EmotioNotes Dataset: Decoding Emotions in Classical Music Through Concert Program Notes

Pratik Khanal(✉) and Patrick J. Donnelly(✉)

Oregon State University, 97330 Corvallis, OR, USA
{khanalp,patrick.donnelly}@oregonstate.edu

Abstract. Understanding emotion in music is an important task in music information retrieval with a wide range of applications in music recommendation, playlist generation, and music therapy. However, the lack of large, public datasets of music and emotion is a major roadblock for researchers. In this paper, we leverage an existing music resource − concert program notes − to create a novel dataset of emotion labels for classical music for research in music emotion recognition. We collect program notes from the New York Philharmonic Society's archive as unorganized documents. We use a large language model, `LLaMA 3.1`, with task-specific prompting to extract program notes for corresponding musical works. From a total of 9,569 documents, we extracted 31,966 program notes and annotated with emotion values determined by the textual content. We use a dataset of emotion values of English words to estimate valence, arousal, and dominance scores for each program note. Finally, we present an emotion recognition model using another LLM, `Longformer`, trained on our dataset to estimate valence, arousal, and dominance scores of the program notes. We achieved Pearson's correlation of 0.877, 0.823, and 0.766 for valence, arousal, and dominance, respectively.

Keywords: Concert program notes · Dataset generation · Large language model · Music emotion recognition

1 Introduction

Music is an integral part of human culture and people throughout the world experience music in a multitude of forms. Music has varying impacts on human emotions, eliciting feelings of happiness, melancholy, intensity, sleepiness, and nostalgia. Computational researchers explore this strong connection between music and emotion in the area of Music Emotion Recognition (MER) [20]. Automatic detection of emotion from music has important applications in mood-based music playlist generation [18] and music recommendation [29]. MER is also pertinent to research in music therapy [17] with potential to help patients with depression, autism, trauma, among other conditions.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
P. Machado et al. (Eds.): EvoMUSART 2025, LNCS 15611, pp. 339–356, 2025.
https://doi.org/10.1007/978-3-031-90167-6_23

However, the lack of large, annotated datasets remains a major obstacle for research in this field. Because of copyright concerns, only a handful of small, public datasets are available [2,11,23,48]. The number of music pieces in such datasets is usually less than 2,000, an insufficient quantity to train a deep neural network for MER tasks. These datasets often sample Western popular music that was commercially successful and rarely feature instrumental music or songs with lyrics in languages besides English. In addition, collecting emotive responses from human participants is cumbersome, time-consuming, and expensive. Furthermore, there are no public dataset addressing emotion in classical music. In contrast to recent popular music, the musical scores for many classical works are available in the public domain. We aim to exploit this advantage along with another existing resource, concert program notes, which are increasingly publicly available [38]. Our contributions in this study are summarized as below.

First, we curate a large dataset of concert program notes using task-specific prompting with a large language model (LLM), `LLaMA 3.1`. Second, we use a dataset of emotion norms [44] to estimate emotional values in terms of valence, arousal, and dominance based on the emotion words that appear in the program notes. Third, we train a transformer model, `Longformer` to make emotion predictions on new program notes. Lastly, we provide a large dataset of concert program notes with emotion annotations to enable further research in music emotion recognition. Our approach is broadly applicable to other tasks requiring efficient text retrieval from noisy and unstructured data sources.

2 Related Work

2.1 Program Notes

Program notes are written descriptions of musical works performed in chamber music performances, orchestra concerts, or an opera, intended to help guide the audience to understand the music and its historical context. Originating in the late 18th century, the first program notes offered brief plot summaries or *argomento* (subject matter) in opera librettos. It was not until the mid-19th century that detailed, analytical program notes began to emerge, corresponding to the rise of public concerts [34]. Notable early examples include Berlioz's printed synopsis for *Symphonie fantastique* (1830) and John Ella's *synopsis analytique* (1845), in which he sought to deepen the audience's musical understanding.

The 19th century saw the development of a more rigorous tradition of writing program notes, with English figures like George Grove and Charles Ainslie Barry contributing thoughtful commentaries on symphonic works [7]. These early program notes provided both analysis and critiques, such as Barry's reflections on Bruckner's Seventh Symphony. By the late 19th century, this practice expanded beyond Britain to continental Europe and America. In these years, program notes began to reflect national tastes and included more extensive analyses, even inspiring thematic guides for certain works, such as the Mahler symphonies.

As classical music concerts expanded globally, program notes became an integral part of the concert-going experience, and by the 20th century, composers

themselves, including Messiaen, Stravinsky, Janáček, and Vaughan Williams, often contributed analytical notes on their own works [45]. The tradition has continued to evolve with technological advancements, with notes becoming more widely available in recorded formats, from LP booklets to CD annotations.

Program notes contain a detailed description of the music: the individual movements, instrumentation, composer's point of view, along with relevant information on art, music history, and contemporary geopolitics [3]. These excerpts are significant not just for offering historical and analytical insights but also for shaping how audiences engage with music, providing context that enhances their listening experience. Many program notes, especially the older ones, describe music with the colorful use of ornate and poetic language which are rich in emotional descriptions. They also describe emotions evoked by different sections or movements of a music work and may subtly allude to emotions by describing characters and events in a story or lyrics. Their evolution reflects broader changes in the relationship between performers, composers, and audiences, highlighting the growing importance of informed musical appreciation. Even though the program notes have been a common staple in concert halls since the early 1900 s, they have often been overlooked and rarely explored in research.

Blom et al. [6] conducted a study with 17 composers to understand role of program notes for listeners and performers. The study concluded that, despite differences in opinions on how the program notes should be written, there is a clear agreement that program notes are useful to relay composers' intentions to the listeners, engage and guide them through the music, and allow them to experience music while shaping their own interpretations. Another study [30] observed that participants enjoyed the music without referencing program notes. However, contrary to the hypothesis that dramatic descriptions of music written with emotive language improve enjoyment more so than structural descriptions, the authors found that it was actually less enjoyed by the participants.

2.2 Optical Character Recognition

The task of extracting text data from image collection in books, artifacts, and historical documents is most often accomplished using techniques such as Optical Character Recognition (OCR) [41] along with a natural language processing (NLP) tool, named entity recognition (NER) [16]. OCR only provides the text in the documents and further processing is usually needed to extract specific sections of the documents. NER is useful to identify and categorize named entities such as person names, organizations, and locations in unstructured documents. Many historical documents, books, and novels lack an organized structure and the extraction of a specific section or other contents from such documents can be challenging. In older documents that are handwritten or unorganized, it is necessary to process the OCR result in order to extract content in an organized manner for better storage and retrieval [32]. We believe that our process of extracting program notes of classical music from OCR text data can be adopted in several fields, including literature, history, and medicine, without requiring the time-consuming tasks of manually finding and organizing the information.



STADIUM CONCERTS REVIEW PAGE 15

A LONDON SYMPHONY

Ralph Vaughn Williams

(Born at Down Ampney on the borders of Gloucestershire and Wiltshire, England, on October 12, 1872)

(Wednesday, July 9)

The first performance of this symphony took place at Queens Hall, London, on March 27, 1914, under the direction of Geoffrey Toye. The British Music Society was responsible for a subsequent performance, allegedly the fourth, wherein a revised version of the work was employed. This was conducted by Albert Coates. The symphony, according to some advices, had been "shortened a good deal, particularly at the closes of the movements. . . ."

Albert Coates directed the piece's first time in America at a concert of the Symphony Society in New York, on December 30, 1920, when the conductor also made his American bow.

It is to Albert Coates that an imaginative exposition of the work's program is due:

"The first movement opens at daybreak by the river. Old Father Thames flows calm and silent under the heavy gray dawn, deep and thoughtful, shrouded in mystery. London sleeps, and in the hushed stillness of early morning one hears 'Big Ben' (the Westminster chimes) solemnly strike the half-hour.

"Suddenly the scene changes (Allegro) one is on the Strand in the midst of the bustle and turmoil of morning traffic. This is London street life of the early hours — a steady stream of foot passengers hurrying, newspaper boys shouting, messengers whistling, and that most typical sight of London streets, the costermonger (Coster 'Arry), resplendent in pearl buttons, and shouting some coster song refrain at the top of a raucous voice, returning from Covent Garden Market, seated on his vegetable barrow drawn by the inevitable little donkey.

"Then for a few moments one turns off the Strand into one of the quiet little streets that lead down to the river and suddenly the noise ceases, shut off as though by magic. We are in the part of London known as the Adelphi. Formerly the haunt of fashionable bucks and dandies about town, now merely old-fashioned houses and shabby old streets, haunted principally by beggars and ragged street urchins.

"We return to the Strand and are once again caught up

(Continued on page 20)

WURLITZER
1859-1941
America's Leading
Musical Instrument House
is Headquarters for
OLD AND NEW
Violins, Violas, Cellos, Double Basses, Best Bows, Strings, Bags, Cases — Artistic Violin and Bow Repairs.
•
Appraisals by our Experts
J. C. FREEMAN and
REMBERT WURLITZER:
$5 verbal; written certificate, $10.
•
WURLITZER
120 West 42nd St., New York
Telephone WIsconsin 7-8140

Victor Records
BRING YOU THE ARTISTRY OF
MISCHA ELMAN
AND OTHER GREAT VIOLINISTS
La Hora Staccato (Dinicu-Heifetz) and Song Without Words (Mendelssohn-Kreisler). 2064. List price 75c.
Minuet in G, No. 2 (Beethoven) and Serenade in G Major (Arensky). 1434. List price 75c.
★
Toscha Seidel — Brahmsiana (Bakaleinikoff) and Albumblatt (Wagner-arr. Wilhelmj). 18014. List price $1.00.
Yehudi Menuhin—Sonata No. 34 in F Major (Mozart) with Hephzibah Menuhin, Pianist. Album M-791*, 4 sides. List price $2.50.
*Also in automatic sequences.
The World's Greatest Artists are on Victor Records
The Gramophone Shop, Inc.
18 East 48th Street 290 Park Avenue

Keep in tune throughout the year with
SHEFFIELD Sealect MILK

Fig. 1. Example page of a program note for Ralph Vaughan William's "A London Symphony", which is split across several pages in the document. The emotion scores for this program are: valence: 0.62, arousal: 0.37, and dominance: 0.62.

2.3 Task-Specific Augmented Generation

LLMs are capable of complex language understanding and question answering, but fine-tuning such models is very resource intensive. Retrieval augmented generation (RAG) [8] is a popular technique to incorporate task-specific knowledge for specialized output generation. There have been research on adding diversity to synthetically generated dataset using RAG [14] and on automated evaluation of synthetic datasets created with RAG [8]. Usually, RAG is an efficient tool used with smart models which are small or have limited context window. With RAG, important information from a given document can be retrieved using similarity search methods which is then augmented to the user prompt along with the task query. This process assists small models to search relevant information from a larger knowledge base given a query and provide appropriate responses based on that added knowledge. In our application, we are using a model with a large context window of 128,000 tokens which allows us to load an entire text document into the context of the model to give out responses in the form of specific program notes for respective music pieces discussed in the program document. Since we retrieve the entire text document without splitting it into smaller chunks, we refer to our method as task-specific augmented generation.

2.4 Annotation of Emotion Scores

Emotion recognition in music is an important topic in music information retrieval. Researchers have attempted to predict emotion from music in different modalities, such as raw audio waveform [21], extracted audio features [33], mel-frequency coefficients [22], social media discourse [15], lyrics [1], and physiological signals [10]. With the continuous improvement of deep learning algorithms, there are opportunities to improve performance in emotion recognition tasks. However, the lack of music emotion data continues to be a serious challenge. First, collecting music audio is often very difficult due to copyright concerns. Second, even after accumulating copyright-free music audio, the task of annotating them with emotion information is a major hurdle. Historically, there have been different forms of emotion annotation such as categorical (e.g., sad, happy, cheerful, angry) [35], or continuous degrees of emotions labeled with valence and arousal values [39]. While the former method seeks to tag music data with emotion categories, the latter attempts to quantify the extent of emotions as a stream of continuous values on a dynamic continuum. The circumplex model was first proposed by Russell to measure emotions in a two-dimensional space of valence and arousal [39]. Valence refers to a measure of pleasantness of a perceived stimulus from a scale of extremely pleasant to unpleasant, while arousal is a measure of excitedness or energy of the stimulus in which the low end of the scale refers to extremely calm and the high end indicates exciting. Dominance, a concept introduced in [31], is the measure of control exerted by the stimuli.

For emotion recognition as continuous values, it is necessary to log the annotations in the form of valence and arousal, but human labeling of emotions to

music is expensive and time-consuming to conduct. As such, techniques to map English words to emotional dimensions are increasingly common in research. Warriner et al. published a large list of English words with affective ratings for valence, arousal, and dominance (VAD) for 13,915 words [44], expanding on the earlier and smaller ANEW dataset [27]. These VAD emotion scores are given on a 9-point scale and the emotion norms for each word is the average of the all participant responses to the particular word. These norms established an understanding of the emotional impact of individual words and can be used to infer emotions of text documents by calculating the average across the words in the text. Given the large vocabulary of emotion annotated words, these norms have become popular in research for annotation of emotion in text data.

In a study designed to understand the mental impact of reading, recalling, and discussing literary fiction, the authors use Warriner's norms to calculate the overall emotion scores from literary works [9]. The authors considered twelve novel-length fictional books, 4,000 short stories, and social media discourse from Reddit and Twitter. In their study, they leveraged the aforementioned norms and concluded that engaging in fiction has positive effects on the mood of the readers. In the music realm, one study [12] used the same approach to infer emotions elicited by songs based on valence and arousal values of emotion words present in a set of tags associated with the song. Another study applied this approach to estimate the emotional content of songs by analyzing the emotion words present in social media comments about specific pieces of music [4]. In our approach, we also implement this method to calculate VAD scores. We create a dataset of emotion scores for program notes which can be used as proxy to understand the overall emotions in the corresponding classical music piece.

2.5 Transformers

The main objective for building a dataset of music works and emotion scores is to explore their applications in music recommendation, automatic music generation, and music therapy. There have been numerous studies on emotion recognition in music in various forms, and text is one of the most approachable forms for analysis and prediction of emotion information. Apart from music, sentiment analysis has broad applications in content filtering, mood based recommendation, and personalized advertisements [43]. In the past, sentiment analysis used text analysis based on tokenization and similarity metrics to find meaningful relationships between different words in a sequence of text. More recently, researchers turned to deep learning approaches to perform sentiment analysis. Initially, recurrent neural networks and long short-term memory networks were used for their ability to process long sequences of text and learn different relationships in the text. However, the introduction of encoder-decoder based deep learning architecture, the transformer [42], provided new approaches to text analysis.

One such architecture, `BERT` [13], has demonstrated its ability to learn important representations and relationships in text. The `BERT` model was trained with

two major objectives, next sentence prediction (NSP) and masked language modeling (MLM). Over the years, numerous BERT-based models have been developed for a variety of applications such as text summarization, sequence classification, question answering, and text sentiment analysis [49]. The primary limitation of BERT-based models is its inability to handle long sequences of text because the attention mechanism scales quadratically with the length of the input. This limits the length of the input to only 512 tokens in most BERT-like models.

Several models, such as Big Bird [47], Longformer [5], LongT5 [19], Reformer [26], and Nystromformer [46], have attempted to linearize the scaling by adjusting the self-attention mechanism of the model. Longformer is the most efficient with an attention mechanism that scales linearly with the input sequence length and can process documents with up to 4,096 tokens. To achieve this, Longformer combines its self-attention mechanism with local windowed attention in which each token attends to a configurable variable number of previous and succeeding tokens along with a task-motivated global attention in which some tokens attend to all other tokens. Longformer was trained only on the MLM objective and not the NSP objective because the RoBERTa [28] model demonstrated that removing NSP loss did not affect performance. Longformer is trained from RoBERTa's pretrained checkpoint with changes mostly in the attention mechanism.

3 Approach

While there are a few online databases of program notes for music, most consider only a small set of musical pieces and consist of program notes written by authors which often lack long descriptions of music scores. In this study, we collect program notes from the archive of New York Philharmonic Society (NY Phil)[1] which boasts a large collection of program documents spanning 180 years. Each document consists of a booklet of program notes for the music performed in that concert. In addition to the program notes, these raw documents also contain advertisements, descriptions of the ensemble, and a list of performers, among other superfluous information. Our approach is divided into four tasks: data gathering, program note extraction, emotion labeling, and emotion prediction.

3.1 Data Gathering

First, we downloaded the original program notes from the NY Phil archive. To date, there are 14,743 program documents available. Each document is a collection of scanned images of the pages in the program booklet. The NY Phil uses OCR to extract the text in each page. First, we compile this noisy text across the many different pages of the program. Next, we leveraged the metadata made available by NY Phil to download the programs. This resulted in a page-wise text

[1] https://archives.nyphil.org/.

of each document along with other metadata such as id, programID, orchestra, season, concert information (event type, location, venue, and date/time), and information about the musical work (composer, title, movement, and soloists).

3.2 Program Note Extraction

Following the collection of raw OCR text data, we attempted to extract corresponding program note text for any given music composition in each program booklet. This proved challenging because of the lack of explicit distinction between program notes of different pieces in the OCR text. The formatting and organization varied greatly across the data. Moreover, the presence of extraneous content such as introduction pages, advertisements, announcements, and content in headers and footers, presented even more challenges. Additionally, because the style and design of the concert program booklets varied depending on the time period in terms of the fonts, writing style, and event types, which made the use of regular expression pattern matching infeasible. Given these challenges, we designed a robust method of text extraction that succeeds for the majority of documents with only occasional failures that required human review.

Our approach leverages the capabilities of an LLM trained on trillions of tokens of publicly available data [40]. Specifically, we use the open-source LLM released by Meta AI `LLaMA 3.1`, which contains 70 billion trainable parameters. Trained on 15 trillion tokens and fine-tuned for use as a chat assistant, `LLaMA 3.1` demonstrates a high level of contextual understanding, reasoning, and question-answering with state-of-the-art performance on standard benchmarks. Chat assistants like `LLaMA 3.1` accept an input prompt and give responses by generating the best probable text output. For our task of extracting the program note text for a specific music piece, the use of an LLM is an efficient way to search the input for the requested program note and provide the response word-for-word. With the increase in context length in `LLaMA 3.1` to 128,000 tokens compared to `LLaMA 3`'s limit of 8,192 tokens, it was possible to feed the entire text in a program note document into the prompt and retrieve the specific program note for any music piece among several others from the same document. We performed prompt-tuning, evaluating different prompts to discover which works the best in achieving the desired outcome. A sample of a prompt that we observed performing the best for the text retrieval is provided as Table 1.

In our approach, we implement task-specific prompt augmented generation. First, for a given document, we create one text document file that consists of the text from all the pages in the metadata for each program document. We use `all-mpnet-base-v2`, a bi-encoder sentence transformer model [36] to encode the text into a word vector embedding which we store in a database in GPU memory using the FAISS library [25]. Next, we augment the entire text into our prompt and instruct the `LLaMA 3.1` model to extract the program notes for a given music composition. To better assist the model in finding the related content, we include the name of the composer, the title of the music score, and the name of the movement of the music score (if present) into the prompt for retrieval

Table 1. The prompt format that was provided to `Llama 3.1` for program note extraction. Specific music score title, composer's name, and movement name are passed in each run to the model for efficient retrieval of the desired program note. The inclusion of these information multiple times was observed to help model better in finding the requested piece of text in the document. At the end, the context is appended to the prompt which is the program document.

Example prompt
Please provide the complete program note for {`composition_name`}, {`movement_name`} by {`composer_name`}. Do not include any starting sentences such as 'Here is the complete program note for ...'. If you do not find any program note description for the requested music composition, please just respond with: 'Not Available'. Respond to this request based on your knowledge and the provided context. Find all the relevant text that is part of a program note for {`composition_name`}, {`movement_name`} by {`composer_name`}. There can be other program notes as well, but focus mainly on the starting of the program notes for {`composition_name`} by {`composer_name`}. To provide you more context, these texts were converted through Optical Character Recognition (OCR). Along with the very important concert program notes, there are some unwanted elements such as advertisements and irrelevant information in between. You need to get rid of such unwanted and irrelevant elements from the text. And sometimes you will find some OCR text conversion errors like invalid characters and misspellings. Please fix those as much as possible. Also, there might be discontinuations due to advertisements, page breaks, and the program note text can be split in different pages which you should identify and merge together. To reemphasize, your primary objective is to accurately extract the whole concert program note text for the {`composition_name`} by {`composer_name`} while disregarding unrelated content. Ensure to preserve the narrative of the description from start till the very end without any summarizing, rewriting, and missing out on relevant parts. Remember to filter out any content that are not part of the concert program notes, such as advertisements or generic announcements, ensuring the focus remains on the program note description. Use the following context to help: {`context`}.

of program notes for each music score in a given program document. Following the instructions provided in the prompt in Table 1, the model removes all the superfluous information and fixes OCR mistakes in its responses. It also organizes the program note into one continuous block of text despite the pagination and interruption by advertisements, headers, footers, and other irrelevant text. Apart from occasional hallucinations, the model responds with a specific message when the requested program note is not available in the program document.

3.3 Emotion Labeling

In this step, we annotate the program notes with emotion scores. To accomplish this, we use emotional values in the 3-dimensional VAD space. Given a program note, we reference each word present in the list of norms [44]. Words are reduced to their root form with lemmatization. For example, words in their inflected forms such as `running`, `runs`, and `ran` are all converted to the lemma `run`. However, if an inflected word is present in the word list itself, we do not reduce to its lemma. We then calculate the sum of emotional values based on the scores for each emotion word present in the program note and determine the normalized average score for that program note in the range of 0 to 1. We use this average to label each program note with valence, arousal, and dominance scores. An example of a program note with calculated emotion scores is given in Fig. 1.

3.4 Emotion Prediction

Now that we have a dataset of program notes with emotion scores, we desire the ability to predict emotion scores for any given program note. For this, we train a deep learning model on our dataset to learn to predict emotional values directly from the text. Program notes in our dataset are of variable sequence lengths ranging from 30 to more than 5,000 tokens. Given the efficiency of handling long input sequences and understanding various relationships between the words in the input sequence, we use `Longformer`. The `LongformerModel` model implementation provided by the HuggingFace NLP library[2] is used. We are motivated to create an emotion recognition model to make emotion predictions from shorter texts or from social media comments which might lack emotion words which hinders our ability to infer emotions using the pre-defined word norms.

In our approach, we first split our entire dataset randomly into train, validation, and test partitions (0.70, 0.15, and 0.15). Before passing the text into a neural network model, it first needs to be converted into tokens and then subsequently converted into tensors so that mathematical computations are possible between the tensor representations of the input tokens. For tokenization, we use the `LongformerTokenizer` from HuggingFace. This splits the input into several words and adds special separator tokens represented by `[SEP]` to separate sentences and appends a classifier token `[CLS]` to the start of the sequence.

During tokenization, we set the maximum length of tokens to 4,096 as expected by the model and truncate the input if it exceeds this limit. Additionally, we pad the tokens to the maximum length to ensure all inputs have the same length before training. The tokenizer returns the attention masks and the input sequence as tensors. Next, we train our model with a batch size of 16, learning rate of 1×10^{-5} and use the mean-squared error as the loss for our regression task. We use the default pre-trained `Longformer` architecture: twelve encoder layers and twelve self-attention heads, to train on the input tensors. In the training process, which runs for five epochs, forward passes of each

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

batch first obtain the hidden representations of the output at the last layer of the model. Then, we take the hidden representation of the `[CLS]` token of each sequence in the batch which we further process using a linear layer and a `tanh` activation function. We further reduce the dimension of the output to 128 and use `LeakyReLU` as the activation function to facilitate a smoother gradient flow. Finally, we apply a linear layer to map the previous 128-dimensional representation to our desired 3-dimensional output of valence, arousal, and dominance. After this, we set the model to evaluation mode and make predictions on the test set before performing a correlation analysis of the predicted values and the calculated emotion values.

4 Results

From our collection of 14,743 program documents, we first filtered out those without program notes. This occurs when the metadata suggests the presence of pieces performed in the concert but the program document contains no text. Filtering these out resulted in 13,849 total documents. From these documents, we extracted 70,298 total music works. However, due to the inherent mismatch between the number of music works in the metadata and the actual number of program note descriptions present in the program document, we removed 32,084 performances for which program notes were not available.

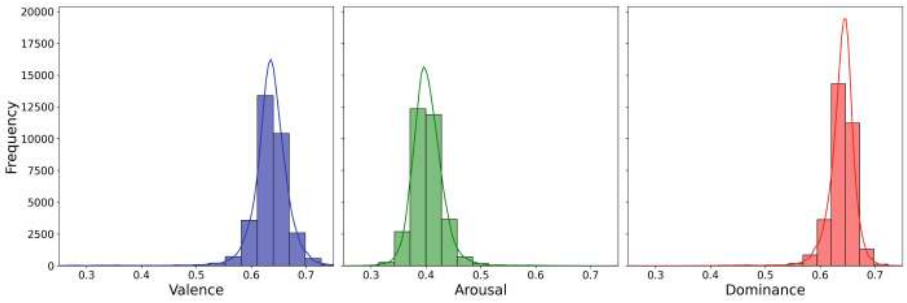


Fig. 2. The distribution of emotion scores across the dataset. Valence has a mean μ=0.634 $\pm$ 0.039; arousal, μ=0.403 $\pm$ 0.030; and dominance, μ=0.639 $\pm$ 0.027.

It is known that LLM models sometimes hallucinate and provide incorrect outputs [24], a phenomenon which we also observed in our process. Although manual inspection of thousands of documents is impractical, after targeted inspections, we found that in many cases a document consists of the music composition information such as title and composer's name, but do not contain the program notes or just have a few trailing texts such as the names of movements and date and place of birth of the composer under the title. When the LLM then extracts erroneous information as program notes, these data points needed to

be excluded. Although our solution resulted in the removal of some very short descriptions of music scores, we made the decision to drop programs with less than 25 words. Sometimes, in the cases of poor OCR conversion of non-English languages, the model provided hallucinated responses in the form of repetitions of music score titles or repeated patterns of random alphabets and numbers until the model reached maximum token generation length which was set to 8,000 tokens. Responses with over 6000 tokens were identified as hallucinations, leading to the removal of 932 entries. Furthermore, we attempted to identify hallucinated responses using keyword matching which resulted in the removal of 351 incorrect outputs which contained some or all parts of the prompt in the output and we observed some instances in which the generated text consisted of some of the metadata from the retrieval process which was augmented into the prompt.

To detect hallucinations where the extracted program notes were not present in the original text but instead were composed by LLM, we performed further manual reviews at random, however the lack of a reliable, fine-grained ground truth makes the technical evaluation of our dataset challenging. Our ground truth, which is the coarse OCR text and each of such text might contain scattered program notes for multiple pieces. Therefore, it is infeasible to use evaluation techniques such as Word Error Rate (WER) and Bilingual Evaluation Understudy (BLEU) [37], which are popular in measuring the performance of LLMs for language translation. For example, BLEU measures the quality of machine translation by comparing the similarity of the words in the sentences of the generated output to the ground truth. Measured between 0 and 1, the higher the BLEU score for a given text output, the closer it is to the ground truth.

In research, a score around 0.6 or 0.7 is considered very good for machine translation tasks. In our case, when we calculate BLEU score for the extracted program note against the coarse OCR text, the resulting scores are variable and difficult to interpret. This is due to the varying length of text in the ground truth which directly affects the BLEU score even when the extracted data is correct. Moreover, the dispersion of program notes in several parts of the OCR text, given the unwanted information in between, will significantly lower the BLEU even if the LLM was able to stitch the parts together correctly. Upon manual inspections for several entries, we carefully selected 0.006 as the threshold for BLEU score below which the entries need to be removed from our dataset. While this threshold was manually selected, we found it is necessary to rid the dataset of incorrect and hallucinated responses as possible. The removal of 2,614 entries with a BLEU score of less than 0.006 resulted in a total of 31,966 program notes from 9,569 different documents. The distribution of emotion scores in our dataset is given in Fig. 2. We concede while there might still be some entries that need to be removed after future evaluations and reviews, our thresholding does not remove any meaningful entries from our dataset.

We evaluate our models performance using Pearsons correlation between our models predictions and the calculated values for valence, arousal, and dominance as seen in Table 2. We observe strong correlations for VAD of 0.877, 0.823, and

Table 2. Pearson's correlation (r), Spearmans's correlation (p), and R-squared score between the predictions and the calculated values in our dataset for VAD.

	r	ρ	R^2
Valence	0.877	0.856	0.751
Arousal	0.823	0.841	0.630
Dominance	0.766	0.754	0.553

0.766 respectively. We also calculate the Spearman's rank correlation, which assesses monotonic relationships between two variables compared to linear relationships assessed by Pearson's correlation. With the absence of obvious outliers in the predictions, Spearman's ρ values are very similar: 0.856, 0.841, and 0.754, respectively, for valence, arousal, and dominance. In regression tasks, it is also important to examine how the variance in the data is explained by the model, using the R^2 metric. A value of 0 means that the model does not explain any of the variability, whereas a value of 1 indicates that the model's predictions perfectly fit the data. We observe high R^2 values for valence with 0.751 and a slightly lower value for arousal which is 0.630. However, our model's ability to predict dominance was lower, with R^2 value of 0.553. Overall, these results imply that our model is able to predict the emotion scores directly from the text.

5 Discussion

Automatic text extraction from large unstructured documents is a challenging task because of the diverse content and varying organization of the documents. Although we attempted to create a system that is as consistent and accurate as possible, we still observed occasional irregularities in the outputs. After experimenting with different word embedding models, open source LLMs, and prompting strategies, we observed the most consistent outputs with the approach presented in this study. While exploring the best query and prompt to help the LLM detect relevant information from the program document, we observed an explicit reference to the work title and composer generated the best results.

Because of variations in the content and organization of the documents across the corpus and given the presence of multiple advertisements and irrelevant information that broke the program notes across pages of the document, we observed varying performance of our model. Even with our diligent attempts using keyword matching and BLEU score to remove incorrect outputs to create a dataset of 31,966 program notes, there may still remain some incorrectly generated program notes. One of our next steps is to identify and remove such items from the dataset. We manually verified hundreds of extracted program notes. In our observation, the majority were accurately extracted by our process.

The occasional errors arise due to several reasons. One of the most important observation in our study was that the model, even with its large context window, could not handle very long sequence of tokens, which was observed for large

program documents with more than 30,000 tokens. However, on a positive note, we observed that in more structured documents with a few superfluous text blocks, the model was successfully able to extract very long program notes with more than 5,000 tokens, which is not possible using other open source models that usually have a context window of 8,192 tokens or less. For example, `LLaMA 3` prematurely stops generating text after its context length of 8,192 tokens is reached. We hypothesize that given the LLM's inability to handle very long sequences of tokens, the model was simply unable to cohere all the information and generated hallucinated outputs which was seen in program documents with mostly non-English text that the OCR incorrectly converted to gibberish. We also observed some limitations given flaws in the original data itself. In our approach, we observed that our system works best in well-organized documents with infrequent occurrences of advertisements or other unrelated content.

We observe a clustering of the distribution of emotion score as seen in Fig. 2. This is because of the bias introduced by the words associated with neutral emotions which pushes the scores closer towards the mean of the distribution. Our approach does not consider the context of the words which can sometimes have different emotional connotations than those in the word list. This may be improved by providing different weights to the neutral and high or low emotion scores when calculating the overall emotion scores for program notes.

The amount of GPU memory required to process large amount of texts using a big model like `LLaMA 3.1` poses a significant resource challenge. Furthermore, the cost of using paid proprietary models is too high when working with millions of tokens of text, which motivated us to find the best open source model for our study. As future work, we will implement RAG with `LLaMA 3.1` with a reranking strategy which will assist in finding the most relevant text portions from the document. We will then retrieve only those portions from the vector database to augment the prompt which will help to reduce the sequence length fed into the model. By evaluating the model's behavior with the updated approach and comparing the results may offer useful insights into its performance. We also intend to use another comparable LLM for program note extraction and identify potential hallucinations by checking for agreement between two outputs.

We create and release this dataset to enable researchers to estimate emotion for music works for which the copyright free audio is not unavailable. Moreover, the data extraction process introduced in this paper can be used in various other fields such as history, literature, and medicine, in which data collection from historical archives is extremely important. With further refinement of the dataset, our next goal is to collect copyright free recordings of classical music for which program notes are available in our database. We believe that with the dual modality of program notes and audio of the same work, we can create a more robust emotion recognition system using joint embeddings.

6 Conclusion

Existing datasets of music and emotion are not large enough to build emotion recognition systems and, given copyright concerns and the overhead of human

annotation, the creation of large datasets is an arduous challenge. In this study, we curate a novel dataset of program notes with emotional annotations. Our novel approach explores the use of program notes to estimate emotions and yield pseudo-annotations for the corresponding classical music works. We use `LLaMA 3.1` in conjunction with a task-specific prompt augmented generation to extract the fine-grained program note texts from coarse text documents generated by OCR. We release this dataset publicly[3] to be used as a proxy to infer emotions elicited by works of classical music. To our knowledge, this is the first dataset of classical music annotated with emotion. Furthermore, this approach for program note data extraction can be applied to similar data extraction contexts.

Additionally, we train a language model to automatically estimate valence, arousal, and dominance directly from long-form text documents. This model can be used by researchers to quickly and automatically annotate blocks of text with emotion values. With this research, we hope to inspire researchers to build new emotion dataset for MER research efficiently and on a large scale.

References

1. Agrawal, Y., Shanker, R., Alluri, V.: Transformer-based approach towards music emotion recognition from lyrics. In: Hiemstra, D., Moens, M.-F., Mothe, J., Perego, R., Potthast, M., Sebastiani, F. (eds.) ECIR 2021. LNCS, vol. 12657, pp. 167–175. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-72240-1_12
2. Aljanaki, A., Yang, Y.H., Soleymani, M.: Developing a benchmark for emotional analysis of music. PLoS ONE **12**(3), 1–22 (2017). https://doi.org/10.1371/journal.pone.0173392
3. Bashford, C.: Writing (british) concert history: the blessing and curse of ephemera. Notes **64**(3), 458–473 (2008). https://doi.org/10.1353/not.2008.0023
4. Beery, A., Donnelly, P.J.: Learning affective responses to music from social media discourse, pp. 93–119. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-44260-5_6
5. Beltagy, I., Peters, M.E., Cohan, A.: Longformer: the long-document transformer. arXiv preprint arXiv:2004.05150 (2020). https://doi.org/10.48550/arXiv.2004.05150
6. Blom, D.M., Bennett, D., Stevenson, I.: The composer's program note for newly written classical music: content and intentions. Front. Psychol. **7** (2016). https://doi.org/10.3389/fpsyg.2016.01707
7. Bower, B.: The routledge companion to applied musicology, chap. George Grove, Programme Notes, and the Dictionary, 46–54. Taylor & Francis (2023)
8. Bradley, M., Lang, P.: Affective Norms for English Words (ANEW): instruction manual and affective ratings (1999). https://www.semanticscholar.org/paper/Affective-Norms-for-English-Words-(ANEW)%3A-Manual-Bradley-Lang/c765eb0a31849361d829b24e173a37bab0919892
9. Carney, J., Robertson, C.: Five studies evaluating the impact on mental health and mood of recalling, reading, and discussing fiction. PLoS ONE **17**(4), 1–32 (2022). https://doi.org/10.1371/journal.pone.0266323

[3] https://github.com/Soundbendor/program_notes_data.

10. Chaturvedi, V., Kaur, A.B., Varshney, V., Garg, A., Chhabra, G.S., Kumar, M.: Music mood and human emotion recognition based on physiological signals: a systematic review. Multimedia Syst. **28**(1), 21–44 (2021). https://doi.org/10.1007/s00530-021-00786-6
11. Chen, Y.A., Yang, Y.H., Wang, J.C., Chen, H.: The AMG1608 dataset for music emotion recognition. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 693–697. IEEE (2015). https://doi.org/10.1109/ICASSP.2015.7178058
12. Delbouys, R., Hennequin, R., Piccoli, F., Royo-Letelier, J., Moussallam, M.: Music mood detection based on audio and lyrics with deep neural net. In: Proceedings of the 19th International Society for Music Information Retrieval Conference, pp. 370–375. ISMIR (2018). https://doi.org/10.5281/zenodo.1492427
13. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of naacL-HLT. vol. 1, pp. 4171–4186. Association for Computational Linguistic (2019). https://doi.org/10.18653/v1/N19-1423
14. Divekar, A., Durrett, G.: SynthesizRR: generating diverse datasets with retrieval augmentation. arXiv preprint arXiv:2405.10040 (2024). https://doi.org/10.48550/arXiv.2405.10040
15. Donnelly, P., Beery, A.: Evaluating large-language models for dimensional music emotion prediction from social media discourse. In: Abbas, M., Freihat, A.A. (eds.) Proceedings of the 5th International Conference on Natural Language and Speech Processing (ICNLSP 2022), pp. 242–250. Association for Computational Linguistics, Trento, Italy (2022)
16. Ehrmann, M., Hamdi, A., Pontes, E.L., Romanello, M., Doucet, A.: Named entity recognition and classification in historical documents: a survey. ACM Comput. Surv. **56**(2), 27:1–27:47 (2023). https://doi.org/10.1145/3604931
17. Feng, H., Mahoor, M.H., Dino, F.: A music-therapy robotic platform for children with autism: a pilot study. Front. Robot. AI **9** (2022). https://doi.org/10.3389/frobt.2022.855819
18. Gaur, S., Donnelly, P.J.: Generating smooth mood-dynamic playlists with audio features and KNN. In: Johnson, C., Rebelo, S.M., Santos, I. (eds.) Proceedings of the 13th International Conference of Evolutionary and Biologically Inspired Music, Sound, Art and Design (EvoMUSART). vol. 14633, pp. 162–178. Springer (2024). https://doi.org/10.1007/978-3-031-56992-0_11
19. Guo, M., et al.: Longt5: efficient text-to-text transformer for long sequences. In: Carpuat, M., de Marneffe, M.C., Meza Ruiz, I.V. (eds.) Findings of the Association for Computational Linguistics: NAACL 2022, pp. 724–736. Association for Computational Linguistics, Seattle, United States (2022). https://doi.org/10.18653/v1/2022.findings-naacl.55
20. Han, D., Kong, Y., Han, J., Wang, G.: A survey of music emotion recognition. Front. Comp. Sci. **16**(6), 1–11 (2022). https://doi.org/10.1007/s11704-021-0569-4
21. He, N., Ferguson, S.: Multi-view neural networks for raw audio-based music emotion recognition. In: 2020 IEEE International Symposium on Multimedia (ISM), pp. 168–172 (2020). https://doi.org/10.1109/ISM.2020.00037
22. Hizlisoy, S., Yildirim, S., Tufekci, Z.: Music emotion recognition using convolutional long short term memory deep neural networks. Eng. Sci. Technol. Int. J. **24**(3), 760–767 (2021). https://doi.org/10.1016/j.jestch.2020.10.009
23. 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 genera-

tion. In: Proceedings International Society for Music Information Retrieval Conferences, pp. 318-325. ISMIR, Online (2021). https://doi.org/10.5281/zenodo.5624519
24. Ji, Z., et al.: Survey of hallucination in natural language generation. ACM Comput. Surv. **55**(12) (2023). https://doi.org/10.1145/3571730
25. Johnson, J., Douze, M., Jégou, H.: Billion-scale similarity search with GPUs. IEEE Trans. Big Data **7**(3), 535–547 (2021). https://doi.org/10.1109/TBDATA.2019.2921572
26. Kitaev, N., Kaiser, L., Levskaya, A.: Reformer: the efficient transformer (2019). https://doi.org/10.48550/arXiv.2001.04451
27. Lewis, P., et al.: Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (eds.) Advances in Neural Information Processing Systems, vol. 33, pp. 9459–9474. Curran Associates, Inc. (2020)
28. Liu, Y., et al.: RoBERTa: a robustly optimized BERT pretraining approach. CoRR **abs/1907.11692** (2019). http://arxiv.org/abs/1907.11692
29. Liu, Z., Xu, W., Zhang, W., Jiang, Q.: An emotion-based personalized music recommendation framework for emotion improvement. Inform. Process. Manage. **60**(3), 103256 (2023). https://doi.org/10.1016/j.ipm.2022.103256
30. Margulis, E.H.: When program notes don't help: music descriptions and enjoyment. Psychol. Music **38**(3), 285–302 (2010). https://doi.org/10.1177/0305735609351921
31. Mehrabian, A., Russell, J.A.: An Approach to Environmental Psychology. The MIT Press, Cambridge (1974)
32. Nguyen, T.T.H., Jatowt, A., Coustaty, M., Doucet, A.: Survey of post-ocr processing approaches. ACM Comput. Surv. **54**(6), 124:1–124:37 (2021). https://doi.org/10.1145/3453476
33. Panda, R., Malheiro, R., Paiva, R.P.: Novel audio features for music emotion recognition. IEEE Trans. Affect. Comput. **11**(4), 614–626 (2020). https://doi.org/10.1109/TAFFC.2018.2820691
34. Pasler, J.: Concert programs and their narratives as emblems of ideology. Int. J. Musicology, 249–308 (1993)
35. Plutchik, R.: A general psychoevolutionary theory of emotion. In: Plutchik, R., Kellerman, H. (eds.) Theories of Emotion, pp. 3–33. Academic Press (1980). https://doi.org/10.1016/B978-0-12-558701-3.50007-7
36. Reimers, N., Gurevych, I.: Sentence-BERT: sentence embeddings using Siamese BERT-networks. In: Inui, K., Jiang, J., Ng, V., Wan, X. (eds.) Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 3982–3992. Association for Computational Linguistics, Hong Kong (2019). https://doi.org/10.18653/v1/D19-1410
37. Reiter, E.: A structured review of the validity of BLEU. Comput. Linguist. **44**(3), 393–401 (2018). https://doi.org/10.1162/coli_a_00322
38. Ridgewell, R.: The concert programmes project: history, progress and future directions. Fontes Artis Musicae, 50–64 (2010)
39. Russell, J.A.: A circumplex model of affect. J. Pers. Soc. Psychol. **39**(6), 1161–1178 (1980). https://doi.org/10.1037/h0077714
40. Touvron, H., et al.: LLaMA: open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023). https://doi.org/10.48550/arXiv.2302.13971
41. Vamvakas, G., Gatos, B., Stamatopoulos, N., Perantonis, S.: A complete optical character recognition methodology for historical documents. In: 2008 The Eighth IAPR International Workshop on Document Analysis Systems, pp. 525–532 (2008). https://doi.org/10.1109/DAS.2008.73

42. Vaswani, A., et al.: Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 6000–6010. NIPS'17, Curran Associates Inc., Red Hook, NY (2017)
43. Wankhade, M., Rao, A., Kulkarni, C.: A survey on sentiment analysis methods, applications, and challenges. Artif. Intell. Rev. **55**(7), 5731–5780 (2022). https://doi.org/10.1007/s10462-022-10144-1
44. Warriner, A.B., Kuperman, V., Brysbaert, M.: Norms of valence, arousal, and dominance for 13,915 english lemmas. Behav. Res. Methods **45**(4), 1191–1207 (2013). https://doi.org/10.3758/s13428-012-0314-x
45. Weber, W.: Concert (ii). Grove Music Online (2001). https://doi.org/10.1093/gmo/9781561592630.article.06240
46. Xiong, Y., et al.: Nyströmformer: a nyström-based algorithm for approximating self-attention. In: Proceedings of the AAAI Conference on Artificial Intelligence vol. 35, no. 1616, pp. 14138–14148 (2021). https://doi.org/10.1609/aaai.v35i16.17664
47. Zaheer, M.: Big bird: transformers for longer sequences. In: Proceedings of the 34th International Conference on Neural Information Processing Systems, pp. 17283–17297. NIPS '20, Curran Associates Inc., Red Hook, NY (2020)
48. Zhang, K., Zhang, H., Li, S., Yang, C., Sun, L.: The PMEmo dataset for music emotion recognition. In: Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval, pp. 135–142. ICMR '18, ACM, New York (2018). https://doi.org/10.1145/3206025.3206037
49. Zhou, C., et al.: A comprehensive survey on pretrained foundation models: a history from BERT to ChatGPT. arXiv preprint arXiv:2302.09419 (2023). https://doi.org/10.48550/arXiv.2302.09419