Original Article

Emotion Classification Using Nature Based Optimization With Transformers And Transfer Learning

Swati Goel

Krishna Engineering College, Ghaziabad, Uttar Pradesh, India Email: goel.swati02@gmail.com
DOI: 10.47750/pnr.2022.13.S10.369

Abstract

This paper presents a methodology that utilizes machine learning models like ResNet50, BERT, and GPT2 for the classification of emotion in text or audio format. Proposed research had included the study of Hindi dataset MER500 and English dataset DEAM which discrete 5 classes of emotion. Models had attained about 85% and 83% for MER500 and DEAP datasets respectively. Ensembled model had also been showcased which had provided the best performance in MER500 dataset whereas BERT for DEAM dataset respectively. Comparative analysis had also showcased about lyrics as a best available form of data to classify the emotion from it. BERT and GPT -2 models are used for the classification of the lyrics. And out of these 2 models BERT had performed best for the 2 datasets. And this shows that, for these 2 dataset emotions are prominent in the lyrics (text features) as compared to audio features. And from the audio, a song and its karaoke are classified using the ResNet50 model. ResNet50 despite pretrained on the ImageNet dataset, transfer learning helped significantly for learning features from the audio. To classify the audio signal, various features were extracted and assembled in multi-dimensional array.

Keywords: MER500, GPT-2, BERT, ResNet50, Emotion, Song

Introduction

Recognition of human emotions from given data and classifying them into discrete types is a huge step in advancement in technology especially automating the process through computers [1]. Rapidly improving the decade engagement of humans with machines from last decade had provided immense technologies which had created a better place for human with more comforts and led us towards artificial intelligence and machine learning approaches [2].

Feedback from a person upon any provided input plays a vital role in understanding the need for a product or any advancement in that particular technology. As humans tend to react by expressing their emotions which can be observed and examined to extract their outputs, replicating the procedure through machines is a challenging task because a single false prediction or classification can lead to various errors in steps carried out upon the result. Observing the performance of machine learning models for various domains in which are implemented showcased excellent accuracy. Considering the trade-off between the accuracy and usage of output which is a classification of emotion can be considered as an area in which technology shall be developed further [3]. By recognition or classification of a person's emotion can become beneficiary in various domains like healthcare for the detection of behaviour or brain-related diseases, in the finance sector for detection of fraud person, etc.

Classification of emotion is one of the most studied fields in the last few years. Various methods are developed for the classification of emotion as well as to discriminate the emotion. VAD score-based method for representation of emotion is the latest and most widely used parameter for measuring emotion [4]. It is a valence, arousal, and dominance score which describes the emotion based on these values. Although, Discrete

classification of emotions like anger, sadness, happiness, etc is also practiced in various research.
For the implementation of the machine learning models, training of the model is needed, and to train its large number of the examined dataset is required. Various forms of data can be utilized to extract emotions including audio signals, text, EEG signals, videos, and images which are few among many other available data forms [5]. Implementation of the models like Decision Tree, Random Forest, Space Vector Machine, C3D classifier, K-nearest neighbour, Convolution Neural Network, LSTMs, etc which are showing out- standing performance for classification, recognition, and many more operations on data. Some widely used datasets on which mainly models are trained are DEAP, IEMOCAP, PMEmo, RAVDESS, RML, SAVEE, etc [6]. On closer examination of some research which showcased the overall accuracies of the model reaching 90% for some models. Moreover, research also provides the importance of the dataset for the model's efficiency. The combination of various models, as well as the usage of multiple forms of input, also increases the performance of the model [7]

1.1 Motivation

Over the last decades, mimicking the process of the human brain and activities by machines had observed a huge advancement in technologies. Artificial intelligence and machine learning concepts had provided us an opportunity to achieve it hence, in this technology we are approaching the detection of human emotion through various types of available data. Considering the importance of human emotion and its importance, if the process is replicated by machine it shall be very useful over a large number of applications like integrating into a video call to detection of potential suicide calls or detecting fraud and extortion. With the available data consisting of voice or music or text, it can be utilized for training an AI model with which a computer shall be able to provide a discrete classification of provided input data. Observing the various researches ongoing on it as well as studying about the various available dataset provided us an opportunity to explore in the Hindi language and provide the model which shall be compatible with the Hindi language also.

1.2 Related Work

Classification of emotions in any given data can be a challenging task. To classify emotion in audio or text it is generally divided into subgroups as of features contained in it [8]. However natural features and synthetic features added to data extracted make it a difficult task to distinguish them. This research showcased the use of a pretrained framework based upon a deep convolutional neural network (DCNN). The various approach is practiced like the use of k- nearest neighbour, support vector machine, multiplayer perceptron, random forest, and decision tree with a dataset as Emo-DB, IEMOCAP, RAVDESS, SAVEE and classify them based on seven emotions. Comparison of each model with their accuracies is showcased which least achieved as 80% with or with- out feature selection technique. Feature extraction was efficiently done through AlexNet as a small number of datasets with labelled classes were available. Considering the availability of huge dataset, approach for small dataset and extracting maximum efficiency had been showcased by [9] in which classification on the basis of the given data and evaluation based upon the whole dataset can be effective but when there is a small amount of data available emotion classification based upon word-level plays a vital role. As currently adopted methods use categorical lexicons which tag the input with the predefined set of words and classify it. In this method, the main drawback is they are only tagged with the specified set of basics emotion which are pre-defined in the model making this method less effective as well as flexible in the practical life scenarios as well as intensity is also ignored which might have played a significant role in classification actual inputs. Research conducted by [10] proposed a word-level distribution vector that consists of a dimensional lexicon incorporated from domain knowledge. Methods link an input word with the fine-grained and more generic taxonomies which had included quantities computed intensities. Further, they are divided into two schemas as implicitly (rule-based conversion strategies) and explicitly (emotional word embedding) in which seven multiclass datasets are tested. Various frameworks showcased different outcomes which include CNN, RNN, transformer-based, and Bert- based but showcased the same flexibility as well as effectiveness in classification using the proposed two schemas. As dataset plays a vital role in the model's performance, author had provided a novel approach which might help in creating labelled and balanced dataset

through music video [11]. Reviewing various research observations states that to classify large or small amounts of data, labelled datasets must be required. Research showcased the construction of the balanced music video dataset which includes various emotions and diverse language, musical instruments, culture, and territory. Testing of the dataset generated is conducted on four multimodal and unimodal convolution neural networks. Firstly, pre-trained fine-tuned unimodal CNN is tested upon unseen data and for classifying music emotion one dimensional model of CNN is also trained with raw waveform input. Observing the performance of the unimodal with various optimizers is evaluated to extract the optimal optimizer which can be implemented on a multimodal structure. As such process is also conducted on the multimodal which classifies music videos with SoftMax classifier as the final classifier which is a late feature fusion strategy. Lastly, all multimodal structures are combined in one predictive model to extract the overall prediction. To overcome the data scarcity problem (overfitting), cross-validation is practiced. Comparative analysis of the performance for each unimodal classifier with multimodal architecture. The predictive model proposed in this research showcased about 0.88 F1 scores whereas 0.987 AUC scores and achieved 88.56% of accuracy. The method proposed suggests the effective classification of high- level human emotions with the CNN model although a small amount of labelled dataset is available for training.

Classification of the available data into emotion categories was performed to extract the outcome of the data provided through various machine learning models [12]. When the data is provided in text and audio form it makes the model difficult to distinguish as input formats are different from each other for the same available data. Research showcased a novel approach in which input data of text and audio is used to classify emotions. It is a deep dual recurrent encoder model which simultaneously utilizes audio as well as text data as input to efficiently understand the speech. In this approach, speech data is broken into text and audio data formats on which proposed RNNs process individually, and the outcome is combined to predict an emotion class under with an input speech data belongs. Dataset is used for the execution of the proposed method termed Interactive Emotional Dyadic Motion Capture (IEMOCAP) which consists of 1636 happy, 1084 sad, 1103 angry, and 1708 neutral speech data in it. Implementation of the model showcased an accuracy that ranges from 68.8% to 71.8%. Discrete classification of an emotion in any provided data shall be an approach but to increase the performance and interrupt the emotion better, author showcased an approach to classify it into its VAD scores [13]. Classification of emotion into various predefined classes can be achieved through various conducted research mentioned above, but to convert into VAD (valence, arousal, and dominance measure which is widely used measurement for emotion classification. In proposed method model is trained by minimizing EMD loss which is between categorial emotion distribution and predicted VAD score distribution. Alongside it can also classify the emotion categories and predict VAD score for given data as text or sentence. A pre-trained model called RoBERTa is used with fine-tuning on three different corpora with categorical labels which were evaluated on EmoBank corpus and its VAD scores are observed. On comparing the proposed method of VAD classifier with categorial emotion classifier, performance on the smaller dataset is comparative ahead. Lastly, training with VAD labels had improved the performance of the proposed model which is also justified by an example. As majority of the studies and researches are based upon the extracted or synthetic datasets, author [14] approached for utilizing the raw data and real data extracted specifically for this research had been showcased. Among various available data for emotion classification instead of text and audio facial expressions as well as physiological plays, higher information can be used for the classification of emotions. To keep the data real research 53 subjects for their 15 different facial expression as well as physiological inputs combining four types of emotions are collected. Emotions were distinguished into four types positive and negative whereas three types of deep-learning models for classification of emotions. For training, the model facial expression, as well as physiological signals, are fed. On conducting the accuracy tests with signal parameters as facial expression model reached 99.99% and for physiological signals, it reached 81.54% whereas combining the model which accepts two input data simultaneously and performs regression for classification of the outcome showcased an accuracy of 86.2%. Different activation functions were also implemented which showcases Results showcase the best performance of the emotional classification with only facial expression rather than multiple inputs as only accuracy is considered. Studies which are combining various forms of data and applying it to machine learning models which might be not the scenario for majority of the real time data sources, with the limited number of data and its format approach to utilizes such data is showcased by [14, 50-51]. Models were observed which utilize the audio as well as visual data for the

classification of the emotion, but in the scenarios where the data available is only in the form of audio and the classification is need to be done onto audio data, it makes the machine learning model vulnerable as chances of the false prediction increases. To mitigate this issue, research showcased the approach to the emotion classification with the audio data as the RAVDESS dataset was taken. Firstly, pre-processing of the audio files was conducted as features like Mel-Frequency Cepstral Coefficients (MFCCs), Log-Mel spectrogram, energy, and pitch were considered for analysis. As data is ready for further processing methods like LHMMs, LSTM, DNNs, and CNNs are implied which classifies the result into 14 classes of emotion comprising of 2 genders with every 7 different emotions. Basically, on observing various models CNN in combination with other models and parameters of layer reached test accuracy up to 86% overall having the 2D structure of 4 layers and Log-Mel spectrogram features. On analysing the outcome of the various model implemented with different structures, it showcased that selection of the input data and choice of audio features plays a vital role in the prediction of the unknown data rather than the complex or bulky machine learning structure.

Implementation of the data onto machine learning model for the prediction of the emotion class with input data as text or audio or video showcased in many researches provided the effective outcome by mentioned methods. As in the real-world implementation, it can be the whole input data in the form of text, audio, video, etc which needed to be simultaneously processed to classify its emotion class. In this research by [15, 52-54], an approach to such a method is showcased in which emotion classification is conducted through the various model which can handle facial expression, speech, text, audio, etc. The Hidden Markov model is responsible for the evaluation of the outcome of each multi-modal model implemented in this research and classifying it into basics seven emotion classes. For Textual data NLP with ml model approach is selected whereas for audio and video data models like GMM, ANN, etc are utilized. To test the proposed method vigorously, multimodal sentences of the input data are constructed which involve more than one emotion in it. The outcome of the various model is converted into mixed vector consisting of all results of input data through models implied in it. Evaluation of the model proposed is done on three values specificity, recall, and precision during its training and testing period showcasing an effective classification of emotion in the multimodal sentence as input data. Although the classification of the emotion from the various type of data available shall be a saturation point where the performance can be achieved at the cost of time. Along with the performance of the model to classify the emotion the processing and the structure of the proposed model plays a vital role in the implementation of the model on large scale. This proposed research by [16] showcasing an optimal model for the classification of emotion based on various types of input data as well as the fast-processing speed of the structure. This advantage had made the method to approach Emotion Recognition in the Wild 2018 challenge [17]. As part of the challenge, the input data is needed to be mapped into any one of the universal six emotions as Disgust, Fear, Sad, Surprise, Anger, and Neutral. The proposed model was a multimodal emotion recognition system that takes input data such as audio, text information, and video. Extraction of bottleneck features from deep neural networks (DNNs) via transfer learning is implemented [18]. Temporal and non-temporal classifiers both are evaluated to obtain the optimal unimodal emotion classification result. Furthermore, Beam Search Fusion (BS-Fusion) is used to calculate emotion possibilities from the results extracted. For testing and training of the model in the challenge, the AFEW database which consists of video clips is used with labels, overall, 773 video clips are fed to the model for training, 383 for validation, and 653 for testing. A model is tested in the Emotion 2018 challenge, it showcased 60.34% of accuracy results from the baseline system and also showcased 1.5% lower accuracy than the winner of the challenge. Implementation of emotion classification upon data such as speech, text, video, etc was proposed by many researchers but the main outcome is the linearity of the input data features while training or testing of the model is a must. When the model is practiced in real-world inputs the assured of the input data features as linear is not assured, which makes the model less effective. To overcome this scenario, research proposed a novel approach to an emotion classification model from speech signal which uses empirical mode decomposition and non-linear features by [19]. A non- linear signal quantifying method which is based on randomness measure also termed as entropy feature is used for the detection of emotions. Initially, input speech data is distinguished into high, mid, and bass frequency domains where the high-frequency entropy measures are directly computed whereas, for mid and base, it is averaged to form a feature vector that possesses randomness of the features for all emotional signals. As a feature vector is obtained it is used to train the classifiers such as linear Discriminants analysis, (LDA), K-nearest neighbour, Naive Bayes, SVM, Gradient Boosting Machine, and Random For- est. Dataset selected

for the model is the Toronto Emotional Speech dataset which showcased peak accuracy on LDA with 93.3% and an F1 score of 87.9%. To classify emotion from the provided text data, the model needs to detect the correlation between the words or label that exists. Current approaches were widely using NLP methods as well as their toolkits which makes it vulnerable to the sector like medical, security, etc where the possibility of false detection plays a vital role in the outcome. Research conducted by [20] showcased a model termed as SpanEmo which cast multi-label emotion classification as span-prediction. This approach helps to learn the emotion recognition model to develop associations between words in sentences and labels. Furthermore, to improve the accuracy of the proposed method, the loss function is introduced which focuses on modelling multiple co-existing emotions in the given input sentence. Dataset of 2018 termed as SemEval2018 which included languages like English, Arabic, and Spanish which demonstrated method's effectiveness. For three different languages, the performance of the model showed the miF1 score lies between 0.662 to 0.724 score.

As video contains a large amount of data for classification of emotion from it by machine learning models as text, speech, audio, images, video, etc can be extracted [10]. To use the extracted information simultaneously with the usage of multiple models is proposed in this research. Basically, from video data, parameters like facial expression, images, audio, video, text, and speech can be extracted to train various models. To reduce the computational cost as well as regularize the training process boosting method is used for audio-video information exchange and separable convolution strategy respectively. Firstly, music and video capture all acoustic and visual findings through multimodal representations followed by computation costs are reduced by the utilization of 2D/3D convolution into separable channels and spatiotemporal interactions. Lastly, overall performance is boosted by using information-sharing methods in multimodal representations which also help in guiding the individual flow of model and data. Various unimodal and multimodal network methods are implemented and tested which best attained the accuracy at 74% and a F1 score of 0.73. [21] observed that the real-time response of the model for the classification of the emotion on the data is poor. When the input data is as image or text with a slow amount of testing, it makes the proposed model effective for classification but in the case of the video stream of a large amount of data it makes it difficult to process. This approach, it had focused on the real time approach where the training of the model can be done once, and due to its light model usage, the input data can be processed at a very high speed which makes it effective for the video stream data. This proposed system consists of logistics regression, and a Stochastic Gradient Descent algorithm for the processing of the data effectively as well as quickly. This proposed real time emotion classification model is trained in an online fashion using an EEG signal stream [22]. DEAP dataset is used to validate the training of the model which is widely used as the benchmark dataset. Results showcased effective performance for the classification of the emotion in real-time from the EEG stream. To compute the emotion effectively instead of discrete classification of the emotion, it is classified based on the VA space emotion model. As implementation of the model in the actual scenario and its error is ranged from 0.6 to more than 0.2. Classifiers like SVM, MLP, DT, and RECS were used and compared with each other for offline as well as the online implementation of the proposed model with an accuracy of 52.83, 55.48, 54.82, and 67.96 respectively. When the selection of the data for training the model, majorly video-based data are selected but when the situation arises where only text or lyrical data is available, author [23] showcased the approach for emotion recognition from audio and lyrics. A Multimodal approach is proposed which firstly classifies the given data into discrete emotions using the MIREX mood classifier is showcased. A MIREX-like audio dataset is used which consists of 903 samples whereas a new multimodal dataset consists of 193 audio, lyrics, and midi samples. While using a multimodal dataset and 19 features only the f-measure attainted 61.1% whereas in using a standard dataset it had reached 44.3% only (F-measure). Results also highlighted the importance of melodic features for improving the efficiency of the model.

1.3 Summary

A detailed study from some related work and research conducted in an emotion classification domain provided me with the importance of emotion classification as well as its immense potential and applications in various sectors. Not only through a video or images but the combination of various inputs to the machine learning model had also been an unexplored domain, especially in the data like [5] EEG signals, text [4], and audio. Effectiveness and availability of the real-world data as well as resources, input data like audio and text shall

be a suitable choice through which a better accuracy, as well as classification, can be attained with the utilization of machine learning models. It also provided me with an overview of the environment in which the model shall be used as well as data that has been collected. The complexity of the data and the model used the main two factors on which the performance of the classification is dependent which led to the selection of the music as our audio data comprises of the lyrical and musical mixture through which the model shall showcase better performance along with its lyrics (text). A combination of lyrical emotion classification, as well as musical emotion classification, may provide us a better and more reliable result with which a discrete emotion of the provided input can be extracted is mainly observed through various research. For natural language processing, BERT and GPT-2 are selected whereas for audio processing, [24] ResNet50 model with attention module.

2 Dataset

2.1 Dataset Acquisition

For this study, two datasets were used for each of the two languages, Hindi and English. MER500 for Hindi language and [25] DEAM Dataset for English songs. Where the MER500 dataset consists of around 100 pieces, with about five categories: Romantic, Happy, Sad, Devotional and Party, DEAM Dataset consists of 1802 songs. This dataset does not contain labels like the MER500 dataset but has continuous annotation between -1 and +1, which is mapped to the same categories from the analysis during the data processing [26].

2.2 Data Processing

We have also utilized the NLP, in which we have extracted lyrics from the songs. Dataset does not contain the poems themselves. Poems have been removed from each dataset using the lyrics extraction model [27].

2.3 Lyrics

Lyrics are classified using the BERT and GPT2 models, which are further explained in the following sections. BERT and GPT2 Require text data to be processed and converted into tokens. A general procedure of text pre-processing includes removing the links and numbers; then tokenisation is performed, splitting the sentence into a sequence or list of words. Where each word is a token, stop words are removed, which are words like the, is, am, are or commonly occurring words. Then stemming of the remaining terms is performed in which each word is converted to its root form; for example, “running” is converted to “run”. Lastly, Lemmatization, in this step, converts given the word to a word with a similar meaning or the closest word in the dictionary of the NLP tool. And then provided to the text embedder which embeds the processed text into numerical form. However, here GPT2 and BERT have their own tokenizer and embedder, once the text is processed it is again converted to string from the processed words. And that text is passed to the tokenizer of the GPT2 and BERT both of one has different tokenizers.

2.4 Music

Since this music is digitally recorded it does not require cleaning, instead, more noise is added to increase the robustness of the model. In all music files, a white noise had been added with random intensities. As the results suggest it improves the robustness of the model, where even if noised music is classified, it classifies with sufficiently good performance. Also, MIDI of the music is treated the same way. And from that music files, additional features were extracted [28].

3 Feature Extraction

The audio signal is processed upon mainly 34 features on which the machine learning model can execute its

function. Zero-Crossing Rate, Spectral Entropy, Spectral Flux, Spectral Centroid, Spectral Spread, Spectral Rolloff, Energy, Entropy of Energy, MFCCs, Chroma Vector, and Chroma Deviation whereas for lyrics, natural language processing is used [29].

3.1 Zero crossing rate

It can be explained as the number of times the signal crosses the zero value or shifts from positive to negative or vice-versa and dividing it by the length of the signal is known as its zero-crossing rate [11]. Mathematically it can be represented as represented as fig. 1

$$z(i) = 1/(2 * W_L) \quad (1)$$

$$\sum_{n=1}^{W_L} abs(sgn(x_i(n)) - sgn(x_i(n-1))) \quad (2)$$

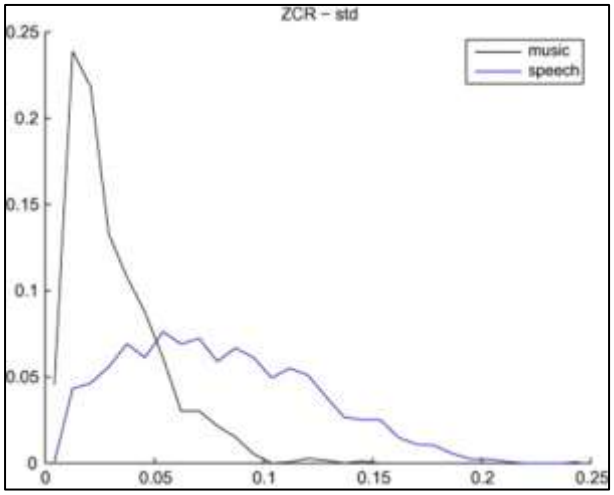

Fig. 1 Zero crossing

Where, W_L stands for the length of windows $sgn(.)$ is the sign function which results 1 in a positive outcome and -1 in a negative outcome. Indicating the noisiness in the signal or discriminating between speech and music or classifying music genres are some of the widely used applications associated with zero-crossing rate. When the music signal is processed for a zero-crossing rate it usually shows a higher number of crossings than a speech-based signal with which we can discriminate between a music and speech signal 1. Energy can be expressed as the sum of squares of signal values of the provided audio signal which showcased the amount of energy a signal possesses [30]. It can be expressed as:

$$E(i) = \sum_{n=1}^{W_L} X_i(n)^2 \quad (3)$$

Where, $E(i)$ is the energy of the signal and W_L is the length of windows. It represents the frame of that particular instance. To normalize the signal, it is usually divided by the frame length of the respective signal and also known as power [31] which is expressed as:

$$E(i) = \frac{1}{W_L}\sum_{n=1}^{W_L} X_i(n)^2 \quad (4)$$

The entropy of Energy: The entropy of energy represents the measure of abrupt changes in the energy level of the provided audio signal [32]. Given audio signal is divided into sub-parts with equal length and its energy is computed and divided by the total energy of each subframe obtained.

$$e_j = \frac{E_{Subframe_j}}{E_{Subframe_i}} \quad (5)$$

In the above equation

$$E_{Subframe_i} = \sum_{k=1}^{K} E_{Subframe_k} \quad (6)$$

Where, $E_{Subframe_i}$ is the total energy of short-term frame k and k is the number of short-term frames of fixed duration, and finally

$$H(i) = -\sum_{j=1}^{K} e_j log_2 e_j \quad (7)$$

Where, $H(i)$ is the entropy of the sequence and k is the number of short-term frames of fixed duration. As this feature showcases the value of the sudden changes in the provided audio signal [31], it is widely used for the detection of the sound like an explosion, gunshot, etc whereas for discrimination of genres, like classical, jazz, and electronics can also conduct with the help of entropy of energy values obtained. Fig. 2, shows the entropy of different genres of music.

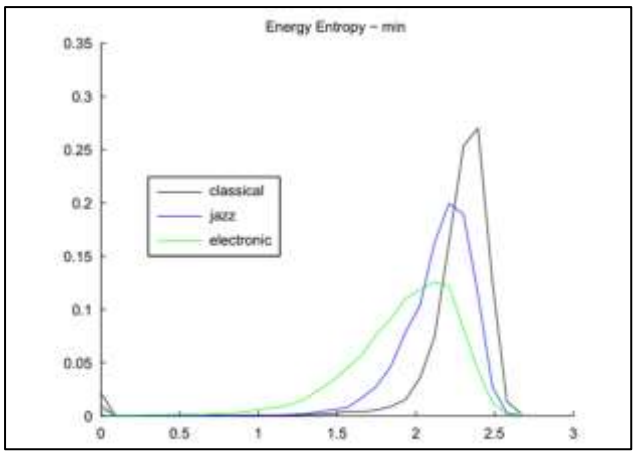

Fig. 2 Energy of different music signal

Spectral Centroid: Spectral Centroid is the widely used audio signal feature as it showcases the central mass of the spectrum in other words, it shows on which band of frequency most energy is concentrated [33]. By calculating the weighted mean of frequencies available in the provided signal spectral centroid is obtained [31]. For the frame f, we can compute its spectral centroid as:

$$Sc_t = \frac{\sum_{k=1}^{N} k.A(k)}{\sum_{k=1}^{N} A(k)} \quad (8)$$

Where, Sc_t is the Spectral Centroid and N is the number of coefficients that are used in the computations while $k.A(k)$ is product of the magnitude of bin number k and the central frequency of that bin. Fig. 3 represents the spectral centroid of the signal.

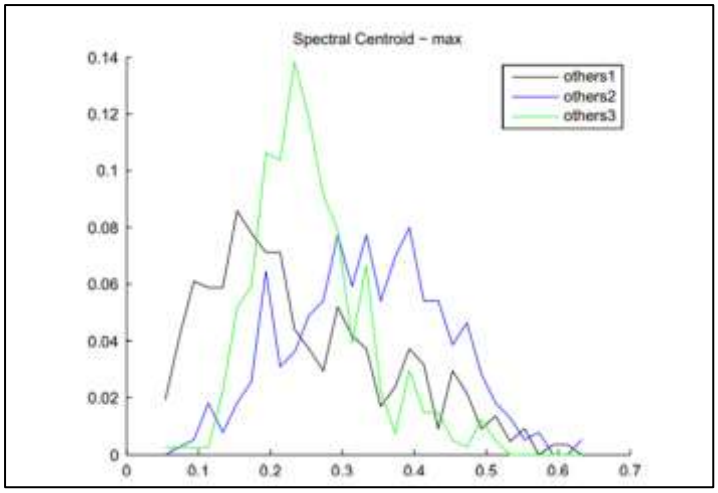

Fig. 3 spectral centroid of the signal

Spectral Spread: After obtaining the spectral centroid of the signal, calculating how the spectrum is distributed around the spectral centroid represents the spectral spread of the given signal [33]. It showcases the variance from the calculated spectral centroid [31]. It can be calculated as:

$$SS_t = \sqrt{\frac{\sum_{k=1}^{N} k - Sc_t A(k)}{\sum_{k=1}^{N} A(k)}} \quad (9)$$

Where, SS_t is the Spectral spread, $k - Sc_t A(k)$ is the distance of the frequency band from the spectral centroid. Spectral centroid and spread are majorly used for the classification of the music genre. In fig 3, other1 is background noise and silence which show cased the comparatively lower spectral centroid with other2 and other3 which are the music of different genres.

Spectral Entropy: The abrupt change in the spectrum is termed spectral entropy [34]. It is similarly calculated as the entropy of energy but in the frequency domain. It can be represented as:

$$H = -\sum_{f=0}^{L-1} n_f log_2 n_f \quad (10)$$

Where, H is the spectral entropy, L is the bins (sub bands), n_f is the normalized spectral energy. This feature is practiced widely for discrimination of the music and speech as signals. Fig. 4 represents the spectral entropy of the signal.

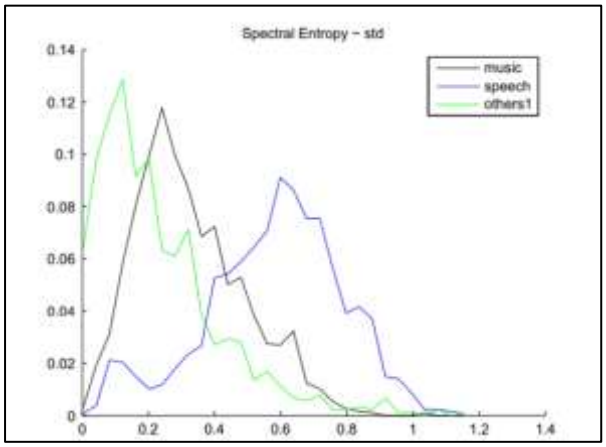

Fig. 4 spectral entropy of the signal

Spectral Flux: Spectral flux can be explained by the measure of spectral change between two successive frames and calculated as the squared difference between the spectra's normalized magnitudes (two successive windows) [31]. By using the mentioned formulae, we can compute spectral flux:

$$Fl_{(i,i-1)} = \sum_{k=1}^{Wf_l}(EN_i(k) - EN_i(k-1))^2 \quad (11)$$

Where,

$EN_i(k) = (\frac{X_i(k)}{\sum_{l=1}^{Wf_l}(X_i(l)^2}$), $Fl_{(i,i-1)}$ is the spectral flux for the i^{th} frame and $EN_i(k)$ is the k^{th} normalized DFT coefficient at i^{th} frame.

It is used for the discrimination between music and speech signal as displayed below in Fig. 5.

Spectral Rolloff: Spectral Rolloff is the frequency for the provided audio signal in which the magnitude distribution of spectrum is concentrated below it (usually it is 90%). Considering C as the adopted percentage, the equation for computing spectral Rolloff is:

$$\sum_{k=1}^{m} X_i(k) = C \sum_{l=1}^{Wf_l} X_i(k) \quad (12)$$

Where, $X_i(k)$ is the Discrete Fourier Transform (DFT) coefficients of the sequence, m is the Rolloff frequency. It represents the spectral shape of the provides an audio signal which can be very helpful for discriminating between voice and unvoiced sounds [35].

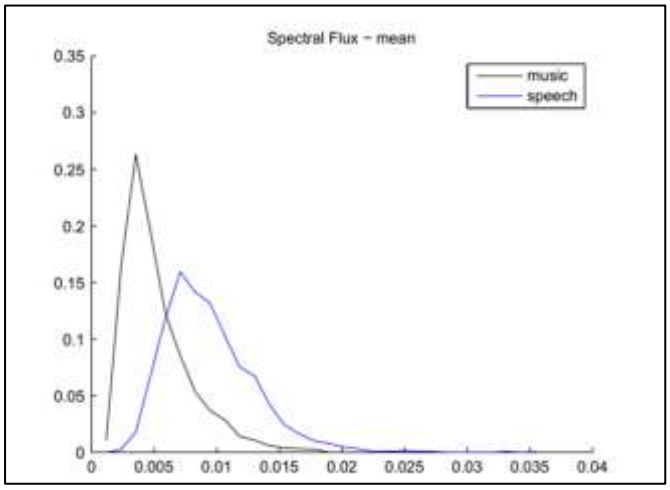

Fig. 5 spectral flux of the signal

MFCC: Mel Frequency Cepstral Coefficient abbreviated as MFCCs is the most popular approach used in the speech processing field [14]. It is a representation of the signal's frequency band distribution according to the mel-scale. It involves majorly 3 steps to extract MFCC feature in the given audio signal which are as: A Discrete Fourier Transform of the signal is obtained. Obtained spectrum is fed into the mel-scale filter which consists of L filters [31]. As it has overlapping triangular frequency responses, the frequency wrapping effect is attempted to confirm it with many specific psychoacoustic observations containing the human auditory system as discriminating neighbouring frequencies is comparatively easier in low-frequency regions. The mathematical representation of the same is as follows,

$$C_l = \sum_{b=1}^{B} \log n_b * \cos\left(\frac{l(b-0.5)\pi}{d}\right) \quad (13)$$

Here b signifies the number of triangular shape-based filter functions ranges from 1 to B, $\log n_b$ represents the

output of the B channel filter bank and l represents the index cepstral coefficient.

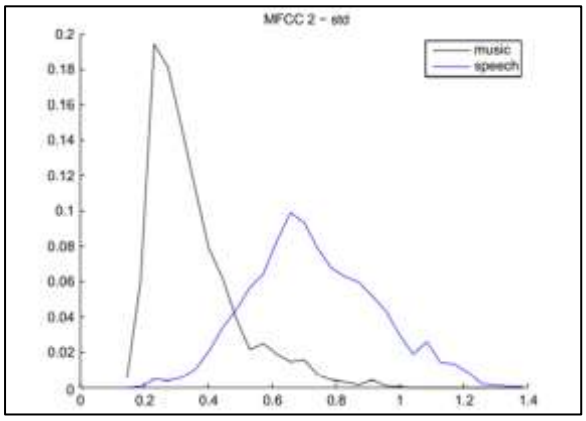

Fig. 6 MFCC of the signal

Chroma vector: The chroma vector feature is focused on music-related analysis and computing. It is a representation of the 12 elements of spectral energy. It is obtained by grouping DFT coefficients (short-term) into 12 bins. 12-equal tempered pitch classes of western type music are included in each bin. The Chroma signifies the 12 semitones which indicate the spectrogram of twelve frequency bins. This feature is used to evaluate the pitch class in minimal duration. The chroma is the circular arrangement of the spectrogram which is warped logarithmically. For calculating the energy at the twelve pitch classes the octave chroma frequency is summed and gathered. Thus, the indication of the musical keys and entire modality are gained using these chroma features.

$$v(t,k) = \Sigma_{n\in S_k} \frac{Z_t(n)}{N_k}, k \in \{0,1,\dots\dots,11\} \quad (14)$$

Where, the elements of chroma feature vector at t^{th} frame is $v(t,k)$. $Z_t(n)$ indicates the logarithmic magnitude of DFT of t^{th} frame, the subset of discrete frequency space for each pitch class is S_k, N_k is the total number of elements.

4 Classification

4.1 Lyric

Text of the audio which is been processed is also been studied to extract the emotion. As lyrics consist of the single language (English) text which can be processed with natural language processing techniques for its sentiment analysis and classification of its emotion. Generally, NLP is implemented in four steps which involve lexical analysis followed by parsing and sematic analysis, furthermore, discourse integration and pragmatic analysis are also conducted [36]. It provides us with the discrete emotion contained in the provided text which shall help compare the results with the audio analysis results. BERT stands for Bidirectional Encoder Representations from Transformers which is a machine learning technique for natural language processing. It uses a mechanism that learns contextual relations between words and sub-words which is known as a transformer. The main difference between the other model and Bert can be spotted as it uses a whole set of words as the inputs and then process the whole inputs words which provides the BERT model to look around every word as well as neighbouring words instead of sequentially.

The process can be explained as I initial step, token embedding takes place which provides an input sentence a token at starting and ending. Furthermore, segment embedding takes place in which markers indicating different sentence is added to each token. Using this step encoder can distinguish between each sentence. For indicating the position of the sentence, the token is provided for it.

A machine understandable format from the raw text input data is been converted which is fed into the training of the model which is broadly divided into two approaches Masked LM (MLM) and Next Sentence Prediction (NSP). In MLM, the provided input is randomly masked out its nearly 15% of the words. Instead of the words, token (mask) is used for its further step to predict the tokenized/masked words by the model is carried out. This

step plays an important role as masking some words in a sentence allows the model to avoid too much focus on the position of the sentence [37]. Alternatively, 80% of words with the masked token, 10% of words randomly masked as well as 10% left unchanged are also been practiced. For Next Sentence Prediction (NSP), it learns to develop a relationship between the words as well a prediction of the next sentence is also performed. Here corpus plays an important role in which the inputs are compared. In our approach, we had fine-tuned the model for next sentence prediction such that, the output is based upon the provided input of the text data and it shall represent the prediction from the five discrete classes of the emotion. By this modification, the output indicated the emotion behind the fed input to the model.

GPT-2 stands for Generative pre-trained transformer which is basically a pre-trained model (transformer) through which visualizing the generation of the words or language, and its output is an auto-regressive in nature. As our main purpose is to process the lyrics data which is in form of text, and analysis it, apart from the processing of the data visualization also helps us to better understand the provided text and its activities associated with it. Foremost advantage for using the GPT is it is a pre-trained as well as open-source model providing the support as well as vast variety of usage like classification of text or emotion or sentiment or predicting the next word. As it is built upon decoder layers, the prediction of the next word based upon the provided input text shall be the task which GPT2 shall providing the highest accuracy and performance. When the input data in the form of text is provided to the GPT2, it undergoes various encoding steps which provide tokens and fed into the pre-trained model of the GPT2. For our emotion classification approach, we had fine-tuned the model with the prediction of the next word from the five classes of the discrete emotion class. By using this feature, we have utilized the GPT2 model for providing the output in the discrete emotion class based on the input text data [38, 43-49].

4.2 Audio

Data consisting of the numerical values of the electronic signal especially the vibration of the air molecules stroked on a microphone is termed as audio signals [39-42]. It plays a vital role in our proposed solution as this data shall provide the model with enough features through which the emotion behind the audio can be classified. ResNet50 is a model which shall be practicing it in our proposed solution. It is 50 layers stacked up together which consists of the 48convolution layer followed by 1 layer of max pool and 1 layer of average pool layer. Comparatively, with other available machine learning models, the skipping approach is also known as a highway network. It provides the skip to the part of the model which shall be degrading the performance and accuracy while training with the particular input. Considering an input x, feed into weight layer wit relu activation followed by another weight layer as the output as F(x) and skipping of the block as x identity, the summation for the final output shall be F(x) + x and it can be pictorial represented as below Fig. 7:

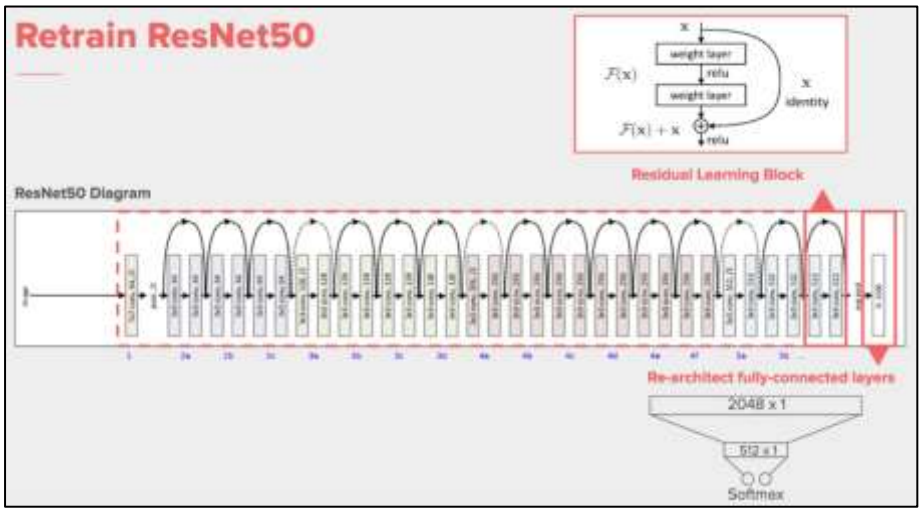

Fig. 7 Pictorial representation of ResNet50

As our audio signal is passed into ResNet50, it goes through these 50 layers of the neural network to classify the discrete emotion. Internally, as represented by a single block, overlaying this type of neural network block on each other making 50 layers and providing the single class of the emotion in the output is the keen interest of the model. The results are further evaluated by comparing them with the produced output of the lyrics.

4.3 Ensemble

In the ensemble model, voting is considered of the 4 models which are created for classifying song, karaoke, and lyric [40]. Since there are 4 models there should be change of having tie, in such scenarios final outcome is selected on the random selection. Random is selected using the Normal distribution, where mean = 0 and the standard deviation is 1. So, threshold is considered a positive and negative values to select either one of the classes. There are also scenarios where all 4 model will have different prediction, because this is a multi-class classification problem and there are 5 classes. In another cases the prediction is carried out on the basis on the max voting. For example, 3 out of 4 model predicts the same outcome then the final outcome is selected. In another scenario if 2 models have predicted same outcome but other 2 had different then in that scenario, outcome which has most count is selected which means the outcome which had been predicted by 2 models. Since there are fewer models than the target class, ensembled model is likely to have lower performance. But in some cases, it may also have the better performance [41].

5 Result and Discussion

As discussed earlier, for the demonstration of the proposed approach, here we have used the 2 datasets and in this section the performance of the model is evaluated using the classification measures: Accuracy, Sensitivity (or precision), and F-1 score. As there are 4 models and the 5th model is the ensembled model. As mentioned earlier, ensemble model is nothing but voting of 4 model. Also, here for each of the 2 dataset these multiple models are for each specific type of the data: Song, Karaoke, and Lyric of the single music. As can be seen from the result highest accuracy can be achieved from the Lyrics of the song, and it's giving the same performance in both the dataset. This indicates that the words which are used in the song are stronger predictor of the emotion. Performance matrix obtained for MER500 and DEAM datasets are be graphically represented in Fig. 8 and Fig. 9.

Table 1 MER500

Performance Matrix	**Song**	**Karaoke**	**Lyrics (BERT)**	**Lyrics (GPT2)**	**Ensembled**
Accuracy	0.81	0.82	0.84	0.78	0.85
Precision	0.80	0.82	0.84	0.78	0.85
Sensitivity	0.81	0.82	0.84	0.78	0.85
F-1 Score	0.81	0.82	0.84	0.78	0.85

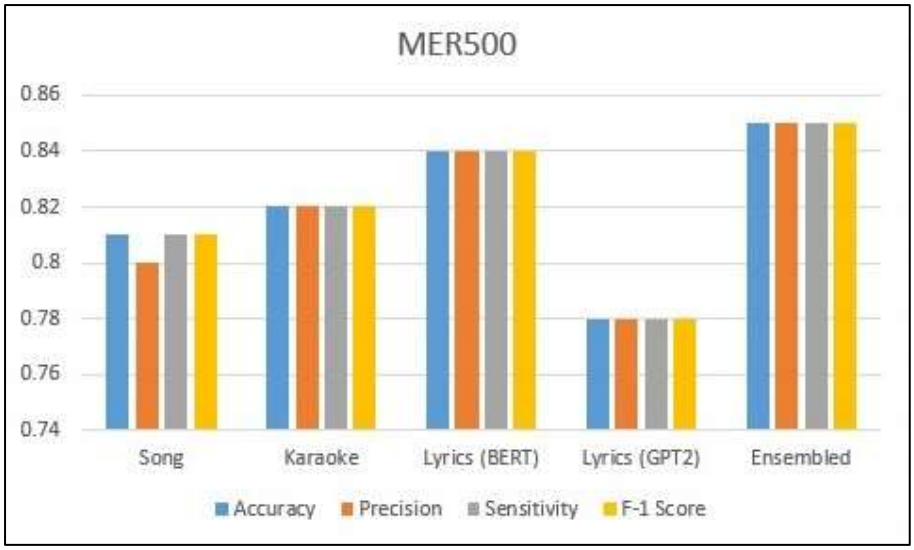

Fig. 8 MER500 Performance Matrix

Table 2 DEAM

Performance Matrix	**Song**	**Karaoke**	**Lyrics (BERT)**	**Lyrics (GPT2)**	**Ensembled**
Accuracy	0.82	0.81	0.84	0.78	0.83
Precision	0.82	0.81	0.84	0.78	0.83
Sensitivity	0.82	0.81	0.84	0.78	0.83
F-1 Score	0.82	0.81	0.84	0.78	0.83

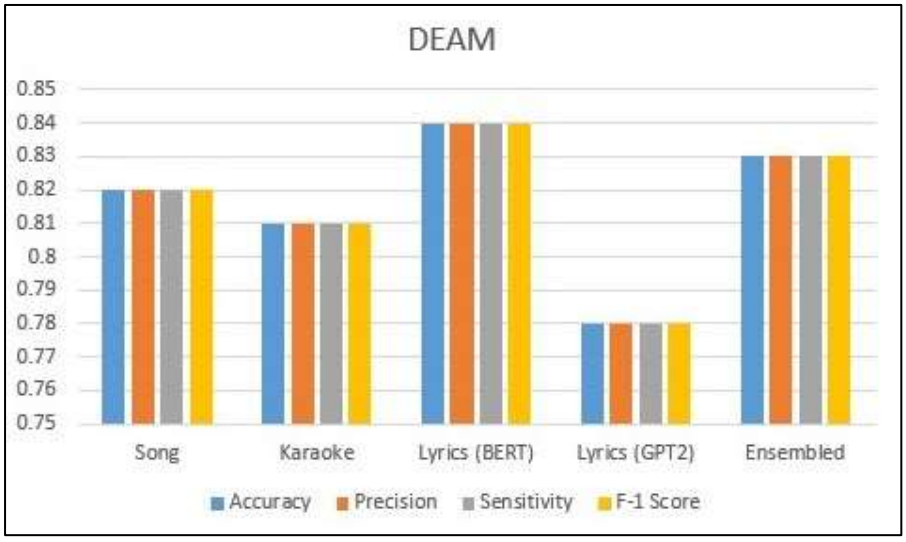

Fig. 9 DEAM Performance Matrix

Table 3 DEAM Testing Accuracy

	296	**345**	**395**	**444**	**494**
Song	0.66	0.77	0.74	0.83	0.87
Karaoke	0.65	0.72	0.76	0.83	0.84

Lyrics (GPT-2)	0.68	0.70	0.78	0.86	0.83
Lyrics (BERT)	0.62	0.73	0.78	0.85	0.82

On closer observation, the values from the above table showcase the progressive nature of accuracy with an increase in the number of samples for training. With only 296 samples, under the song category the model had attained about 66% of accuracy whereas the samples rise, the accuracy of the model had also risen to 87% when 494 samples are used. It indicates that the proposed model can attain higher accuracy if the training dataset shall be available in a larger amount. As this dataset is labelled Hindi 500 songs. As the deficiency of data produced a limitation, usage of the DEAM dataset which is much larger than MER500 is practiced but it is for English songs.

Table 4 MER 500 Testing Accuracy

	1079	**1259**	**1440**	**1620**	**1802**
Song	0.64	0.72	0.80	0.8	0.83
Karaoke	0.65	0.70	0.77	0.83	0.81
Lyrics (GPT-2)	0.64	0.71	0.77	0.82	0.80
Lyrics (BERT)	0.62	0.74	0.77	0.82	0.80

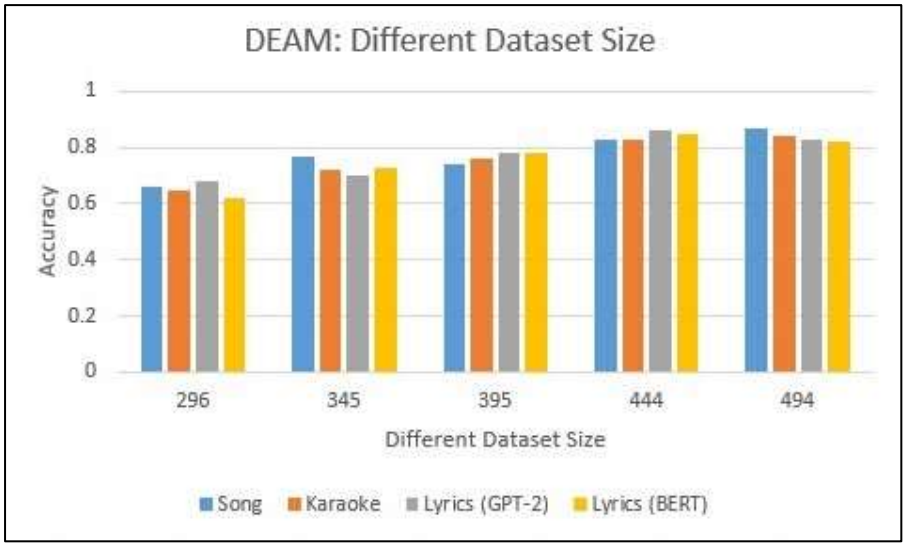

Fig. 10 DEAM Accuracy with Different Dataset Size

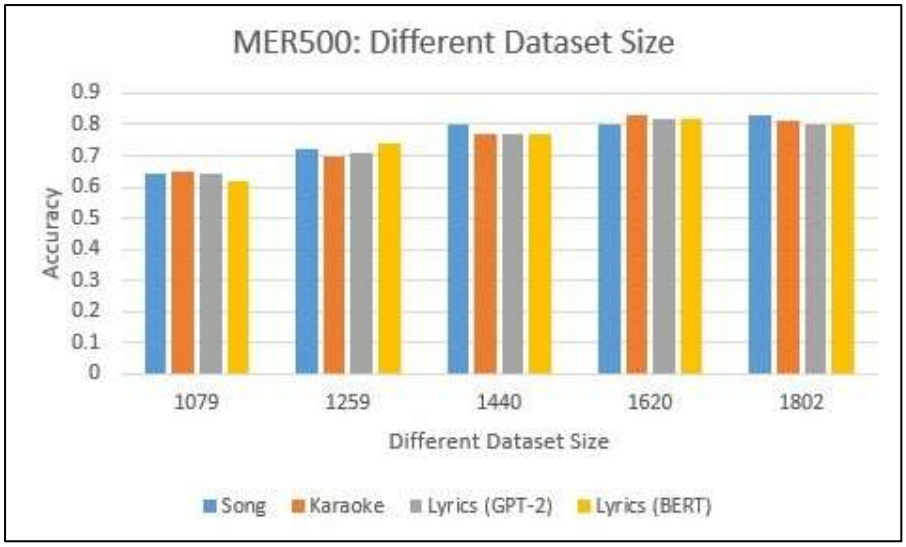

Fig. 11 MER 500 Accuracy with Different Dataset Size

Using DEAM dataset, for comparing the results with MER500, 5 splits of DEAM dataset had been tested. With the each testing the accuracy of dataset had been increased and the ratio of the data had been kept same in order to study the comparative results. Ranging from 1079 samples to 1802 samples the model's accuracy had been observed rising. Similar behaviour had been also observed for MER500 dataset which provided the limitation of dataset with model's accuracy. Graphical representation of the accuracy with different size of datasets had been showcased in Figure 10 and Figure 11 for DEAM and MER500 datasets respectively. From the above results we can conclude that if the dataset and number of samples are available in large amount, higher accuracy of the model can be achieved for classifying the discrete emotion from song or karaoke or lyrics. It is also observed that with the help of lyrics, better results classification results are attained as lyrics is a more important data than other for classifying emotion.

On comparing the various BERT-based classification models and their respective studies it can be observed that apart from the data used for classification, classes (outcome elements) are also important which affects the accuracy of the model. When the label rose from 8 to 343 the accuracy of the model can be observed to decrease. Size of the data as well as output of the model's performance best suitable example can be observed in the table 5 where only of 4 categories are in the output had provided the models with 80% to 90% of the classification accuracy. Furthermore, if the binary classification is executed as showcased in table 5 which is showcasing about 93.87% of accuracy. After observing various models and their research based on the BERT classifier, their results are compared with the observation in our research, although the usage of the dataset was different, the outcome ranges between 60% to 80% whereas our model had also showcased similar results for the classification of emotion.

Table 5 Comparison of TEXT model

Research Paper Title	**Model**	**Dataset**	**Label**	**Observation**
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [42]	BERT base and BERT large	GLUE, SQuAD v1.1 and v2.0, and SWAG	Binary classification in GLUE	Between 80 to 93% models had showcased accuracies based upon their feature.
				Ranging from

DocBERT: BERTfor Document Classification [43]	BERT base and BERT large	Reuters, AAPD, IMDB, and Yelp'14	46 topics for Reuters	55.6% to 90.7% accuracies were attained on validation of data.
BERT for Joint Intent Classification and Slot Filling [44]	Joint BERT and Joint BERT + CRF	Snips and ATIS	10 classes	98.6% and 97.0% for snips by Joint BERT whereas 97.5% and 96.1% by Joint BERT + CRF.
PatentBERT: Patent Classification with Fine-Tuning a pre-trained BERT Model [45]	DeepPatent and PatentBERT	IPC + Title + Abstract and CPC + Claim	632 labels	73.88% and 84.26% were reached by the models respectively.
Comparing BERT against traditional machine learning text classification [46]	BERT	IMDB	Binary Classification	Reached upto 93.87% of accuracy.
BAE: BERT-based Adversarial Examples for Text Classification [47]	BERT based model	IMDB, Amazon, Yelp, and MR are the dataset used.	Four categories	For sentiment classification, about 80 to 90% of accuracy had been reached in different cases.
Enriching BERT with Knowledge Graph Embeddings for Document Classification [48]	BERT model with 12 hidden layers.	Books datasets consisting about 14548 books.	8 and 343 labels	Achieved an F1- score of 87.20 for 8 labels whereas, detailed classification using 343 labels yield an F1-score of 64.70

6 Conclusion

Classifying the emotion from the given data in discrete parameters (type of emotion) had been focused in our proposed solution in which we had utilized ResNet50, BERT and GPT2. MER500 dataset which is collection of hindi songs with 500 sample and DEAM dataset which is available in English are been utilised for training and testing the proposed models. Various features with which emotion can be classified are also studied. Observing the results justifies that with the increase in number of samples in datasets, the accuracy is also increased as well as lyrics had showcased better accuracy than songs or karaoke data type. For MER500 the

model had attained about 85% whereas
83% of accuracy had been achieved by ensembled model. As described earlier, which the increase in the dataset samples and including the different types of classes of emotion with well labelled dataset can produce significant rise in the proposed model's accuracy. By eliminating the pre-processing (filtration) of noise from the dataset, we had showcased the robustness of the model. As robustness plays an important role when the actual implementation of the model is practised, in future we shall also dig deep in this arena of robustness by superimposing artificial noise on the dataset.

References

[1] Muther, T., Syed, F.I., Lancaster, A.T., Salsabila, F.D., Dahaghi, A.K., Negahban, S.: Geothermal 4.0: Ai-enabled geothermal reservoir development-current status, potentials, limitations, and ways forward. Geothermics 100, 102348 (2022)

[2] Munawar, H.S., Mojtahedi, M., Hammad, A.W.A., Ostwald, M.J., Waller, S.T.: An ai/ml-based strategy for disaster response and evacuation of vic- tims in aged care facilities in the hawkesbury-nepean valley: A perspective. Buildings 12, 80 (2022)

[3] Siam, A.I., Soliman, N.F., Algarni, A.D., El-Samie, A., Fathi, E., Sedik, A.: Deploying machine learning techniques for human emotion detection. Computational Intelligence and Neuroscience 2022 (2022)

[4] Buechel, S., Hahn, U.: Emobank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. arXiv preprint arXiv:2205.01996 (2022)

[5] Chen, J., Ro, T., Zhu, Z.: Emotion recognition with audio, video, eeg, and emg: A dataset and baseline approaches. IEEE Access 10, 13229–13242 (2022)

[6] Jia, N., Zheng, C., Sun, W.: A multimodal emotion recognition model integrating speech, video and mocap. Multimedia Tools and Applications,
1–22 (2022)

[7] Van, L.T., Le, T.D.T., Xuan, T.L., Castelli, E.: Emotional speech recognition using deep neural networks. Sensors 22, 1414 (2022)

[8] Amjad, A., Khan, L., Chang, H.-T.: Effect on speech emotion classification of a feature selection approach using a convolutional neural network. PeerJ Computer Science 7, 766 (2021)

[9] Li, Z., Xie, H., Cheng, G., Li, Q.: Word-level emotion distribution with two schemas for short text emotion classification. Knowledge-Based Systems
227, 107163 (2021)

[10] Pandeya, Y.R., Lee, J.: Deep learning-based late fusion of multimodal information for emotion classification of music video. Multimedia Tools and Applications 80, 2887–2905 (2021)

[11] Yoon, S., Byun, S., Jung, K.: Multimodal speech emotion recognition using audio and text, pp. 112–118 (2018)

[12] Park, S., Kim, J., Ye, S., Jeon, J., Park, H.Y., Oh, A.: Dimensional emo- tion detection from categorical emotion. arXiv preprint arXiv:1911.02499 (2019)

[13] Oh, S., Kim, D.-K.: Comparative analysis of emotion classification based on facial expression and physiological signals using deep learning. Applied Sciences 12, 1286 (2022)

[14] Venkataramanan, K., Rajamohan, H.R.: Emotion recognition from speech. arXiv preprint arXiv:1912.10458 (2019)

[15] Caschera, M.C., Grifoni, P., Ferri, F.: Emotion classification from speech and text in videos using a multimodal approach. Multimodal Technologies and Interaction 6, 28 (2022)

[16] Lian, Z., Li, Y., Tao, J., Huang, J.: Investigation of multimodal features, classifiers and fusion methods for emotion recognition. arXiv preprint arXiv:1809.06225 (2018)

[17] Li, B., Weng, Y., Song, Q., Sun, B., Li, S.: Continuing pre-trained model with multiple training strategies for emotional classification. In: Proceed- ings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis, pp. 233–238 (2022)

[18] Das, S., Lønfeldt, N.N., Pagsberg, A.K., Clemmensen, L.H.: Towards transferable speech emotion representation: On loss functions for cross- lingual latent representations. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6452–6456 (2022). IEEE

[19] Krishnan, P.T., Raj, A.N.J., Rajangam, V.: Emotion classification from speech signal based on empirical mode decomposition and non-linear features. Complex Intelligent Systems 7, 1919–1934 (2021)

[20] Alhuzali, H., Ananiadou, S.: Spanemo: Casting multi-label emotion classification as span-prediction. arXiv preprint arXiv:2101.10038 (2021)

[21] Nandi, A., Xhafa, F., Subirats, L., Fort, S.: Real-time emotion classi- fication using eeg data stream in e-learning contexts. Sensors 21, 1589 (2021)

[22] Sakalle, A., Tomar, P., Bhardwaj, H., Iqbal, A., Sakalle, M., Bhardwaj, A., Ibrahim, W.: Genetic programming-based feature selection for emotion classification using eeg signal. Journal of Healthcare Engineering 2022 (2022)

[23] Panda, R.E.S., Malheiro, R., Rocha, B., Oliveira, A.P., Paiva, R.P.: Multi-modal music emotion recognition: A new dataset, methodology and comparative analysis, pp. 570–582 (2013)

[24] Madhavi, M., Gujar, I., Jadhao, V., Gulwani, R.: Facial emotion classifier using convolutional neural networks for reaction review. In: ITM Web of Conferences, vol. 44, p. 03055 (2022). EDP Sciences

[25] Balan, O., Moise, G., Petrescu, L., Moldoveanu, A., Leordeanu, M., Moldoveanu, F.: Emotion classification based on biophysical signals and machine learning techniques. Symmetry 12(1), 21 (2019)

[26] Tokuhisa, R., Inui, K., Matsumoto, Y.: Emotion classification using mas- sive examples extracted from the web. In: Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), pp. 881–888 (2008)

[27] Tanna, D., Dudhane, M., Sardar, A., Deshpande, K., Deshmukh, N.: Sen- timent analysis on social media for emotion classification. In: 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), pp. 911–915 (2020). IEEE

[28] Tong, G.: Music emotion classification method using improved deep belief network. Mobile Information Systems 2022 (2022)

[29] Feng, T., Hashemi, H., Annavaram, M., Narayanan, S.S.: Enhancing privacy through domain adaptive noise injection for speech emotion recognition. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7702–7706 (2022). IEEE

[30] ATİLA, O., ŞENGÜR, A.: Automatic speech emotion recognition using machine learning and iterative neighborhood component analysis

[31] Giannakopoulos, T., Pikrakis, A.: Introduction to Audio Analysis: a MATLAB® approach. Academic Press, ??? (2014)

[32] Burkhardt, F., Wagner, J., Wierstorf, H., Eyben, F., Schuller, B.: Nkul- uleko: A tool for rapid speaker characteristics detection. In: Proceedings of LREC (2022)

[33] Yang, Z., Nayan, K., Fan, Z., Cao, H.: Multimodal emotion recognition with surgical and fabric masks. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4678–4682 (2022). IEEE

[34] Yan, Y., Shen, X.: Research on speech emotion recognition based on aa- cbgru network. Electronics 11(9), 1409 (2022)

[35] Elbarougy, R., El-Badry, N.M., ElBedwehy, M.N.: An improved speech emotion classification approach based on optimal voiced unit

[36] He, N., Ferguson, S.: Music emotion recognition based on segment- level two-stage learning. International Journal of Multimedia Information Retrieval, 1–12 (2022)

[37] Li, S., Zhao, K., Yang, J., Jiang, X., Li, Z., Ma, Z.: Senti-exlm: Uyghur enhanced sentiment analysis model based on xlm. Electronics Letters (2022)

[38] Troiano, E., Oberlander, L., Wegge, M., Klinger, R.: x-envent: A cor- pus of event descriptions with experiencer-specific

emotion and appraisal annotations. arXiv preprint arXiv:2203.10909 (2022)

[39] Huang, Y., Song, R., Giunchiglia, F., Xu, H.: A multitask learning frame- work for abuse detection and emotion classification. Algorithms 15(4),
116 (2022)

[40] Desai, S., Kshirsagar, A., Sidnerlikar, A., Khodake, N., Marathe, M.: Leveraging emotion-specific features to improve transformer performance for emotion classification. arXiv preprint arXiv:2205.00283 (2022)

[41] Iyer, A., Das, S.S., Teotia, R., Maheshwari, S., Sharma, R.R.: Cnn and lstm based ensemble learning for human emotion recognition using eeg recordings. Multimedia Tools and Applications, 1–14 (2022)

[42] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)

[43] Adhikari, A., Ram, A., Tang, R., Lin J.: Docbert: Bert for document classification. arXiv preprint arXiv:1904.08398 (2019)

[44] Chen, Q., Zhuo, Z., Wang, W.: Bert for joint intent classification and slot filling. arXiv preprint arXiv:1902.10909 (2019)

[45] Lee, J.-S., Hsiang, J.: Patentbert: Patent classification with fine-tuning a pre-trained bert model. arXiv preprint arXiv:1906.02124 (2019)

[46] Gonzalez-Carvajal, S., Garrido-Merchan, E.C.: Comparing bert against traditional machine learning text classification. arXiv preprint arXiv:2005.13012 (2020)

[47] Garg, S., Ramakrishnan, G.: Bae: Bert-based adversarial examples for text classification. arXiv preprint arXiv:2004.01970 (2020)

[48] Ostendorff, M., Bourgonje, P., Berger, M., Moreno-Schneider, J., Rehm, G., Gipp, B.: Enriching bert with knowledge graph embeddings for document classification. arXiv preprint arXiv:1909.08402 (2019)

[49] Swati Goel, Monika Agarwal, “Cyber Security Technique for Internet of Things Using Machine Learning” published as a Book chapter in Holistic Approach to Quantum Cryptography in Cyber Security (1st ed.). (2022) CRC Press. https://doi.org/10.1201/9781003296034