Emotion4MIDI: A Lyrics-Based Emotion-Labeled Symbolic Music Dataset

Serkan Sulun[1,2(✉)], Pedro Oliveira[2], and Paula Viana[1,3]

[1] Institute for Systems and Computer Engineering, Technology and Science (INESC TEC), Porto, Portugal
paula.viana@inesctec.pt
[2] Faculty of Engineering, University of Porto, Porto, Portugal
serkan.sulun@inesctec.pt, up201707038@edu.fe.up.pt
[3] ISEP, Polytechnic of Porto, School of Engineering, Porto, Portugal
pmv@isep.ipp.pt

Abstract. We present a new large-scale emotion-labeled symbolic music dataset consisting of 12 k MIDI songs. To create this dataset, we first trained emotion classification models on the GoEmotions dataset, achieving state-of-the-art results with a model half the size of the baseline. We then applied these models to lyrics from two large-scale MIDI datasets. Our dataset covers a wide range of fine-grained emotions, providing a valuable resource to explore the connection between music and emotions and, especially, to develop models that can generate music based on specific emotions. Our code for inference, trained models, and datasets are available online.

Keywords: Sentiment analysis · Symbolic music · Emotion classification · Music dataset

1 Introduction

Music has long been a powerful medium for emotional expression and communication [16]. The emotional response that music elicits has been studied by scholars from various fields such as psychology [19], musicology [15], and neuroscience [17]. Especially with the advent of deep learning, there has been an increasing interest in developing machine learning algorithms to automatically analyze and generate music that can evoke specific emotions in listeners [3].

Symbolic music—or MIDI (Musical Instrument Digital Interface) as it is used interchangeably—is represented as a sequence of notes and is a popular choice for

This work has been funded by National Funds through the Portuguese funding agency, FCT—Fundação para a Ciência e a Tecnologia, within project LA/P/0063/2020. Serkan Sulun received the support of fellowships from "la Caixa" Foundation (ID 100010434) with the fellowship code LCF/BQ/DI19/11730032 and FCT—Fundação para a Ciência e a Tecnologia with the fellowship code 2022.09594.BD. Corresponding author: Serkan Sulun (serkan.sulun@inesctec.pt)

N. Moniz et al. (Eds.): EPIA 2023, LNAI 14116, pp. 77–89, 2023.
https://doi.org/10.1007/978-3-031-49011-8_7

machine learning models due to its compact and structured representation. Large raw MIDI datasets [30,31] enable unsupervised training of deep neural networks to automatically generate symbolic music. Similar to language modeling, these networks learn to predict the next token i.e. the next note, and at inference time, generate output autoregressively, one token at a time.

However, a human composer's creative process does not simply involve mechanically writing one note after another; it often includes high-level concepts such as motifs, themes and, ultimately, emotions [24]. To train deep neural networks to generate music based on emotions, large datasets of symbolic music annotated with emotional labels are required. Although there are some publicly available datasets with emotional labels, they are relatively small and do not cover a wide range of emotional states [33].

To address this issue, we present a new large-scale emotion-labeled symbolic music dataset created by analyzing the lyrics of the songs. Our approach leverages the natural connection between lyrics and music, established through emotions. To this end, we first trained models for emotion classification from text on GoEmotions [5], one of the largest text datasets with 28 fine-grained emotion labels. Using a model that is half the size of the baseline model, we obtained state-of-the-art results on this dataset. Later, we applied this model to the lyrics of songs from two of the biggest available MIDI datasets, namely Lakh MIDI dataset [30] and Reddit MIDI dataset [31]. Ultimately, we created a symbolic music dataset consisting of 12 k MIDI songs labeled with fine-grained emotions. We hope that this dataset will encourage further research in the field of affective algorithmic composition and contribute to the development of intelligent music systems that can understand and evoke specific emotions in listeners.

The remaining of this paper has the following structure: after having introduced our aim and the overall results in Sects. 1 and 2 presents the current state of the art on the most relevant topics for this work, namely text emotion classification and the existing emotion-labeled symbolic music datasets. Section 3 will delve into the proposed solution describing all the implemented steps, while results are presented and discussed in Sect. 4. Finally, we conclude by pointing out some possible future work in Sect. 5.

2 Related Work

2.1 Text Emotion Classification

Emotion classification from text—or sentiment analysis, as used interchangeably in the machine learning literature—allows us to automatically identify and/or quantify the emotion expressed in a piece of text, such as a review, social media post, or customer feedback [23]. Identifying the underlying emotion in text is useful in various fields such as customer service [10], finance [25], politics [14], and entertainment [1].

Machine learning methods have significantly advanced the state of the art in text emotion classification for the past two decades. However, the earliest works in this field relied on hand-crafted features, such as frequently used n-grams

[27], or adjectives and adverbs that are associated with particular emotions [35]. Nonetheless, the advent of deep learning has made it computationally feasible to process raw inputs without extracting features manually, leading to better performance [18]. Recurrent Neural Networks and their improved variants such as Long Short-Term Memory were initially used [22] but were later replaced by the transformer model [34], which is the current state of the art in natural language processing (NLP) tasks.

Fine-tuning pretrained models on specific tasks has been shown to produce better performance. The GPT (generative pretraining) model is a large transformer that was pretrained on the task of next token prediction and then was fine-tuned on specific NLP tasks, resulting in state-of-the-art performance [29]. The BERT (Bidirectional Encoder Representations from Transformers) model improved upon these results by employing masked token prediction as its pretraining task [6].

2.2 Emotion-Labeled Symbolic Music Datasets

MIDI (Musical Instrument Digital Interface) is a symbolic music format widely used to represent musical performances and compositions in the digital domain. MIDI files contain only the musical information, such as the notes, tempo, and dynamics, without the sound itself, like a "digital music sheet". Compared to audio formats, MIDI files have a smaller size and dimensionality, which makes them more manageable and suitable for modeling with deep neural networks [3].

The majority of existing literature on symbolic music generation relies on a non-conditional approach. In other words, these methods are trained on raw MIDI data without any explicit labels, allowing them to generate new music that is similar to the examples in the training dataset [12]. Some approaches, however, leverage low-level features within the data to create music in a conditional way [11]. For instance, they might use short melodies, chords, or single-instrument tracks as a basis for generating corresponding melodies. While such methods could be considered as "conditional", they do not make use of specific labels and are thus unable to capture high-level factors such as emotions or genres.

Using emotion as the specific high-level condition gives rise to the field of "affective algorithmic composition" (AAC) [36]. However, the development of machine learning AAC models is currently limited by the lack of large-scale symbolic music datasets with emotion labels. Some existing datasets include VGMIDI, which contains 204 piano-based video game soundtracks with continuous valence and arousal labels [8], Panda et al., which includes 193 samples with discrete emotion labels [26], and EMOPIA, which consists of 387 piano-based pop songs with four emotion labels [13]. Unfortunately, due to their small sizes, these datasets are insufficient for training deep neural networks with millions of parameters. Sulun et al. addressed this issue by labeling 34 k samples with continuous valence and arousal labels [33]. Though initially designed for audio samples, these labels were matched to their corresponding MIDI files to train emotion-based symbolic music generators that produced output music with emotional coherence. While this study exploited the correspondence between audio

and symbolic music, there has been no utilization of the correspondence between lyrics and symbolic music to acquire high-level semantic labels.

3 Methodology

This section outlines the steps we followed to achieve our goal of creating a symbolic music dataset with emotion labels. Specifically, we begin by describing the model utilized for emotion classification, followed by a discussion of the training process, and conclude with an overview of how the model was applied to song lyrics to extract the corresponding emotion labels.

3.1 Model

We employ DistilBERT as the backbone of our model [32], which is a condensed and compressed variant of the BERT (Bidirectional Encoder Representations from Transformers) model [6], achieved through knowledge distillation [4,9]. DistilBERT utilizes fewer layers than BERT and learns from BERT's outputs to mimic its behavior. Our model consists of 6 layers, with each layer containing 12 attention heads and a dimensionality of 768, yielding a total of 67 M parameters. To facilitate multi-label classification, we have customized the output layer while adding a sigmoid activation layer at the end. The output layer's size is determined by the number of labels present in the training dataset, which can be either 7 or 28.

3.2 Training

The first step towards our aim of building an emotion-labeled symbolic music dataset is to train the model to perform multi-label emotion classification based on text input.

Dataset We trained our model using the GoEmotions dataset [5]. This dataset consists of English comments from the website *reddit.com*, which were manually annotated to identify the underlying emotions. It is a multi-label dataset, which means that each comment can have more than one emotion label. The dataset comprises 27 emotions and a "neutral" label. The labels are further grouped into 7 categories, including the six basic emotions identified by Ekman (joy, anger, fear, sadness, disgust, and surprise) as well as the "neutral" label [7]. The dataset has a total of 58 k samples, which were split into training, validation, and testing sets in the ratio of 80, 10, and 10%, respectively. Given the number of labels and its size, GoEmotions is one of the largest emotion classification datasets and has the highest number of discrete emotion labels [20].

Training and Evaluation Metrics We trained our models using binary cross-entropy loss. For evaluation, we used precision, recall, and F1-score, with macro averaging. The decision cutoff was set at 0.3, meaning that predictions with a value of 0.3 or greater are considered positive predictions and others negative.

Implementation Details We trained two models to classify a given text into 7 and 28 labels. We used a dropout rate of 0.1 and a gradient clipping norm of 1. The batch size was set to 16 for the model with 7 output labels and to 32 for the model with 28 output labels. We applied a learning rate of $5e-5$ for the former and $3e-5$ for the latter. We used early stopping considering the F1-score on the validation dataset, which corresponded to training for 10 epochs for both models. We implemented the models using Huggingface library [37] with Pytorch backend [28] and trained them using a single Nvidia GeForce GTX 1080 Ti GPU.

3.3 Inference

After training the models for text-based emotion classification, we used it in inference mode, using the song lyrics from the MIDI files as inputs. This allowed us to create a MIDI dataset labeled with emotions.

Datasets We used two MIDI datasets that are publicly available and were created by gathering MIDI files from various online sources: the Lakh MIDI dataset consisting of 176 k samples [30] and the Reddit MIDI dataset containing 130 k samples [31]. We filtered the datasets by selecting MIDI files that contain lyrics in the English language with at least 50 words. This filtering process resulted in a total of 12509 files, consisting of 8386 files from the Lakh MIDI dataset and 4123 files from the Reddit MIDI dataset. During inference, we utilized the two pretrained models, feeding the entire song's lyrics, using a truncation length of 512.

4 Results

In this section, we will first present the emotion classification performance of our trained models. Then, we will introduce the emotion-labeled MIDI dataset, which we created by analyzing the sentiment of the song lyrics using our trained models.

4.1 Emotion Classification on the GoEmotions Dataset

We evaluated the performance of our trained models on the test split of the GoEmotions dataset and compared our results with the baseline presented in the original paper [5]. Similar to the original paper, we report our results for scenarios using two sets of labels, with 7 and 28 emotions. For each label, we reported the precision, recall, and F1-scores along with the macro-averages. It is important to mention that, as the dataset is imbalanced, macro-averaging is more appropriate than micro-averaging, as it was also used in the original paper. We note that the baseline model is BERT and has twice the size of our model [6].

The trade-off between precision and recall is determined by the cutoff value. Therefore, we emphasize higher F1-scores because they provide a more balanced perspective by taking the harmonic mean of precision and recall, and are much less sensitive to the cutoff value. Although the original paper did not state the cutoff value, we achieved the best F1-score and similar performance to the original paper on the 7-label dataset using a cutoff value of 0.3. For consistency, we used the same value for the 28-label dataset. We present our results on the dataset with 28 and 7 labels in Tables 1 and 2, respectively.

Table 1. 7-label classification results

	Precision		Recall		F1-score	
	Baseline	Ours	Baseline	Ours	Baseline	Ours
Anger	0.50	0.50	0.65	0.67	**0.57**	**0.57**
Disgust	0.52	0.57	0.53	0.49	**0.53**	0.52
Fear	0.61	0.57	0.76	0.73	**0.68**	0.64
Joy	0.77	0.75	0.88	0.89	**0.82**	**0.82**
Neutral	0.66	0.63	0.67	0.75	0.66	**0.68**
Sadness	0.56	0.57	0.62	0.67	0.59	**0.61**
Surprise	0.53	0.59	0.70	0.62	**0.61**	**0.61**
Macro-average	0.59	0.60	0.69	0.69	**0.64**	**0.64**

Based on the F1-scores, our model performs comparably to the baseline on the 7-label dataset. Specifically, our model has a better performance on 2 labels, worse on 2 labels, and the same on 3 labels, as well as for the macro-average. On the 28-label dataset, our model surpasses the baseline with only a lower performance on 2 labels, equal performance on 4 labels, and better performance on the remaining 22 labels. Furthermore, our model demonstrates an improvement of 0.04 in terms of the macro-average.

We hypothesize that a smaller model, such as ours (DistilBERT), may perform better than a larger baseline model (BERT) in certain settings, such as when there are a limited number of training samples or a high output/target dimensionality, as in the case of the 28-label dataset. In these scenarios, models are more prone to overfitting, as has been previously observed [38]. Additionally, the original paper [32] demonstrates that the DistilBERT model outperforms BERT on the Winograd Natural Language Inference (WNLI) dataset [21].

Table 2. 28-label classification results

	Precision		Recall		F1-score	
	Baseline	Ours	Baseline	Ours	Baseline	Ours
Admiration	0.53	0.65	0.83	0.75	0.65	**0.70**
Amusement	0.70	0.72	0.94	0.91	0.80	**0.81**
Anger	0.36	0.53	0.66	0.49	0.47	**0.51**
Annoyance	0.24	0.40	0.63	0.31	0.34	**0.35**
Approval	0.26	0.39	0.57	0.38	0.36	**0.39**
Caring	0.30	0.37	0.56	0.46	0.39	**0.41**
Confusion	0.24	0.52	0.76	0.42	0.37	**0.47**
Curiosity	0.40	0.47	0.84	0.62	**0.54**	0.53
Desire	0.43	0.66	0.59	0.42	0.49	**0.51**
Disappointment	0.19	0.39	0.52	0.22	**0.28**	**0.28**
Disapproval	0.29	0.39	0.61	0.41	0.39	**0.40**
Disgust	0.34	0.64	0.66	0.39	0.45	**0.48**
Embarrassment	0.39	0.72	0.49	0.35	0.43	**0.47**
Excitement	0.26	0.43	0.52	0.47	0.34	**0.45**
Fear	0.46	0.60	0.85	0.76	0.60	**0.67**
Gratitude	0.79	0.88	0.95	0.92	0.86	**0.90**
Grief	0.00	0.00	0.00	0.00	**0.00**	**0.00**
Joy	0.39	0.59	0.73	0.61	0.51	**0.60**
Love	0.68	0.78	0.92	0.85	0.78	**0.81**
Nervousness	0.28	0.45	0.48	0.43	0.35	**0.44**
Neutral	0.56	0.61	0.84	0.76	**0.68**	**0.68**
Optimism	0.41	0.56	0.69	0.52	0.51	**0.54**
Pride	0.67	0.83	0.25	0.31	0.36	**0.45**
Realization	0.16	0.39	0.29	0.14	**0.21**	**0.21**
Relief	0.50	0.00	0.09	0.00	**0.15**	0.00
Remorse	0.53	0.59	0.88	0.86	0.66	**0.70**
Sadness	0.38	0.57	0.71	0.60	0.49	**0.59**
Surprise	0.40	0.56	0.66	0.50	0.50	**0.53**
Macro-average	0.40	0.53	0.63	0.50	0.46	**0.50**

4.2 Labeled MIDI Dataset

We used our trained models to analyze the song lyrics of the Lakh and Reddit MIDI datasets, resulting in an augmented dataset that contains the file paths to 12509 MIDI files and their corresponding predicted probabilities for emotion labels. To provide more flexibility to the users, we did not apply a threshold to the predicted probabilities, allowing the entire dataset to be used as is. We generated

two CSV (comma-separated values) files containing the 7 and 28 emotion labels as columns, with the 12509 MIDI file paths as rows. Our code for inference, trained models, and datasets are available online.[1]

For demonstration purposes, we provide transposed versions of the tables, using only 3 samples, shown in Tables 3 and 4. We note that the values do not necessarily add up to one, due to the nature of multi-label classification.

For further demonstration and ease of analysis, we provide excerpts from the lyrics of each of the three sample songs in Listing 1.1, along with the emotions having predicted probabilities higher than 0.1 in descending order. It is noteworthy that having a dataset with 28 emotion labels allows for a more nuanced representation of emotions. For instance, when we examine this dataset, the song "Imagine" is predicted to have "optimism" as its top emotion, whereas "Take a Chance on Me" is predicted to have "caring" as its top emotion. However, both songs are predicted to have "joy" as their top emotion in the dataset with only seven labels.

We also present the number of samples containing each emotion in our datasets in Fig. 1. In these figures, we excluded the "neutral" label. We also considered emotions with a prediction value higher than 0.1 as positive labels, meaning that those emotions are present for a given sample.

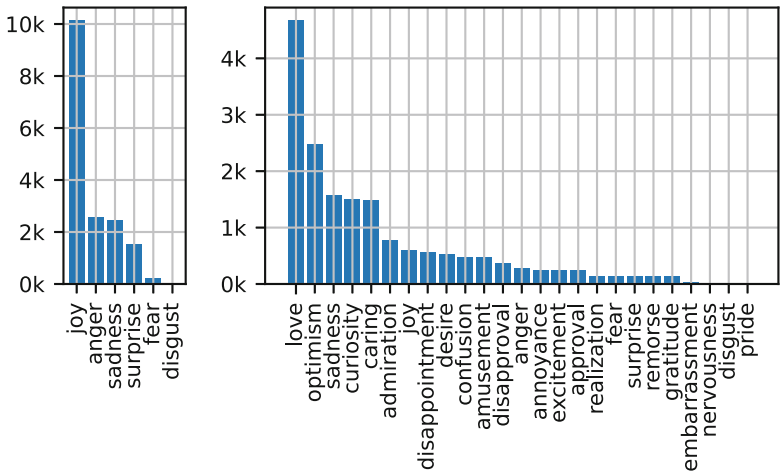

Fig. 1. The number of samples containing each emotion in our 7-label (left) and 28-label (right) datasets. The "neutral" label is excluded. Emotions with a prediction value higher than 0.1 are considered positive labels, meaning that those emotions are present for a given sample.

[1] https://github.com/serkansulun/lyricsemotions.

Listing 1.1. Sample entries with excerpts from lyrics, and emotions with a predicted value higher than 0.1.

```
File path: lakh/5/58c076b72d5115486c09a7d9e6df1029.mid
Artist - Title: John Lennon - Imagine
Lyrics:
    Imagine there's no heaven.
    It's easy if you try.
    No hell below us.
    Above us, only sky.
    Imagine all the people.
    Livin' for today.
7-label predictions:
    joy:            0.8072
    neutral:        0.1953
28-label predictions:
    optimism:       0.7554
    neutral:        0.2954

File path: reddit/A/ABBA.Take a chance on me K.mid
Artist - Title: ABBA - Take a Chance on Me
Lyrics:
    If you change your mind, I'm the first in line.
    Honey, I'm still free.
    Take a chance on me.
    If you need me, let me know, gonna be around.
    If you've got no place to go, if you're feeling down.
7-label predictions:
    joy:            0.8948
    neutral:        0.1420
28-label predictions:
    caring:         0.6169
    neutral:        0.4288
    optimism:       0.1423
    love:           0.1079

File path: reddit/P/PRESLEY.Are you lonesome tonight K.mid
Artist - Title: Elvis Presley - Are You Lonesome Tonight
Lyrics:
    Are you lonesome tonight?
    Do you miss me tonight?
    Are you sorry we drifted apart?
    Does your memory stray to a bright summer day,
    When I kissed you and called you sweetheart?
7-label predictions:
    sadness:        0.7372
    surprise:       0.5465
28-label predictions:
    curiosity:      0.6502
    sadness:        0.1767
    remorse:        0.1491
    confusion:      0.1029
```

Table 3. Sample entries from the 28-label dataset.

	John Lennon—Imagine	ABBA—Take a Chance on Me	Elvis Presley—Are You Lonesome Tonight
Admiration	0.0021	0.0091	0.0048
Amusement	0.0051	0.0012	0.0027
Anger	0.0025	0.0018	0.0053
Annoyance	0.0024	0.0020	0.0075
Approval	0.0026	0.0809	0.0072
Caring	0.0067	**0.6169**	0.0601
Confusion	0.0070	0.0035	0.1029
Curiosity	0.0332	0.0141	**0.6502**
Desire	0.0482	0.0472	0.0055
Disappointment	0.0044	0.0016	0.0199
Disapproval	0.0019	0.0030	0.0048
Disgust	0.0007	0.0003	0.0009
Embarrassment	0.0006	0.0002	0.0045
Excitement	0.0130	0.0049	0.0011
Fear	0.0026	0.0026	0.0035
Gratitude	0.0007	0.0017	0.0059
Grief	0.0008	0.0016	0.0085
Joy	0.0025	0.0040	0.0018
Love	0.0021	0.1079	0.0193
Nervousness	0.0007	0.0017	0.0094
Neutral	0.2954	0.4288	0.0757
Optimism	**0.7554**	0.1423	0.0060
Pride	0.0010	0.0013	0.0006
Realization	0.0023	0.0040	0.0045
Relief	0.0004	0.0033	0.0011
Remorse	0.0005	0.0012	0.1491
Sadness	0.0011	0.0027	0.1767
Surprise	0.0107	0.0005	0.0020

Table 4. Sample entries from the 7-label dataset. Due to space limitations, the file paths are replaced with the artist and song names and are as the following: John Lennon—Imagine: "`lakh/5/58c076b72d5115486c09a7d9e6df1029.mid`" (artist and title obtained using Million Song Dataset [2]), ABBA - Take a Chance on Me: "`reddit/A/ABBA.Take a chance on me K.mid`", Elvis Presley—Are You Lonesome Tonight: "`reddit/P/PRESLEY.Are you lonesome tonight K.mid`"

	John Lennon—Imagine	ABBA—Take a Chance on Me	Elvis Presley—Are You Lonesome Tonight
Anger	0.0051	0.0146	0.0272
Disgust	0.0003	0.0009	0.0045
Fear	0.0005	0.0024	0.0131
Joy	**0.8072**	**0.8948**	0.0477
Neutral	0.1953	0.1420	0.0782
Sadness	0.0013	0.0069	**0.7372**
Surprise	0.0754	0.0053	0.5465

5 Conclusion and Future Work

In this work, we first trained models on the largest text-based emotion classification dataset, GoEmotions, in both 7-label and 28-label variants [5]. We achieved state-of-the-art results using a model half the size of the baseline. We then used these trained models to analyze the emotions of the song lyrics from the two largest MIDI datasets, Lakh MIDI dataset [30] and Reddit MIDI dataset [31]. This analysis resulted in an augmented dataset of 12509 MIDI files with emotion labels in a multi-label format, using either 7 basic-level or 28 fine-grained emotions. We made the datasets, inference code, and trained models available for researchers to use in various tasks, including symbolic music processing, natural language processing, and sentiment analysis.

In our future work, we plan to further narrow the considerable gap between symbolic music and emotion. In particular, we aim to create superior models that can automatically compose music that is based on emotions or user-provided input. We believe that incorporating emotions is vital in composing music, hence it can help to push the boundaries of computational creativity, bringing it one step closer to human-like performance.

References

1. Almeida, J., Vilaça, L., Teixeira, I.N., Viana, P.: Emotion identification in movies through facial expression recognition. Appl. Sci. **11**(15) (2021)
2. Bertin-Mahieux, T., Ellis, D.P.W., Whitman, B., Lamere, P.: The million song dataset. In: Proceedings of the 12th International Society for Music Information Retrieval Conference, pp. 591–596 (2011)
3. Briot, J., Hadjeres, G., Pachet, F.: Deep Learning Techniques for Music Generation. Springer, Berlin (2020)
4. Buciluǎ, C., Caruana, R., Niculescu-Mizil, A.: Model compression. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 535–541 (2006)
5. Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., Ravi, S.: GoEmotions: a dataset of fine-grained emotions. In: 58th Annual Meeting of the Association for Computational Linguistics (ACL) (2020)
6. 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, pp. 4171–4186 (2019)
7. Ekman, P.: Are there basic emotions? Psychol. Rev. **99**(3), 550–553 (1992)
8. Ferreira, L., Whitehead, J.: Learning to generate music with sentiment. In: Proceedings of the 20th International Society for Music Information Retrieval Conference, pp. 384–390 (2019)
9. Hinton, G., Vinyals, O., Dean, J.: Distilling the Knowledge in a Neural Network. arXiv:1503.02531 (2015)
10. Hu, M., Liu, B.: Mining and summarizing customer reviews. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 168–177 (2004)

11. Huang, C.Z.A., Cooijmans, T., Roberts, A., Courville, A.C., Eck, D.: Counterpoint by convolution. In: Proceedings of the 18th International Society for Music Information Retrieval Conference, pp. 211–218 (2017)
12. Huang, C.Z.A., Vaswani, A., Uszkoreit, J., Simon, I., Hawthorne, C., Shazeer, N., Dai, A.M., Hoffman, M.D., Dinculescu, M., Eck, D.: Music transformer: generating music with long-term structure. In: 7th International Conference on Learning Representations (2019)
13. Hung, H.T., Ching, J., Doh, S., Kim, N., Nam, J., Yang, Y.H.: EMOPIA: A multimodal pop piano dataset for emotion recognition and emotion-based music generation. In: Proceedings of the 22nd International Society for Music Information Retrieval Conference, pp. 318–325 (2021)
14. Iyyer, M., Enns, P., Boyd-Graber, J., Resnik, P.: Political ideology detection using recursive neural networks. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pp. 1113–1122 (2014)
15. Juslin, P.N.: Communicating emotion in music performance: a review and a theoretical framework. In: Music and Emotion: Theory and Research, Series in Affective Science, pp. 309–337. Oxford University Press, New York, NY, US (2001)
16. Juslin, P.N., Sloboda, J.A.: Music and Emotion. Elsevier, Academic (2013)
17. Koelsch, S.: Brain correlates of music-evoked emotions. Nat. Rev. Neurosci. **15**(3), 170–180 (2014)
18. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Commun. ACM **60**(6), 84–90 (2017)
19. Krumhansl, C.L.: Music: a link between cognition and emotion. Curr. Dir. Psychol. Sci. **11**(2), 45–50 (2002)
20. Kusal, S., Patil, S.A., Choudrie, J., Kotecha, K., Vora, D.R., Pappas, I.O.: A review on text-based emotion detection—techniques, applications, datasets, and future directions (2022). ArXiv:abs/2205.03235
21. Levesque, H., Davis, E., Morgenstern, L.: The winograd schema challenge. In: Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning (2012)
22. Li, D., Qian, J.: Text sentiment analysis based on long short-term memory. In: 2016 First IEEE International Conference on Computer Communication and the Internet (ICCCI), pp. 471–475 (2016)
23. Liu, B., Zhang, L.: A survey of opinion mining and sentiment analysis. In: Mining Text Data, pp. 415–463. Springer, Berlin (2012)
24. Meyer, L.B.: Emotion and Meaning in Music. University of Chicago Press (2008)
25. Nguyen, T.H., Shirai, K., Velcin, J.: Sentiment analysis on social media for stock movement prediction. Expert Syst. Appl. **42**(24), 9603–9611 (2015)
26. Panda, R., Malheiro, R., Rocha, B., Oliveira, A., Paiva, R.P.: Multi-modal music emotion recognition: A new dataset, methodology and comparative analysis. In: International Symposium on Computer Music Multidisciplinary Research (2013)
27. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? sentiment classification using machine learning techniques. In: Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing, pp. 79–86 (2002)
28. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems, vol. 32, pp. 8024–8035. Curran Associates, Inc. (2019)

29. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving Language Understanding by Generative Pre-Training. OpenAI (2018)
30. Raffel, C.: Learning-Based Methods for Comparing Sequences, with Applications to Audio-to-MIDI Alignment and Matching. Ph.D. thesis, Columbia University (2016)
31. Reddit MIDI dataset. https://www.reddit.com/r/WeAreTheMusicMakers/comments/3ajwe4/the_largest_midi_collection_on_the_internet/
32. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter (2019). ArXiv:abs/1910.01108
33. Sulun, S., Davies, M.E.P., Viana, P.: Symbolic music generation conditioned on continuous-valued emotions. IEEE Access **10**, 44617–44626 (2022)
34. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems, pp. 5998–6008 (2017)
35. Whitelaw, C., Garg, N., Argamon, S.: Using appraisal groups for sentiment analysis. In: Proceedings of the 14th ACM International Conference on Information and Knowledge Management, pp. 625–631 (2005)
36. Williams, D., Kirke, A., Miranda, E.R., Roesch, E., Daly, I., Nasuto, S.: Investigating affect in algorithmic composition systems. Psychol. Music **43**(6), 831–854 (2015)
37. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M.: Huggingface's transformers: State-of-the-art natural language processing (2019). ArXiv:abs/1910.03771
38. Yu, Z., Yu, J., Fan, J., Tao, D.: Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In: IEEE International Conference on Computer Vision, pp. 1839–1848. IEEE Computer Society (2017)