ICIC Express Letters
Volume **17**, Number **2**, February **2023**
 ISSN 1881-803X
pp. **181–189**

MULTIMODAL APPROACH FOR EMOTION RECOGNITION USING FEATURE FUSION

Kenny[1] and Andry Chowanda[2]

[1]Computer Science Department, BINUS Graduate Program – Master of Computer Science
[2]Computer Science Department, School of Computer Science
Bina Nusantara University
JL. K. H. Syahdan No. 9, Kemanggisan, Palmerah, Jakarta 11480, Indonesia
kenny005@binus.ac.id; achowanda@binus.edu

Received February 2022; accepted May 2022

Abstract. *Emotion recognition has been a challenge. Multimodality approach in emotion classification has been used in many research to improve the recognition performance. Nevertheless, there is a lack of understanding between how the multimodality affects the performance of the model. This paper uses IEMOCAP as dataset and creates several unimodal model and multimodal model resulting in combination of the top unimodal model for emotion recognition with feature fusion method which merges features from different models. After evaluating the models, this paper analyzes the connection of every unimodality involved and its implication to multimodality built. This paper also applies audio augmentation to reducing overfitting in model's prediction. The top result of multimodal model consisting of 3 modalities achieves F1 score of 71.25% and the model consisting of 2 modalities achieves F1 score of 76.5%.*
Keywords: Emotion recognition, Multimodal classification, Deep learning, Text classification, Image classification, Audio classification, Audio augmentations

1. **Introduction.** Emotion is a very interesting subject and has been researched since the 19th century by experimental psychologists [1]. Classification of emotion in artificial intelligence has been improving over the years, starting with the usage of neuroimaging, Autonomic Nervous System (ANS), facial expression – Facial Action Coding System (FACS), and Speech Emotion Recognition (SER) [1]. Most of the recent classification implements deep learning classification. There is several feature extraction and layer combination in a model for images, textual, and speeches dataset. Although the usage of a single modality is able to create a prediction model with good accuracy, there are approaches that can be used to improve the overall performance of the model. With the rapid advancement of computational power, multi-modality is introduced. Multi-modality is an approach that makes use of multiple inputs of different types instead of one modality. There are some scenarios which utilize multi-modality to improve the robustness of a model, such as detecting a person's emotion in a recorded/real-time, their facial expressions, gestures, speeches, and the textual information spoken can be observed. Many research has proven that the implementation of multi-modality approach increases the overall performance of the model such as [2] and [3]. Although multi-modal approach increases model's performance, there is still a lack of knowledge of the best combination of modality and each modality role in the improvement.

This research aim is to study the importance of each modality and the best combination that can provide the best accuracy for the model. It uses The Interactive Emotional Dyadic Motion Capture (IEMOCAP) [4] as the dataset. The IEMOCAP data used in

DOI: 10.24507/icicel.17.02.181

this research are the speech audio file, motion capture features (head, hand, and rotated), and text transcription of speech. This research uses Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), and Bidirectional Long Short-Term Memory (Bi-LSTM) model for the motion capture modality. For the text modality, this paper uses Bidirectional Encoder Representations from Transformers (BERT) base, Robustly optimized BERT approach (RoBERTa), and DistilBERT. CNN, Bi-LSTM, and a mix of CNN and Bi-LSTM architecture are used for speech modality. In the late stage of the research, a multi-modality model is presented by a combination of the feature fusion of the best model from each modality and applies a classification dense layer.

The main contributions of the paper are as follows: 1) discover the value of each modality used in multimodal and the combination; 2) propose combinations of feature-fusion model that is created by the combination of 3 different modalities (facial, text, and speech). With this paper, the reader can understand more about the role of unimodal in multimodal model and the factors that influence the multimodal model performance. The remainder of the paper is presented as follows: the second chapter of this paper talks about the recent work of related research, the third chapter talks about the flow and the setup of the experiment, the fourth chapter talks about the experiment's result and discussion, and the fifth chapter talks about the research conclusion and possibility of future works.

2. **Recent Work.**

2.1. **Unimodal emotion recognition.**

The early method of Natural Language Processing (NLP) emotion classification can be seen from traditional machine learning. [5] and [6] proposed a classifier with Random Forest (RF) algorithm and Support Vector Machine (SVM) for classification, respectively. [7] proposed the CNN model which consists of one or many combinations of convolutional layer that serve as feature extractor and pooling layer, which are an example of deep learning application. [8] also proposed an LSTM model which is built with the purpose to reduce the loss of important information during the training. Lastly, a pretrained model with a big amount of data that only need slight modification to be used called BERT which is currently state of the art was proposed by [9]. BERT model solves one of the problems of the traditional approach of NLP, which is the needs of many dataset. In the Facial Emotion Recognition (FER), there are conventional methods that extract geometric features and appearance features from the coordinate within the facial images. [10] proposed an FER with geometric features as the input and Hidden Markov Model (HMM) as the classifier. With the advancement of computer vision, there are many approaches to recognizing emotion using deep learning. With that said, [11] used another approach that applies existing low-level feature as input which applies modified Local Directional Pattern (LDP) to the deep learning model that can produce higher level feature. There is also a broadly used deep learning model called convolution neural network which utilize Convolutional layers which processes the images to obtain high level feature. One of the approaches that can utilize CNN can be seen in [12]. The traditional machine learning in Speech Emotion Recognition (SER) consists of extracting acoustic features from the utterance such as Mel-frequency cepstral coefficient and pitch and uses a classifier such as SVM and HMM [13]. [14] proposed an SER model that utilizes Mel Frequency Spectrum Coefficients (MFCC) and Mel Energy Spectrum Dynamic Coefficients (MEDC) with SVM as the classifier. [15] created a deep learning model that combined LSTM and CNN layer, which outperforms the traditional SVM method. [16] also proposed a neural network model that utilizes CNN layer and Bi-LSTM layer to predict valence and arousal with waveform and spectrograms as input.

2.2. **Multimodal emotion recognition.** Research on multimodal approaches on emotion recognition has been progressing in the past years. Some of the researches classify concatenated feature from different modalities which are called early fusion which is a method implemented in [17]. There is other approach that uses the output of unimodal to do another classification in another classification layer called late fusion, which can be seen in [18]. Although it has been proven in many researches that multimodal performs better than unimodal, one of the fields that can be improved in the multimodal is to know more about the importance of the modality involved and the origin of the performance boost in multimodal.

3. **Proposed Method.** The motion captures modality uses the combined feature of head, hand, and rotated facials feature. This research uses a simple LSTM model, Dense model, CNN model, with the addition of simple Bi-LSTM model as possible improvement as feature extractor, which are utilized in [19]. The proposed models are used because they are still many applications in the recent research and they are specifically useful for this research requirement and purposes, which is emotion recognition. LSTM model consists of an LSTM layer with 256 units followed with a 128-unit Dense layer with activation function Rectifier Linear Unit (ReLU). The Dense model consists of a Dense layer with 256 units followed with a 128-unit Dense layer with ReLU activation. The CNN model consists of 3 Convolutional layers with kernel size 3, Stride 2, and the filters 32, 64, and 64, respectively. Every Convolutional layer is followed with ReLU activation and a dropout of 0.2. The output of the last convolutional layer is followed with a 128-unit Dense layer with ReLU activation. The Bi-LSTM model consists of an LSTM layer with bidirectional applied followed with a 128-unit Dense layer with ReLU activation. For the text modality, this paper proposes the utilization of 3 different BERT models, which are BERT [20], RoBERTa [21], and DistilBERT [22]. All BERT models start with an input layer, which accept token embeddings and mask embeddings. After the input layer, the next layer consists of pretrained transformers model from each BERT version, with the trainable of every layer disabled. One of the model structure examples can be seen in Figure 1.

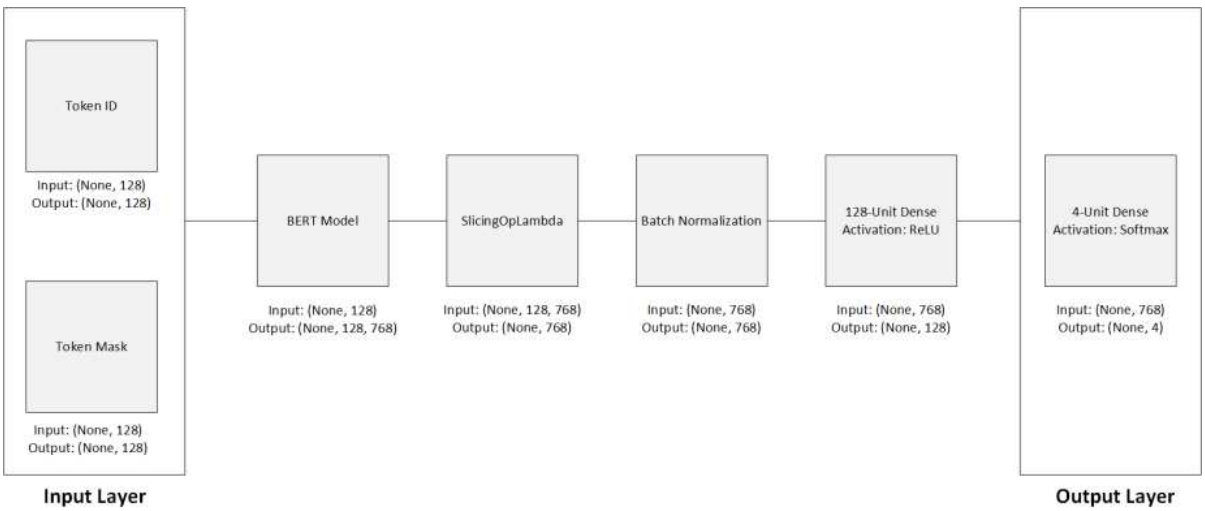

Figure 1. BERT based proposed model

The speech modality in this research proposes 3 models which are a simple Bi-LSTM model, CNN model, and mixed CNN-Bi-LSTM model. The Bi-LSTM model consists of a 512-unit Bi-LSTM layer, followed by 128-unit Dense layer with activation function Rectifier Linear Unit (ReLU). The CNN model consists of 5 Convolutional layers with kernel size 3, Stride 2, and the filters 32, 64, and 64, respectively. Every Convolutional

layer is followed with ReLU activation and a dropout of 0.2 and the output of the last convolutional layer is followed with a 128-unit Dense layer with ReLU activation. The mixed CNN-Bi-LSTM model is a model that consists of 5 convolutional layers, followed by a 512-unit Bi-LSTM layer. The output of the Bi-LSTM layer is then followed by a Dense layer with 128 units. The multimodal models are built from the best performance model from each modality via feature fusion technique which combines multiple models on feature level. With feature fusion, it is easier to add or remove unimodal model from the proposed multimodal models. The multimodal model's early layer consists of the same layer as the used unimodal model until its feature extraction layer. The output features from each unimodal are then concatenated and counted as multimodal model's feature. Figure 2 shows an example of multimodal model combined from text and motion capture.

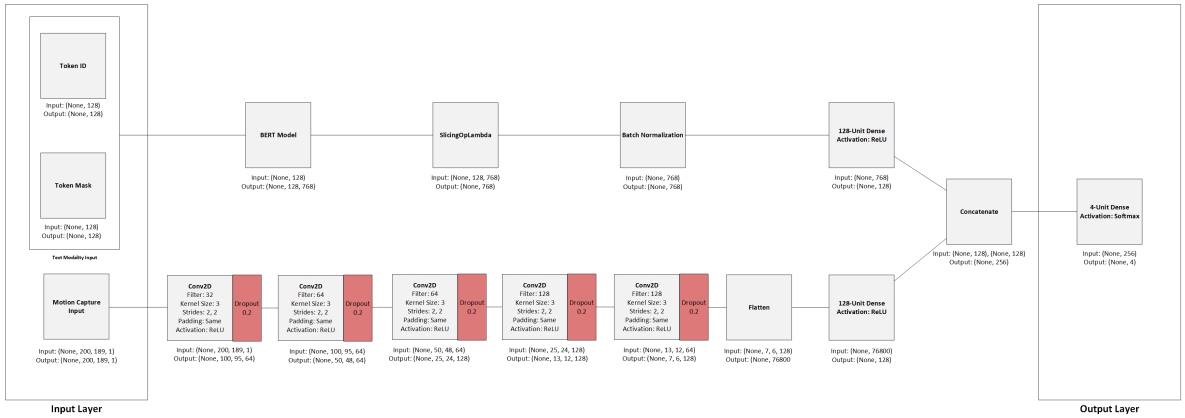

FIGURE 2. Combination of text BERT and motion capture CNN

4. **Experimental Results.** There are 3 phases in this experiment. The first phase consists of setup of the experiments environment and dataset. The 2nd phase consists of training and testing the performance for the unimodalities models. This is then followed by the 3rd phase, which is training and testing the performance of the model consisting of the best performing unimodalities. All models in this research have the same hyperparameter applied to ensuring the fairness of the result for the comparison purpose. This research was carried out in the Google Colab environment, runtime environment of GPU Tesla K80 with specification of 12.69 GB RAM and 78.2 GB of Disk. The hyperparameters used in this research are an epoch of 50, batch size of 32 because of memory limitation, and a starting learning rate of 0.001 which is the default value of Adam optimizer.

4.1. **Data pre-processing.** IEMOCAP dataset consists of 5 sessions in which every session represents recording of a dialog between a male and female actor in both scripted and improvised scenarios. The recording session is then divided into utterances which are annotated by multiple annotators. The utterances are evaluated based on 10 possible emotions (angry, happy, sad, neutral, frustrated, excited, fearful, surprised, disgusted, and other).

This research first processes the available information and then reconfigures it into a table with columns of start time, end time, wavefile name, and annotated emotions. After that, the data are further filtered by specific emotions (angry, happy, sad, neutral, and excited) in consideration of the data's balance following method proposed by [23]. Then, the data with excited emotion are replaced into happy to further increase the data's balance. The result of pre-processing is 5,521 total utterances with 1,102 angry, 1,627 happy, 1,084 sad, and 1,708 neutral utterances. The top view result of the process can be seen in Table 1.

Table 1. Pre-processed data

Start time	End time	File name	Emotion
4.81	10.06	Ses01F_script03_2_F000	Happy
15.48	19.55	Ses01F_script03_2_F001	Happy
22.31	26.13	Ses01F_script03_2_F002	Happy
37.5	43.15	Ses01F_script03_2_F004	Neutral
46.75	48.7825	Ses01F_script03_2_F005	Neutral
100.2315	102.31	Ses01F_script03_2_F013	Angry

This research handles the motion capture features following the approach used in [19]. In text modality, this research applies BERT model's tokenizer to generating token embeddings and mask embeddings with a maximum length of 128 from the words of every utterance. Although the word is the same, the value of the token embeddings can be different following the pretrained BERT model's dataset. The example of token and mask can be seen in Table 2.

Table 2. BERT base tokenized example

Words	Token (max length 18)	Mask (max length 18)
It'll be good. Wow, that's great.	[101 1,135 112 1,325 1,129 1,363 119 11,750 117 1,115 112 188 1,632 119 102 0 0 0]	[1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0]
I knew they would, your mother anyway.	[101 146 1,450 1,152 1,156 117 1,240 1,534 4,050 119 102 0 0 0 0 0 0 0]	[1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0]

In speech data, this research uses Mel-spectrograms, a spectrograms with the Mel scale serves as y axis and time as x axis for the features. The first step starts finding the tolerated max length, which is 16s. After that, the audio was appended with less 16s with silence audio to match up the duration. Audios longer than 16s are fixed into the first 16s to match other's audios length. After all the audios have matched 16s, this research extracts the Mel-spectrograms feature with Librosa library from every audio. The examples of Mel-spectrograms feature can be seen in Figure 3.

This research also prepares another set of audio data with purpose to reduce the overfitting that happens when we evaluated the proposed model. The augmentation done to

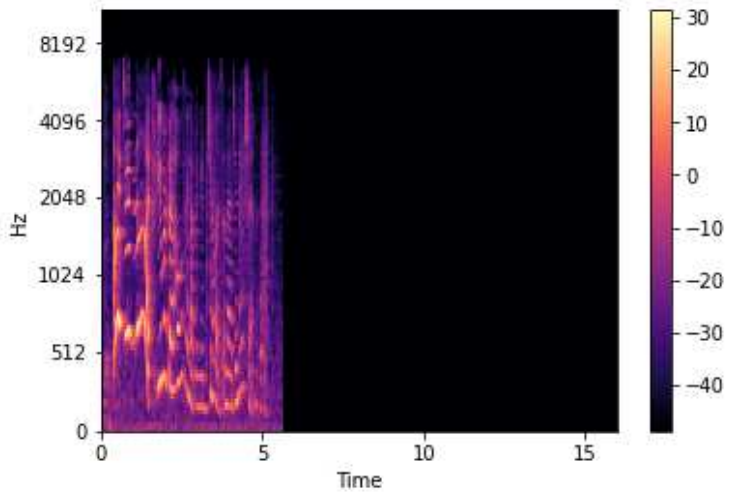

Figure 3. (color online) Padded Mel-spectrogram's feature

the audio file is an addition of Gaussian noise of a range of 0.001-0.015 amplitude, time stretch between 0.8 to 1.2 multiplier, a shift of pitch from -4 semitones to 4 semitones, and a shift of fraction between -0.5 to 0.5.

4.2. **Unimodal model evaluation.** The 2nd phase evaluates a total of 22 unimodal models, which consists of 4 different motion capture parts (hand, head, rotated, and combined) that are trained in CNN, Bi-LSTM, Dense, and LSTM model. BERT base, RoBERTa, and DistilBERT are for text modality, and Bi-LSTM, CNN, and combination of Bi-LSTM and CNN are for audio modality. During the experiment, the differences between accuracy and validation accuracy on audio are big, so this research also proposes the usage of augmented audio data on audio model.

Based on Table 3, motion capture's best performing model is CNN with the best F1 score of 0.6625. In text, every BERT model has a competitive performance, with the biggest difference of 0.02 F1 score. The best performing model is the BERT base with F1 score of 0.675. For the audio model, there are big differences between the best and the worst performing model with the difference of 0.3575 F1 score. The best performing model in audio is CNN with F1 score of 0.5375 for non-augmented audio data and 0.5425 for augmented audio data, while the worst performing model is the Bi-LSTM model, which shows that the plain Bi-LSTM model is not suited for the audio processing. The effect of applying the augmented data shows that there is a minor increase in accuracy and major increase in validation accuracy in CNN related model which shows that the overfitting has been reduced.

4.3. **Multimodal model evaluation.** In the 3rd phase, every multimodal combination of best performing model is designed and evaluated. With all unimodal models evaluated, the best models chosen for the 3rd phase are the CNN for motion capture, BERT base for the text, and CNN for the audio.

Based on experiment result in Table 4, the result of model 4 which consists of 2 modalities performs the best F1 score with 0.765. The difference between every model which consists of 2 modalities can be explained in the unimodality it parted of. Models 3 and 4 have a difference in audio and motion capture modality, which in the unimodal model the performance of motion capture model is better than the audio model. Both model 3 and model 4 also outperform model 2 which consists of 2 of the worst performing unimodal model, audio and motion capture, respectively. Another viewpoint that can be seen is that models 2, 3, and 4 have better validation accuracy and some of them has fewer validation loss compared to model 1 which consists of 3 modalities. It shows that the model with the best modality does not always achieve the best performance. The performance result of multimodal model is greatly implicated by the model it consists of, which can be seen in the difference between model 1 and model 2. Although the model 2 achieves a little better validation accuracy compared to model 1, the validation loss achieved by model 1 by adding the best unimodality which is text is greatly reduced. This research also compares the model with the research of [19] in the form of model 5 as the baseline. The comparation itself is not fair, because there is a difference in hyperparameters and split ratio. Model 5 is trained in a ratio of $77 : 22$ and batch size of 64, while models 1 to 4 are trained in the ratio of $8 : 2$ and batch size of 32. The comparation results show that this research multimodal model performs better in training and worse in predicting, although the difference is not significant.

Table 5 shows the evaluation results for the models which use the augmented audio dataset. Compared to the non-augmented dataset, the most significant change is model 1, which has an increase of accuracy around 0.01, reduction of loss by around 0.04, increase of validation accuracy by around 0.05, and an increase of validation loss around by 0.01. The F1 score also has an increase of 0.05. Model 2 and Model 3 do not have any significant change on the performance after the application of augmented audio. This result shows

Table 3. Unimodal evaluation result

Motion capture hand	Accuracy	Loss	Validation accuracy	Validation loss	F1 score
LSTM	0.6458	0.8128	0.526	1.1904	0.55
Dense	0.7684	0.5879	0.5147	1.4814	0.525
Bi-LSTM	0.7072	0.6973	0.5124	1.4018	0.525
CNN	**0.7358**	**0.6636**	**0.5509**	**1.2279**	**0.5775**
Motion capture head	Accuracy	Loss	Validation accuracy	Validation loss	F1 score
Dense	0.718	0.9968	0.3688	5.0209	0.3325
LSTM	0.9824	0.0054	0.3688	3.644	0.365
CNN	0.7574	0.6078	0.3835	1.7882	0.3725
Bi-LSTM	**0.984**	**0.0369**	**0.3688**	**3.6755**	**0.38**
Motion capture rotation	Accuracy	Loss	Validation accuracy	Validation loss	F1 score
Dense	0.3024	1.3672	0.3201	1.3622	0.1225
LSTM	0.44	1.2296	0.44	1.2173	0.25
Bi-LSTM	0.4151	1.237	0.4412	1.2232	0.275
CNN	**0.7381**	**0.6369**	**0.5452**	**1.2138**	**0.5475**
Motion capture combined (Head + Hand + Rotation)	Accuracy	Loss	Validation accuracy	Validation loss	F1 score
Bi-LSTM	0.5317	1.1055	0.5192	1.1293	0.48
LSTM	0.5286	1.0982	0.4876	1.1387	0.495
Dense	0.5798	6.486	0.5317	7.0407	0.4975
CNN	**0.8414**	**0.4124**	**0.655**	**1.1344**	**0.6625**
Text	Accuracy	Loss	Validation accuracy	Validation loss	F1 score
RoBERTa	0.6318	0.8776	0.6452	0.8796	0.6475
DistilBERT	0.8675	0.3662	0.6606	1.0519	0.66
BERT base	**0.8786**	**0.3256**	**0.6715**	**1.0945**	**0.675**
Audio	Accuracy	Loss	Validation accuracy	Validation loss	F1 score
Bi-LSTM	0.32	1.3526	0.3294	1.3564	0.18
CNN + Bi-LSTM	0.9767	0.0604	0.4226	3.7153	0.4075
CNN	**0.9789**	**0.0603**	**0.5294**	**2.6747**	**0.5375**
Audio augmented	Accuracy	Loss	Validation accuracy	Validation loss	F1 score
Bi-LSTM	0.308	1.3593	0.314	1.3622	0.13
CNN + Bi-LSTM	0.9708	0.774	0.4851	3.1189	0.485
CNN	**0.9821**	**0.0734**	**0.5348**	**4.127**	**0.5425**

that the expected result of decrease in overfitting is achieved in model 1. The increase of performance in multimodal model is very big compared to the minor increase in unimodal model. The new performance results of model 1 outperform the model 5 that is used as baseline.

5. **Conclusion and Future Work.** The purpose of this research is to find out the characteristic of multimodal and the relation between the multimodal model and the unimodal model behind it. This research creates several unimodal models for motion

Table 4. Multimodal evaluation result

No	Multimodal	Accuracy	Loss	Validation accuracy	Validation loss	F1 score
1	Text + motion capture + audio	0.9728	0.0776	0.6561	1.9717	0.66
2	Motion capture + audio	0.9848	0.0452	0.6724	3.1287	0.6775
3	Text + audio	0.9758	0.0738	0.695	1.7082	0.7
4	**Text + motion capture**	**0.9389**	**0.1703**	**0.7638**	**0.9509**	**0.765**
5	Text + motion capture + audio [19]	0.9666	0.0836	0.6731	2.1394	–

Table 5. Multimodal evaluation result with augmented audio dataset

No	Multimodal	Accuracy	Loss	Validation accuracy	Validation loss	F1 score
1	Text + motion capture + audio	0.9898	0.0324	0.7077	2.0518	0.7125
2	Motion capture + audio	0.9721	0.0898	0.6615	1.8232	0.665
3	Text + audio	0.9776	0.0674	0.6869	1.7712	0.6975

capture, text, and audio modality to be evaluated. The top performing model from each modality is then combined to build the multimodal model. The experiment results show that the performance of every unimodal model implicates the multimodal model. There is also another finding that more modality number contributed to the multimodal model does not always improve the performance of the model, although the performance of every multimodal model is better than the top performing unimodal model. The experiment also shows that augmenting audio data reduce the model overfitting, which have little effect on unimodal model and quite more significant effect in some of the multimodal model.

Although there are some conclusions achieved in this research, there are still limitations in this research that can still be improved in the future. The unimodal models used in this research can still be improved, whether modified or totally changed that may give new result. There is also a possibility to introduce a new data preprocessing method and new ways to evaluate the model.

REFERENCES

[1] T. Thanapattheerakul, K. Mao, J. Amoranto and J. H. Chan, Emotion in a century: A review of emotion recognition, *Proc. of the 10th International Conference on Advances in Information Technology*, pp.1-8, 2018.

[2] C. T. Duong, R. Lebret and K. Aberer, Multimodal classification for analysing social media, *arXiv.org*, arXiv: 1708.02099, 2017.

[3] S. Yoon, S. Byun and K. Jung, Multimodal speech emotion recognition using audio and text, *2018 IEEE Spoken Language Technology Workshop, SLT 2018 – Proceedings*, pp.112-118, DOI: 10.1109/SLT.2018.8639583, 2019.

[4] C. Busso et al., IEMOCAP: Interactive emotional dyadic motion capture database, *Language Resources and Evaluation*, vol.42, no.4, pp.335-359, 2008.

[5] P. Vora, M. Khara and K. Kelkar, Classification of tweets based on emotions using word embedding and random forest classifiers, *International Journal of Computer Applications*, vol.178, no.3, pp.1-7, DOI: 10.5120/ijca2017915773, 2017.

[6] M. M. Saad, N. Jamil and R. Hamzah, Evaluation of support vector machine and decision tree for emotion recognition of Malay folklores, *Bulletin of Electrical Engineering and Informatics*, vol.7, no.3, pp.479-486, 2018.

[7] K. Shrivastava, S. Kumar and D. K. Jain, An effective approach for emotion detection in multimedia text data using sequence based convolutional neural network, *Multimedia Tools and Applications*, vol.78, no.20, pp.29607-29639, 2019.

[8] M. H. Su, C. H. Wu, K. Y. Huang and Q. B. Hong, LSTM-based text emotion recognition using semantic and emotional word vectors, *2018 1st Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia 2018)*, 2018.

[9] F. A. Acheampong, H. Nunoo-Mensah and W. Chen, Transformer models for text-based emotion detection: A review of BERT-based approaches, *Artificial Intelligence Review*, vol.54, pp.5789-5829, DOI: 10.1007/s10462-021-09958-2, 2021.

[10] R. Agarwal, S. Mishra, N. Kohli and M. Rahul, Facial expression recognition using geometric features and modified hidden Markov model, *International Journal of Grid and Utility Computing*, vol.10, no.5, pp.488-496, DOI: 10.1504/ijguc.2019.10022683, 2019.

[11] M. Z. Uddin, M. M. Hassan, A. Almogren, M. Zuair, G. Fortino and J. Torresen, A facial expression recognition system using robust face features from depth videos and deep learning, *Computers and Electrical Engineering*, vol.63, pp.114-125, 2017.

[12] H. Ma and T. Celik, FER-Net: Facial expression recognition using densely connected convolutional network, *Electronics Letters*, vol.55, no.4, pp.184-186, 2019.

[13] S. K. Pandey, H. S. Shekhawat and S. R. M. Prasanna, Deep learning techniques for speech emotion recognition: A review, *2019 29th International Conference Radioelektronika, RADIOELEKTRONIKA 2019 – Microwave and Radio Electronics Week (MAREW 2019)*, DOI: 10.1109/RADIOELEK.2019.8733432, 2019.

[14] Y. Chavhan, M. L. Dhore and P. Yesaware, Speech emotion recognition using support vector machine, *International Journal of Computer Applications*, vol.1, no.20, pp.6-9, DOI: 10.5120/431-636, 2010.

[15] G. Trigeorgis et al., Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network, *IEEE International Conference on Acoustics, Speech and Signal Processing – Proceedings*, pp.5200-5204, DOI: 10.1109/ICASSP.2016.7472669, 2016.

[16] Z. Yang and J. Hirschberg, Predicting arousal and valence from waveforms and spectrograms using deep neural networks, *Proc. of the Annual Conference of the International Speech Communication Association*, pp.3092-3096, DOI: 10.21437/Interspeech.2018-2397, 2018.

[17] J. Williams, S. Kleinegesse, R. Comanescu and O. Radu, Recognizing emotions in video using multimodal DNN feature fusion, *Proc. of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)*, pp.11-19, 2018.

[18] Y. R. Pandeya and J. Lee, Deep learning-based late fusion of multimodal information for emotion classification of music video, *Multimedia Tools and Applications*, vol.80, no.2, pp.2887-2905, DOI: 10.1007/s11042-020-08836-3, 2021.

[19] S. Tripathi, S. Tripathi and H. Beigi, Multi-modal emotion recognition on IEMOCAP dataset using deep learning, *arXiv.org*, arXiv: 1804.05788, 2018.

[20] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, *Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol.1, pp.4171-4186, 2019.

[21] Y. Liu et al., *RoBERTa: A Robustly Optimized BERT Pretraining Approach*, http://arxiv.org/abs/1907.11692, 2019.

[22] V. Sanh, L. Debut, J. Chaumond and T. Wolf, *DistilBERT, A Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter*, http://arxiv.org/abs/1910.01108, 2019.

[23] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera and D. Manocha, M3ER: Multiplicative multimodal emotion recognition using facial, textual, and speech cues, *Proceedings of the AAAI Conference on Artificial Intelligence*, vol.34, no.2, pp.1359-1367, DOI: 10.1609/aaai.v34i02.5492, 2020.