Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# LyBERT: Multi-class classification of lyrics using Bidirectional Encoder Representations from Transformers (BERT)

**Revathy V Rajendran** ( ✉ revathyvrajendran@gmail.com )
Hindustan Institute of Technology and Science
**Anitha S Pillai**
Hindustan Institute of Technology and Science
**Fatemah Daneshfar**
University of Kurdistan

**Research Article**

**Keywords:** Emotion classification, transformer approach, music lyrics recommendation, BERT, Music Information Retrieval, Emotion detection from lyrics

**Posted Date:** April 5th, 2022

**DOI:** https://doi.org/10.21203/rs.3.rs-1501499/v1

**License:** 🅭 🅯 This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

# Abstract

Recent developments in music streaming applications and websites have made the music emotion recognition task continually active and exciting. Recognizing the music mood has many advantages. One of them is that it helps find out the prominent feeling or state of mind it brings to a listener. Identifying a listener's taste is the primary motive of currently available music emotion recognition systems such as music streaming systems (YouTube), music recommender systems (Spotify), automatic playlist generation, etc. Being a subdomain of music information retrieval (MIR), for the past years, several challenges of music emotion recognition have been studied and solved by researchers. Music emotion recognition's significant challenges include data accessibility, data volume, recognizing emotionally relevant features, etc. Several researchers have proved that emotionally relevant features can be identified by analyzing multiple features and the semantic features of lyrics and effects of audio signals create music emotion. Then one can initiate the music recognition task from a lyrical perspective because it has semantic features and audio features such as valence and arousal. The challenging part is the availability of these features that influence the music's emotion. The lyrical features relevant for identifying four emotions (happy, sad, relaxed, and angry) were learned with the help of state-of-the-art algorithms. After that, those features were used to predict the feelings of Music4All dataset lyrics from the Music Emotion Recognition (MER) dataset. This work tries to identify the importance of these features through transfer learning and combining pre-trained model BERT. After transfer learning, the BERT model is then applied to the dataset to improve the model's accuracy. The overall accuracy achieved is 92%.

# 1 Introduction

Music is a special kind of art that plays several roles in the lives of human beings. Music adds social and cultural elements by influencing humans in diverse ways. The central part of economic and social lives that affect human growth is happening in digital space. High-speed internet and multimedia applications have become part of daily life. The proliferation of smartphones and mobile internet has brought several music streaming applications at an accelerating rate. Music has emotional connectivity with an individual's life. Music streaming websites and music information retrieval has become an active research field.

Emotions are pretty common among humans during communication. They help recognize moods and feelings, which are an inevitable part of humans and their personalities. Music has a significant role in evoking feelings and affecting social activities and interactions. From ancient times music has played a substantial role in each day of a human as a language that brings emotions into their deep mindset. Thus, we can say music can be recalled as a medium to communicate emotion. In a musical environment, humans express emotional moods such as happiness, sadness, aggressiveness, tenderness, etc.

Music mood recognition is the discovery process wherein the emotions of a musical piece are identified through various means, including audio and lyrical text analysis. A lot of research work on music classification is based on audio signals and the features of the music. Even though music audio and lyrics are firmly connected, the brain operates the music features, specifically the semantic features of the lyrics and the audio signals' harmonic (tunes) features, separately [1]. Thus, the investigation of music emotion can start from the view of lyrics as it contains prominent features that represent emotionally dependent and relevant information

Also, a new era of music recommendation systems often uses information retrieval methods, leveraging more collaborative approaches, specifically including listeners' history to improve recommendations. Such recommender systems may suffer sparsity problems which is an issue due to the missing of an unpopular song recommendation. These systems may fail to recommend new songs or songs that lack user interaction repeatedly. As a result, it biases listening patterns and might forget to recommend less widely known tracks (and therefore have sparser user listening history) [1]. In simple words, music retrieval systems need a deep understanding of what music is all about. Deep knowledge of music can help find similar songs in a better way. Here the vital part of the music, such as lyrics and audio, play an inevitable role. Emotions behind the lyrics and audio of the music bring deep light to learning the emotional theme of a song and can better relate to the user's emotional preference over it.

This paper seeks to investigate which method is suitable to label the text data and which model is the most promising and efficient for emotion classification. Lyrics by themselves are deep information about a song that can reveal meta-data features such as the type, sentiment, and message of a piece simply by reading its lyrics. This work aims to automate the same. For this work, a classifier is built that can predict whether a song is happy, sad, relaxed, or angry based on its lyrics. Such a classifier can easily find its use in the music industry. For example, it can automatically classify and form playlists in music players or recommender systems.

This paper proposes an improvised approach to label lyrics data based on emotionally relevant features and pre-trained word-embeddings. This work shows the importance of pre-trained models and their utilization for performing transfer learning from a corpus dataset to the lyric dataset. The proposed model can be re-trained to perform multi-class text classification and improve performance. The pre-trained model is suitable for predicting songs' emotions, and this prediction can be applied to solve several challenges in Music Information Retrieval (MIR) tasks.

# 2 Literature Review

The authors performed a multi-class emotion classification of lyrics and created a labeled emotion dataset [17]. This work used the BERT model to run on the out-of-the-domain dataset and used that model to classify songs based on joy and sadness, but the BERT model's performance was not better when compared to Naïve Bayes approach [17]. Their study concluded that when datasets from different domains other than song lyrics containing emotions from conversation and tweets cannot be used to predict emotion labels of song lyrics [17]. In [11], the authors identified several features to classify the emotions of lyrics. The authors selected the songs that one could demonstrate according to Russell's Model (Fig. 1.). Major features considered for extracting the music emotions are Content-Based Features (CBF), Song-Structure-Based Features (StruBF), Stylistic-Based Features (StyBF), and Semantic-Based Features (SemBF)[11]. The authors also constructed a lyrics emotion dataset of 771 song lyrics from the AllMusic platform into four emotions which are happy, angry, sad, and relaxed [11]. To pre-train the proposed BERT model in this paper and identify the emotions of lyrics of the Music4All dataset, the dataset [11].

BERT is used widely in sentiment classification tasks. Authors have done an Algerian dialect texts classification task using several deep learning techniques [2]. They applied contextual embedding for the analysis using the BERT model. Alberti et al. have applied BERT for the first time on the Natural Questions corpora [3] for question answering, outperforming other approaches used in the study. Auxiliary sentence creation is another trend in aspect-based sentiment analysis. Chi Sun et al. has created auxiliary sentences using BERT from aspects and applied sentence pair classification tasks such as question-answer creation and Natural language Inference (NLI) and achieved higher accuracy and f1 score than the LSTM approach [4]. BERT also showed satisfactory results with sentiment analysis.

Review Reading Comprehension (RRC) is another recently found application of BERT [5]. The authors of [5] have done a detailed study of Machine Reading comprehension (MRC), which has been extremely useful for creating the ReviewRC dataset for customer support by creating many questions and answering them from customers' reviews. They built a post-training model using BERT and applied the model to enhance the domain and review aspect extraction and sentiment classification [5]. Fine-tuning part of the implementation work of this paper was done after the post-training of the model, which makes the entire approach effective for the tasks mentioned above [5]. Context-based sentiment analysis is also a field of NLP where BERT can be used. In [6], the authors created an improved BERT-based approach that focuses on the target context of the text and its aspects. The authors [7] proposed a BERT-DAAT model applied for cross-domain sentiment analysis. BERT-DAAT model is a context-aware BERT-based model fed with sufficient knowledge of the target domain and post-trained to work enough to become a domain-aware BERT model [7]. Adversarial training is also included with post-training, and the proposed model makes its application promising in several areas such as named entity recognition, question answering, and reading comprehension [7]. Classification of scientific articles is yet another exemplary application of BERT, which can help the research community and researchers. Such work is done by Ambalavanan, A.K. and Devarakonda, M.V for the biomedical domain [8]. Question generation is a type of text generation task. Authors of [9] proposed a text generation model based on BERT, which generates questions from the context fed into it and generates the targeted answer. BERT is also good in applications integrating with label Semantics for multi-label classification through adjustive attention [13]. The authors of [13] proposed a Hybrid BERT model that incorporates Label semantics via Adjustive attention (HBLA), which simultaneously reveals the semantic dependencies of the label and search space. This approach fuses the word representation of search spaces retrieved through label graph embedding. The authors also introduced a naive attention mechanism that can calculate the semantic relation between a word and a label [13]. This study has proven the use of BERT in label semantics, and the results were outstanding compared to previous state-of-the-art approaches.

## 3 Background

# 3.1. The Data Tagging Process

"Data is the most valuable fuel." This saying is widely spread for the importance and relevance of the sample data during the machine learning process. It has affected several business analytics tasks such as future forecasting, prediction, etc. Data has such a precious position in the current world because several applications and approaches to processing data for essential duties will be a significant milestone[31]. The starting point of the recent data boom occurred when organizations started to store the data produced out of various business tasks. The data has become essential and is essential to perform daily processes in the organization. Data analysis and processing, which was limited to private purposes such as for business or organizational goals, became even more critical for multiple purposes as the internet developed.

From searching for information to automatic playlist generation, suggesting the route to a destination, people generate different forms of data using apps and the internet that comes them as handy. The volume of data is drastically changing as this data is generated in milliseconds or microseconds. The public is more open to communicating their views, concerns, or complaints through social media platforms, blogs, chatting platforms, live video streaming applications, etc.

The data was large enough to handle using traditional applications such as database management systems and data warehouse technology for one decade. But today, these applications are inefficient. The voluminous amount of data is not helpful until or unless it gives some insights. That is how the data becomes useful. Owing to this fact, extracting information from the data, and bringing insights are the core activities of business analytics and machine learning.

The data becomes more valuable when A.I. technologies extract and reveal hidden insights. Machine learning algorithms are soiled by training on labeled data. Choosing suitable soil to cultivate a particular crop is essential. It is not easy to find the right way to decide the fair-weather or soil conditions for growing a crop. Finding the appropriate data features and tagging them for preparing data samples for machine learning algorithms is more complicated.

Deep learning techniques are advanced nowadays and can exploit millions of parameters to generate robust predictions. The more data is available as input, the better deep learning models can create more training data samples and learn more appropriately about the data. Thus, deep learning models tend to be greedy for data.

Several approaches are there to feed a hungry model with sufficient and relevant data samples. A few decades ago, experts tried to manually label the data by analyzing the data, conducting surveys or studies [31]. This method is not viable or practicable nowadays as it is very time-consuming and requires more resources or an expensive way of labeling the data [31].

Preparing training data samples is a significant constriction for machine learning problems. Some of the methods to solve this problem are as follows:

**Crowdsourcing services**: Websites like AllMusic (Popular for diversified music data and metadata) or Twitter (diversified tweets) provide high-quality or well-defined organized information for an efficient labeling process. Crowdsourcing services provide quality tools to perform labeling from websites that offer sources for labeling data [31]. The labeling quality of these services can be good as they rely on performing several approaches such as data cleaning to retrieving relevant information[31].

**Live or offline applications:** Gaming apps or online games or other websites, or mobile apps attract people to interact with a pop-up menu or conversations to seek or identify some samples and ask to fill or click the object to get a response. This method is a random approach and sometimes compels users to do this and allows the user to proceed with that application or website. Another instance of a spontaneous response is asking to fill in a captcha or identify an object or part of a picture while users need to login into a web service. All the responses from users are used as a piece of source information for labeling the data[31].

**Crawling**: Crawling data from resourceful websites or online data sources with or without the help of APIs can also help retrieve a sufficient or enormous amount of data is also an available option for labeling the data[31].

**Transfer learning**: A source dataset within the same domain can be selected to perform modeling and then use the knowledge of that model to predict the labels for the target dataset[31].

## 3.2. From emotionally dependent lyrics to emotion detection

Natural language processing (NLP) is the area of machine learning and artificial intelligence, which employs several linguistic and computational algorithms that can process and learn human language interpretations either in the form of text or speech/audio, or both. Some notable research contributions include emotion detection, sentiment, and polarity detection. Emotion detection and categorization have been studied from a dimensional and discrete space. The discrete area of emotions is derived from that feelings result from human beings' biological processes and culture. According to Paul Ekman's study, the most common and expressive emotions based on the facial expression of the human body's physical reactions are surprise, anger, fear, disgust, sadness, and happiness [27].

This paper is more focused on the dimensional space of emotions. The dimensional space of this methodology comes from text, audio, and video, and the features represent the emotion behind the context. The models that represent different dimensions to interpret human emotion from any media, such as conversation, texts, audio-visual streams, and many more, can be mapped using dimensional models. The literature contains several models that can help identify human emotions based on multi-modal dimensions.

In 1897, Wilhelm Max Wundt, a well-known German physiologist and the father of modern psychology, described the human emotions that can be represented in three dimensions [7]. The dimensions which are the basis of most dimensional space models are valence and arousal. Some of the most popular dimensional models for emotion are the Circumplex model, the vector model, Plutchik's model PAD emotional state model [34]. This paper is based on the Circumplex model proposed by James Russell [2]. This model considers valence and arousal dimensions, and the emotions are distributed according to those dimensions. The feelings are labeled using low to high values of valence and arousal features and then plotted onto the horizontal and vertical axis, respectively, as shown in Fig. 1.

# 3.3. NLP Word Embedding models

The music corpus considered for this work is purely lyrics, and it does not contain any numerical data. It is important to maintain a standard representation structure to understand the lyrics and their meaning for machine learning models. The vector format is the representation form used for each word of the text data; instead, here it is lyrics. The computation of vectors is the next step for which multiple methodologies are used in NLP. Some of the popular word representation methodologies are Bag of Words (BoW) [28], Word2Vec [30], TF-IDF [29], BERT embeddings [32]. These representation approaches generate diversified vector representations and dimensions. The models discussed in this work focus on BERT embedding. BERT architecture works by accepting text sequences and converting them internally to tensors of size (3, 128). BERT's tokenizer performs this task and is then described in more detail.

# 3.3.1 The BERT Model:

Introduced and developed by Google in 2018, the BERT model is a model which has been a significant milestone in the history of Natural Language Processing[32]. The BERT is a short name for Bidirectional Encoder Representations from Transformers. As the name implies, the BERT model is a part of the Transformer family and is purely working based on encoding mechanisms[32]. But the architecture of BERT is like a Blackbox. It should be fed with BERT tokens, and it outputs vectors in specific shapes for each word provided as a token[32]. Tokens are fed in a sequential format, and each sentence is differentiated by the model using two tokens like delimiters. One such delimiter token is [SEP], which represents a separator token that helps distinguish between two sentences[32]. Another delimiter token is [CLS], which represents the start of a sentence (see Fig. 2).

## 4 Materials And Methods

## 4.1 Dataset

The Music4All dataset [38] is a newly designed dataset with multi-modal data, apt for several music applications such as playlist generation, recommender system, emotion detection. It is a large dataset that contains diverse information about songs, such as genres, tags, lyrics, and audio clips. This database is genuinely efficient for solving several challenges the music information retrievals (MIR) community faces. Some of the challenges are the limited number of datasets required to perform a variety of research development, the volume of music data, accessibility, lack of diversified features for multi-modal song classification tasks, and many more. Because the dataset is large enough, the lyrics and audio clips are available in more than one lakh. Also, 15K user's listening history, 16 metadata features are attractive features of this dataset. Each song lyric is saved as a text file under a file name, a unique song_id around the dataset. To prepare the data for the work, around 28,988 text files of lyrics were merged and kept under the song file name as the I.D. of the file to create a CSV file. The preprocessing step started with cleaning the merged lyrics and metadata files. Afterward, the lyrics CSV file was combined with corresponding metadata features available in another CSV file in the dataset. The resultant data file consisted of 28,988 rows and 12 columns. Among 28,988 rows, only 21978 rows were taken for experimenting. The rest of the rows were skipped as they do not belong to the English language.

## 4.2 Data preparation

As the primary goal of this work is music mood recognition from lyrics and labeling the data using emotion classes, the first step is selecting the features related to lyrics. The feature emotionally dependent on the lyrics data available in the dataset was valence only. Music4All dataset's valence column [value between 0-1] can be used to classify the lyrics to moods happy and sad by considering the valence value >0.5 as happy lyrics otherwise sad lyrics. So, following this method and labeling the rows 1 and 0(happy and sad) according to valence values, the resultant rows marked as 0 and 1 were 17231 and 11757 in the count, respectively.

## 4.3 The LyBERT architecture

The LyBERT, the proposed model, is shown in Fig. 3. The first part of the LyBERT model corresponds to transfer learning. The transfer learning module acquires the knowledge of the source domain MER dataset and then transfers the knowledge to the BERT model. Transferring knowledge, in this paper, mainly means that the emotion labels of the source lyrics dataset are learned and then used to predict the emotions of the target lyrics dataset. Once the transfer learning is completed, the BERT is trained on the same dataset to re-predict the emotional labels of the lyrics. The main aim here is to label the emotions of the Music4All lyrics dataset.

## 4.3.1 Emotion classification

The Music4All dataset [38] does not include an emotion label. The first step is to find out emotionally relevant features for classification. The first trial for the process was applying and utilizing the valence feature, which is emotionally relevant and uses binary classification. This approach is explained as follows,

**Binary classification**: HAPPY and SAD songs using valence feature [33] as an initial step towards the implementation started with state-of-the-art binary classification models such as Naïve Bayes, Support vector machine, Random Forest, and Extreme Gradient Boosting. Among the state-of-the-art models, SVM and Random Forest outperformed other models (Table. 1).

**Table I**

Results of various state-of-the-art models

| Model | Embedding | Accuracy | Precision | Recall | F1_Score |
|---|---|---|---|---|---|
| SVM [35] | Count Vectors | 0.67 | 0.47 | 0.62 | 0.53 |
| SVM [35] | TF-IDF Vectors | 0.67 | 0.46 | 0.62 | 0.52 |
| RF [36] | Count Vectors | 0.66 | 0.37 | 0.65 | 0.49 |
| RF [36] | Word Level TF-IDF | 0.67 | 0.4 | 0.64 | 0.48 |
| XGBoost [37] | Count Vectors | 0.65 | 0.37 | 0.61 | 0.46 |
| XGBoost [37] | Word Level TF-IDF | 0.66 | 0.39 | 0.63 | 0.48 |

Table 1 shows the results of the lyrics binary (happy or sad) classification task carried out using several state-of-the-art algorithms such as Support Vector Machine (SVM) with Count Vectors, SVM with TF-IDF Vectors, Random Forest with Count Vectors: Random Forest with WordLevel TF-IDF, Xgboost with Count Vectors, and Xgboost with WordLevel TF-IDF. SVM and Random Forest gave the same accuracy, around 67%. SVM has slightly higher accuracy than Random Forest. The SVM with count vectorizer achieved a precision of 0.47, recall of 0.62, and F1 score of 0.53. The SVM model with TF-IDF vectorizer embedding achieved identical scores for all performance metrics. Random Forest with count vectorizer embedding achieved 66 percent accuracy and precision; recall and F1 scores were 0.37, 0.65, and 0.49. The Random Forest with word-level TF-IDF performs like the SVM model because the accuracy is the same; the precision, recall, and f1score are 0.4, 0.64, and 0.48. Here the recall is better for Random Forest. XGBoost Count Vectors and Word Level TF-IDF give comparable results with 65% and 66% accuracy. The XGboost Word Level TF-IDF precision, recall, and F1-Score are slightly higher than XGBoost Count Vectors. Then according to the results of Table 1. SVM algorithm is more suitable for binary classification of the lyrics.

## Multi-class classification

The valence column was the emotionally relevant feature related to the text (lyrics) and audio in the Music4All dataset [38]. Energy, danceability, mode, and popularity features are more related to audio emotion. Moreover, the Music4All dataset [38] does not have labels for the emotional classification of lyrics. For this reason, they were not considered for binary classification. And because of that, to perform lyric classification with more emotions, there was a requirement to find an approach that can be applied to predict the emotion label of the lyrics and perform multi-class lyrics emotion classification.

The multi-class classification approach consists of more than two classes or outputs. In this paper, multi-class classification is performed for emotions of lyrics such as Happy (class 0), Angry (class 1), Sad (class 2), and Relaxed (class 3). The multi-class classification is performed by first predicting the emotion classes of music4all lyrics data using the classes of the MER dataset. Classifying the Music4All lyrics into four quadrants of Russell's Circumplex model in Fig. 1 requires more features to label the dataset. Initially, the transfer learning method was carried out for labeling the dataset. Transfer learning helped learn the lyrical features from the MER dataset and then transferred the knowledge to the Music4All dataset to label the songs' emotions (Fig. 3).

MER dataset has annotated 771 lyrics (nearly balanced), which are categorized into four quadrants of Russell's Circumplex model (as shown in Fig.1.) [11]. MER dataset [11] contains the data coming under all the features of four quadrants happy, angry, sad, and relaxed (encoded as 0, 1, 2, and 3). The reason for selecting this dataset for predicting the Music4All dataset is that the source dataset is classified based on a robust and diverse range of emotionally relevant features of a lyric. The authors of the dataset have considered several lyrical features[11]. The Content-Based Features (CBF) represent the most relevant features to the lyrics and specifically mean what the lyrical content is all about. The next feature is stylistic-based Features (StyBF) that relate to the lyrics' writing style in a specific language, slang words (which represent some text features that are mainly associated with several emotional words used in particular genres of lyrics). Another feature available was Song-Structure-Based Features (StruBF) which represents the features in the text of the lyrics is structured. Semantic-Based Features (SemBF) were also considered to represent the features with specific semantic factors [11]. These features have inspired this research work to apply the MER dataset to find the emotions of Music4All lyrics.

After building a random forest model on the MER dataset, which has given the best accuracy of (67%) among other state-of-the-art models, predictions were made on the music4All dataset and labeled the dataset into four quadrants based on the same Russel's model. The predicted model was then submitted to predict the emotions again using the best-known transformer model BERT uncased.

The transfer learning from the MER dataset to the music4All dataset predicts the emotion labels. The prediction using the same domain dataset and re-training of the BERT model used to predict the emotion labels are significant parts of the proposed model that performs significantly during emotion detection. And this combination (BERT embeddings +transferred knowledge from the same domain as illustrated in Fig. 3) lies as the important contributions of this work. Figure IV shows the distribution of classes in the music4All dataset.

Figure 5 shows the percentage of lyrics in each emotion class labeled through transfer learning. The labeled dataset is now tested for pre-trained model BERT Large uncased, which gives higher accuracy than the state-of-the-art models. The results with evaluation accuracy Bert-uncased achieved is 91.5% through two-fold approaches transfer learning from MER corpus dataset and pre-training method.

## 5 Experimental Results

The important steps in LyBERT based training, evaluation and prediction process are as shown below:

# 5.1 BERT tokenization

This process includes converting the lyrics sentences to the format of BERT tokens. Here is a tokenized (sequences to be at a maximum of 128 tokens long) sample of the first training set observation of lyrics looks like:

['within', 'heart', 'every', 'man', 'symbol', 'deep', 'truly', 'all', '##des', '##cend', '##ing', 'power', 'unfortunately', 'still', 'asleep', 'may', 'put', 'hands', 'eyes', 'gleam', 'never', '##ending', 'much', 'turn', 'inside', 'conceal', '##s', 'understanding', 'five', 'pointed', 'grey', 'star', 'car', '##ven', 'sign', 'aryan', 'race', 'five', 'pointed', 'grey', 'star', 'car', '##ven', 'fore', '##hand', 'evil', 'face']

Here is a sample of the token representation shown above, which is converted to the input features which BERT can interpret:

Sample Lyrics: "within heart every man symbol deep truly all descending power unfortunately still asleep may put hands eyes gleam never ending much turn inside conceals understanding five pointed grey star carven sign aryan race five pointed grey star carven forehand evil face"

________________ ________________

Tokens :  ['within', 'heart', 'every', 'man', 'symbol', 'deep', 'truly', 'all', '##des', '##cend', '##ing', 'power', 'unfortunately', 'still', 'asleep', 'may', 'put', 'hands', 'eyes', 'gleam', 'never', '##ending', 'much', 'turn', 'inside', 'conceal', '##s', 'understanding', 'five', 'pointed', 'grey', 'star', 'car', '##ven', 'sign', 'aryan', 'race', 'five', 'pointed', 'grey', 'star', 'car', '##ven', 'fore', '##hand', 'evil', 'face']

________________

Input IDs :  [101, 2306, 2540, 2296, 2158, 6454, 2784, 5621, 2035, 6155, 23865, 2075, 2373, 6854, 2145, 6680, 2089, 2404, 2398, 2159, 24693, 2196, 18537, 2172, 2735, 2503, 19819, 2015, 4824, 2274, 4197, 4462, 2732, 2482, 8159, 3696, 26030, 2679, 2274, 4197, 4462, 2732, 2482, 8159, 18921, 11774, 4763, 2227, 102, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]

________________

Input Masks :  [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]

_______________

Segment IDs :  [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]

## 5.2. The LyBERT’s hyperparameters’ configuration

The LyBERT's experiment was carried out using different specifications. In this work, BERT-Base, Uncased model developed by Google with 12-layer was used. Moreover, the entire code implementation was carried out using Google Colab. According to the specifications given by the Google research community, the following values were used to initialize the hyperparameters and initialize the TPU configurations while using Google Colab with TPU. The batch size was set as 32, 64, and 128. Other specifications are as follows:

Learning rate=  2e-5, 3e-5 and 5e-5

Number of training epochs = 3.0

Warm up proportion= 0.1

## *LyBERT Model configurations*

Save checkpoints steps = 300

Save summary steps = 100

The overfitting issue is prevented by adding a dropout layer.

## 5.3 LyBERT results on the validation dataset

Table 2 shows the average accuracy of running the LyBERT model 10 times using different learning rates and batch sizes. The average accuracy score is high as 92%. The precision, recall, and F1 score are the same at 0.96. The validation data categorized according to emotion class is illustrated below in figure 12:

The evaluation results showing accuracy, precision, and f1 score on the validation dataset are shown in Table 2.

**Table 2**

BERT_uncased_L-12_H-768_A-12/1 model implementation

| **Overall Accuracy** | Precision | Recall | F1-score |
|---|---|---|---|
| **91.69** | 0.96 | 0.96 | 0.96 |

The evaluation on validation dataset got the following outcomes: false_negatives: 46.0, false_positives 41.0, true_negatives': 558.0, and 'true_positives': 1112.0

## 5.4 Multi-class classification performance metrics

The multi-class classification results of BERT are shown in Table 3.

**Table 3**

Multi-class classification results

| **Classes** | **precision** | **recall** | **f1-score** | **support** |
|---|---|---|---|---|
| **0** | 0.82 | 0.83 | 0.82 | 1620 |
| **1** | 0.76 | 0.73 | 0.75 | 665 |
| **2** | 0.77 | 0.77 | 0.77 | 1583 |
| **3** | 0.72 | 0.67 | 0.69 | 525 |
| ***macro avg*** | 0.76 | 0.76 | 0.76 | 4393 |
| ***weighted avg*** | 0.78 | 0.78 | 0.78 | 4393 |
| ***accuracy*** | 0.78 | 0.78 | 0.78 | 4393 |

Table 3 shows the detailed multi-class classification results. The results shown in Table 3 are the mean value of the results for the ten experiments conducted under batch size 32 and learning rate 5e-5. The experiments conducted under other configurations by changing learning rate and batch size, but there was the results for all metrics were low or the same for the classes happy, angry and sad. In the batch size 32 and learning rate 5e-5, the class relaxed gave improved accuracy of 72. But in other batch sizes and learning rates, the accuracy for relaxed was 69%, and precision was also 69%. Classes 0, 1, 2, and 3 correspond to Happy, Angry, Sad, and Relaxed emotions. The precision, recall, f1-score, and support of

Happy emotion are 0.82, 0.83, 0.82, and 1620, respectively. The precision, recall, f1-score, and support of Angry emotion are 0.76, 0.73, 0.75, and 665, respectively. The precision, recall, f1-score, and support of Sad emotion are 0.77, 0.77, 0.77, and 1583, respectively. The precision, recall, f1-score, and support of Relaxed emotion were 0.72, 0.67, 0.69, and 525, respectively.

The confusion matrix in the figure 7 shows the actual and predicted values of the four emotion classes used for the work. The cells which are arranged diagonally represent the true positive (T.P.) values of the classes concerned. For the emotion class, Happy, the T.P., is 1300, Angry has a T.P. value is 480, Sad has a T.P. is 1200, and relaxed has a T.P. of 360. Similarly, the True Negative (T.N.) for Happy is calculated as 2433 (480+63+27+100+360+1200+79+96+28), False Positive is calculated as  286 (60+180+46), and the F.N. value is calculated as 283(57+190+36). Happy and Sad emotion labels have more appeared in the test dataset when compared to relaxed and angry. Owing to this, the performance of LyBERT in *angry* and *relaxed* emotions is 76% and 67% less.

When the classes are merely or severely unbalanced, the Precision-Recall curve can be visualized to understand and calculate the quality of predictions created on the data. Precision measures the significance of the outcome, whereas recall is an indicative measure of how often relevant outcomes are returned. The precision-recall curve depicts the tradeoff between precision and recall for various thresholds. A large area under the curve indicates good recall and precision, with high precision indicating a low false-positive score and high recall indicating a low false-negative score. High recall and precision scores show that the classifier produces accurate (high precision) outcomes, and most of the outcomes (high recall) produced are positive. A model with recall with a high score and precision at low can return an enormous number of outcomes but predicted labels would be inaccurate in contrast to training labels. A model with a low recall score and precision at high can return fewer outcomes and the majority of predicted labels accurate in contrast to training labels.

The plot shown in figure 8 illustrates the quality of predictions for the LyBERT model. The class with a large area is class 0, which represents happy lyrics because the precision-recall scores are high for that class. The class with the second largest area is class 2, which represents sad lyrics and has higher precision-recall scores than classes 1 and 3. The class with the third-largest area is class 1, which represents angry lyrics. The class with the least area is class 3, which means that it has lower precision and recall scores when compared to the rest of the three classes.

Figure 9 shows a basic ROC (Receiver operating characteristic) curve plot of all the four classes of emotions. According to that curve, the true positive –false positive curve of class 0 has a large area under it. The decreasing order of area under the curve of the rest three classes that show the quality outcomes are class 1, class 2, and class 3. The area is measured and shown in the next plot shown in figure 10, the black diagonal dash line in the middle is the average line where TPR (True Positive Rate) = FPR (False Positive Rate).

Any line closer to the top-left area is a good prediction of a multi-class classification model. If a line is closer to the TPR=FPR line in the middle, it is not a good performance indicator. So, for example, Consider the yellow line; it is a better predictor than the red line because it is slightly higher than the red. But the blue line is not better than the yellow line even though it is higher because it is more towards the right side. Therefore,  it is also necessary to check AUC called Area Under Curve and the ROC curve. This area is the total area under the ROC line. In this way, there is no need to measure which line is better by looking at it. By checking the area values and according to the graph, the light blue (Cyan, class 0) line is the best predictor with an area = 0.86. AUC and ROC curves suggest that the proposed lyrics emotion classification approach is a promising research area that can be further improved through transfer learning and sentence embedding. The following is a sample of randomly selected lyrics for emotion detection figure 11:

The prediction results of the sample lyrics in figure 11 are shown in figure 12.

The percentage of predicted emotion labels for each class is shown in the pie diagram below (Figure 13):

The bar diagram in figure 14 below shows the distribution of predicted emotion labels for each emotion class. The highest number of counts is predicted for class happy; the second highest is for class sad; the third-highest is for class angry, and the least number of counts for class relaxed. The number of samples in the subset of the dataset considered for this work contain the following amount of data: Class 0: n=7859 (35.781%), Class 1, n=3541 (16.122%), Class 2 n=8014 (36.487%), Class=3, n=2550 (11.610%). Since this work aimed to predict the lyrics' emotion labels, the subset as a whole was experimented with using the LyBERT model and have maintained the classes with high count(happy and sad) and low count(angry and relaxed) as it is.

# 6 Discussion

The BERT classification results show various scores across the different classes of emotion. For the class *Happy* (0), all the four-performance metrics (precision, recall, F1 score, and support) have high scores (0.82, 0.83, 0.82, and 1620). This class has the highest accuracy. Similarly, classes *angry* (1) and *Sad* (2) have approximately identical scores. The lowest performance scores are for the *relaxed* class (3). This is because the test sample, as well as the training sample, consists of imbalanced classes. The main aim of this work was to classify the lyrics using emotion classes, even though the results achieved show variations for those classes having the minor training sample, which is not much concern. In [38], the emotion behind the lyrics has not been appropriately recognized. Their results were not satisfactory and only considered the valence and energy features for the mood classification. In this work, the emotion behind lyrics is also identified based on emotionally relevant features through transfer learning and integrated the knowledge with the BERT embeddings to more understand the sentiment behind songs. The LyBert's validation experiments achieved overall high accuracy results and high f1 score (96%), high recall(96%), high true_negatives (558), true_positives ( 1112.0), and low scores for false negative(46) and false positive(41). These results show that the LyBERT model is well-performing on the validation dataset with 600 happy and sad lyrics, 300 angry lyrics, and 220 relaxed songs, as shown in figure 11. But in the case of the test data, even though the ratio of happy and sad songs are the same, there is much difference in angry and relaxed song counts, as shown in figure 14. Owing to this, the accuracy achieved in the test dataset for the emotions sad and relaxed are less. The main reason for this is imbalanced classes in the dataset. But the results have significantly improved compared to [38] and emotion detection is a perfect application of this emotionally labeled dataset of lyrics.

# 7 Conclusion And Future Enhancement

The literature review shows that BERT is a promising approach for text-based applications as they are giving better results despite the domain of application. In this paper, a lyric classification model is built based on emotions. The results obtained are satisfactory compared to existing work on the same dataset [38]. In this work, the application of BERT embeddings significantly improved the accuracy of all emotions on the dataset. The accuracy of each class ranges from (67 to 82%). Owing to this, integrating multiple music corpus or datasets for lyrics according to the same emotional plane can improve the results shown in this work. So this work is a beginning, and more techniques to build more efficient transfer learning approaches will be explored and experimented with in the future. Also, this work is supposed to improve the performance of the lyrics-based music recommender system. It will help address several issues in the collaborative and content-based recommender systems. In the future, it is planned to explore different emotional planes to investigate more emotions in the lyrics. The main extension for the work is to improve the prediction results by using some emotional corpus and making the balanced apply to different related tasks. Another extension to this work in the future is further investigating emotion-genre dependence.

# Declarations

### Ethical Approval and Consent to participate

Not Applicable.

### Human and Animal Ethics

Not Applicable

### Consent for publication

Not Applicable

### Availability of supporting data

Not Applicable

### Competing Interests

The authors have no competing interests to declare relevant to this article's content.

### Funding

No funds, grants, or other support were received.

### Authors' contributions

Revathy V R made substantial contributions to the design of the work; or the acquisition, analysis, or interpretation of data. Anitha S.Pillai and Fatemeh Daneshfar revised it critically for important intellectual content and approved the version to be published.

### Acknowledgements

We acknowledge the authors, namely Igor Andre Pegoraro Santana, Fabio Pinhelli, Juliano Donini, Leonardo Catharin, Rafael Biazus Mangolin, Yandre Maldonado e Gomes da Costa, Valeria Delisandra Feltrim, and Marcos Aurelio Domingues from the Department of Informatics (DIN), State University of Maringa (UEM), Maringa, PR, Brazil who permitted to access to the Music4All dataset for this research work.

## References

1. Balakrishnan, Anusha, and Kalpit Dixit. "DeepPlaylist: Using Recurrent Neural Networks to Predict Song Similarity." stanford Univ (2014): 1-7.
2. Russell, J.A., 1980. A circumplex model of affect. Journal of personality and social psychology, 39(6), p.1161.
3. An Experimental Study on Sentiment Classification of Algerian Dialect Texts Leila MOUDJARIa,_, Karima AKLI-ASTOUATIa, aRIIMA Laboratory (https://www.riima.usthb.dz/node/2), USTHB, Algiers, Algeria. https://www.usthb.dz/en/
4. Alberti, C., Lee, K. and Collins, M., 2019. A bert baseline for the natural questions. arXiv preprint arXiv:1901.08634.
5. Sun, C., Huang, L., and Qiu, X. (2019). Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. 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), pages 380–385, Minneapolis, Minnesota. Association for Computational Linguistics.
6. Hu Xu1, Bing Liu1, Lei Shu1 and Philip S. Yu, BERT Post-Training for Review Reading Comprehension and Aspect-based Sentiment Analysis Department of Computer Science, University of Illinois at Chicago, Chicago, IL, USA,Institute for Data Science, Tsinghua University, Beijing, China
7. Wu, Z. and Ong, D. C. (2020). Context-guided bert for targeted aspect-based sentiment analysis. arXiv preprint arXiv:2010.07523.
8. Du, C., Sun, H., Wang, J., Qi, Q. and Liao, J., 2020, July. Adversarial and domain-aware bert for cross-domain sentiment analysis. In Proceedings of the 58th Annual Meeting of the Association for

Computational Linguistics (pp. 4019-4028).

9. Ambalavanan, A.K. and Devarakonda, M.V., Using the contextual language model BERT for multi-criteria classification of scientific articles. Journal of biomedical informatics, 112, p.103578.
10. Chan, Y.H. and Fan, Y.C., 2019, November. A recurrent BERT-based model for question generation. In Proceedings of the 2nd Workshop on Machine Reading for Question Answering (pp. 154-162).
11. Malheiro, R., Panda, R., Gomes, P., Paiva, R.P.: Emotionally-relevant features for classification and regression of music lyrics. IEEE Transactions on Affective Computing 9(2), 240{254 (2016).
12. Liu, Yi, Jiahuan Lu, Jie Yang, and Feng Mao. "Sentiment analysis for e-commerce product reviews by deep learning model of Bert-BiGRU-Softmax." Mathematical Biosciences and Engineering: MBE 17, no. 6 (2020): 7819-7837.
13. Garg, S.; Ramakrishnan, G. BAE: BERT-based Adversarial Examples for Text Classification. arXiv 2020, arXiv:cs.CL/2004.01970.
14. Cai, L., Song, Y., Liu, T. and Zhang, K., 2020. A Hybrid BERT Model That Incorporates Label Semantics via Adjustive Attention for Multi-Label Text Classification. Ieee Access, 8, pp.152183-152192.
15. Chang, W.C., Yu, H.F., Zhong, K., Yang, Y. and Dhillon, I., 2019. Taming Pretrained Transformers for Extreme Multi-label Text Classification. arXiv preprint arXiv:1905.02331.
16. Croce, D., Castellucci, G. and Basili, R., 2020, July. Gan-bert: Generative adversarial learning for robust text classification with a bunch of labeled examples. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 2114-2119).
17. Edmonds, D. and Sedoc, J., 2021, April. Multi-Emotion Classification for Song Lyrics. In Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (pp. 221-235).
18. Agrawal, Y., Shanker, R.G.R. and Alluri, V., 2021. Transformer-based approach towards music emotion recognition from lyrics. arXiv preprint arXiv:2101.02051.
19. Vystrčilová, M. and Peška, L., 2020, June. Lyrics or Audio for Music Recommendation. In Proceedings of the 10th International Conference on Web Intelligence, Mining and Semantics (pp. 190-194).
20. Jin, D., Jin, Z., Zhou, J. T., & Szolovits, P. (2020). Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment. Proceedings of the AAAI Conference on Artificial Intelligence, 34(05), 8018-8025
21. Choi K. (2021) Bimodal Music Subject Classification via Context-Dependent Language Models. In: Toeppe K., Yan H., Chu S.K.W. (eds) Diversity, Divergence, Dialogue. iConference 2021. Lecture Notes in Computer Science, vol 12645. Springer, Cham.
22. Michaela Vystrčilová and Ladislav Peška. 2020. Lyrics or Audio for Music Recommendation? In Proceedings of the 10th International Conference on Web Intelligence, Mining and Semantics (WIMS 2020). Association for Computing Machinery, New York, NY, USA, 190–194.
23. Yao, L., Mao, C., and Luo, Y., "KG-BERT: BERT for Knowledge Graph Completion", <i>arXiv e-prints</i>, 2019.

24. Z. Gao, A. Feng, X. Song and X. Wu, "Target-Dependent Sentiment Classification With BERT," in IEEE Access, vol. 7, pp. 154290-154299, 2019, doi: 10.1109/ACCESS.2019.2946594.
25. Zeng, Z., Xiao, C., Yao, Y., Xie, R., Liu, Z., Lin, F., Lin, L. and Sun, M., 2021. Knowledge Transfer via Pre-training for Recommendation: A Review and Prospect. Frontiers in big Data, 4.
26. Sun, F., Liu, J., Wu, J., Pei, C., Lin, X., Ou, W. and Jiang, P., 2019, November. BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In Proceedings of the 28th ACM international conference on information and knowledge management (pp. 1441-1450).
27. Ekman, Paul (January 1992). "Facial Expressions of Emotion: New Findings, New Questions". Psychological Science. 3 (1): 34–38. doi:10.1111/j.1467-9280.1992.tb00253.x. S2CID 9274447.
28. Zhang, Y.; Jin, R.; Zhou, Z.H. Understanding bag-of-words model: A statistical framework. Int. J. Mach. Learn. Cybern. 2010, 1, 43–52. [CrossRef]
29. Qaiser, S.; Ali, R. Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents. Int. J. Comput. Appl. 2018, 181. [CrossRef]
30. Goldberg, Y.; Levy, O. word2vec Explained: Deriving Mikolov et al.'s Negative-Sampling Word-Embedding Method. arXiv 2014, arXiv:1402.3722.
31. Çano, E., 2018. Text-based sentiment analysis and music emotion recognition. *arXiv preprint arXiv:1810.03031*.
32. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv 2018, arXiv:1810.04805.
33. Revathy V.R and Anitha S.P. "Binary Emotion Classification of Music using Deep Neural Networks" Lecture Notes in Networks and Systems. Proceedings of the 13th International Conference on Soft Computing and Pattern Recognition (SoCPaR 2021). Vol. 417
34. Plutchik, R., 2001. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. American scientist, 89(4), pp.344-350.
35. Cortes, C. and Vapnik, V., 1995. Support-vector networks. Machine learning, 20(3), pp.273-297.
36. L. Breiman. Random forests. Maching Learning, 45(1):5–32, Oct. 2001.
37. Chen, T. and Guestrin, C., 2016, August. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).
38. Santana, I.A.P., Pinhelli, F., Donini, J., Catharin, L., Mangolin, R.B., Feltrim, V.D. and Domingues, M.A., 2020, July. Music4all: A new music database and its applications. In 2020 International Conference on Systems, Signals and Image Processing (IWSSIP) (pp. 399-404). IEEE.

# Figures

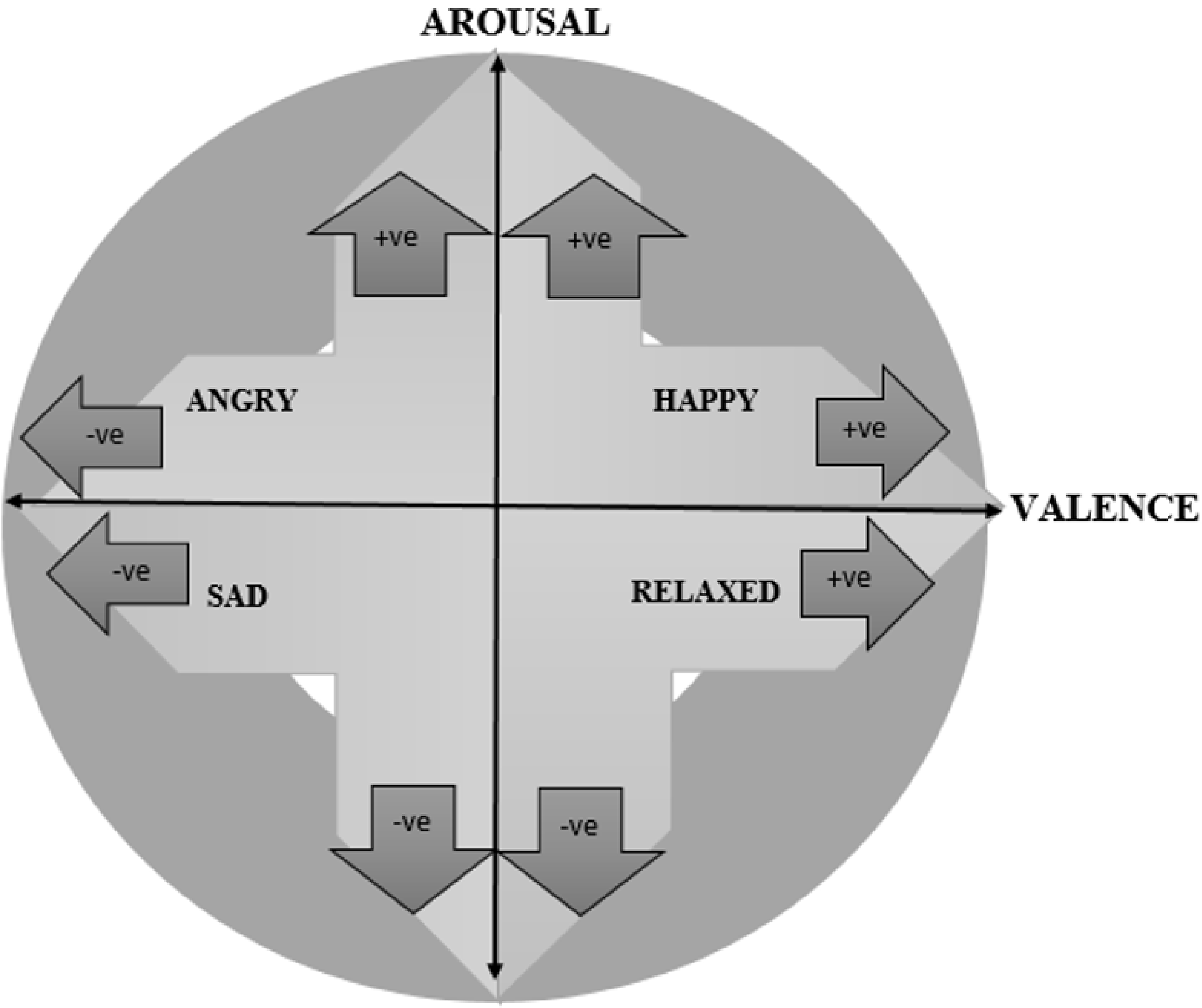
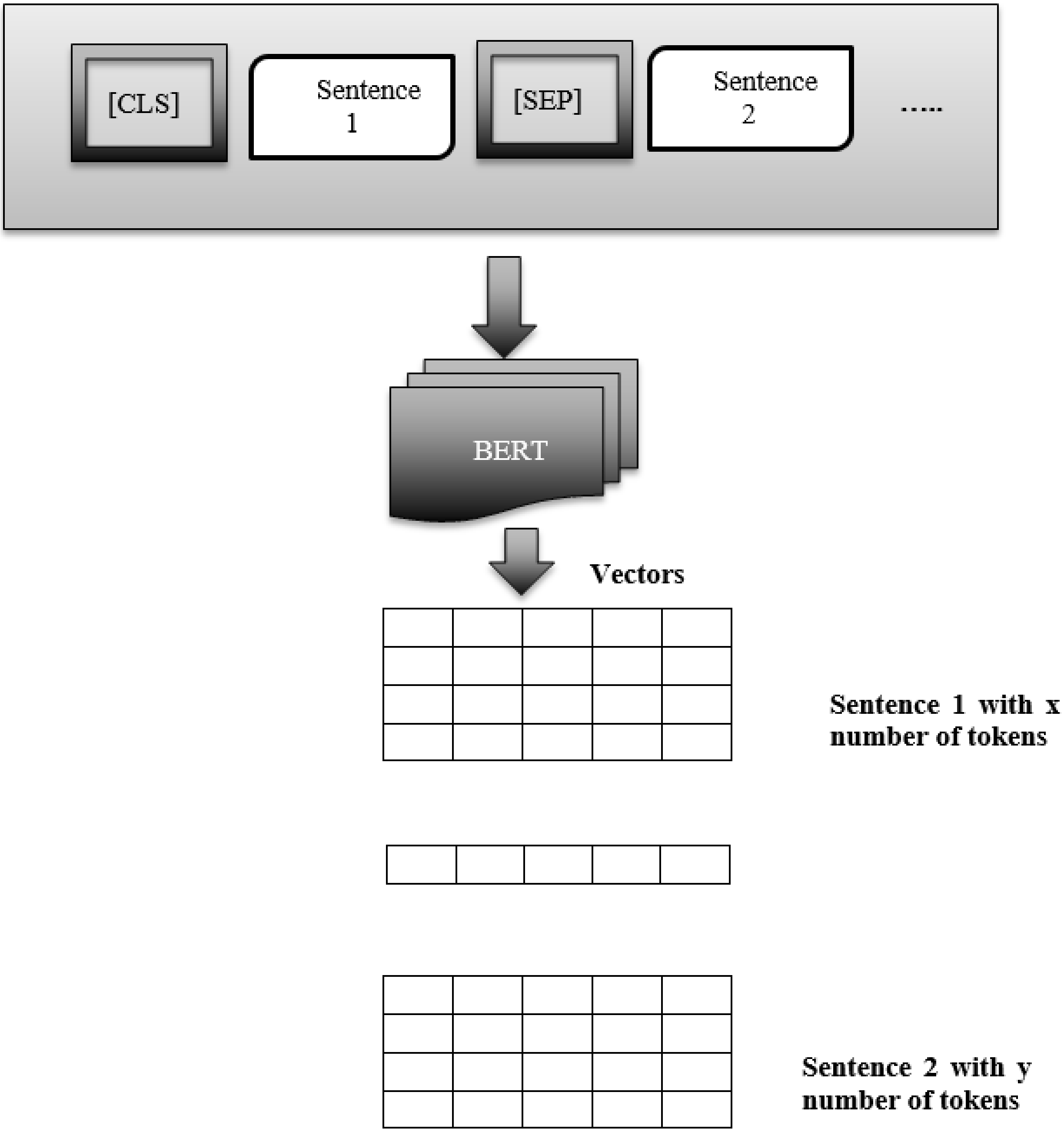


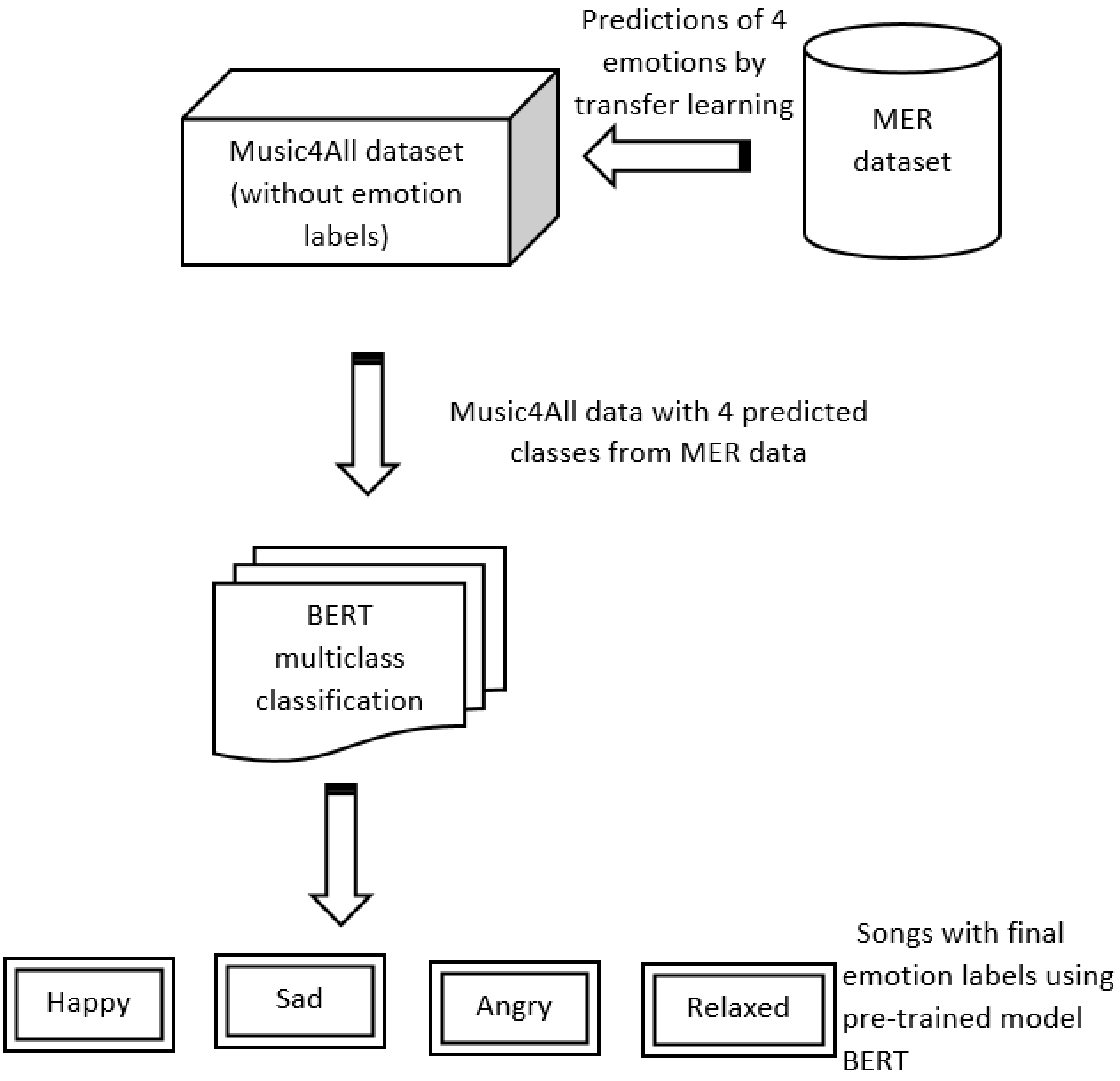
Figure 1

The dimensional model of James Russell [2]


[CLS]
Sentence 1
[SEP]
Sentence 2
.....
BERT
Vectors
Sentence 1 with x number of tokens
Sentence 2 with y number of tokens

### Figure 2

Sample BERT tokenizing illustration [32]


Predictions of 4 emotions by transfer learning
MER dataset
Music4All dataset (without emotion labels)
Music4All data with 4 predicted classes from MER data
BERT multiclass classification
Happy
Sad
Angry
Relaxed
Songs with final emotion labels using pre-trained model BERT

### Figure 3

The proposed BERT model architecture

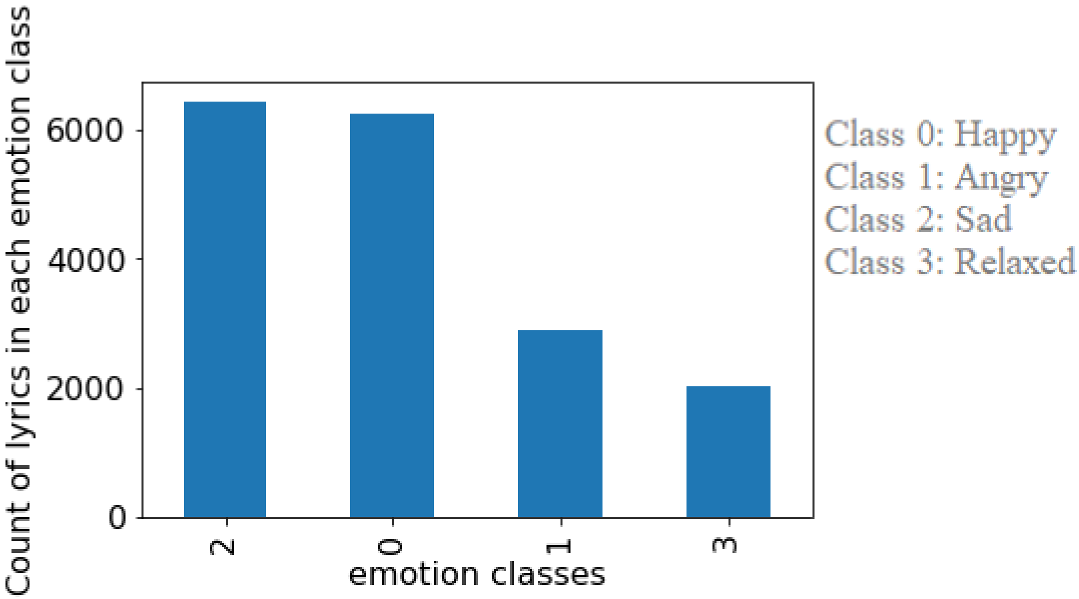


**Figure 4**

Distribution of emotion classes predicted through the transfer learning approach

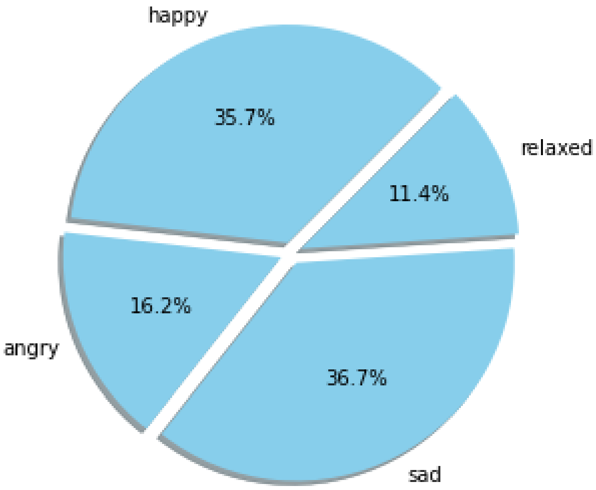


**Figure 5**

Percentage of lyrics in each emotion class

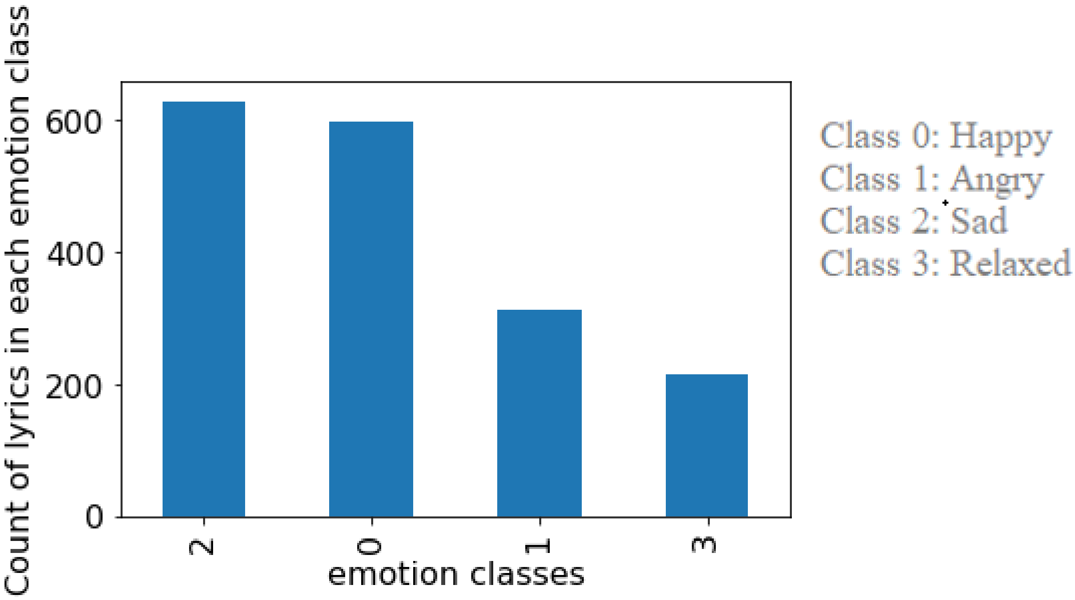


**Figure 6**

Validation data categorized with emotion classes

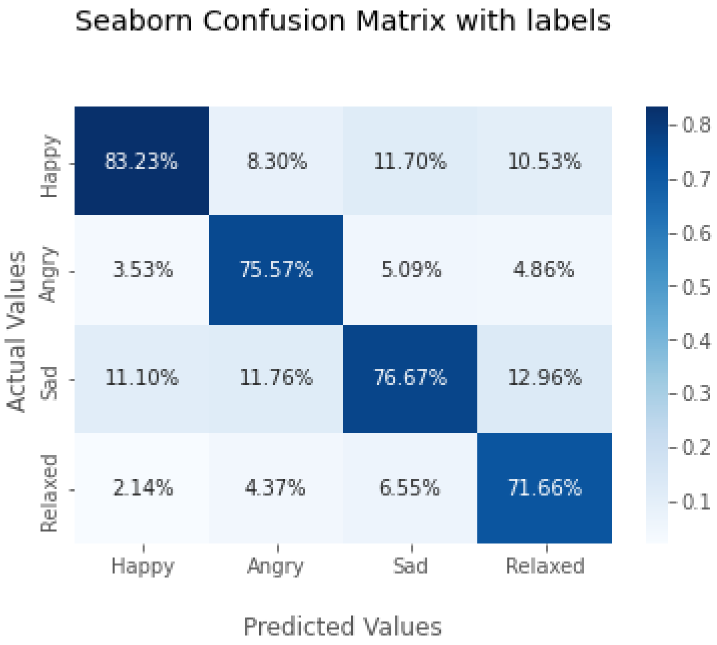


**Figure 7**

Confusion matrix

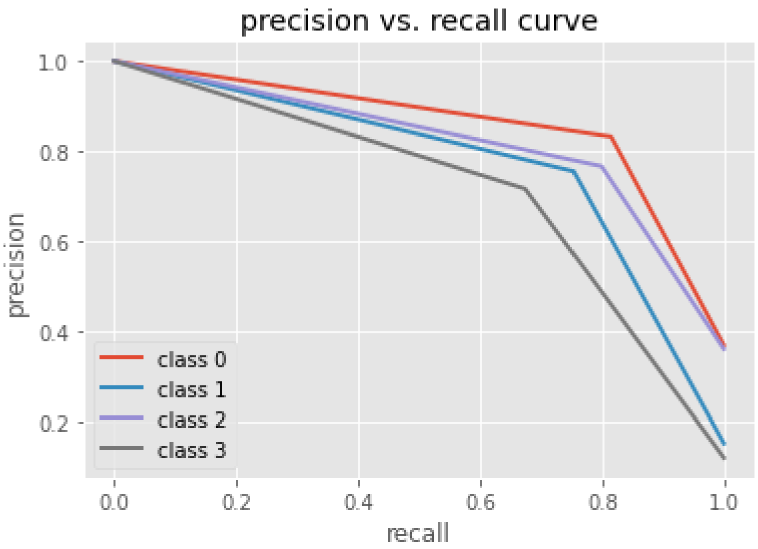


**Figure 8**

Precision-Recall curve

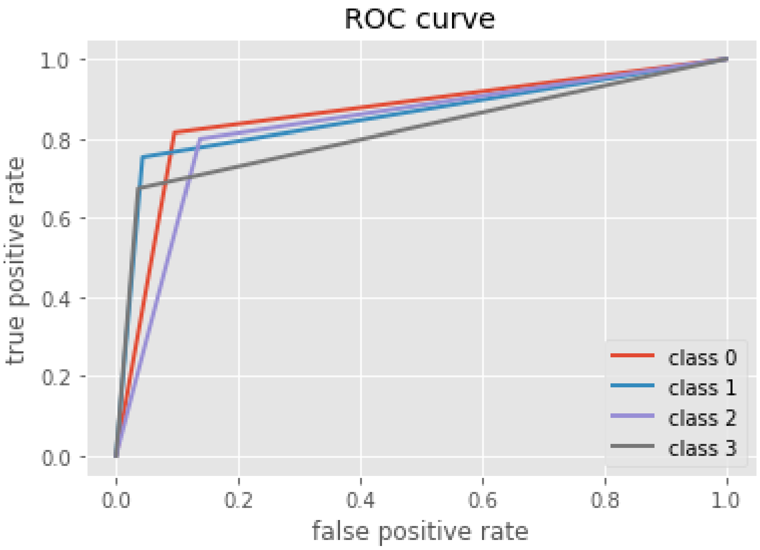
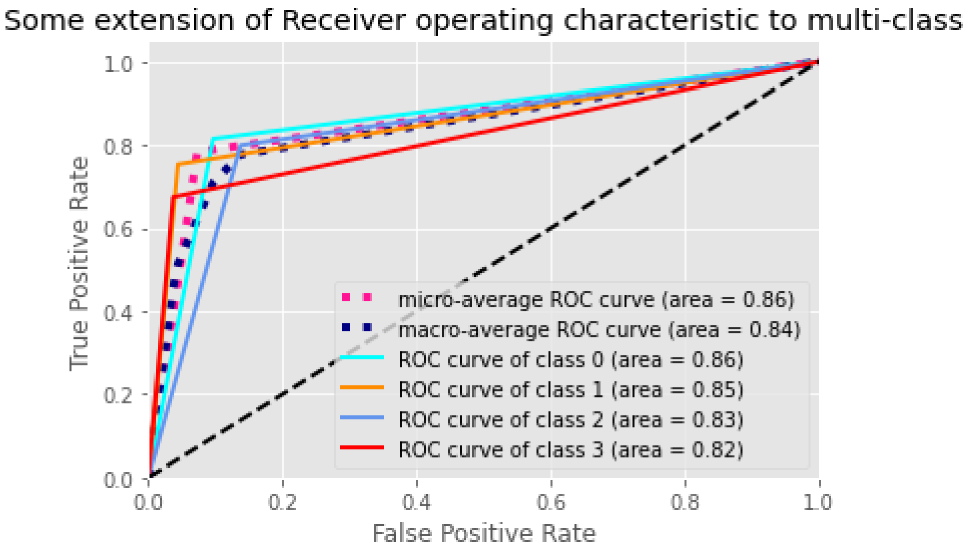


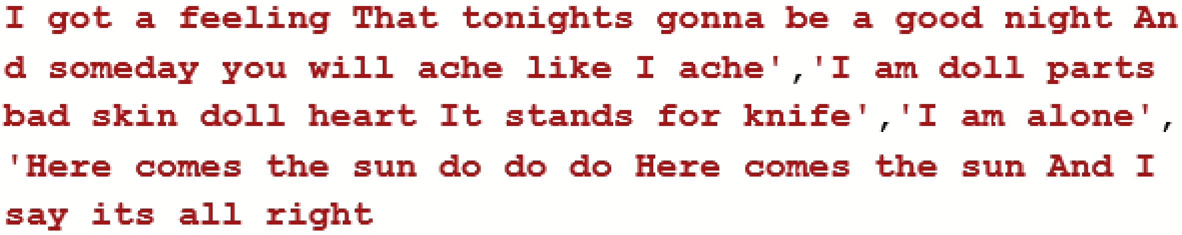
**Figure 9**

ROC curve



### Figure 10

ROC curve with micro and macro average



```
I got a feeling That tonights gonna be a good night An
d someday you will ache like I ache','I am doll parts
bad skin doll heart It stands for knife','I am alone',
'Here comes the sun do do do Here comes the sun And I
say its all right
```

### Figure 11

A sample lyrics for prediction from the music4All dataset [38]

```
[('I got a feeling That tonights gonna be a good night
And someday you will ache like I ache',
array([-0.9834268 ,
-3.9297834,-1.3124819 ,-1.08983125], dtype=float32),
  3,
  'Happy'),
 ('I am doll parts bad skin doll heart It stands for
knife',
  array([-4.021492  , -3.9906926 , -2.0321531 , -
0.18328136], dtype=float32),
  3,
  'Relaxed'),
 ('I am alone',
  array([-3.4526794 , -2.186017  , -1.337868  , -
0.52159786], dtype=float32),
  3,
  'Relaxed'),
 ('Here comes the sun do do do Here comes the sun And
I say its all right ',
  array([-4.577334  , -4.404998  , -3.588032  , -
0.05145232], dtype=float32),
  3,
  'Relaxed')]
```

## Figure 12

The prediction score for each emotion in each line of music lyrics

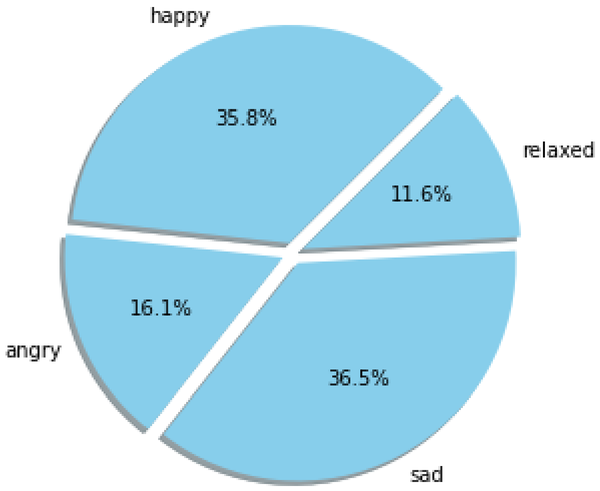


**Figure 13**

Percentage of lyrics in each emotion class predicted by LyBERT

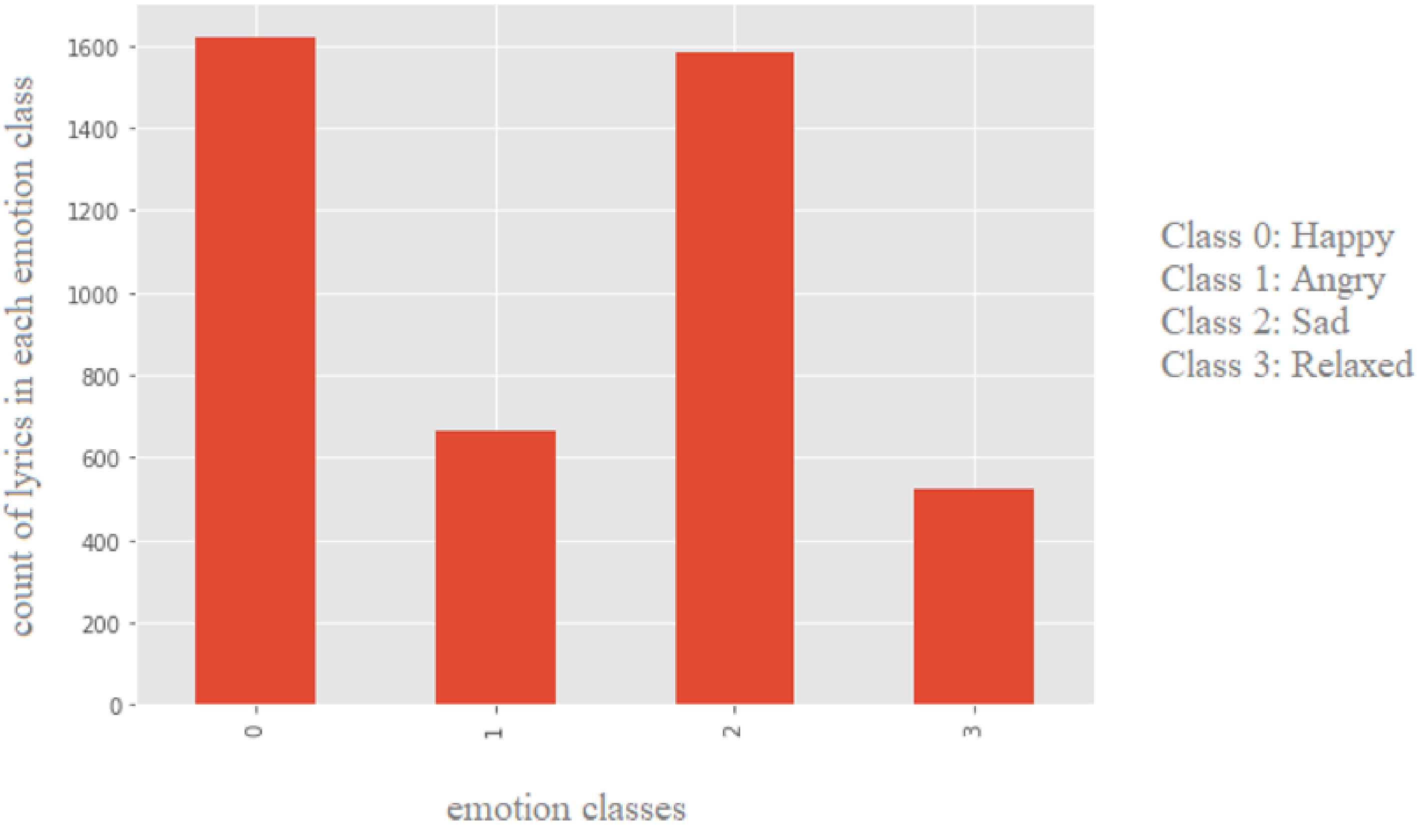


**Figure 14**

Distribution of the LyBERT's predicted emotion labels for each class.