Emotion Recognition on Multimodal with Deep Learning and Ensemble

David Adi Dharma, Amalia Zahra
Computer Science Department-BINUS Graduate Program-Master of Computer Science
Bina Nusantara University, Jakarta, Indonesia 11480

***Abstract*—Emotion Recognition on multimodal dataset is a difficult task, which is one of the most important tasks in topics like Human Computer Interaction (HCI). This paper presents a multimodal approach for emotion recognition on dataset MELD. The dataset contains three modalities, audio, text, and facial features. In this research, only audio and text features will be experimented on. For audio data, the raw audio is converted into MFCC as an input to a bidirectional LSTM, which will be built to perform emotion classification. On the other hand, BERT will be used to tokenize the text data as an input to the text model. To classify the emotion in text data, a Bidirectional LSTM will be built. And finally, the voting ensemble method will be implemented to combine the result from two modalities. The model will be evaluated using F1-score and confusion matrix. The unimodal audio model achieved 41.69% of F1-score, while the unimodal text model achieved 47.29% of F1-score, and the voting ensemble model achieved 47.47% of F1-score. To conclude this research, this paper also discussed future works, which involved how to build and improve deep learning models and combine them with ensemble model for better performance in emotion recognition tasks in multimodal dataset.**

***Keywords*—Emotion recognition; deep learning; ensemble method; transformer; natural language processing**

I. Introduction

Humans, along with technological developments pour their emotions or feelings either through some media such as text, photos, audio, or video recordings. Human emotions are complex, which make it difficult to be studied or predicted, and it takes a high level of intelligence to be able to recognize the emotions expressed by people in the current media [1]. Due to the complexity of human emotions, the variety of human's feelings, and the media where they convey their emotions, the AI model's learning has evolved to multimodal datasets, where existing media, audio, video, text, biological information, can improve the model's ability to classify emotions more accurately [2]. This emotion recognition task is involved in some study subjects, which are Natural Language Processing (NLP) and Machine Learning (ML).

The algorithms used to classify emotions are also being actively researched and developed. The same algorithm used for classification can have different results, depending on the dataset used. The datasets used for classification are media such as text, or images, or EEG (Electroencephalogram) signals, as well as sound. Examples of Machine Learning (ML) method that was used for classifying emotions are SVM (Support Vector Machine), KNN (K-Nearest Neighbor), and Bayesian Network in the research [3]. Then the emotion classification method developed to neural networks-based model which are Deep Learning (DL) such as Recurrent Neural Network (RNN) in [4] research, Deep Neural Network (DNN) in [5], DialogueRNN which is RNN-based model in the research by [6], LSTM (Long Short-Term Memory) and CNN (Convolutional Neural Network) conducted in the [7] experiment. Lastly, a Bidirectional LSTM-CNN model was used to learn context and classify emotions in the [8] and [9] works.

Transformer which is proposed in 2017 [10] is the neural network architecture. This model outperformed any RNN and LSTM model that were popular in NLP. Thus, Transformer develops into BERT in 2019 [11]. BERT stands for Bidirectional Encoder Representation of Transformer, and until today BERT has been widely used to extract contextual information from texts. BERT also used in emotion recognition task on the text dataset, in [11] research published in 2022.

An ensemble learning [12] is a technique that combines base predictors model and improves it to become more outstanding predictor. There are few kinds of ensemble learnings which are called bagging, boosting, and stacking [13]. In the [14], there is ensemble learning called voting, which choose the highest probability value as the final classification result. In this experiment, the voting ensemble learning will be performed.

There are some datasets that are widely used in multimodal emotion classification research, such as IEMOCAP (The Interactive Emotional Dyadic Motion Capture) and MELD (Multimodal EmotionLines Dataset). MELD was made and published in 2018 by [15]. These works have used IEMOCAP dan MELD for building and developing models that are used to classify emotion in multimodal. This paper built a model using MELD dataset, which contains 7 labels of emotions, such as anger, disgust, sadness, joy, neutral, surprise and fear. All data in the MELD dataset are in English. There are three modalities that are provided the dataset, video which are audio and facial data, and textual data.

The purposes of the experiment are to extract emotion features in multimodal dataset and build a model that could recognize the emotion that is learned from the features. This experiment also intends to evaluate the model built and compare the evaluation result with the existing models.

There are two tasks that will be conducted in this research, Speech Emotion Recognition (SER) and Text Emotion Recognition (TER). Based on the state-of-the-art mentioned above, BERT will be built to tokenize the textual dataset, then a Bidirectional LSTM model is built to classify the emotion for the text dataset. On the other hand, a Bidirectional LSTM model will be used for SER task. Finally, a voting ensemble learning model will be applied as the ensemble learning. This paper will utilize the advantages of models that were mentioned above for improving emotion recognition model's performance.

The remainder of the paper is structured as follows: Section II will review related research for emotion recognition for both unimodal and multimodal dataset, Section III will provide proposed method in this research, Section IV on experiment scenario, discussion on Section V, and finally, conclusion and future works are provided on Section VI.

II. Related Works

Research related to emotion classification has been developed using various combinations of algorithms with various datasets. Datasets can be in the form of text, sound, images, to multimodal such as a combination of text with sound and video recordings (image and sound). Research conducted in [4] uses the IEMOCAP dataset as the multimodal dataset, where the researcher takes audio and text data. In the audio dataset, the features are extracted using the tools openSMILE for audio and NLTK for text. In the classification of emotions, the results of the extraction of these features are processed by an RNN model. Another study by [16] used the OMG (One-Minute Gradual-Emotion Recognition) dataset and divided the dataset into 2 classes, namely arousal and valence. To extract video data features, researchers used OpenFace and VGG Face, to extract audio data features using OpenSMILE, and for text using Lexicon.

Then another study using the AFEW (Acted Facial Expression in the Wild) dataset conducted in [2] and compared several feature extraction methods, such as TF-IDF (Term Frequency-Inverse Document Frequency) and WV (Word Vectors) for text data, C3D and VGG for visual datasets, then MFCC, SoundNet, and VGG for audio data. Then each data type, namely audio, text, and visual is classified using 3 methods, namely RF (Random Forest), SVM (Support Vector Machine), and LSTM (Long Short-Term Memory).

Subsequent research conducted in [6] proposed a model called DialogueRNN, which was used to classify emotions from the IEMOCAP and AVEC datasets. In this study, audio data features were extracted using OpenSMILE tools, 3D-CNN to extract visual features, and CNN to extract text features. Further research by [17] used the IEMOCAP dataset by taking audio and text data. Text data is extracted using Word2Vec, while audio dataset is extracted to new low-level features and then processed into high-level features or contextual features using CNN-LSTM. The method used for emotion classification is Deep Neural Network (DNN) which consists of four layers, namely input layer, hidden layer 1, hidden layer 2, and softmax layer.

Research by [18] uses the IEMOCAP dataset and retrieves audio and text data. The method used to classify is divided into two, namely, several Machine Learning (ML) models and several Deep Learning (DL) models. The ML methods used are Random Forest (RF), Gradient Boosting (XGB), Support Vector Machine (SVM), Multinomial Naïve Bayes (MNB), and Logistic Regression (LR). Meanwhile, the DL method used is Multi-Layer Perceptron (MLP) and Long Short-Term Memory (LSTM).

Another study conducted by [5] used the IEMOCAP dataset, taking audio and text data from IEMOCAP. Text data feature extraction using GloVe, while audio data is handcrafted. Researchers combined CNN (Convolutional Neural Network) with LSTM (Long Short-Term Memory) to classify emotions from audio data, and Bi-LSTM (Bidirectional Long Short-Term Memory) to classify text data. Then in the classification of multimodal data, the method used is DNN (Deep Neural Network) by combining high-level features from 2 data, namely audio and text, then using a softmax layer to perform the classification.

Furthermore, research conducted by [19] used the MELD dataset where the data taken were audio and text data. Audio data is converted to MFCC first, while text data is processed using BERT. Then the audio and text data are entered into the GRU and then forwarded to the proposed model, namely MMFA (Multi-Level Multi-Head Fusion Attention).

The next study conducted in [17] used the MELD dataset and took the audio and text data as multimodal data, where the proposed model is GNN (Graph Neural Network). Research by [20] uses the MELD dataset and uses all data modalities, namely audio, text, and visual. The proposed model is HFGCN (Hierarchical Fusion Graph Convolutional Network). HFGCN provides output in the form of utterance-level features which are then entered into a fully connected layer for emotional classification.

In the [21] research in 2018, Bi-LSTM and LSTM are combined with Attention Layer to classify emotions, while the datasets are crawled from the internet. After bi-LSTM and LSTM layers are being utilized, a soft voting model is deployed as the ensemble strategy to get the final emotion. In [11] used ResNet for emotion recognition in audio modality and used BERT based model to emotion recognition in text modality. Then final score is being calculated using weighted score formula mention in [11].

In the [15] along with MELD dataset, a Bidirectional LSTM model was built to classify the emotions. This bidirectional LSTM will be included as a baseline model for the unimodal emotion model in this experiment. Also, in the [22], there are some bidirectional architectures that were experimented on. The best architecture will be included as the baseline model in this experiment. The unimodal model that will be built is the modified version from these two baseline models. Then, for the multimodal classification, voting will be performed from the predictions which are generated by these baseline models as the benchmarks.

III. RESEARCH FRAMEWORK

The topic of this research is to recognize emotion through classification from the model that will be built with multimodal dataset. Many Machine Learning and Deep Learning models have been researched to do emotion recognition tasks in the multimodal dataset. Common multimodal datasets such as IEMOCAP, AFEW, OMG, and MELD are being widely used to train the emotion recognition model. Most research uses bimodal such as audio and text dataset, there are also few research that use audio, facial, and text features. There are many combinations of methods in multimodal dataset, but to find the most effective one for emotion classification, it will need a lot of experiments.

This research proposed deep learning models to recognize emotions in the two dataset modalities in the MELD dataset. This research will be done in supervised learning as MELD has already been labeled to train deep learning models. For audio features, bidirectional LSTM (Long-Short Term Memory) will be used to classify emotions. Then, as for text modality, BERT will be deployed to extract contextual features and classify the emotions. Lastly, voting ensemble learning is also being experimented in research.

As shown in Fig. 1, this research starts from choosing the topic of the research, and then studies all the literature related to the topic. After that the method and deep learning model is chosen. The next part will be preparing the dataset, preprocessing the data, and extracting the features from the dataset. Then, in the next part of the research, the model will be tuned, then train the model for each modality, which are audio and text. After that, voting ensemble learning is applied. At last, after all models are trained, the models will be evaluated using evaluation metrics such as F1-score and confusion matrix, then analyze the experiment result, and conclude the experiment result in the paper.

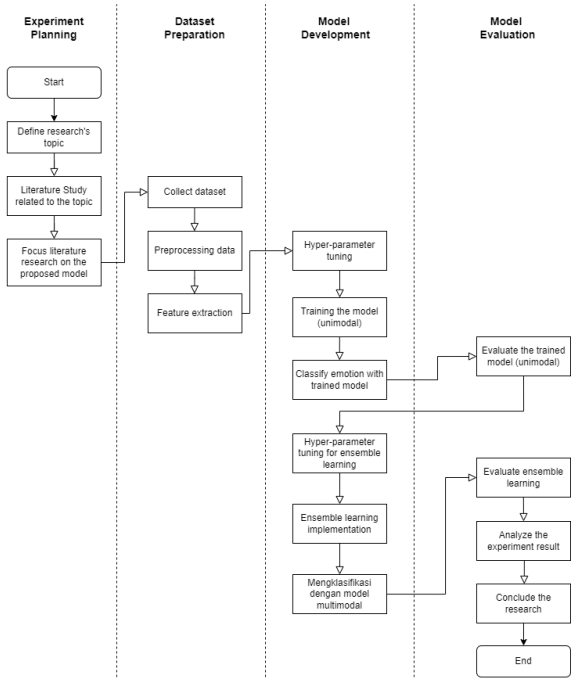

Fig. 1. Experiment scenario.

A. Preparing Dataset

In this research, MELD dataset will be used to train and build the model. MELD [15] contains 13,000 utterances from 1,433 dialogs in the TV series called "Friends". The dataset is divided into three parts, train, test, and dev. Every utterance from 1,433 dialogs already has emotion label and sentiment label. The dataset contains three modalities, which are audio, facial, and text modality. There are seven emotion labels annotated in the dataset, Joy, Sadness, Fear, Anger, Surprise, Disgust, and lastly Neutral. There are also three annotations for sentiment analysis, such as neutral, positive, and negative. But in this research, only emotion labels will be used as the label to train the model.

The distribution of the emotion in the dataset is shown the Fig. 2. The train part contains the most data, while test and dev parts contain less data respectively. Fig. 3 below will show how the emotion labels are divided in percentage using pie chart. Neutral label is dominant with 47% label from all emotion label that exist in the MELD dataset. Then followed by Joy with 17%, Anger and Surprise with 12% each, Sadness is 7%, Disgust is 3% and Fear with only 2% from the dataset.

As the dataset is greatly imbalance, which will affect the model's performance, the emotion labels in this experiment will be converted to only five emotion labels, which are neutral, anger, joy, sadness, and surprise. Fear and disgust are excluded because of the lack of data. As for the train data, the labels distribution will be divided into 683 data for neutral, anger, joy, and surprise, and sadness label. The total training data for the model will be 3415. The data distribution bar chart is represented in Fig. 4.

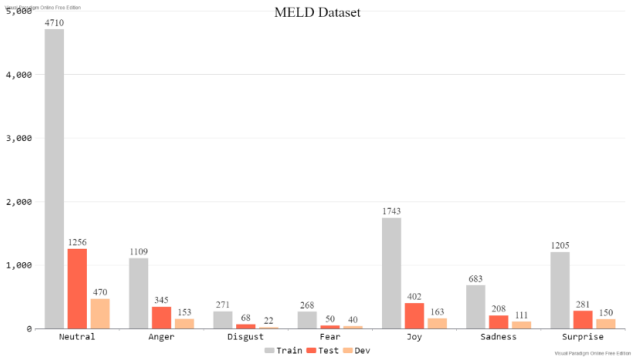

Fig. 2. MELD label distribution.

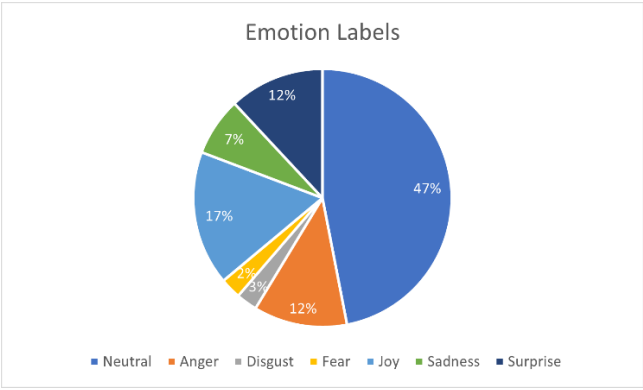

Fig. 3. Labels percentage.

Fig. 4. Training data distribution.

B. Feature Extraction

There are two feature extraction methods in this research. For text data, raw texts are tokenized with Distil-BERT model. As for the audio modality, the audio file will be extracted into MFCC [23]. Feature extractor methods are being used to convert the raw data so that the data can be learned by the model. After features are extracted from the dataset, there will be unimodal emotion recognition to classify emotion from given features. As for text data, an embedding layer will be applied before Bidirectional LSTM to understand the context of the text.

IV. Theory and Methods

A. Emotion Recognition

Emotion recognition is one of the topics that has been widely researched along with the development of AI (Artificial Intelligence), the main reason being that the application of emotion classification is carried out on many difficult AI tasks, such as creating dialogues, understanding user behavior, and multimodal interactions [15]. The classification of emotions in a conversation can be used to make a suitable response by analyzing the emotions of the user.

Human emotions become the most important aspect to perform and develop natural human-machine interaction (HMI). While AI is one of the most developed topics in recent years, it'll need data from several modalities. Also, there are many methods to be explored to make AI learn effectively.

B. MFCC

The mel-frequency cepstral coefficient (MFCC) is an interpretation of an audio file in the form of a sequence of numbers [23]. The sequence of numbers is obtained from dividing raw audio files into frames, then performed some steps that are shown in Fig. 5. Lastly, the MFCC array is transposed using arithmetic formula. These MFCC numbers represented human voices amplitude.

C. LSTM

LSTM (Long Short-Term Memory) is one of the Deep Learning models derived from the RNN (Recurrent Neural Network) model which has special properties [24]. LSTM is the proposed solution model after finding the shortcomings of the traditional RNN model, where the RNN has a vanishing gradient and an exploding gradient, causing the RNN to fail to capture long-term dependencies, and as a result, the prediction accuracy of the RNN model decreases.

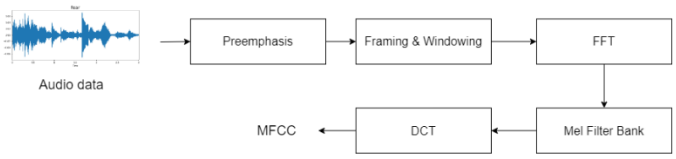

Fig. 5. MFCC process.

The LSTM network architecture treats the hidden layers contained in the neural network as memory units, where the memory units collected in one recurrent hidden layer are called memory blocks. These memory blocks can be used to memorize temporal state from the neural network so that LSTM can remember the correlation between features in a sequence of time.

D. BERT

BERT (Bidirectional Encoder Representations from Transformers) is a new Transformer model for language representation designed to understand the contextual relationship between words from unlabeled text [25]. BERT also penetrated the start-of-the-art because it can see the context of a sentence from two directions, from left to right and from right to left. The BERT model architecture consists of a multi-layer bidirectional Transformers encoder.

BERT can be used to process not only embedding in text data, but also perform classification. The total Transformers blocks contained in the BERT baseline model are 12 blocks, with 768 hidden units, and 12 self-attention heads. The corpus used for pre-training in BERT is Books Corpus (800 million words) and English Wikipedia (2,500 million words).

V. Proposed Method

The proposed method for this research is shown in Fig. 6, Fig. 7, and Fig. 8. There are two models that are used to classify emotion from each modality. For the audio data, Bidirectional LSTM is used to recognize emotion in audio dataset, where the model learns audio features that were already extracted into MFCC.

This SER model's architecture is based on Bidirectional LSTM that was experimented on [15] and [22]. The baseline models have two layers of Bidirectional LSTM. These twp layers can learn the MFCC features very well, better than 1 layer or 3 or more layers, based on the first experiment by the authors. Bidirectional LSTM that are too complex couldn't perform well with the MFCC features, so the Bidirectional LSTM adjusted into two layers. The LSTM unit in the layers are 256 and 128 units, respectively. Then, Dense Layer and Dropout is applied to reduce the complexity of the output from Bidirectional LSTM layers. The SER model architecture is shown in Fig. 6.

On the other hand, Distil-BERT pretrained model is utilized to tokenize the raw words from the text dataset. Then, the tokenized words are fed into an embedding layer. This embedding layer is trained with training dataset to understand

the context from the text data. The output from the embedding layer is inputted to 3-layered Bidirectional LSTM and classifies the emotion.

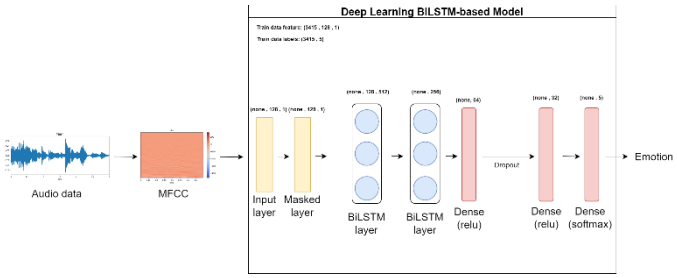

Fig. 6. Proposed SER deep learning model.

This TER model's architecture is based on Bidirectional LSTM that was built in [15] and [22]. There are two layered Bidirectional LSTM that were applied on both experiments. Also, in the [22] there are six layers of BiLSTM and two Dense Layers. But the performance of the 6-layered BiLSTM is worse than the 2-layered BiLSTM. So, the 2-layered BiLSTM can be the baseline model of this experiment. Then, 2-layered BiLSTM is adjusted into 3-layered BiLSTM which can perform better than the 2-layered. Then, a flatten layer and three layers of Dense layers are added to reduce the complexity of the output from the Bidirectional LSTM layers. The TER model architecture is shown in the following Fig. 7.

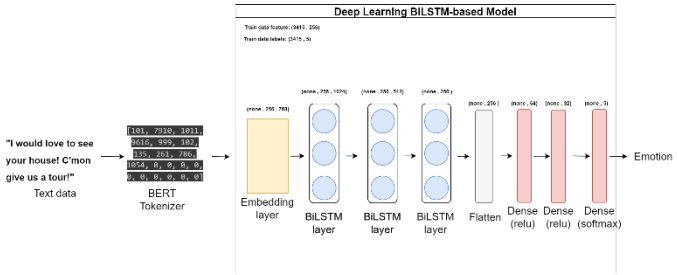

Fig. 7. Proposed TER deep learning model.

Lastly, a voting ensemble method is deployed to classify the final emotion. The voting ensemble model is shown in the following Fig. 8. First, the unimodal models, SER model and TER model are trained. After that, each model, both SER and TER model predict using the features of the test data. The output would be on one-hot encoded format, because of the softmax activation function in the last layer in both models. This softmax function produces numbers that could be interpreted as probability. For example, the prediction generated from the first model is ([0.654 0.346]). The second model generates ([0.781 0.219]). Because the probability from the second model which is 0.781 is higher than 0.654, the voting will choose prediction from the second model.

So, every prediction that is generated from the unimodal models will be compared, which model produces higher probability. If the SER model generates output with higher probability, then the final emotion label will be selected from the SER model's prediction. On the contrary, if TER model generates output with higher probability, then the emotion label will be selected from TER model.

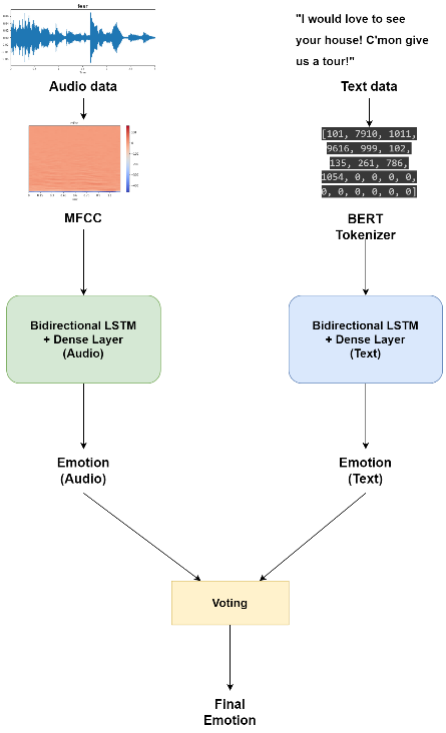

Fig. 8. Proposed voting ensemble learning.

The experiment is implemented using Keras library with the help of librosa library as the audio feature extractor and BERT tokenizer from transformers library built by the Hugging Face. The librosa library is utilized to extract raw audio files into MFCC while the BERT model was used as the word tokenizer in the experiment.

The details of the bidirectional LSTM and deep learning layers which is the proposed method for classifying emotions from audio data, text data, and the voting result from the two modalities are written in the following tables. The details of the SER model are shown in Table I.

On the other side, the construction bidirectional LSTM and deep learning layers for the textual emotion recognition (TER) are written in the following Table II.

TABLE I. BIDIRECTIONAL LSTM LAYERS FOR SER

Bidirectional LSTM				
Layers	**Units**	**Dropout**	**Output Size**	**Activation**
Input layer	1	-	(None, 128, 1)	-
Masking	1	-	(None, 128, 1)	-
Bidirectional LSTM	256	0.4	(None, 128, 512)	tanh
Bidirectional LSTM	128	0.3	(None, 256)	tanh
Dense	64	-	(None, 64)	ReLU
Dropout	0	0.2	(None, 64)	-
Dense	32	-	(None, 32)	ReLU
Dense	5	-	(None, 5)	Softmax

TABLE II. BIDIRECTIONAL LSTM LAYERS FOR TER

Bidirectional LSTM				
Layers	**Units**	**Dropout**	**Output Size**	**Activation**
Embedding		-	(None, 256, 768)	-
Bidirectional LSTM	512	0.1	(None, 256, 1024)	tanh
Bidirectional LSTM	256	0.1	(None, 256, 512)	tanh
Bidirectional LSTM	128	-	(None, 256)	tanh
Flatten		-	(None, 256)	-
Dense	64	-	(None, 64)	ReLU
Dense	32	-	(None, 32)	ReLU
Dense	5	-	(None, 5)	Softmax

VI. RESULT AND DISCUSSION

Results from the experiment provided in Table III presented model's performance on the single modality and multimodality. Since the MELD has already been divided into train and test datasets, the model is being built using train data. The validation dataset takes 15% randomly from the train dataset to show the learning progress of each epoch.

The Bidirectional LSTM for the SER task trained with 30 epochs with batch size 32. While the TER task trained with only 20 epochs using the same batch size, which is 32. Both proposed SER and TER model used the l2 kernel regularizer where the kernel regularizer is set to $1x10^{-4}$. This regularizer is added to reduce or prevent the models that were built in the experiment from overfitting. The optimizer used to train both SER and TER model is Adam. For the TER model, the learning rate was set to $1x10^{-5}$. On the SER model, the learning rate was set to $1x10^{-4}$ because the number of epochs to train SER model is higher compared to TER model. These hyperparameters are manually tuned until receive the best result.

The experiment results are compared to some research papers shown in the following Table III. The benchmark models are trained using the same amount data as the proposed model with the same hyperparameters. The proposed method could surpass the baseline models with small gaps in decimals. If the f1-score is converted from decimal into percent, the proposed model does not surpass more than 1%. The gap between voting performed on model built by [15] and the proposed model approximately 1.5%.

From the evaluation result that is shown in Table III, proposed SER model, TER model, and the vote result based on SER and TER models could surpass the F1-score benchmarks. The emotion labels are strongly related with the context spoken by the speaker. This shows that the embedding layer that has been trained could understand contextual features better compared to MFCC, so that the TER model could perform better in recognizing the emotions given in the dataset.

As for the baseline TER model that was built in [15], an embedding layer is added, so the baseline model could learn the contextual features from tokenized word like model that was built in [22]. Both embedding layers are trained with the same parameter from the proposed model, so that the architecture of Bidirectional LSTM can be evaluated and compared fairly.

The details of the evaluation metrics such as precision, recall, and f1-score for each emotion label are provided in the following Tables IV, V, and VI. Table IV contains the details from SER model, while Table V contains the details from TER model and voting result, respectively.

The metrics that are shown in Table IV above interpret the details of how the proposed SER model predicted each label from the test dataset. As the test data mainly contain neutral emotion, the precision, recall, and f1-score for neutral emotion are good. But, for other emotions such as joy and surprised, the model performed poorly on predicting both emotions.

TABLE III. EXPERIMENT RESULT ON MELD DATASET BASED ON F1-SCORE

Modality	**Approach**	**F1-score**
Audio	BiLSTM [15]	0.408
	BiLSTM [22]	0.402
	BiLSTM (Proposed)	0.417
Text	BiLSTM [15]	0.443
	BiLSTM [22]	0.464
	BiLSTM (Proposed)	0.473
Multimodal	Voting from 2 modalities [15]	0.459
	Voting from 2 modalities [22]	0.468
	Voting from 2 modalities (proposed)	0.475

TABLE IV. EVALUATION METRICS SCORE FOR PROPOSED SER MODEL

Emotion	**Precision**	**Recall**	**F1-score**	**Support**
Anger	0.297	0.183	0.226	345
Joy	0.196	0.067	0.1	402
Neutral	0.521	0.705	0.599	1256
Sadness	0.145	0.255	0.185	208
Surprised	0.145	0.039	0.062	280

TABLE V. EVALUATION METRICS SCORE FOR PROPOSED TER MODEL

Emotion	**Precision**	**Recall**	**F1-score**	**Support**
Anger	0.310	0.446	0.366	345
Joy	0.344	0.236	0.280	402
Neutral	0.746	0.584	0.655	1256
Sadness	0.132	0.284	0.179	208
Surprised	0.476	0.489	0.482	280

The metrics that are shown in Table V interpret the details of how the proposed TER model predicted each label from the test dataset. Just like SER model, the TER model could predict neutral emotion well enough. The difference between the SER and the TER model is how well the TER model predicted surprised emotion, compared to the SER model. Overall, the TER model performed better except in predicting sadness emotion label, but only by inches.

TABLE VI. EVALUATION METRICS SCORE FOR PROPOSED VOTING ENSEMBLE

Emotion	Precision	Recall	F1-score	Support
Anger	0.313	0.452	0.370	345
Joy	0.363	0.246	0.293	402
Neutral	0.745	0.582	0.654	1256
Sadness	0.132	0.288	0.181	208
Surprised	0.477	0.488	0.482	280

The metrics that are shown in Table VI above interpret the details of the performance from two modalities TER and SER combined. Overall, the F1-score in the models is not as good as F1-score on the other dataset, for example research by [14] on WASSA2018 dataset. The main reason of the overall scores only in range 40% of F1-score is because the dataset is greatly imbalanced. To cope the imbalanced label from the train dataset, undersampling was performed so that the training data is much fewer compared to all training data, which is 3415 compared to 9989. So, the models only trained with one third of the full dataset.

Secondly, some of the data also does not contain keywords that are related to the emotion, for example, in the dataset, there is 1 word that is used in different emotions, which is "Hey." This word is used in joy, anger, and neutral expression in the dataset. It creates confusion so the model can't recognize the emotion in each sentence or utterance.

VII. CONCLUSION

This study experimented with a deep learning architecture which is bidirectional LSTM for doing unimodal emotion classification. This study also includes voting ensemble learning to combine two different modalities of the input information.

Experimental results on the MELD dataset demonstrated the good result of the proposed method. The performance of the proposed method could improve the performance from the previous state-of-the-art strategies both [15] and [22] around 1% performance improvement in F1-score on both SER and TER model. However, both unimodal models and the multimodal model performed poor in predicting some emotions such as sadness and joy.

For future studies, other input modalities such as different physiological measurements, or sentiment, or maybe context of the dialogue should also be added and included to the dataset, so that the input for the deep learning model become more complex and richer. Also, the facial features from the video dataset could be utilized as a feature to predict the human's emotions.

The imbalanced dataset also could be tackled by upsampling the dataset, especially on some emotion labels that only have a little data, which leaves some spaces to work in the future. Therefore, more multimodal datasets should be evaluated in future work.

ACKNOWLEDGMENT

The present paper was supported by the University of Bina Nusantara as thesis research to complete the graduate program. The writers are genuinely thankful and honored to those who help with the work on this paper and hopefully can produce many other great research or projects in the future.

REFERENCES

[1] S. H. Park, B. C. Bae, and Y. G. Cheong, "Emotion recognition from text stories using an emotion embedding model," in Proceedings - 2020 IEEE International Conference on Big Data and Smart Computing, BigComp 2020, Feb. 2020, pp. 579–583. doi: 10.1109/BigComp48618.2020.00014.

[2] Z. Lian, Y. Li, J. Tao, and J. Huang, "Investigation of Multimodal Features, Classifiers and Fusion Methods for Emotion Recognition," 2018. [Online]. Available: https://arxiv.org/abs/1809.06225.

[3] A. T. Sohaib, S. Qureshi, J. Hagelbäck, O. Hilborn, and P. J. Jerči´c, "Evaluating Classifiers for Emotion Recognition Using EEG," 2013.

[4] S. Yoon, S. Byun, and K. Jung, "Multimodal Speech Emotion Recognition using Audio and Text," SLT, pp. 112–118, 2018.

[5] L. Cai, Y. Hu, J. Dong, and S. Zhou, "Audio-Textual Emotion Recognition Based on Improved Neural Networks," Math Probl Eng, vol. 2019, 2019, doi: 10.1155/2019/2593036.

[6] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria, "DialogueRNN: An Attentive RNN for Emotion Detection in Conversations," Nov. 2018, [Online]. Available: http://arxiv.org/abs/1811.00405.

[7] M.-H. Su, C.-H. Wu, K.-Y. Huang, and Q.-B. Hong, LSTM-based Text Emotion Recognition Using Semantic and Emotional Word Vectors, ACII Asia. 2018.

[8] J. L. Wu, Y. He, L. C. Yu, and K. Robert Lai, "Identifying Emotion Labels from Psychiatric Social Texts Using a Bi-Directional LSTM-CNN Model," IEEE Access, vol. 8, pp. 66638–66646, 2020, doi: 10.1109/ACCESS.2020.2985228.

[9] C. Li, Z. Bao, L. Li, and Z. Zhao, "Exploring temporal representations by leveraging attention-based bidirectional LSTM-RNNs for multi-modal emotion recognition," Inf Process Manag, vol. 57, no. 3, May 2020, doi: 10.1016/j.ipm.2019.102185.

[10] A. Vaswani et al., "Attention Is All You Need," Jun. 2017, [Online]. Available: http://arxiv.org/abs/1706.03762.

[11] S. Padi, S. O. Sadjadi, D. Manocha, and R. D. Sriram, "Multimodal Emotion Recognition using Transfer Learning from Speaker Recognition and BERT-based models," Feb. 2022, [Online]. Available: http://arxiv.org/abs/2202.08974.

[12] K. Pham, D. Kim, S. Park, and H. Choi, "Ensemble learning-based classification models for slope stability analysis," Catena (Amst), vol. 196, Jan. 2021, doi: 10.1016/j.catena.2020.104886.

[13] A. Verma and S. Shikha Mehta, A Comparative Study of Ensemble Learning Methods for Classification in Bioinformatics. 2017.

[14] Q. Zhou, Z. Zhang, and H. Wu, "BiLSTM-Attention and LSTM-Attention via Soft Voting in Emotion Classification," in WASSA 2018 - 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, Proceedings of the Workshop, 2018, pp. 149–155. doi: 10.18653/v1/P17.

[15] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations," Oct. 2018, [Online]. Available: http://arxiv.org/abs/1810.02508.

[16] D. Deng, Y. Zhou, J. Pi, and B. E. Shi, "Multimodal Utterance-level Affect Analysis using Visual, Audio and Text Features," May 2018, [Online]. Available: http://arxiv.org/abs/1805.00625.

[17] Y. Gu, S. Chen, and I. Marsic, "Deep Multimodal Learning for Emotion Recognition in Spoken Language," 2018.

[18] G. Sahu, "Multimodal Speech Emotion Recognition and Ambiguity Resolution," Apr. 2019, [Online]. Available: http://arxiv.org/abs/1904.06022.

[19] N. H. Ho, H. J. Yang, S. H. Kim, and G. Lee, "Multimodal Approach of Speech Emotion Recognition Using Multi-Level Multi-Head Fusion Attention-Based Recurrent Neural Network," IEEE Access, vol. 8, pp. 61672–61686, 2020, doi: 10.1109/ACCESS.2020.2984368.

[20] Z. Lian, J. Tao, B. Liu, J. Huang, Z. Yang, and R. Li, "Conversational emotion recognition using self-attention mechanisms and graph neural networks," in Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2020, vol. 2020-October, pp. 2347–2351. doi: 10.21437/Interspeech.2020-1703.

[21] S. Tang, Z. Luo, G. Nan, Y. Yoshikawa, and I. Hiroshi, "Fusion with Hierarchical Graphs for Mulitmodal Emotion Recognition," Sep. 2021, [Online]. Available: http://arxiv.org/abs/2109.07149.

[22] B. Tris Atmaja, K. Shirai, and M. Akagi, "Speech Emotion Recognition Using Speech Feature and Word Embedding," 2019.

[23] M. G. de Pinto, M. Polignano, P. Lops, and G. Semeraro, Emotions Understanding Model from Spoken Language using Deep Neural Networks and Mel-Frequency Cepstral Coefficients. 2020 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS), 2020.

[24] Z. Zhao, W. Chen, X. Wu, P. C. Y. Chen, and J. Liu, "LSTM network: A deep learning approach for Short-term traffic forecast," IET Intelligent Transport Systems, vol. 11, no. 2, pp. 68–75, Mar. 2017, doi: 10.1049/iet-its.2016.0208.

[25] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Oct. 2018, [Online]. Available: http://arxiv.org/abs/1810.04805.