from Song Lyrics.

Hu, X.; and Downie, J. S. 2010. When Lyrics Outperform Audio for Music Mood Classification: A Feature Analysis. In *ISMIR*, 619–624. Citeseer.

Hu, X.; and Yu, B. 2011. Exploring The Relationship Between Mood and Creativity in Rock Lyrics. In *ISMIR*, 789–794. Citeseer.

Interiano, M.; Kazemi, K.; Wang, L.; Yang, J.; Yu, Z.; and Komarova, N. L. 2018. Musical trends and predictability of success in contemporary songs in and out of the top charts. *Royal Society open science*, 5(5): 171274.

James, W. 1884. What is an emotion? *Mind*, 188–205.

Jansson, A.; Bittner, R. M.; Ewert, S.; and Weyde, T. 2019. Joint singing voice separation and f0 estimation with deep u-net architectures. In *2019 27th European Signal Processing Conference (EUSIPCO)*, 1–5. IEEE.

Jansson, A.; Humphrey, E. J.; Montecchio, N.; Bittner, R. M.; Kumar, A.; and Weyde, T. 2017. Singing Voice Separation with Deep U-Net Convolutional Networks. In *ISMIR*.

Korzeniowski, F.; Nieto, O.; McCallum, M.; Won, M.; Oramas, S.; and Schmidt, E. 2020. Mood classification using listening data. *arXiv preprint arXiv:2010.11512*.

Landis, J. R.; and Koch, G. G. 1977. The measurement of observer agreement for categorical data. *biometrics*, 159–174.

Laurier, C.; Grivolla, J.; and Herrera, P. 2008. Multimodal music mood classification using audio and lyrics. In *2008 Seventh International Conference on Machine Learning and Applications*, 688–693. IEEE.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Li, X.; Bing, L.; Zhang, W.; and Lam, W. 2019. Exploiting BERT for End-to-End Aspect-based Sentiment Analysis. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, 34–41. Hong Kong, China: Association for Computational Linguistics.

Ma, T.; Yao, J.-G.; Lin, C.-Y.; and Zhao, T. 2021. Issues with Entailment-based Zero-shot Text Classification. In *59th ACL-IJCNLP*, 786–796. Online: Association for Computational Linguistics.

McVicar, M.; Freeman, T.; and De Bie, T. 2011. Mining the Correlation between Lyrical and Audio Features and the Emergence of Mood. In *ISMIR*, 783–788.

Medina, Y. O.; Beltrán, J. R.; and Baldassarri, S. 2020. Emotional classification of music using neural networks with the MediaEval dataset. *Personal and Ubiquitous Computing*, 1–13.

Mori, K.; and Iwanaga, M. 2014. Pleasure generated by sadness: Effect of sad lyrics on the emotions induced by happy music. *Psychology of Music*, 42(5): 643–652.

Panda, R.; Malheiro, R. M.; and Paiva, R. P. 2020. Audio features for music emotion recognition: a survey. *IEEE Transactions on Affective Computing*.

Panda, R. E. S. 2019. *Emotion-based Analysis and Classification of Audio Music*. Ph.D. thesis, 00500:: Universidade de Coimbra.

Pang, B.; and Lee, L. 2008. Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, 2(1–2): 1–135.

Parisi, L.; Francia, S.; Olivastri, S.; and Tavella, M. S. 2019. Exploiting synchronized lyrics and vocal features for music emotion detection. *arXiv preprint arXiv:1901.04831*.

Qu, Y.; Shanahan, J.; and Wiebe, J., eds. 2004. *AAAI Spring Symposium on Exploring Attitude and Affect in Text*, SS-04-07. AAAI Press. ISBN 978-1-57735-219-8.

Saari, P.; Eerola, T.; Fazekas, G.; Barthet, M.; Lartillot, O.; and Sandler, M. B. 2013. The Role of Audio and Tags in Music Mood Prediction: A Study Using Semantic Layer Projection. In *ISMIR*, 201–206.

Sachs, M. E.; Damasio, A.; and Habibi, A. 2015. The pleasures of sad music: a systematic review. *Frontiers in human neuroscience*, 9: 404.

Schlaifer, R.; and Raiffa, H. 1961. *Applied statistical decision theory*. Wiley Classics Library.

Schmidt, E. M.; and Kim, Y. E. 2011. Modeling Musical Emotion Dynamics with Conditional Random Fields. In *ISMIR*, 777–782. Citeseer.

Sousou, S. D. 1997. Effects of Melody and Lyrics on Mood and Memory. *Perceptual and Motor Skills*, 85(1): 31–40. PMID: 9293553.

Stratton, V. N.; and Zalanowski, A. H. 1994. Affective Impact of Music Vs. Lyrics. *Empirical Studies of the Arts*, 12(2): 173–184.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Wang, A. Y.; Anderson, I. J.; and McCurry, P. H. 2021. Systems and methods for determining descriptors for media content items. US Patent App. 16/732,176.

Wang, H.-C.; Syu, S.-W.; and Wongchaisuwat, P. 2021. A method of music autotagging based on audio and lyrics. *Multimedia Tools and Applications*, 80(10): 15511–15539.

Williams, A.; Nangia, N.; and Bowman, S. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *NAACL, Volume 1 (Long Papers)*, 1112–1122. Association for Computational Linguistics.

Yang, D.; and Lee, W.-S. 2009. Music Emotion Identification from Lyrics. In *2009 11th IEEE International Symposium on Multimedia*, 624–629.

Yin, D.; Meng, T.; and Chang, K.-W. 2020. SentiBERT: A Transferable Transformer-Based Architecture for Compositional Sentiment Semantics. In *58th ACL*, 3695–3706. Online: Association for Computational Linguistics.

Yin, W.; Hay, J.; and Roth, D. 2019. Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach. In *EMNLP-IJCNLP*, 3914–3923.

Zaanen, M.; and Kanters, P. 2010. Automatic Mood Classification Using TF*IDF Based on Lyrics. In *ISMIR*.

Zhang, L.; Wang, S.; and Liu, B. 2018. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4): e1253.