Bulletin of Electrical Engineering and Informatics
Vol. 12, No. 1, February 2023, pp. 355~364
ISSN: 2302-9285, DOI: 10.11591/eei.v12i1.4231

Multimodal music emotion recognition in Indonesian songs based on CNN-LSTM, XLNet transformers

Andrew Steven Sams, Amalia Zahra
Department of Computer Science, BINUS Graduate Program, Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia

Article Info

Article history:

Received Jun 11, 2022
Revised Sep 6, 2022
Accepted Oct 6, 2022

Keywords:

CNN-LSTM
Indonesian song dataset
Mel spectrogram
Multimodal
Music emotion recognition
Stacking ensemble method
XLNet transformers

ABSTRACT

Music carries emotional information and allows the listener to feel the emotions contained in the music. This study proposes a multimodal music emotion recognition (MER) system using Indonesian song and lyrics data. In the proposed multimodal system, the audio data will use the mel spectrogram feature, and the lyrics feature will be extracted by going through the tokenizing process from XLNet. Convolutional long short term memory network (CNN-LSTM) performs the audio classification task, while XLNet transformers performs the lyrics classification task. The outputs of the two classification tasks are probability weight and actual prediction with the value of positive, neutral, and negative emotions, which are then combined using the stacking ensemble method. The combined output will be trained into an artificial neural network (ANN) model to get the best probability weight output. The multimodal system achieves the best performance with an accuracy of 80.56%. The results showed that the multimodal method of recognizing musical emotions gave better performance than the single modal method. In addition, hyperparameter tuning can affect the performance of multimodal systems.

Corresponding Author:

Andrew Steven Sams
Department Computer Science, BINUS Graduate Program, Master of Computer Science
Bina Nusantara University
West Jakarta, Indonesia
Email: andrew.sams@binus.ac.id

1. INTRODUCTION

Music is a cultural expression that has gone through a long history and is used to express a feeling or event. In today's era, music has become a lifestyle that can be listened to anywhere and anytime or a helpful tool to reduce stress after a long day at work. In 2021, the world still suffered from the pandemic COVID-19. Peoples prefer to do their activities at home and work at home to decrease the spread of the pandemic. These activities boost the music industry because people tend to listen to music to kill their boredom while doing their activities at home. Based on statistical data released by business of apps shows how Spotify performed during the pandemic and how pandemic COVID-19 may affect the music industry. The Spotify perfomance will be divide to two figures, Figure 1(a) shows Spotify's profit during the COVID-19 pandemic era, and Figure 1(b) shows the people that start streamed music on Spotify.

Figure 1(a) shows Spotify's profit rose around 16%, from 2.5 billion to 2.7 billion revenues during the COVID-19 pandemic era, on Figure 1(b) shows up to 400 million people start streamed music on Spotify, which shows the Spotify peak performance in 2021. The crisis may accelerate underlying trends in the music industry. This fact shows the need to stream music, which has grown from 9% to 47% in just six years. Music has various genres, including jazz, rock, pop, blues, RnB, metal, classic, country, hip-hop, and many

more. The exciting thing about music is that it can carry emotions and let the listeners feel the same emotions, affecting their mood [1], involving 2,400 people who stated, “even music with sad melodies can calm someone's feelings”. Also, this study found that people tend to prefer listening to sad songs when they feel a painful event such as a breakup or are grieving the loss of a loved one. Songs can express our inner feelings, produce goosebumps, bring us to tears, share an emotional state with a composer or performer, or trigger specific memories. Interest in a deeper understanding of the relationship between music and emotion has motivated researchers from various areas of knowledge for decades [2]. This proves that music can affect or improve a person's mood when given to people in the right mood. Music contains much human emotional information. Also, Panda *et al.* [3] gave an example of music containing multiple emotional features. The emotional information can be extracted and can help identify basic emotions contained in the song through the study of music emotion recognition (MER) [4]. Kumar and Saxena [5], they study the behavior of human with bipolar probability which also count the survivability of the subject, MER study also happen to study the behavior of human that listen to the music or the creator of the music by studying the emotion that consists off the music.

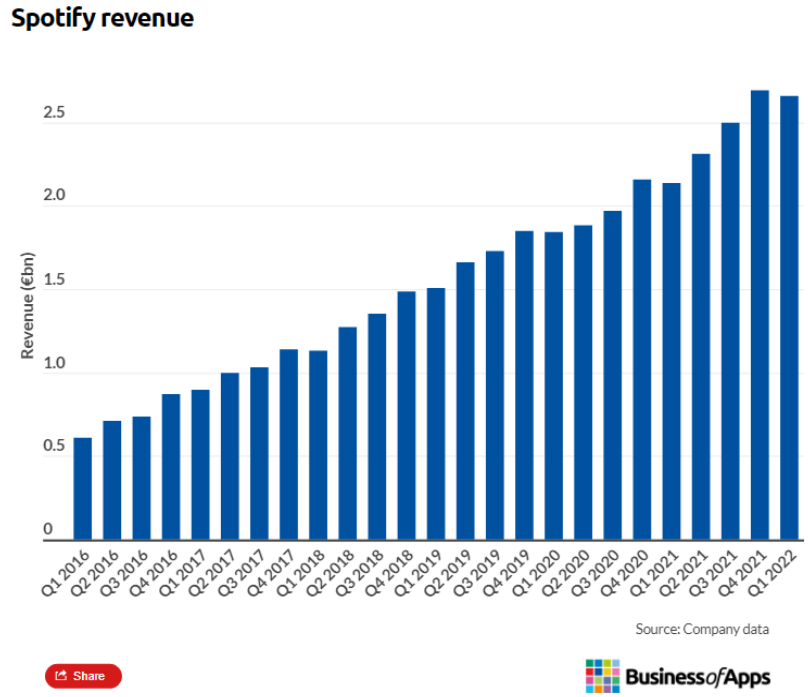
The study of MER has recently gained attention, and several studies have served as the foundation for this study. The study about lyrics-based MER achieved 94.7% accuracy using a deep neural network (DNN) and the XLNet transformer model [6]. While [7] achieved 100% accuracy using convolutional long-short term memory deep neural network (CLDNN). The architecture consists of convolutional neural network (CNN), long short term memory (LSTM), and DNN layers. With audio modal [8], using architecture stacked CNN, bidirectional gated recurrent unit (BiGRU) achieved root mean square error (RMSE) of 0,01. A multimodal study by [9] uses CNN-LSTM to process 2D features, DNN to process 1D features, the stacking method to ensemble results, and artificial neural network ANN as the meta classifiers to achieve 78.2% average accuracy. With different methods [10], using LSTM on audio, bidirectional encoder representations from transformers (BERT) transformer on lyrics, and using late fusion subtask merging (LFSM) method achieve 79.62%. The problem common in the MER studies is the limitation to achieving higher accuracy in detecting emotion just by using one music information variable (like audio, lyrics, or video) [11]. Through the good results of MER, researchers try to achieve more significant results by combining two or more datasets called the multimodal method.

The proposed method CNN-LSTM uses the same architecture as the simple LSTM as there are LSTM layers and dropout layers respectively throughout the entire architecture [12]. Talafha *et al.* [13] used the LSTM layer as a decoder on RNN architecture, the proposed method uses the CNN layer and CNN-LSTM layer as a decoder on LSTM architecture [14], [15]. After each CNN layer, [16] uses the max pool layer to reduce the overfitting/underfitting. With the same goal, the proposed method CNN-LSTM uses a dropout layer after each CNN layer. The order of the layer in the proposed method are CNN layer as the input layer and forget layer on LSTM architecture also, the CNN layer is the state of art in decoding images and other 2 dimensions or more data and is considered capable to decode the Mel spectrogram data used in this paper. The second layer is the CNN-LSTM layer or known as the ConvLSTM2D layer in the keras library, the CNN-LSTM layer is used as the output layer on the LSTM architecture. The last layer is the dropout layer function to reduce overfitting or underfitting.

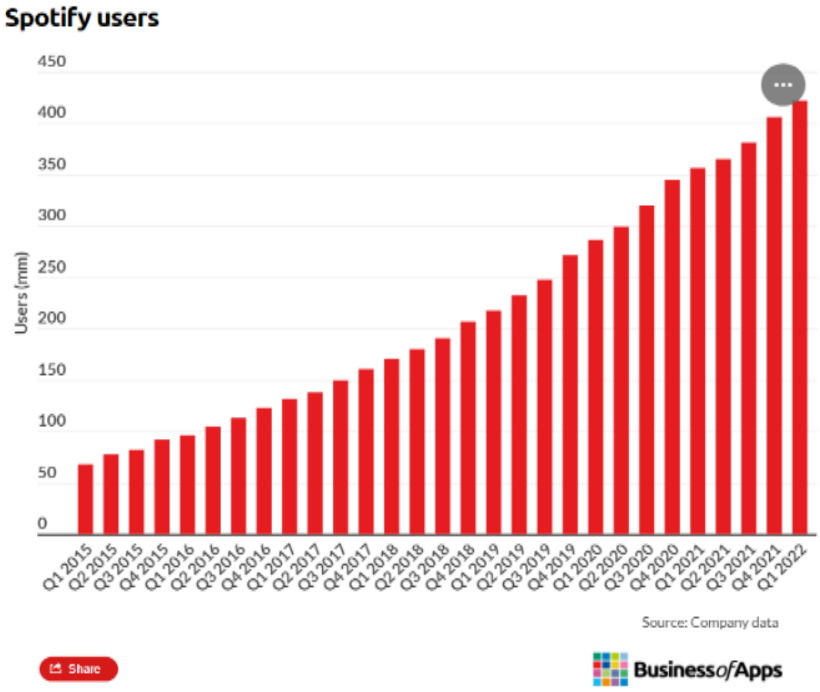
After researching journals and papers about MER, this paper was inspired by the [10] method, which combines the deep learning model for audio and transformers for the lyrics. The model used is the best results study on audio MER and lyrics MER. Hence, the best result model, CNN-LSTM, which is also used as the audio classifier on [9], [14], [15], and the XLNet transformers [6], is expected to achieve a better result. The stacking method for ensembled the results, and simple ANN as the meta-classifier was used for this experiment. This paper proposes a multimodal method for audio and lyrics data. The song dataset consists of 476 audio and lyrics from Indonesian pop songs. The dataset will be grouped by positive, neutral, and negative. The other student will share and annotate each song with Google form. The output emotion such as audio_positive, audio_negative, audio_neutral, text_positive, text_negative, text_neutral. There are two main contributions this paper can propose such as:

Combine both lyrics and audio features from the Indonesian song dataset to fit the lack of classification performance. We still see the potential in this experiment because the previous study still uses only one feature. Combining these two features into one or multimodal method will create opportunities for this experiment to achieve higher results. The model used in this experiment is CNN-LSTM for the audio classifier, XLNet transformers for the lyric classifier, and ANN for the meta-classifier. Lastly, to solve the problem of two heterogeneous outputs [7], we proposed the stacking ensemble method, which fuses emotion output from different modalities by averaging both outputs and feeding them to the SoftMax layer. Besides solving the heterogeneity problems, the fusion method significantly improves the accuracy of this multimodal MER research by achieving both audio and lyrics musical information.

In summary, this paper proposes the solution of achieving higher results by using the multimodal method, the modal used in this paper are audio and lyrics. Different with the model that will be used in this paper is the best result achieved model from the literature. CNN-LSTM will be used as the audio classifier and XLNet transformers as the lyric classifier. To ensemble both classifiers, a neural network is used in this paper as the meta-classifier. The flow of this experiment will be shown in chapter 2.



(a)



(b)

Figure 1. Spotify performance report for the 4th quarter of 2021 (a) Spotify profit has risen to approximately 2.7 billion and (b) 400 million users have used Spotify

2. METHOD

This section discussed the proposed CNN-LSTM model as the audio classifier, XLNet transformers as the text classifier, and the ensemble method. For this experiment, the total audio dataset was 476 Indonesian songs in .wav format, while the lyric dataset was 476 Indonesian song lyrics. Both datasets have been annotated by crowdsourcing method like [17], [18] and divided into positive, neutral, and negative subgroups. Then the audio data got trimmed and cleaned from the blank sounds, while the text got cleaned from special characters, punctuation, and got lowercase. After the data got cleaned, the flow of audio MER includes the data extracted for the Mel spectrogram features. The input data is split into 50% training data, 25% testing data, and 25% validation data. Then input the audio data to train and validate the CNN-LSTM model, after the train is done the model is tested with the rest of the input data. The output is the matrix of emotion labels with the size of 476×3 and saved with the pickle library for the next process. While the text MER flow includes inputting the text to test the pre-trained model XLNet transformers which has been pre-trained with the Indonesian language. The output is the matrix of emotion labels with the size of 476×3 and saved with the pickle library for the next process.

The fusion method includes importing the text MER output and audio MER output from the pickle library. Next, concatenate both output results to a matrix 476×6. After concatenating the data, the data got split into 70:30 training and testing data. Then input the 333×6 training data matrix to the ANN networks to train the model. After training the model, the model gets tested with the rest of the data. The test result is a 143×6 matrix that shows the result for each label, the last thing to sum up the results is to choose the best result for each row of the test matrix, so the result can be concluded. The final result is the concluded emotion label for each song, which results in a 143×1 matrix. Figure 2 shows the flowchart of how the proposed multimodal system works. Figure 2 explains that each data will be trained in a separate classifier and concatenated afterward. Then both outputs will be trained and predicted with the meta classifier, and the highest probability value will be picked as the final results.

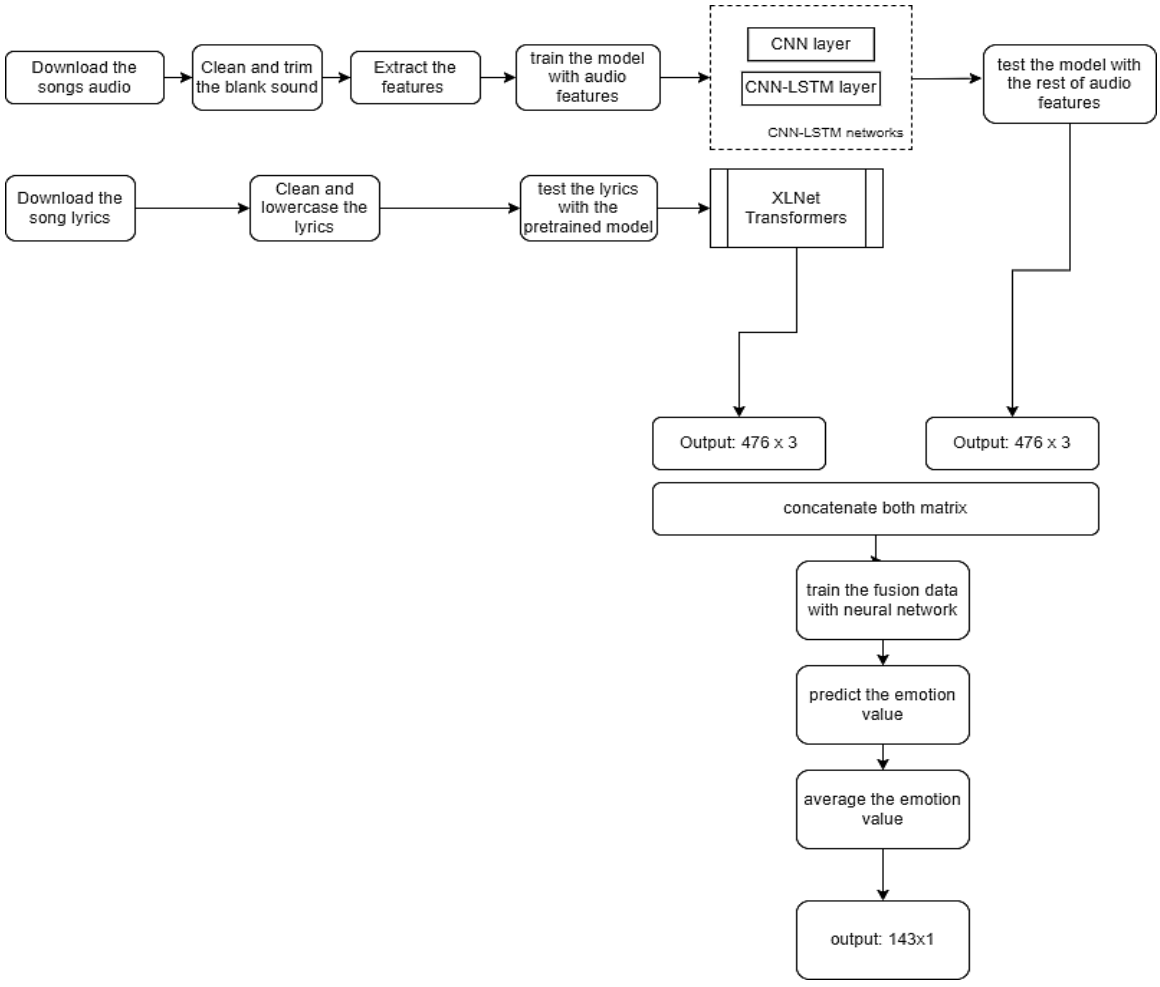


Figure 2. The multimodal system flowchart

2.1. CNN-LSTM

CNN-LSTM is a deep learning model that combines the CNN architecture with the LSTM layer and is designed to solve sequential prediction problems using spatial input such as images or videos [19], [20]. The LSTM layer could learn the temporal information in the feature [12], while CNN is useful for analyzing image data. The combination of CNN and LSTM was used to analyze images with time-domain information (e.g., songs, and videos) [21]. The Mel-spectrogram is one of the audio features that turns time domain information into a Mel-scale with the image's shape. The CNN-LSTM model is suitable for this experiment because the feature that will be used on audio data is Mel-spectrogram, an image. The CNN-LSTM architecture is illustrated in Figure 3. The architecture is made by stacking the convolution and CNN-LSTM layers, respectively.

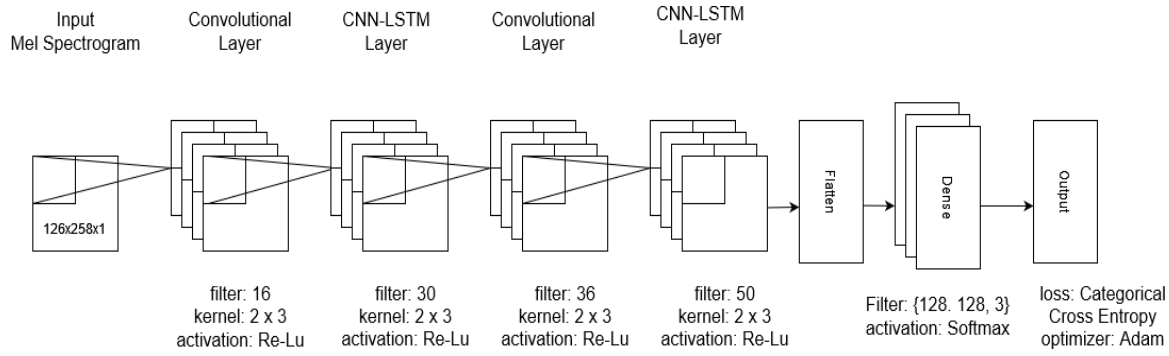


Figure 2. The architecture of the proposed CNN-LSTM

The architecture consists of a convolutional layer, a CNN-LSTM layer, 3 dense layers, and the output layer. A 2×3 kernel was used for the convolutional and CNN-LSTM layers. We used rectified linear unit (ReLU) for the convolutional, CNN-LSTM layers and the dense layer for the activation function. For the output layer, SoftMax was used as the activation function. The Adam optimizer and the categorical cross-entropy loss function were also used. The network was implemented using the Keras framework.

To conduct the audio-based MER experiment, we need to download the dataset from Google drive using the Pydrive library. Then, we need to transform the audio signal into the Mel-spectrogram using the Librosa library. Several studies also use Mel-spectrogram and achieved great results [4], [10], [22], [23]. After all the song data was turned into a spectrogram, all the data was grouped into positive, neutral, and negative and saved to a folder then uploaded to Google drive. Next, split the dataset into 60% training, 20% test, and 20% validation data. The next thing is building the model by using the Keras library. The model was built by stacking the CNN layer (with 8 and 32 filters) and CNN-LSTM layer (with 16 and 64 filters) sequentially. All the layers used ReLU activation and dropout layer 0.8 dropout layer help reduce the overfitting. It was then stacked with 128 and 92 filters dense layer as the output layer. The output layer used SoftMax activation. An early stopping monitor and model checkpoint was used to save the best result to save and stop the model if it achieves the best result. The parameter that used: optimizer=Adam, loss function=categorical cross-entropy, batch size=4, epoch=11. After achieving the best result from the training process, the model predicts the test dataset that was prepared early, consisting of a 109 Mel-spectrogram. The output of model prediction is the probability value and the true prediction value. The next step is saving the result with the pickle library for later late fusion.

2.2. XLNet transformers

XLNet transformers is an auto-regressive language model based on transformer XL and outputs the token's joint probability [24]. The probability is calculated with the permutation of word tokens in a sentence. This is different from BERT, which uses [mask] on a random token as the learning method. XLNet predicts the word sequentially by calculating all possible permutations of each token. This way allows the model to get more textual information. The XLNet is advanced in any NLP task because these transformers use the two-stream attention system [25].

There are several steps for the text-based MER, but before that, the lyrics must be preprocessed. The preprocess such: clear the symbol or special characters, inlining the word, lowercase the capital word, and annotating the lyrics with positive, neutral, and negative. The next thing to do is list the artist, title, annotation, and lyrics in the comma separated values (CSV) file [26]. Afterward, the CSV files were

uploaded to Google drive to ease the experiments. Next, download the CSV files from Google drive using the Pydrive library and read the data using the Pandas library. The feature that will be used is the emotion label (e.g., positive, neutral, and negative). Then, import the XLNet transformer tokenizer and model using the transformer library. The imported XLNet model is from hugging face, titled 'malay-huggingface/xlnet-base-bahasa-cased. Next, feed the lyric to the tokenizer, so it returns the encodings, attention mask, and input id. After that, split the data to train, validate, and test data with the ratio of 50:25:25. The next thing to do was create the training loop, resize the input shape, and save the model if it achieved the best accuracy. The hyperparameter used in experiment such as: batch size=4, epochs=3, optimizer adamW, learning rate=$3e^{-5}$. After achieving the training output, feed the test data to the model for the prediction task. The prediction task should be finished in about 10 minutes, and the output of the prediction is a probability and the true prediction value. After achieving the great result on prediction, the model prediction and the probability were saved with the pickle library for multimodal fusion usage.

2.3. Stacking ensemble method

Stacking is one of the ensemble method techniques. These techniques aim to apply the output of a multiclass model to generate a new model, this method is seen in these studies [27]–[32]. The ensemble method acts as a meta or sub-classifier for the basic classifier's output, where the basic classifier's output has been trained for different tasks or features [33], [34]. The stacking method has achieved excellent performance for image and text classification tasks where these statements also fulfill their task [9]. Also, the stacking ensemble method is stable and accurate. The model program is simple to implement, and there is no need to adjust the previously constructed single modal classification model. This paper used the SoftMax layer as the meta-classifier for the experiment. According to Shinohara *et al.* [35] study results, LSTM and recurrent neural network (RNN) architectures were functional model dependencies on a short or few-second scale. They also used SoftMax in one of the hidden layers as an activation function. In conclusion, the basic neuron network layer with SoftMax activation on the output is suitable as the meta-classifier for our multiclass classifier network. A multimodal stacking ensemble model, as shown in Figure 4.

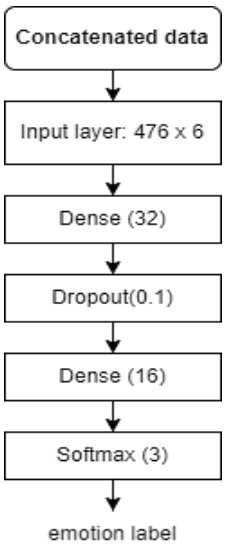


Figure 3. The architecture for the meta-learner

The audio and text classifier output will be fused by concatenating both outputs into one input vector. The fusion vector will be used as input for the meta-classifier using the neuron networks model. The first step of the stacking method is to download the audio and lyrics output locally by using the pickle library. Next, concatenate the probability and true prediction value into one vector. This vector will act as the input for the meta-classifier, which uses the neural network. The neural network uses two dense layers and the output layer, which uses SoftMax as the activation function. Next, create the training loop and feed the fused data to the model for the training process. Lastly, predict the data and measure the performance. The output is varied from 0 to 1 for each label. To achieves the linear output, the highest value of the output will be present as the final output.

3. RESULTS AND DISCUSSION

For this research experiment, Google Collab is used as a development environment and is run using the hosted server. The experiment would be divided into the audio-based MER, text-based MER, and fusion method. The hyperparameter is tuned to achieve the best result. The batch size, epoch, and learning rate are the hyperparameter used for tuning the text classifier. The batch size, epoch, and the number of channels used for tuning the audio classifier. Lastly, the meta classifier used the number of channels. Tables 1-3 summarize the result of hyperparameter tuning on the audio classifier, text classifier, and fusion method.

Table 1. The result of hyperparameter tuning on text-based MER

No.	Batch size	Learning rate	Validation accuracy (%)
1.	8	$3e^{-5}$	57.88
2.	8	$3e^{-4}$	52.38
3.	8	$3e^{-3}$	44.76
4.	8	$3e^{-2}$	39.64
5.	4	$3e^{-5}$	58.39
6.	4	$3e^{-4}$	50.17
7.	4	$3e^{-3}$	43.92
8.	4	$3e^{-2}$	51.07

Table 2. The results of hyperparameter tuning on audio-based MER

No.	Filter	Batch size	Validation accuracy (%)
1.	8,16,32,64	1	55.56
2.	16,32,64,128	1	48.15
3.	16,32,32,64	1	66.67
4.	8,16,16,32	1	55.56
5.	8,16,32,64	4	51.85
6.	16,32,64,128	4	55.56
7.	16,32,32,64	4	44.4
8.	8,16,16,32	4	44.4

Table 3. The result of hyperparameter tuning on fusion method

No.	Filter	Validation accuracy (%)
1.	16,8	78
2.	32,16	80.56
3.	32,64	79.66
4.	64,128	75.2

In Table 1, the text-based MER experiment using the XLNet model with hyperparameter settings: batch size=4, epoch=3, and learning rate=$3e^{-5}$ reached 58.88%. With hyperparameter tuning, the XLNet model increased accuracy by about 10%, increasing accuracy from 39.64% to 58.8%. In Table 2, experiments on the CNN-LSTM model using batch size=4, epoch=15, and the number of channels 8, 16, 16, and 32 settings achieved the highest results with an accuracy of 66.67%. The results of hyperparameter tuning on CNN-LSTM and XLNet transformers increase, although the increase in accuracy is not much because the dataset is still relatively small. The tuning hyperparameter is also applied to the ANN model. The hyperparameter used in the meta-classifier is the number of channels. The number of channels was chosen because this hyperparameter is quite easy to adjust, and the tuning results are immediately visible. In Table 3, the best result of the hyperparameter tuning ensemble method is 80.56%, with channel numbers 32 and 16. The order of layers used in the neural network model is the dense filter, dropout layer, dense filter, and output layer. The optimizer used is stochastic gradient descent (SGD).

The conclusion from the hyperparameter tuning results, the ensemble method significantly increases results with hyperparameter tuning from the audio model and the text model. The highest result achieved a validation accuracy of 82.40%, which is classified as very good compared to this paper experiments with audio and text classifiers, which did not reach above 80%. This also proves that the multimodal method has achieved higher results than the audio and text classifiers. The only thing holding the experiment back is that the dataset is too small, which causes the model not to learn well. Table 4 will also show the best result of each experiment.

Table 4 shows the best result after getting the hyperparameter tuning. The proposed audio classifier gets 43.12% in test accuracy, lyrics get 58.88% in test accuracy, and the fusion method gets 80.56% in test accuracy. The fusion method achieves much higher results than audio classifiers or text classifiers. As in the [36] study, multi-modality outperforms single since the former has access to a better latent space

representation. Tables 5-7 show the comparison result for each experiment of this paper with the relevant MER experiments research.

Table 4. The best result of each experiment

No.	Method	Val accuracy (%)	Test accuracy (%)
1.	CNN-LSTM (Audio modal)	44.4	43.12
2.	XLNet Transformers (Text modal)	58.39	58.88
3.	Fusion Method	82.40	80.56

Table 5. The comparison result of this paper experiments with the audio MER research

No.	Research	Modal	Model	Dataset	Test accuracy (%)
1.	[22]	Audio	CNN	774 songs	72.40
2.	[14]	Audio	CNN-LSTM	361 Indonesian songs	58.33
3	[7]	Audio	CNN, LSTM, DNN	124 Turkish traditional songs	100
4.	This paper	Audio	CNN-LSTM	476 Indonesian songs audio	43.12

Table 6. The comparison result of this paper experiments with the lyrics MER research

No.	Research	Modal	Model	Dataset	Test accuracy (%)
1.	[37]	Lyrics	BiLSTM	2189 Lyrics	91.00
2.	[6]	Lyrics	DNN + XLNet	2595 lyrics +180 songs	94.78
3.	[3]	Lyrics	SVM	900 lyrics	74.8
4.	This paper	Lyrics	XLNet Transformers	476 Indonesian song lyrics	58.88

Table 7. The comparison result of this paper experiments with the multimodal MER research

No.	Research	Modal	Model	Dataset	Test accuracy (%)
1.	[9]	Multimodal	DNN + CNN-LSTM	2000 songs + 2000 lyrics	78.20
2.	[10]	Multimodal	LSTM + BERT	1200 songs + 1200 lyrics	79.62
3.	[23]	Multimodal	Deep convolution network	20000 songs	79.48
4.	[29]	Multimodal	CNN-LSTM	1000 chinese songs	78.2
5.	[31]	Multimodal	CNN + LSTM	1162 minnan songs	83
6.	This paper	Multimodal	Proposed fusion method	476 songs + 476 lyrics	80.56

From Tables 5 and 6, it can be concluded that MER text experiments consistently achieve higher accuracy than MER audio. This happens because the complexity of Mel-spectrogram data from audio is much higher. From Table 5, the highest results for audio were achieved by research from [22] got an accuracy of 72.4%. In Table 6, the highest result for the lyric mode was achieved by [6], who obtained an accuracy of 94.78%, and In Table 7, this paper achieved a similar multimodal result compared to the other multimodal research with an accuracy of 80.56%, got slightly lower than [31] with the 83% accuracy.

For audio modal, this paper achieved an accuracy of 43.12% and has not been able to outperform the research of [22] with an accuracy of 72.4%. This is due to the lack of an audio dataset and an unbalanced model. Likewise, the lyric modal of this paper, with an accuracy of 58.88%, has not been able to outperform the research of [6] with an accuracy of 94.78%. This is due to the possibility that the XLNet model which was pre-trained was in Malay, considering that the dataset used was Indonesian, and the XLNet model was trained in Malay. Meanwhile, for multimodal, this study obtained high results with an accuracy of 80.56% which can be seen in Table 7. This happens because this study proposes to use a state of art model with the highest results from other research studies; besides that, hyperparameter tuning plays a vital role as a determinant of the final accuracy of the multimodal system even though the results are not maximized due to dataset limitations.

In conclusion, the method that this research proposes has achieved quite good results even though the results obtained are less than optimal due to the small dataset. However, this study succeeded in enhancing and enriching the research of [14], which uses the same dataset with an accuracy of 58.33% to 80.56%. This method also got similar results to current multimodal MER studies.

4. CONCLUSION

This study introduces multimodal emotion recognition in Indonesian songs that use audio and lyric modes. The model used in the audio-based MER is the CNN-LSTM model, while the lyrics-based MER uses

the XLNet model. The stacking ensemble method combines the output audio and text classifier results. The neural network is also used in this research function as a meta classifier that aims to obtain linear prediction results.The multimodal method aims to enrich and improve the accuracy of the audio-based MER which used the CNN-LSTM. After carrying out the testing process, the fusion method achieved an accuracy of 80.56% which resulted in a significant increase in results compared to the single modal MER study. The results of this study are also competitive with the other multimodal MER method. Several areas must be improved in future research, such as try exploring other methods to improve the multimodal network, improving the text model with features such as part of speech (POS), or trying to improve the audio model with CNN-LSTM networks. Asking the expert help to annotate emotions in music datasets or crowdsourcing methods, the last and most important is to add more audio data and lyrics data also, make sure the data is balanced on each label.

ACKNOWLEDGEMENTS

The author wanted to thank the Research and Technology Transfer Office (RTTO) Department at Bina Nusantara University for financially supporting this paper. Also, thanks to Amalia Zahra for giving advice and knowledge as the supervisor.

REFERENCES

[1] T. Eerola and H. R. Peltola, "Memorable experiences with sad music-reasons, reactions and mechanisms of three types of experiences," *PLoS ONE*, vol. 11, no. 6, 2016, doi: 10.1371/journal.pone.0157444.

[2] J. S. G-Cañón *et al.*, "Music Emotion Recognition: Toward new, robust standards in personalized and context-sensitive applications," *IEEE Signal Processing Magazine*, vol. 38, no. 6, pp. 106–114, Nov. 2021, doi: 10.1109/MSP.2021.3106232.

[3] R. Panda, R. Malheiro, and R. P. Paiva, "Novel Audio Features for Music Emotion Recognition," *IEEE Transactions on Affective Computing*, vol. 11, no. 4, pp. 614–626, Oct. 2020, doi: 10.1109/TAFFC.2018.2820691.

[4] B. Xie, M. Sidulova, and C. H. Park, "Article robust multimodal emotion recognition from conversation with transformer-based crossmodality the title fusion," *Sensors*, vol. 21, no. 14, Jul. 2021, doi: 10.3390/s21144913.

[5] Y. K. A. Kumar and A. K. Saxena, "Stochastic modelling of transition dynamic of mixed mood episodes in bipolar disorder," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 12, no. 1, Feb. 2022, doi: 10.11591/ijece.v12i1.pp620-629.

[6] Y. Agrawal, R. G. R. Shanker, and V. Alluri, "Transformer-Based Approach Towards Music Emotion Recognition from Lyrics," in *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 – April 1, 2021, Proceedings, Part II*, Berlin, Heidelberg, Mar. 2021, pp. 167–175, doi: 10.1007/978-3-030-72240-1_12.

[7] S. Hizlisoy, S. Yildirim, and Z. Tufekci, "Music emotion recognition using convolutional long short term memory deep neural networks," *Engineering Science and Technology, an International Journal*, vol. 24, no. 3, pp. 760–767, Jun. 2021, doi: 10.1016/j.jestch.2020.10.009.

[8] R. Orjesek, R. Jarina, M. Chmulik, and M. Kuba, "DNN Based Music Emotion Recognition from Raw Audio Signal," in *2019 29th International Conference Radioelektronika (RADIOELEKTRONIKA)*, Apr. 2019, pp. 1–4, doi: 10.1109/RADIOELEK.2019.8733572.

[9] C. Chen and Q. Li, "A Multimodal Music Emotion Classification Method Based on Multifeature Combined Network Classifier," *Mathematical Problems in Engineering*, vol. 2020, p. e4606027, Aug. 2020, doi: 10.1155/2020/4606027.

[10] G. Liu and Z. Tan, "Research on Multi-modal Music Emotion Classification Based on Audio and Lyirc," in *2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, Jun. 2020, vol. 1, pp. 2331–2335, doi: 10.1109/ITNEC48623.2020.9084846.

[11] L. Parisi, S. Francia, S. Olivastri, and M. S. Tavella, "Exploiting Synchronized Lyrics and Vocal Features for Music Emotion Detection," *arXiv*, Jan. 15, 2019, doi: 10.48550/arXiv.1901.04831.

[12] M. A. Priatna and E. C. Djamal, "Precipitation prediction using recurrent neural networks and long short-term memory," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 18, no. 5, Oct. 2020, doi: 10.12928/telkomnika.v18i5.14887.

[13] B. Talafha, A. Abuammar, and M. Al-Ayyoub, "Atar: Attention-based LSTM for Arabizi transliteration," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 11, no. 3, Jun. 2021, doi: 10.11591/ijece.v11i3.pp2327-2334.

[14] A. A. Wijaya, I. Yasmina, and A. Zahra, "Indonesian Music Emotion Recognition Based on Audio with Deep Learning Approach," *Advances in Science, Technology and Engineering Systems Journal Journal (ASTES)*, vol. 6, no. 2, pp. 716–721, 2021, doi: 10.25046/aj060283.

[15] J. S. Murthy, S. G. Matt, S. K. H. Venkatesh, and K. R. Gubbi, "A real-time quantum-conscious multimodal option mining framework using deep learning," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 11, no. 3, Sep. 2022, doi: 10.11591/ijai.v11.i3.pp1019-1025.

[16] S. Bekhet, A. M. Alghamdi, and I. F. Taj-Eddin, "Gender recognition from unconstrained selfie images: a convolutional neural network approach," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 12, no. 2, Apr. 2022, doi: 10.11591/ijece.v12i2.pp2066-2078.

[17] K. Zhang, H. Zhang, S. Li, C. Yang, and L. Sun, "The PMEmo Dataset for Music Emotion Recognition," in *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, New York, NY, USA, Jun. 2018, pp. 135–142, doi: 10.1145/3206025.3206037.

[18] Y.-H. Yang, Y.-C. Lin, Y.-F. Su, and H. H. Chen, "A Regression Approach to Music Emotion Recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 448–457, Feb. 2008, doi: 10.1109/TASL.2007.911513.

[19] T.-Y. Kim and S.-B. Cho, "Predicting residential energy consumption using CNN-LSTM neural networks," *Energy*, vol. 182, pp. 72–81, Sep. 2019, doi: 10.1016/j.energy.2019.05.230.

[20] W. Lu, J. Li, Y. Li, A. Sun, and J. Wang, "A CNN-LSTM-Based Model to Forecast Stock Prices," *Complexity*, vol. 2020, p. e6622927, Nov. 2020, doi: 10.1155/2020/6622927.

[21] Y.-H. Yang, Y.-C. Lin, H.-T. Cheng, I.-B. Liao, Y.-C. Ho, and H. H. Chen, "Toward Multi-modal Music Emotion Classification," in *Advances in Multimedia Information Processing-PCM 2008*, Berlin, Heidelberg, 2008, pp. 70–79, doi: 10.1007/978-3-540-89796-5_8.
[22] T. Liu, L. Han, L. Ma, and D. Guo, "Audio-based deep music emotion recognition," *AIP Conference Proceedings*, vol. 1967, no. 1, p. 040021, May 2018, doi: 10.1063/1.5039095.
[23] G. Tong, "Multimodal Music Emotion Recognition Method Based on the Combination of Knowledge Distillation and Transfer Learning," *Scientific Programming*, vol. 2022, p. e2802573, Feb. 2022, doi: 10.1155/2022/2802573.
[24] W. Rahman *et al.*, "Integrating Multimodal Information in Large Pretrained Transformers," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, Jul. 2020, pp. 2359–2369, doi: 10.18653/v1/2020.acl-main.214.
[25] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized Autoregressive Pretraining for Language Understanding." arXiv, Jan. 02, 2020, doi: 10.48550/arXiv.1906.08237.
[26] C. V. Nanayakkara and H. A. Caldera, "Music Emotion Recognition With Audio and Lyrics Features," *International Journal of Digital Information and Wireless Communications (IJDIWC)*, vol. 6, no. 4, pp. 260–273, 2016.
[27] M. Sharafi, M. Yazdchi, R. Rasti, and F. Nasimi, "A novel spatio-temporal convolutional neural framework for multimodal emotion recognition," *Biomedical Signal Processing and Control*, vol. 78, p. 103970, Sep. 2022, doi: 10.1016/j.bspc.2022.103970.
[28] P. Singh, R. Srivastava, K. P. S. Rana, and V. Kumar, "A multimodal hierarchical approach to speech emotion recognition from audio and text," *Knowledge-Based Systems*, vol. 229, p. 107316, Oct. 2021, doi: 10.1016/j.knosys.2021.107316.
[29] L. Zhang and Z. Tian, "Research on Music Emotional Expression Based on Reinforcement Learning and Multimodal Information," *Mobile Information Systems*, vol. 2022, p. e2616220, Jun. 2022, doi: 10.1155/2022/2616220.
[30] P. Tzirakis, J. Chen, S. Zafeiriou, and B. Schuller, "End-to-end multimodal affect recognition in real-world environments," *Information Fusion*, vol. 68, pp. 46–53, Apr. 2021, doi: 10.1016/j.inffus.2020.10.011.
[31] Z. Xiang, X. Dong, Y. Li, F. Yu, X. Xu, and H. Wu, "Bimodal Emotion Recognition Model for Minnan Songs," *Information*, vol. 11, no. 3, Mar. 2020, doi: 10.3390/info11030145.
[32] C. Wu, F. Wu, Y. Chen, S. Wu, Z. Yuan, and Y. Huang, "Neural Metaphor Detecting with CNN-LSTM Model," in *Proceedings of the Workshop on Figurative Language Processing*, New Orleans, Louisiana, Jun. 2018, pp. 110–114, doi: 10.18653/v1/W18-0913.
[33] K. Song and Y.-S. Kim, "An Enhanced Multimodal Stacking Scheme for Online Pornographic Content Detection," *Applied Sciences*, vol. 10, no. 8, Jan. 2020, doi: 10.3390/app10082943.
[34] S. Chung, J. Lim, K. J. Noh, G. Kim, and H. Jeong, "Sensor Data Acquisition and Multimodal Sensor Fusion for Human Activity Recognition Using Deep Learning," *Sensors*, vol. 19, no. 7, Jan. 2019, doi: 10.3390/s19071716.
[35] V. Shinohara, J. Foleiss, and T. Tavares, "Comparing Meta-Classifiers for Automatic Music Genre Classification," in *Anais do Simpósio Brasileiro de Computação Musical (SBCM)*, Sep. 2019, pp. 131–135, doi: 10.5753/sbcm.2019.10434.
[36] Y. Huang, C. Du, Z. Xue, X. Chen, H. Zhao, and L. Huang, "What Makes Multi-modal Learning Better than Single (Provably)." *arXiv*, Oct. 26, 2021, doi: 10.48550/arXiv.2106.04538.
[37] J. Abdillah, I. Asror, and Y. F. A. Wibowo, "Emotion Classification of Song Lyrics using Bidirectional LSTM Method with GloVe Word Representation Weighting," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 4, no. 4, Aug. 2020.

BIOGRAPHIES OF AUTHORS

Andrew Steven Sams is a student currently pursuing a double degree (Bachelor's and Master's) in Computer Science at Bina Nusantara University. He has been a faculty member since 2017. His research interests include artificial intelligence, signal processing, and speech technology. He can be contacted at email: andrew.sams@binus.ac.id.

Amalia Zahra is a lecturer at the Master of Information Technology, Bina Nusantara University, Indonesia. She received her bachelor's degree in computer science from the Faculty of Computer Science, University of Indonesia (UI) in 2008. She does not have a master's degree. Her Ph.D. was obtained from the School of Computer Science and Informatics, University College Dublin (UCD), Ireland in 2014. Her research interests cover various fields in speech technology, such as speech recognition, spoken language identification, speaker verification, speech emotion recognition, and so on. Additionally, she also has an interest in natural language processing (NLP), computational linguistics, machine learning, and artificial intelligence. She can be contacted at email: amalia.zahra@binus.edu.