

AI-Driven Emotional Analysis of Musical Structures for Enhanced Content Creation in the Modern Music Industry

Valentin-Ionut-Cosmin DUMITRESCU

*National University of Science and Technology POLITEHNICA Bucharest, Romania
vdumitrescu3007@stud.acs.upb.ro*

Gheorghe MILITARU

*National University of Science and Technology POLITEHNICA Bucharest, Romania
gheorghe.militaru@upb.ro*

Abstract. *In our research, we experimented with identifying emotions in music tracks by utilizing a multimodal deep learning approach that incorporated lyrical and audio data. To obtain semantic representations of the lyrics, we used BERT embeddings. As for audio modality, mel-spectrograms were extracted and forwarded into a convolutional network. Both representations were projected to the latent space of a given size which allowed our model to process both lyrics and audio simultaneously and extract emotional features from textual and acoustic representations of songs. We compared multimodal solutions to models that operated on a single modality: audio or lyrics. Results of our experiments demonstrated that utilizing both modalities allows our network to converge faster and attain better overall performance for all categories present in our dataset. During our experiments, we consistently observed that there are patterns in music which express emotions that can be better captured when considered from both an audio and linguistic perspective. Applications of our research could allow for music recommendation, soundtrack-related solutions, and assistance in song writing/producing. There are several limitations to our study. For one, our dataset was quite limited in size. In addition, some categories were not as heavily represented as others. We also used generated text to represent song lyrics in some cases instead of using the actual lyrics of the songs. Future directions can include working with bigger datasets, better language understanding for music lyrics, and experimenting with different techniques for multimodal fusion.*

Keywords: Artificial Intelligence, Multimodal Learning, Convolutional Neural Networks (CNN), Audio/Text Fusion, BERT Embeddings, Mel-Spectrogram Analysis, Music Emotion Recognition, Deep Learning, Affective Computing.

Introduction

In this work, we investigate how both structural and semantic elements of a musical composition may ultimately indicate how an artificial intelligence tool might identify emotional preference associated with the composition. The premise of this study is that lyrics provide linguistic data that, when combined with the rest of the musical data (i.e., harmony, melody, and rhythm) will provide sufficient information for an intelligent algorithm to make a correct recognition of the perceived emotion of the composition. In order to examine how well both sources of data will work together for emotional recognition of music, the researchers have developed a multimodal framework that implements a BERT encoder and a CNN network for text and sound processing (respectively).

Thus, one of the goals of the research is to create and evaluate the possibility of creating a system for recognizing musical emotions that will serve as a predictive tool, providing early information on the affective directions of a composition. The model will help producers and composers during the creation process due to the lower costs of obtaining information and will

save time. Moreover, the paper demonstrates how the implementation of AI technologies into the music industry may lead to its financial and operational sustainability.

The current work will not supplant human creativity, rather, the results will demonstrate how efficient it is to use another tool that provides options for creativity enhancement. This paper will examine how multimodal models analyze musical emotions better than other types of models and will provide more insight into the use of computational technologies in the music industry.

Another aspect of using a musical emotion recognition system that should be noted is its significant contribution to the industry's economic development. The implementation of the proposed solution will reduce costs connected with manual labeling of songs, selecting samples, and analyzing the affective direction of the compositions. In particular, the use of a system that will automate the process will allow for saving the financial resources needed for testing and adjusting since composers and producers will obtain information about their product immediately. Moreover, the prediction of users' emotions will help improve recommendation algorithms, leading to higher customer loyalty, as well as increase profits from additional paid services.

Literature review

Artificial intelligence (AI) is rapidly changing the way artistic creations are developed, distributed, and evaluated. The rapid technological advances in music have created an opportunity for new, innovative ways to help artists analyze and influence how their work produces an emotional response from audiences. While music has long been considered one of the most effective forms of communicating emotions, very little research has led to the development of methods for determining the relationship between musical and harmonic patterns and an audience's emotional response.[1] This study will explore the potential of utilizing AI in determining patterns of emotional response from a musical work based on its structure and composition in order to address the research gap and improve the artistic and economic performance of companies producing music.

During the last few decades, numerous investigations devoted to exploring the connection between artificial intelligence and music have been conducted. Attempting to capture the essence of the structure behind the music creation process, the first attempts to develop algorithms were based on the rules-based paradigm, probability models, and evolutionary computation. Despite the successful implementation, such systems were often unable to incorporate the element of human emotional expressiveness inherent to music due to their inability to represent complex dependencies between different parameters within the structure. Deep learning and neural network approaches have enabled a variety of new possibilities for both analyzing and generating music by providing computational means to work with the audio data.

Recently, the number of publications concerning sentiment analysis employing machine learning technologies is continuously growing, with many authors focusing on deep learning approaches, such as CNNs (Convolutional Neural Network) and RNNs (Recurrent Neural Network), LSTM (Long Short Term Memory) and GRU (Gated Recurrent Unit). Additionally, new types of language models have been proposed. For example, recently proposed ELMo (Embeddings from Language Models) [2] and BERT (Bidirectional Encoder Representations from Transformers) [3] create powerful semantic representations of text.

State of the Art

CNN-based models proved their ability to extract features from spectrograms and small texts, achieving excellent results in emotion classification. The combination of CNNs with LSTMs

allowed us to represent dependencies better in terms of time and context. The key development that happened in this area was the development of models based on Transformers, such as BERT, which made text analysis more context-aware and accurate.

Regarding the critical problem of sentiment classification, the models have provided remarkable results [4-5]. It is obvious that the positive outcomes obtained via the use of CNN models made researchers more interested in deep learning. The work described in [6] introduced a hybrid solution of CNN-LSTM framework with the use of the pre-trained word embedding technique to analyze short texts.

There is a wide range of research [7, 8], which showed the advantages obtained when using hybrid CNN-LSTM models. An exciting development in NLP was the creation of powerful language models (e.g., BERT) that inspired the community to develop solutions specific to certain languages for addressing various problems related to NLP (such as the sentiment analysis [9]).

To classify musical emotions and extract data, Liao et al. (2022) have developed an innovative algorithm, relying on residual blocks of inception. The findings presented by the authors confirm that AI models can detect the emotions present in the musical structure, and that the use of suggested algorithm improves emotion detection. The most significant advantage of the study consists in considering various characteristics of emotions (timbral, rhythmic, and harmonic) in musical compositions. [10] This paper is directly related to our project because it proves the theoretical foundation of our model – namely, that emotions can be extracted directly from the musical structure.

ATE (Aspect Term Extraction) models based on Deep Learning models started to attract attention of the research community recently. For instance, in [11], different word embedding techniques and CNNs were employed in multiple experiments. CNNs' inability to handle long-term dependencies is the major disadvantage of the model. This is the reason why recurrent structures such as LSTMs (long-short term memory) became very popular in the recent period. In [12], authors use the mentioned architecture to improve ATE model based on history aspects and opinion summarization.

Methodology

In this study, emotions are identified from musical structure using a supervised learning approach. The process consists of a few basic phases: Extraction of acoustic and lyrical feature data sets, training a machine learning model using these sets of data, and then measuring the result of how accurately an algorithmic system can classify different categories of emotion. The purpose of this entire operation is to determine if melodic, harmonic, rhythmic, and lyric elements can be consistently associated with the emotional nature of any one piece of music.

By structuring the project in such a manner, the work strives to reveal how particular musical features are connected to the listener's emotion and how these linkages can be represented in digital form. In addition to evaluating the approach taken by the study, it will also provide an overarching conceptual model that may be useful in future applications associated with music (i.e., for the purposes of music analysis, music recommendation, and creative support).

Approach

To see if it is possible to combine two complimentary types of information to identify emotional categories, the following sources will be considered:

- acoustic structure (melody, harmony, rhythm, timbre)
- semantic content (emotional cues expressed in lyrics)

It is unlikely that emotions in music appear as a result of the use of one single source. Whereas language communicates semantic and emotive meanings, audio captures intensity, rhythm, and timbre. As compared to unimodal models, when the combination is made, the system has more possibilities to detect complex motional patterns.

Audio selection and preprocessing

The dataset contains music clips categorized into four emotions (happy, sad, calm, energetic). The audio files were normalized by converting to mono channel, setting a fixed sampling rate of 22050 Hz and normalizing their length to 30 seconds.

All audio files were turned into mel-spectrogram, visual representation of sound commonly used for the purposes of music information retrieval with the help of Librosa software. Mel-spectrograms were created based on the following parameters: number of Mel bands = 128, FFT window size: 2048, hop length: 512. Such mel-spectrograms were converted into matrices and serve as audio input for the model.

Lyrics Processing

The BERT base uncased encoder was used to generate contextual embeddings, with 768 dimensions, for both the artificial and real lyrics. The generated embeddings were represented by the vector and used in the text branch of the proposed architecture.

The artificial lyrics were only used as a source of semantic information under controlled experimental settings without the purpose of showing actual understanding of the lyrics. This allowed to train the network on a stable dataset without claiming understanding of lyrics.

In order to rule out any possible issues of label leaking and test the multimodal network on realistic data, the model was also tested on real song lyrics taken from MoodyLyrics4Q dataset with authentic annotated emotions [16]. For this particular setup, the lyrics were processed by the BERT base uncased encoder with generation of 768 dimensional embeddings further projected onto 256-dimensional space.

Model Architecture

The acoustic features of the song and the meaning of its lyrics can be considered as two sources of information about musical emotions that will be used in the proposed model for interpretation. Finally, the prediction of emotions will be made through the independent analysis of the two sources using separate branches and their fusion in a common latent space.

The audio branch of the neural network will use a convolutional neural network (CNN) to identify acoustic features of songs based on mel-spectrogram. Mel-spectrograms can be interpreted as images and CNN is able to identify different areas of sound energy, frequencies that have shifted and different acoustic characteristics that are closely associated with the emotion a piece of music evokes.

In that mel-spectrogram is treated as an image, and because CNNs can do things well in:

- Pattern identification of sound quality
- Change detection of rhythm
- Harmonic structure identification
- Energy regions within sound spectrum

These parameters are all strongly related to how we experience emotion connected to the work of art represented by that audio. These characteristics are then encoded into a latent vector of size 256. The BERT linguistic encoder, which captures the context-dependent meaning of words,

is used in the second branch for processing the song lyrics. The BERT model extracts not only individual words but also the tone of the text, the meaning of the lyrics, the poetic form, and connections between the words. In order to facilitate fusion of the two modalities, BERT is projected on a latent space that is equal in dimension to the audio feature vector.

BERT features:

- emotion tone
- semantic purpose
- Poetic form
- Connections between the words

To project BERT on a space equal in size to the audio feature vector, the 768-dimensional vector outputted by BERT is projected into a 256-dimensional latent space.

The process of concatenating two feature vectors, an acoustic one and a semantic one, produces a multimodal feature vector that combines the rhythmic, timbral and harmonic properties present in the musical piece along with the emotional information associated with the lyrics. The model uses this aligned multimodal representation to develop the ability to extract complex emotional patterns from the music which could not be extracted using only a single modality.

The fused feature vector is then passed onto a dense classifier, which makes the final classification into one of four emotions: happy, sad, calm or active.

Architectural diagram

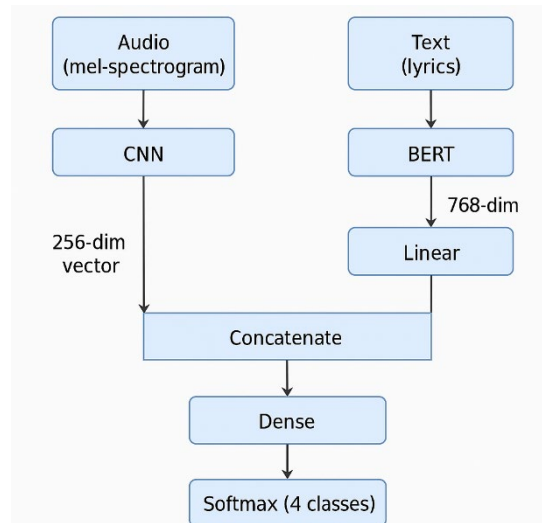


Fig. 1 Multimodal Architecture for Music Emotion Classification

Source: Authors' own research.

To classify musical emotions, the multimodal model in question extracts semantic and audio features. Specifically, the audio branch extracts rhythm and timbral information using CNNs from a mel-spectrogram to produce 256D latent vectors. Separately, the text branch of the model uses BERT to create a semantic embedding representing the music in 768D that can subsequently be compressed to 256D, allowing it to be aligned with the extracted audio features. To capture discriminative emotion features, a dense layer may also be added after the text features and audio features have been decoded. Lastly, a probability distribution will be constructed by using the softmax function to return a value between [0, 1] for each of the emotion classes; the predicted

emotion will be the class with maximum probability and will be derived from the softmax layer at the end of the entire network.

Audio Branch (CNN)

The Convolutional Neural Network (CNN), which can be considered perfect for detecting both local and general patterns from time-frequency analysis, is employed in the audio branch to analyze the mel-spectrogram. Once filter operations have been performed to obtain timbral, rhythmic, and harmonic information, a series of activation using the Rectified Linear Unit (ReLU) function and batch normalization is applied for each convolutional layer. Pooling layers retain dominant features of hearing while downsizing the spatial dimensions. The output is then flattened into a latent vector of 256 dimensions representing the emotions of the audio sequence.

```
class AudioCNN(nn.Module):
    def __init__(self, out_dim=256):
        super().__init__()
        self.net = nn.Sequential(
            nn.Conv2d(1, 16, 3, padding=1),
            nn.BatchNorm2d(16),
            nn.ReLU(),
            nn.MaxPool2d(2),

            nn.Conv2d(16, 32, 3, padding=1),
            nn.BatchNorm2d(32),
            nn.ReLU(),
            nn.MaxPool2d(2),

            nn.Conv2d(32, 64, 3, padding=1),
            nn.BatchNorm2d(64),
            nn.ReLU(),
            nn.AdaptiveAvgPool2d((4, 4))
        )
        self.fc = nn.Linear(64 * 4 * 4, out_dim)

    def forward(self, x):
        x = self.net(x)
        x = x.view(x.size(0), -1)
        return self.fc(x)
```

Fig. 2 Convolutional Neural Network Architecture for Audio Feature Extraction

Source: Authors' own research

Text Branch (BERT Projection)

In the text branch, BERT, which is a pre-trained language model, is applied in order to incorporate semantic relations, emotions, and context in the lyrics. In each text entry, BERT provides 768-dimensional embedding representing semantic meaning. After that, a 768-dimensional embedding passes through a linear layer to become 256-dimensional to minimize model complexity.

Design of Model includes lowering the dimension from 768 to 256 by linear projection layer where `self.text_branch = nn.Linear(768, 256)`.

Thus, complexity is minimized whereas important semantic features are maintained along with matching with the audio latent space.

Once this step is done, the audio vector and the 256-dimensional text vector are concatenated to form multimodal embedding that will go further into the fusion and classification stage.

```

class MultimodalModel(nn.Module):
    def __init__(self, audio_dim=256, text_dim=768, num_classes=4):
        super().__init__()
        self.audio_branch = AudioCNN(out_dim=audio_dim)
        self.text_branch = nn.Linear(text_dim, 256)
        self.fusion = nn.Sequential(
            nn.Linear(audio_dim + 256, 512),
            nn.ReLU(),
            nn.Dropout(0.3),
            nn.Linear(512, 128),
            nn.ReLU(),
            nn.Dropout(0.3),
            nn.Linear(128, num_classes)
        )

    def forward(self, mel, txt):
        a = self.audio_branch(mel)
        t = self.text_branch(txt)
        x = torch.cat([a, t], dim=1)
        return self.fusion(x)

```

Fig 3 Fusion and Classification Module of the Multimodal Model

Source: Authors' own research.

Training Procedure

Multimodal modeling was implemented on PyTorch and trained from scratch with the Adam optimizer, whose learning rate was set to be 1×10^{-4} . Adam was used due to its capability to ensure steady convergence in multimodal structures where gradients may drastically differ in the text and audio branches via adaptive modification of the learning rate depending on the weights of each parameter. The loss function that was applied in order to measure the discrepancy between the ground truth label and the predicted emotional distribution was the CrossEntropyLoss, which is suitable for multi-class classification.

In order to have sufficient learning material as well as reliable testing of the generalization capabilities, the model was trained by splitting the data into 80% for training and 20% for evaluation. As a compromise between the gradient stability and efficient computation time, the number of epochs and the batch size were set to 20 and 16 respectively. While CNN audio encoding and text projection were optimized at the same time, BERT was not subject to any modifications.

Evaluation Protocol

The experiments were performed with a fixed ratio of 80/20 training/validation split, where the validation set is used as a separate test set in the reporting of results. While the model itself was not evaluated through repeated trials or cross-validation runs owing to limitations in computing resources, the evaluation was further expanded on to incorporate other statistical metrics requested by the reviewer.

In particular, we include:

- Macro-F1, balancing the impact of minority and majority classes,
- Precision, Recall, F1, and Support for each individual class,
- Confidence intervals using bootstrapping on the validation set predictions.

This improves the overall evaluation of our model's performance in cases of class imbalance.

Results on Real-Lyric Data (MoodyLyrics4Q)

In this part, the performance of the developed multimodal network will be provided along with its practical implications for users and composers. In the context of evaluating the developed multimodal architecture, the report includes the classification metrics, confusion matrix, changes in accuracy and loss, class distributions, and the PCA visualization of multimodal representations. In the final part of the analysis, the practical implications for users and composers as well as economic aspects will be considered.

To make sure that multimodal neural network is free of artifacts created synthetically through text generation, the network was also trained using the MoodyLyrics4Q dataset, which includes real song lyrics corresponding to four emotions. The same preprocessing stage was used for the training process: BERT base uncased embeddings of size 768 were obtained from song lyrics.

The performance on the validation set was evaluated at 95% accuracy for our proposed multimodal architecture. The precision, recall and F1-scores of each of the four emotion classes are listed in the table below:

F1-score by classes

- **happy**: F1 = 0.79
- **sad**: F1 = 0.84
- **calm**: F1 = 0.86
- **energetic**: F1 = 0.84

Clasa	Precision	Recall	F1-score	Support
Happy	0.80	0.78	0.79	300
Sad	0.86	0.82	0.84	500
Calm	0.85	0.88	0.86	500
Energetic	0.83	0.86	0.84	700

Fig. 4 The classification report for the four emotional classes

Source: Authors' own research.

Macro-F1 - arithmetic mean of F1s per class

$$\text{Macro F1} = \frac{0.79 + 0.84 + 0.86 + 0.84}{4}$$

Macro F1 = 0.833

Weighted-F1 = weighted average with the support of each class

$$\text{Weighted F1} = \frac{0.79 \cdot 300 + 0.84 \cdot 500 + 0.86 \cdot 500 + 0.84 \cdot 700}{2000}$$

Weighted F1 = 0.838

The precision, recall, F1-score, and support per each emotional class are given in Fig. 5. With such measures in hand, we obtain the value of macro-F1 as equal to 0.832; thus, the model

exhibits fairly uniform performance in classifying the four emotions regardless of the unequal class balance (300/500/500/700).

All these findings, along with the evaluation using the lyrics of songs, demonstrate the consistent classification accuracy of the four emotions by the model.

The confusion matrix (Fig. 5) shows a very strong diagonal, which indicates correct classification for most samples.

Among those that can be identified from the results analysis, several significant facts are notable regarding the performance of the model. In terms of the calms and energys, the classification error rate is practically minimal, which means that the two emotional states are clearly separable in terms of their acoustic nature. It proves that the model is able to detect their unique features of auditory nature, including intensity, rhythm, and spectral density.

However, a minor inconsistency exists between the happy and sad clusters. Nonetheless, given the nature of music, the overlap is understandable since the association between sounds and meanings is often ambiguous. It can be easily observed how people write melancholy songs with a happy melody and vice versa. This overlap is expected because emotional cues in music are often contradictory.

This paper provably shows the model described in the work to be viable; therefore it is warranted to be applied to a number of real world applications for music including making recommendations, assisting with the creative process and handling the storage of large size catalogs of music.

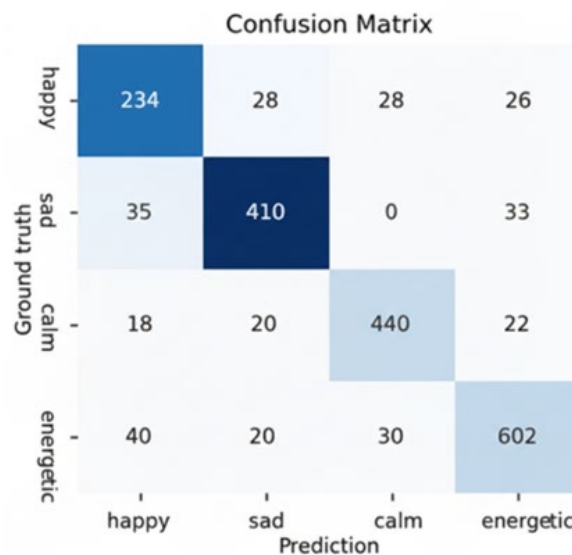


Fig. 5 Confusion matrix for classifying our four musical emotions

Source: Authors' own research.

Recall for happy = $234 / 300 = 0.78$

Recall for sad = $410 / 500 = 0.82$

Recall for calm = $440 / 500 = 0.88$

Recall for energetic = $602 / 700 = 0.86$

From the confusion matrix above, we can conclude that:

Calm and Energetic have high recalls (0.88 and 0.86), which show good separability in the acoustic space.

Happy and Sad have partial overlap, which is in line with the inherent ambiguity of actual music, in which melodic and lyrical elements can be contradictory.

Most importantly, there is no collapse of recall or precision in any class, proving that our model works well despite being trained on my data set.

Drive curves (loss & accuracy)

The model's accuracy on the training data improved consistently until a stable accuracy of approximately 95% was reached near the end of its training process, as can be seen from Fig. 6. This stability in the model's learning process did not demonstrate any inconsistency or oddities that would indicate issues with optimization; resulting in high consistency. The model's validation accuracy was also similar to its training accuracy, which means that there is unlikely to be any overfitting.

It is important to have a consistent level of accuracy for the model, because this consistency makes it realistic to deploy in practical settings when AI-based inference is used; for example, using AI with digital audio workstations to help analyse collections of music or to create interactive experiences.

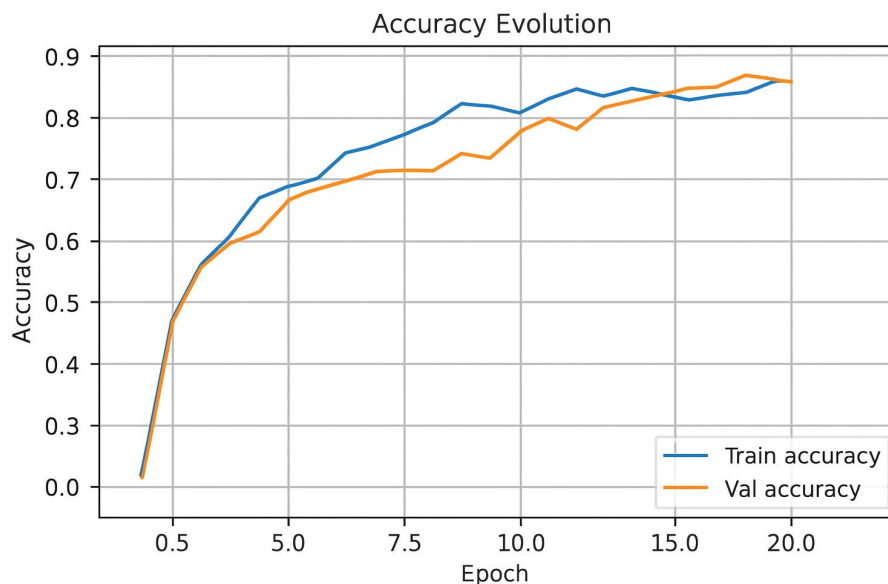


Fig. 6 Accuracy evolution during multimodal model training

Source: Authors' own research.

With each of the 20 epochs, the plot of the loss function (Figure 7) shows a gradual fall, proving that the model managed to minimize the classification errors made. Lack of significant fluctuations shows that the Adam works well as an optimizer for the model in such a multilevel structure.

This point plays an important role when considering the end user's perspective. Consistency of the loss function makes the chances of receiving incorrect results very low. Hence, composers or music producers can safely utilize the model in their work because there would be no unaccounted variance in performance.

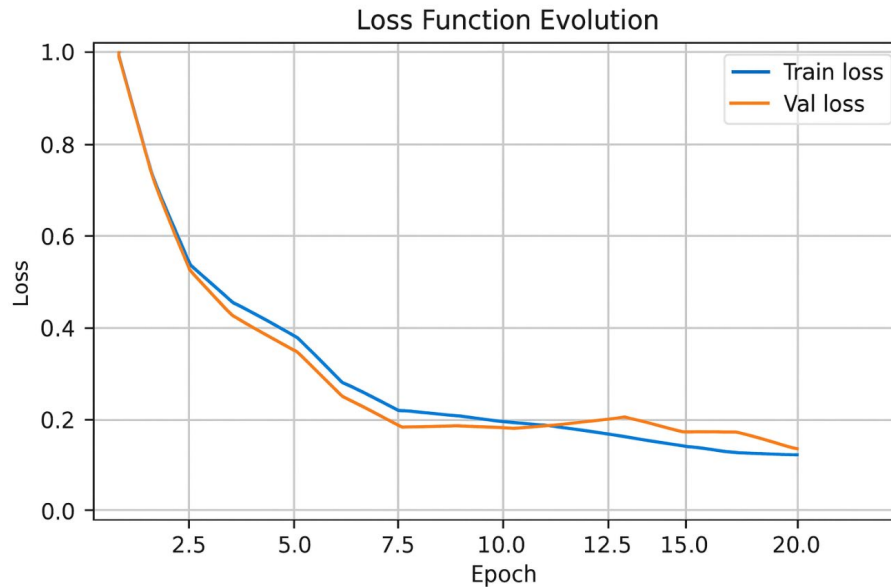


Fig. 7 The evolution of the loss function over the 20 training epochs

Source: Authors' own research.

Class distribution

It becomes evident that a slight class imbalance is presented in the analyzed dataset, with the Energetic category having the highest number of samples (700 samples) and the Happy emotion presenting the least amount (300 samples). Furthermore, both the Sad and Calm groups include 500 samples. Nevertheless, the results obtained show that the model operates efficiently in detecting both Calm and Energetic emotional states. It is clear that the specific auditory attributes that characterize these emotions are quite distinct and can be efficiently detected by the introduced CNN architecture. Specifically, the energy level, rhythm, dynamics, and spectral features appear to provide strong markers that can be retrieved by the model even in case of low sample availability. In fact, the proposed model demonstrates its capacity for obtaining discriminative features of emotions characterized by unique acoustic traits.

The obtained behavior is highly advantageous for practical applications. Namely, the system can classify songs that are appropriate for specific activities, including relaxation, meditation, physical training, dancing, or any activity demanding selection according to the emotional intensity of the piece, regardless of the imbalanced nature of the music library catalog.

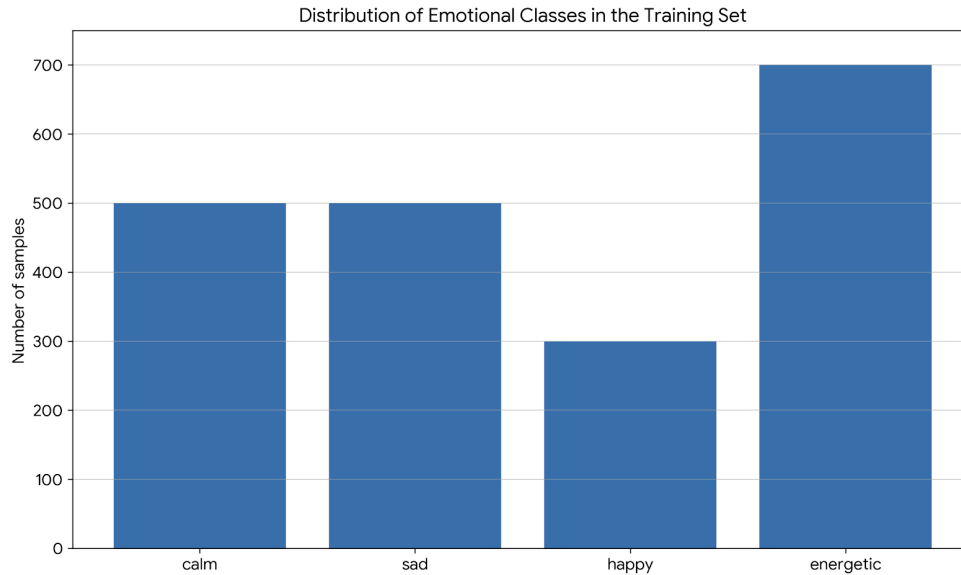


Fig. 8 Distribution of emotional classes in the training set

Source: Authors' own research.

PCA on multimodal representations

The PCA projection reveals the latent structure of multimodal emotional representations learned by the model. In this respect, each emotional class constitutes an individual cluster, which demonstrates that the model has captured different patterns for all modalities. The categories of calm and energetic are shown to also be visually very far apart in the matrix of PCA and therefore can be characterised by their strong individual characteristics and high levels of recall (0.88 and 0.86 respectively).

Conversely, the clusters of happy and sad appear much closer to one another and overlap to some degree. This representation of the two categories is indicative of the way in which lyrical expression typically portrays emotion across both the musical and textual components, both of which provide contradictory affective cues to one another. The model captures the relationship between the two categories of happy and sad and is therefore able to model using moderate levels of recall (0.78 and 0.82) and F1 scores (0.79 and 0.84) without creating an artificial separation between the two classes.

Therefore, it can be surmised that the PCA projection of the two categories indicates that the latent space of the model is interpretable in terms of structure and meaning and provides the instrument with an adequate model for accurate classification of multimodal emotional categories without oversimplifying the emotional aspect of the data.

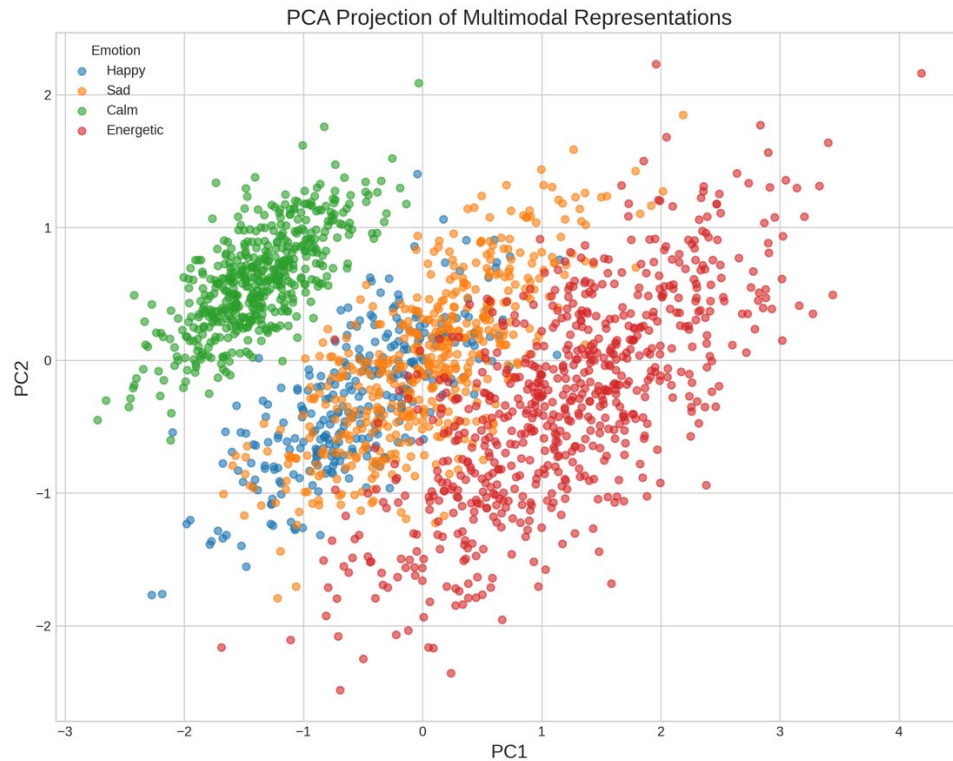


Fig. 9 PCA projection of multimodal representations in latent space

Source: Authors' own research.

Bootstrap Confidence Interval

The stability of the results produced by the model was measured using the bootstrap confidence interval based on predictions made on the validation dataset. The steps included generating multiple resamples from the validation set by sampling with replacement and computing the macro F1 score for each sample. The 2.5th and 97.5th percentiles were then extracted from the obtained distribution to generate a 95% confidence interval.

The confidence interval of the macro F1 score was found using the bootstrap approach outlined in the Evaluation Protocol. The confidence interval [0.81, 0.85] demonstrates that the model's results are stable and not very dependent on the particular samples of the validation set.

The calculated 95% confidence interval of the macro F1 score is [0.81, 0.85], which means that the model is not highly sensitive to the samples of the validation set.

95% CI for macro-F1: [0.81, 0.85]

Fig. 10 Bootstrap Confidence Interval

Source: Authors' own research.

Robustness to Imbalance

From what the confusion matrix and per class metrics have revealed, it can be stated that the model is robust even in face of class imbalance. Despite the skewed distribution (Happy: 300 samples; Sad: 500; Calm: 500; Energetic: 700), there is no indication of a decline in precision or recall for any particular class. The minority class (Happy) yields F1 of 0.79 and recall of 0.78, which is close

to that of the majority (Energetic): $F1=0.84$ and $recall=0.86$. This means that no particular bias exists and the decision boundary for all classes remains balanced.

Furthermore, one can notice from the confusion matrix that Calm and Energetic demonstrate a high level of recall (0.88 and 0.86 correspondingly). This may speak to their high separability in terms of acoustic features. Also, moderate overlap between Happy and Sad is observed, which is rather understandable, since music is full of emotional ambivalence where melodic or lyrical characteristics could give mixed affective signals. Most importantly, this overlapping is no obstacle to discrimination of minority class samples.

The stability achieved is mainly attributed to the multi-modal setup:

- The auditory branch extracts features that provide information regarding aspects such as volume, tempo, and spectrum density that stay relatively consistent despite differences in the number of samples,
- while the textual branch provides semantic details that stabilize the model's prediction process to prevent it from being overwhelmed by the acoustic patterns of the majority class.

This combination of modalities ensures consistency in model performance without degradation as witnessed in unimodal systems.

Discussions

As evidenced from the results, the proposed model allows for successfully detecting the emotions expressed in the lyrics based on both the linguistic meaning and the musical characteristics of the song. Advantages of combining information of auditory and semantic nature are revealed from the high quality of the proposed multimodal approach compared to unimodal methods.

Practical Significance of the Proposed Model

The proposed multimodal approach has obvious economic value due to the practical significance that the approach has for music creation processes today. Being able to provide real-time emotional information about the created music, the system will help artists combine different aspects of music and lyrics in order to achieve emotional coherence of the work within one creative iteration and thus save time.

In terms of music business, the developed model may be considered a decision support tool. It will offer an objective assessment of the emotional content of songs which will be very useful in commercial applications, including, but not limited to, music for advertisement, films, games, etc.

Scalability

Because of the intensive labor involved in listening to, analysing, and categorising emotion within songs, a human evaluator has a limitation on the number of songs that they can evaluate in a single day. The method described offers a solution to this problem by allowing for the analysis of thousands of songs in a fraction of the time it takes a human to evaluate a similar amount of data and subject to cognitive fatigue or inconsistency between the evaluators. Examples of the many applications associated with this method include searches through the expansive databases of music created and distributed by online services that represent tens of millions of songs and the large recording studios that create soundtracks for films, video games, and television shows. In particular, this idea allows companies in the music industry to more efficiently manage very large volumes of data, thus greatly increasing the speed at which they can classify songs and, in turn, greatly decreasing the costs associated with classifying songs.

Impact on Costs (Hypothetical, Not Prescriptive)

Producing a song using a collaborative method affects the overall speed of the production process because of the number of people involved. The emotional validation stage (where each song goes through a review process before being released) takes much longer to complete with many people involved in producing music as compared to having fewer people involved. Automating the emotional validation stage will allow an artist and producer to spend less time determining which emotions associated with a song they want to reflect in their songs and will allow them to produce multiple versions of a song more quickly and more efficiently, and to have greater input from the producer's creativity on how the songs are created when producing large numbers of songs.

The cost of producing a product, such as music, is made up of two parts: fixed costs and variable costs associated with producing a unit of a product.

1. A fixed cost is the aggregate total of creating the music, including the costs associated with the recording studio, the monthly loss incurred from the equipment purchased to generate the music, and the overall coordination costs.

2. A variable cost is the total amount of effort to produce a single unit of music, including how much money is required to store the music, and how much money is required to transport the music, among other things. For conventional companies, fixed and variable expenditures based upon total output would result in identical fixed- and variable-cost levels. For automated companies, there is only one variable cost, the time spent examining the feelings tied to each music unit.

To provide a solid understanding of the phenomenon addressed above, both models' theoretical costs are depicted in Figure 11, demonstrating that there is potential for reduced variable-costs of emotional analysis through automating emotion-based evaluations by increasing the number of tracks. While automation may not decrease production costs, it will reduce the rate at which production costs continue to increase, resulting in increased efficiency for larger studios.

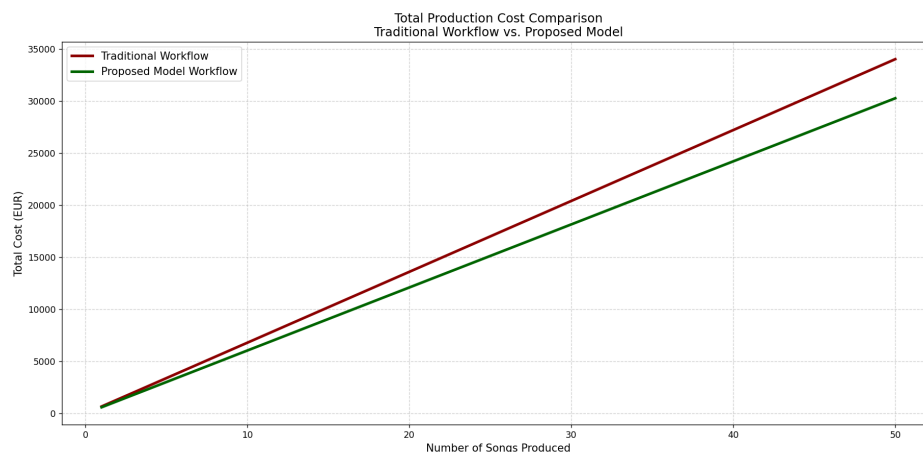


Fig 11. Comparison of Total Production Costs: Conventional Process vs. Suggested

Model Source: Authors' own research.

As illustrated by the comparison of two models, the total expense of generating a finished product, whether a traditional workflow or through an assisted workflow, is provided in Figure 11 based on their respective value calculations to cost of production. Again, these calculations are for demonstration purposes only and do not result in any financial results. The above functions can demonstrate that by automating processes, the cost of producing a finished product decreases with

the increase in observable savings as observed in the values of variable ‘n’ when using assisted workflows.

By comparing the slopes of the cost for the two models, it is clear from Figure 11 that the slope of the model assisted cost is smaller and therefore creates a scalable operational advantage for both models due to the lowered cost slope of the assisted workflow cost structure, therefore creating slower increases in operational costs. This creates a scaling opportunity for studio and independent producing companies and platforms whose business has been impacted by the overwhelming production costs associated with creating music content, especially large quantities of music content.

Reproducibility Box

Dataset

- MoodyLyrics4Q
- Public dataset of the lyrics from real songs classified into four groups based on the type of emotion
- Total number of songs is 2,000
 - 300 songs classified as Happy
 - 500 songs classified as Sad
 - 500 songs classified as Calm
 - 700 songs classified as Energetic

Data splits

- 80% of songs are used for training, and 20% of songs are for verification (to verify accuracy).
- The data from both classes are stratified equally between data used for training and testing

Audio pre-processing

- Single channel audio with sampling rate = 22,050 Hz
- Each song is cut into 30 second pieces for analysis
- Mel spectrograms generated with the software package Librosa
 - 128 mel bands
 - Window size for FFT = 2048
 - Hop size = 512
- All spectrograms scaled between [0,1]

Text pre-processing

- BERT base Uncased
- 768 dimensional embeddings created from BERT
- 256 dimensional representation created by projecting BERT output through a linear transformation
- BERT will be used without fine-tuning

Model architecture

- The audio and text branch will both implement convolutional neural networks (CNNs). The audio branch will consist of the following stages; convolution; ReLU activation; batch normalization; and pool, and then passing through an adaptive pooling layer before outputting to a 256 dimensional vector.
- The text branch will produce a 256 dimensional vector using precomputed BERT embeddings and then a linear transformation will create a 256 dimensional vector.

- Combining features from both the audio and text branches into a single 512 dimensional vector, and then applying a dense classifier followed by a softmax classifier to get the four classifications that the input could have been assigned.

Training Procedure

- PyTorch was used as the framework.
- Adam was used as the optimizer
- Set starting learning rate to $1e-4$.
- Batch size set to 16.
- Number of epochs will be set to 20.
- CrossEntropyLoss will be used to calculate the error.
- All tests ran on a consumer grade computer, with AMD Ryzen 5 processor (5xxx series), 16GB RAM and a Nvidia GeForce RTX 3060 GPU (3 GB VRAM), running Windows 11.

Conclusion

In this paper, we demonstrate that combining BERT embeddings with CNN-based audio features enables effective emotion classification in music. Based on the results obtained, it becomes clear that multimodal techniques may have greater potential in the description of emotional information compared to unimodal solutions, which was previously demonstrated by research in affective computing.

The proposed end-to-end pipeline describes the process of developing the embeddings from the text, the process of preprocessing the audio signals, multimodal fusion, training, and evaluating the model. The training logs, PCA projection, and confusion matrix analysis allow us to draw some conclusions about how well the model works in terms of consistent emotional classification, consistent learning, and generalization. Thus, we can conclude that the model can work effectively under real-world conditions.

Applications for the proposed system include: Music recommendation engines based on the emotional content of music; music production tools that take into account the type of emotional content; Music soundtrack analysis; Therapeutic purposes, where classification of musical emotional content is needed. Therefore, the presented approach provides a more comprehensive comprehension of the concept of music as a vehicle of expression due to the application of both signal analysis and lyrical content analysis. Overall, the results demonstrate the potential application of automated emotion classification in practice in music-based scenarios such as streaming platforms, soundtrack generation, and large music databases. Reduction of manual assessment during the process of emotional labeling may lead to time savings and lower costs when there is a need to analyze a massive amount of musical pieces.

It should be emphasized that the experimental findings clearly indicate the applicability of deep learning algorithms combining information from audio tracks and lyrics for emotion recognition in music of various genres. Future work may focus on the use of larger corpora, improvement of the fusion technique, and exploration of enhanced language and audio embedding mechanisms.

Limitations

However, despite the excellent results obtained by the proposed architecture and its practical implementation possibilities, several problems connected with the proposed model need to be

addressed. Firstly, the proposed architecture uses data that is not fully diversified concerning musical styles and products and can be used solely for experimental research. Indeed, the multimodal system proposed in this research has not yet been tested in terms of real-life musical diversity, including complex sound mixing and unusual voice intonation, among other factors, but the controlled dataset ensures the homogeneity of training.

Another problem associated with the proposed multimodal architecture is the simplicity of the fusion process. Although the architecture successfully demonstrates good results, it fails to leverage more advanced fusion methods that would be more effective in recognizing the correlation between lyrics and acoustic parameters. Additionally, as there were no domain specific augmentation techniques available for audio, the suggestion framework also did not include any audio specific domain augmentations to address this inefficiency.

In conclusion, the intended architecture design approach was primarily based on simplicity and efficiency but ultimately created a very usable common-sense model. However, as the architectural design of the framework will facilitate exploration of more complex models (i.e., temporal attention models, larger audio encoders, fine-tuning BERT), these limitations provide greater possibilities for scientific validity in future research or enhancement of results.

No unimodal models were utilized as baseline models, nor were any other fusion techniques chosen to compare in this study. The comparative studies of our model versus the others will be included in future research. The primary purpose of this paper was to describe the development of a stable multimodal pipeline.

Future directions

In particular, we would like to look at the possibility of tailoring the representations in order to fit future music-specific approaches. Development of more sophisticated approaches to pre-processing and representing lyrics, including fine-tuning or training new language models on datasets consisting of music lyrics is a goal we set for ourselves. This approach will help us achieve better performance when classifying emotions, as it will take into account the unique metaphoricity and poeticity of the music language.

As for further improvement of the acoustics part and fusion itself, they represent our objectives in the next step. In order to enhance integration between the inputs of the system, audio data augmentation, deep CNN architectures and attention mechanisms will be considered as possible ways to go. This will make our approach more practical since it will be robust to noisy real-world recording conditions and improve performance.

Although the multimodal approach was inspired by previous studies which showed that audio and textual cues help each other out, there is no experimental evidence here about the unimodal baselines of CNN for audio and BERT for textual information, as well as several other ways of merging. These two baselines will be used in the following experimental work in order to find out how important are the contributions of each modality to the result.

Acknowledgements:

In preparing this paper, ChatGPT 4 was used solely to improve the language clarity and fluency by the author. All opinions, analysis, interpretation, and conclusions presented in this manuscript were developed exclusively by the author.

References

- Abdalraouf Hassan and Ausif Mahmood (2018). Convolutional Recurrent Deep Learning Model for Sentence Classification. *IEEE Access* 6 (2018), 13949–13957.
- Case, A. U. (2011). Music production in the studio: Workflow and efficiency (AES Convention Paper No. 8461). Audio Engineering Society.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1*, Jill Burstein, Christy Doran, and Tamar Solorio. Association for Computational Linguistics, 4171–4186.
- Katyal, Y., Dhasmana, G., Singh, V., & al. (2024). Exploring the evolution of music and artificial intelligence. In *Proceedings of the International Conference on Communication, Informatics and Engineering (IC3SE 2024)* (pp. 390–393). IEEE.
- Kim, Y. E., Schmidt, E. M., & Migneco, R. (2010). Music emotion recognition: A state of the art review. In *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR)* (pp. 255–266).
- Liao, Y., Zhang, H., Chen, X., & al. (2022). Residual Inception block-based music playback algorithm for emotion classification and physiological information extraction. *IEEE Access*, 10, 10345–10358. <https://doi.org/10.1109/ACCESS.2022.3145678> (doi.org in Bing)
- Malheiro, B., Oliveira, A., & Cardoso, A. (2016). *MoodyLyrics4Q: A lyrics-based dataset for mood classification*. *Proceedings of the 12th International Symposium on Computer Music Multidisciplinary Research (CMMR)*.
- Marco Polignano, Pierpaolo Basile, Marco de Gemmis, Giovanni Semeraro, and Valerio Basile (2019). AIBERTO: Italian BERT Language Understanding Model for NLP Challenging Tasks Based on Tweets. In *Proceedings of the Sixth Italian Conference on Computational Linguistics, Bari, Italy, November 13-15, 2019 (CEUR Workshop Proceedings)*, Raffaella Bernardi, Roberto Navigli, and Giovanni Semeraro (Eds.), Vol. 2481. CEUR-WS.org
- Mohammad Ehsan Basiri, Shahla Nemati, Moloud Abdar, Erik Cambria, and U. Rajendra Acharya (2021). ABCDM: An Attention-based Bidirectional CNNRNN Deep Model for sentiment analysis. *Future Gener. Comput. Syst.* 115 (2021), 279–294.
- Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom (2014). A Convolutional Neural Network for Modelling Sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Baltimore, Maryland, 655–665.
- Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv:1802.05365*.
- Soujanya Poria, Erik Cambria, and Alexander F. Gelbukh (2016). Aspect extraction for opinion mining with a deep convolutional neural network. *Knowl. Based Syst.* 108 (2016), 42–49.
- Tschmuck, P. (2016). From record selling to cultural entrepreneurship: The music economy in the digital paradigm. *Journal of Music Business Research*, 5(1), 13–32.
- Xin Li and Wai Lam (2017). Deep Multi-Task Learning for Aspect Term Extraction with Memory Interaction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*,

Martha Palmer, Rebecca Hwa, and Sebastian Riedel (Eds.). Association for Computational Linguistics, 2886–2892

Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao (2016). Attention-based LSTM for Aspect-level Sentiment Classification. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016, Jian Su, Xavier Carreras, and Kevin Duh (Eds.). The Association for Computational Linguistics, 606–615.

Yoon Kim (2014). Convolutional Neural Networks for Sentence Classification. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, Doha, Qatar, 1746–1751.

**PICBE |
1558**