BIMODAL EMOTION RECOGNITION USING TEXT AND SPEECH WITH DEEP LEARNING AND STACKING ENSEMBLE TECHNIQUE

STEPHEN WILLIAM[1], AMALIA ZAHRA[2]

[1,2] Computer Science Department, BINUS Graduate Program, Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480

E-mail: [1] stephen.william@binus.ac.id, [2] amalia.zahra@binus.edu

ABSTRACT

Understanding human emotion means communicating with our fellow humans on a deeper level. Unfortunately, understanding human emotion is not as easy as it sounds. As humans, we express our emotions in different ways, be it using our tone of speech, words of choosing, or facial expressions. To just pick one over the many ways we express emotion and draw a conclusion would mean that we lose out on information causing us to arrive at the wrong conclusion. This research shows that fusing two modalities, text and speech, with a stacking ensemble method, shows leap and bounds improvements in its accuracy when compared to the unimodal approach. Tested on the IEMOCAP dataset with a 4 emotions subset of anger, happy, sad, and neutral, the text model of BERT (Bidirectional Encoder Representations from Transformer) managed to achieve an accuracy score of 65.4%, and the audio model of CNN (Convolutional Neural Network) + Bi-LSTM (Bidirectional Long Short-Term Memory) with the implementation of LFLB (Local Feature Learning Block) managed to achieve an accuracy score of 60.6%. These results were then combined into one with a stacking ensemble method and achieved an accuracy of 75.181%.

Keywords: *Audio Processing, Deep Learning, Ensemble Technique, Emotion Recognition, Natural Language Processing*

1. INTRODUCTION

Emotion has been a rather abstract concept for a long time, with many able practitioners classifying the elicited emotions into their understanding of the concept such as Ekman's six basic emotions and Parrot's six basic emotions. Abstract as it may be, to understand emotion means to communicate with our fellow humans on a deeper level than we already are, thus we can infer their needs and the actual meaning of their words and not just take them at face value. The study that teaches machines to recognize the humans' elicited emotion is called Emotion Recognition, and these systems can help in many different aspects of life, such as depression detection [1], gaining insights into the citizens' point of view during an election [2], and even detecting stress during an emergency call to help determine the legitimacy of the call [3].

But teaching machines so that the machine can recognize emotion is not that easy. Humans express their emotions in a lot of different ways, the words they chose, the volume of their voice, and the facial expression that they are expressing. To make a better system, we can make use of these channels that humans use to express their emotions and make a multimodal system rather than just use one of the channels, be it text, audio, or facial expression. A search on Google Scholar yields results as shown in Figure 1 with the query being "emotion recognition" concatenated with each

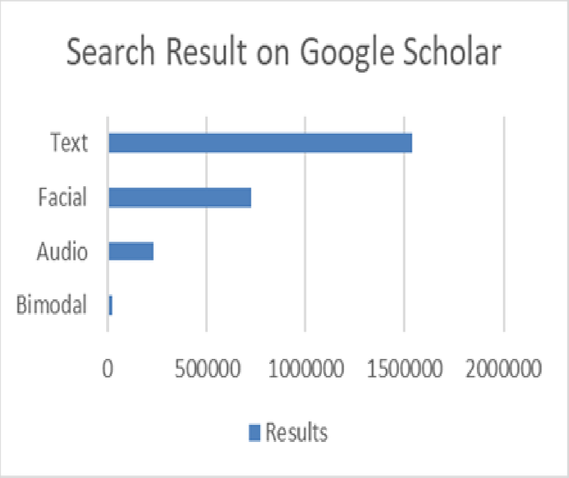


Figure 1: Search Result on Google Scholar Using the Corresponding keywords

keyword shown in the figure. It shows that bimodal emotion recognition results in the least number of results even though results show that the bimodal model achieves a better result than that of unimodal models [4][5].

This research will present a new bimodal system that makes use of BERT as its text model, and a combination of CNN and Bi-LSTM with the implementation of LFLB for its audio model. Then, these individual models will be fused into one bimodal model with the help of the stacking ensemble technique. The research's main contribution is the audio model that implements the LFLB technique with an improvement Bi-LSTM as opposed to the older LSTM, and the usage of stacking ensemble, an ensemble technique that has never been used in this field of study before. The designed architecture is then trained and tuned with 5-fold cross-validation and managed to achieve a text recognition accuracy of 65.4%, audio recognition accuracy of 60.6%, and an ensemble accuracy of 75.181%.

2. RECENT WORK

A study on the field of audio emotion recognition was done by the use of a CNN-LSTM network that implemented a novel technique that they called Local Feature Learning Block or LFLB for short [6]. Their experiment begins by preprocessing the raw audio file into computer-readable features, in this case, they have chosen mel spectrogram as their feature, as it carries the information needed to make the deduction. As Mel spectrogram is a feature that can be represented in the form of an image, CNN will be able to work at its best to help learn the local dependencies between each pixel, whereas the LSTM network will then be used to help learn the long-term dependencies in the image. This method yields a result of 52.14% on the IEMOCAP (Interactive Emotional dyadic Motion Capture Database) dataset [7].

In the field of textual emotion recognition, a transfer learning model that goes by the name BERT (Bidirectional Encoder Representations of Transformers), was released in 2018 [8]. It is a new transformer-based model that the Google Brain team developed upon the transformer encoder-decoder architecture. It is a modified version of the transformer architecture [9], that works on the idea of rather than having an encoder-decoder architecture, it ditches the decoder and stacks twelve of the encoder upon each other and implemented a self-attention mechanism to help gather up the contextual information needed. It outperforms all the specifically tailored solutions for each NLP task on the GLUE benchmark [10] as a generalized solution that only needs to be fine-tuned in the targeted downstream.

Ensemble methods are statistical and computational learning procedures reminiscent of the human social learning behavior of seeking several opinions before making any crucial decisions [11]. Signals from different modalities often carry complementary information about different aspects of an object, event, or activity of interest [12]. From this data, we can surmise that different modalities can carry different information, such as speech will carry with it the intonation of its speaker and video will carry the speaker's facial expression. By fusing these different modalities, we can create a model that is better at predicting the answer to a question when compared to single modality models. This fusing of the modality can be achieved through means of ensemble methods.

Stacked generalization [13] is an ensemble of a diverse group of models that introduces the concept of a meta-learner. The meta-learner is a higher-level model that will learn the features that are predicted from lower-level models. The step for implementing this technique are as follows: 1). Split the dataset into training and testing sets. 2). Train the low-level learner on the training set. 3). Use the predictions generated by the low-level learner to train the meta-learner. Figure 2 depicts the general flow of the stacking ensemble technique. In the figure presented below, there are many models used that are trained with the new data. Then, these models will output their predictions that will be used to train the meta-learner. Finally, the meta-learner will generate its final prediction based on its training.

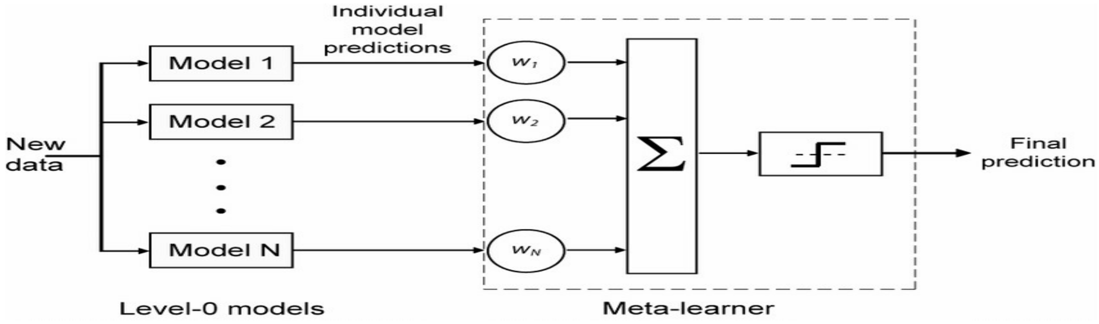


Figure 2: Stacking Ensemble

Moving over to the fields of bimodal recognition, the IEmoNET model was released in 2019 [5]. IEmoNet is a very flexible system that is divided into 5 parts, which is: pre-processing, ASR system, TER system, SER system, and classification. These 5 systems can be optimized individually without dependencies on the other system. It is a model that made use of the IEMOCAP dataset and is trained on two modalities, text, and audio. It uses the BERT model and also a CNN-LSTM architecture to train their respective text and audio model. For the audio model, it made use of the ISO9 feature set [14]and it does not implement the LFLB technique mentioned earlier. They then combine the two results and get a result of 73.5%.

Other than IEmoNet, a model called STSER was released in 2020 [15]. The model also uses text and audio features for its features and is also tested on the IEMOCAP dataset. A smodel will be representing the speech information that will be used later. It takes advantage of CNN, Bi-LSTM, and the attention mechanism in general. A tmodel that represents the textual feature is created using ALBERT (A Lite BERT). In the end, the models will be fused using a multi-scale fusion strategy and average ensemble technique to improve overall accuracy. This model is tested on the IEMOCAP dataset and has outperformed the previously mentioned IEmoNet on the unweighted accuracy score, with a score of 72.05%.

From the review that has been done, we can now articulate a problem statement as follows: What needs to be done to improve the overall accuracy of the dual-modality model? This research will aim to answer the said question by implementing a stacked generalization method for the ensemble model, as opposed to the more traditional average or voting, and by improving the unimodal model accuracy itself.

3. RESEARCH METHODOLOGY

3.1 Data Gathering

While there are many datasets available for use of emotion recognition, there are not many that have good quality. After some review, the dataset was narrowed down to two: the RAVDESS emotional dataset [16] and the IEMOCAP dataset [7]. In the end, the IEMOCAP dataset was chosen for its robustness, abundance amount of research done with it, and also the fact that the IEMOCAP dataset is a semi-natural database, as opposed to the RAVDESS simulated database. With the research focusing on creating a bimodal text and audio model, the RAVDESS dataset cannot be used for its simulated nature, in which an actor or actress acts out a certain scenario in which the emotion elicited was given to them beforehand. Not only that, the RAVDESS dataset only uses one sentence for all emotions elicited, whereas the IEMOCAP dataset is an acted-out scenario that is divided into a scripted scenario and an improvised scenario. Another factor is the robustness and number of research that has been done with the IEMOCAP dataset [17][18][19].

3.2 Dataset

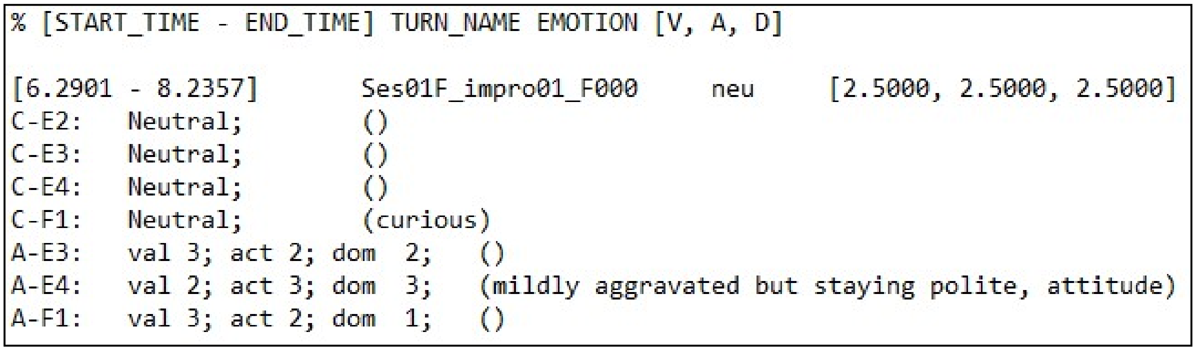
The chosen IEMOCAP dataset is a semi-natural dataset, this means that this dataset is comprised of acted scenarios where the actors are not given a script that is down to every word, but a scenario of how the scene should go. It has over 12 hours of emotional audio dialogues which are acted by 10 professional actors, 5 being male and the other half female. The data are split into two types of dialogue which is categorized to improvised and scripted. For the improvised dialogues, the actors are asked to improvise from a hypothetical scenario that has been designed to elicit different emotions. The actors were given scripts with clear emotional content to read for the scripted category.

The acted dialogues were then evaluated by a group of evaluators using the 6-basic emotion plus frustration, excited, and neutral on an utterance level (a sentence or a word spoken by the actors on each turn). Each utterance is evaluated by three evaluators, these evaluations were then used to determine the ground truth of an utterance. An evaluation is accepted as ground truth only if a minimum of two evaluators agreed on the emotion elicited in the utterance.

Figure 3 depicts the example of labelled data from the IEMOCAP dataset. The START_TIME – END_TIME depicts the start time and end time of the utterance on a single audio file. The TURN_NAME denotes the current speaker utterance on the specified timeframe on a single audio file. The emotion denotes the agreed-upon emotion that has a minimum of two evaluators agreeing on a single emotion. The last V, A, and D denotes the valence, activation, and dominance which will not be used in this research.

3.3 Data Pre-Processing

For the text dataset, the first step is getting the total number of data that is available for us in the IEMOCAP dataset. After the total data is obtained, the discrepancy in the data distribution becomes visible. The discrepancy is shown in **Error! Reference source not found.**. It is visible from the data distribution that there is a data discrepancy between the available data, such as neutral with 1708 data, and fearful which is only a meager 40 total data. To handle the data discrepancy, previous research that also used the IEMOCAP dataset and classifies the emotion into four classes, Angry, Happy, Sad, and Neutral, with the happy class being the merged result of the happy and excited class, was followed [5][15]. The total data following the process tallies to 5,531.



```
% [START_TIME - END_TIME] TURN_NAME EMOTION [V, A, D]

[6.2901 - 8.2357]      Ses01F_impro01_F000     neu     [2.5000, 2.5000, 2.5000]
C-E2:   Neutral;        ()
C-E3:   Neutral;        ()
C-E4:   Neutral;        ()
C-F1:   Neutral;        (curious)
A-E3:   val 3; act 2; dom  2;   ()
A-E4:   val 2; act 3; dom  3;   (mildly aggravated but staying polite, attitude)
A-F1:   val 3; act 2; dom  1;   ()
```

Figure 3: IEMOCAP Sample Data

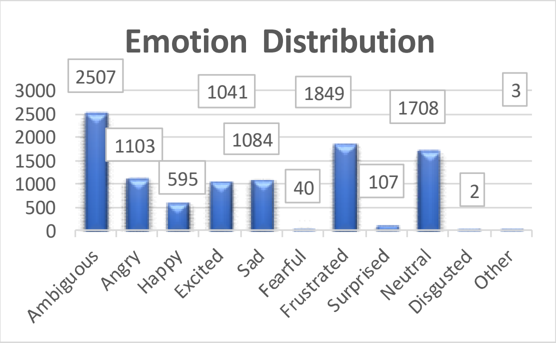


Figure 4: IEMOCAP Dataset Emotion Distribution

For the text dataset, we applied a lowercase to all text and removed all the special characters in the hope that the model will not be learning any of these as an emotion distinguishing feature. Then we will use the BERT tokenizer to change the sentences to something that the model understands.

For the audio file, a normalization will be done. Since the audio data have different duration, we will need to normalize them to get cleaner data that will not confuse our model when we use it to train and test. The audio file will firstly be padded or truncated until the file is 8 seconds long. If it is longer than 8 seconds, it will be stopped at 8 seconds, but if it is less than 8 seconds, the file will be padded with silent audio. The example of a truncated file is shown in Figure 6 and the example of the padded file is shown in Figure 7.

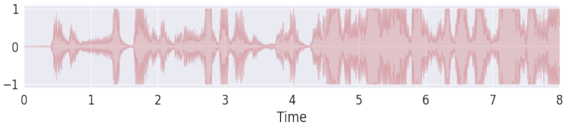


Figure 6: Truncated Audio File

Then, the data will be converted to a mel-spectrogram with the help of the Python librosa library [20]. Parameters used for the conversion will be following previous research on the same SER field with the Fast Fourier Transform (FFT) window set to 2.048 along with the hop length to 512 and Mel frequency bin to 128 [6]. The resulting mel-spectrogram is shown in Figure 5.

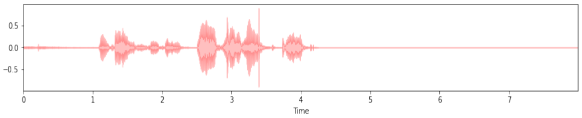


Figure 7: Padded Audio File

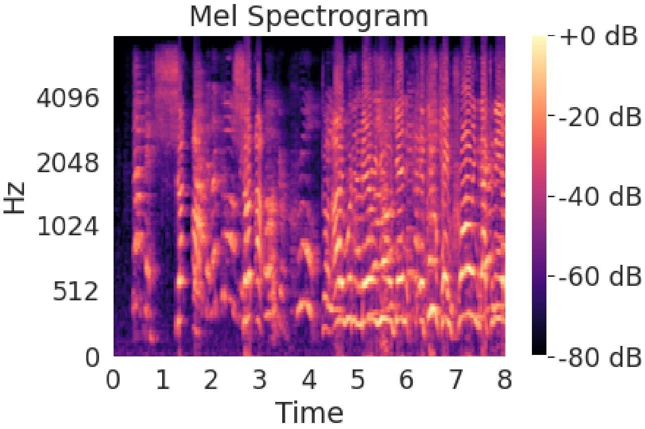


Figure 5: Mel-Spectrogram

3.4 Modeling

This research proposes a bimodal model for audio and textual emotion recognition, with the architecture presented in **Error! Reference source not found.**. The BERT model that is already available in Tensorflow Hub will be used as the Text Emotion Recognition model and modified by adding a drop-out layer to avoid overfitting. For the Speech Emotion Recognition, we will be using the model that was proposed by Zhao [6] and improves it by using Bidirectional LSTM [21], rather than just LSTM. This is done in the hope that the model will be able to predict better after they learn left-to-right dependencies and right-to-left dependencies.

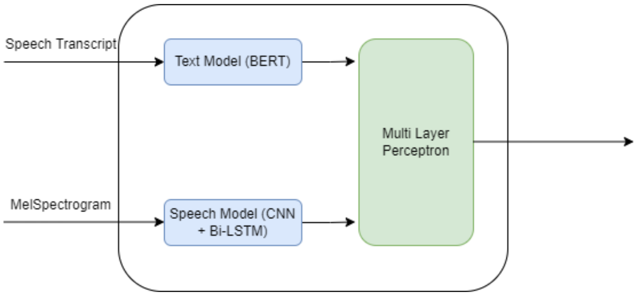


Figure 8: Proposed Architecture

The proposed model is presented in Table 1. 4 LFLBs stacked upon each other will be used, with the first LFLB kernel size and stride being 2 x 2, and the rest being 4 x 4. For the ensemble model, an implementation of the stacking ensemble technique will be used. The technique will allow us to stack the prediction result of the text architecture and the audio architecture into one result, and the result will be used to train our chosen ensemble model, which in this case is a multi-layer perceptron. A multi-layer perceptron is chosen because there is no need for a complex model, as the resulting dataset itself is not that complex.

Table 1: CNN and Bi-LSTM Architecture

Layer	Kernel Size	Stride	
Conv2D	3 x 3	1 x 1	LFLB x 4
BatchNormalization	-		
Activation			
MaxPooling2D	2 x 2 for first 4 x 4 for the rest	2 x 2 for first 4 x 4 for the rest	
BidirectionalLSTM	-		
Dense			

4. EXPERIMENTS & RESULTS

The experiment will begin by splitting the data to get the training set and the validation set for both the unimodal and the bimodal model. The data splitting will follow the scenario shown in Figure 9. All experiments are done with a 5-fold validation to ensure stability, with the total 5,531 data broken down to 85% for the unimodal dataset, and 15% for the bimodal dataset. Then, the 85% unimodal dataset is further broken down to an 80:20 ratio for training and testing purposes. Since we will be using the result of the unimodal prediction with its test dataset, there will be no further need to break down the bimodal dataset.

The experiment was conducted several times with different settings used for each experiment to find the most optimal settings for the architecture. The result for the unimodal textual 5-fold evaluation is presented in the below table, with its comparison against other research.

From the result shown in Table 2, it is

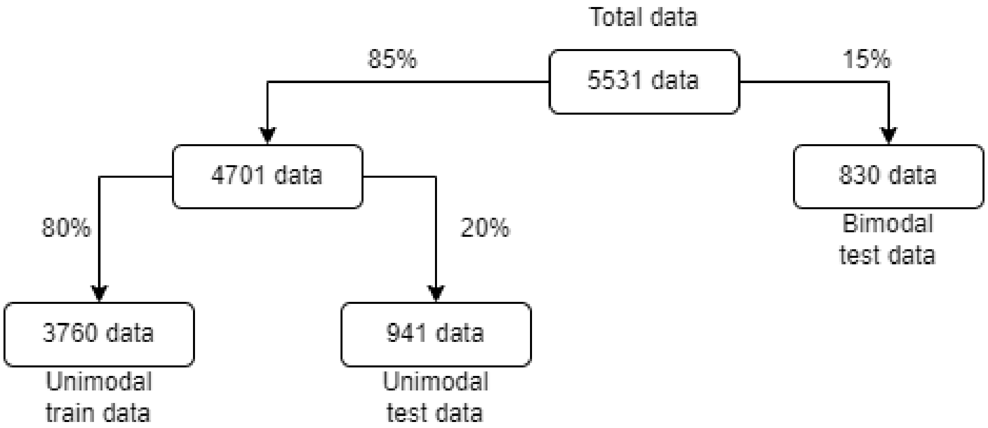


Figure 9: Data Splitting Scenario

visible that the text model proposed in this paper does not have the best results when compared to the state-of-the-art model that was reviewed. This is fine because this research intends to leverage the stacking ensemble method to increase the final overall accuracy of the model during its bimodal experiment.

Table 2: Text Model Comparison

Model	Accuracy
Proposed Text Model	**65.4%**
IEmoNet – BERT [5]	70.9%
IEmoNet – XLNet [5]	69.4%
MDREA - TRE [22]	63.5%
LSTM [23]	64.8%

he result of the experiment for the 5-fold validation of the audio model is presented in

below along with its comparison against other research. This time, it is visible that our model is one of the best performing out of all the model that was compared, this is due to the implementation of LFLB along with Bi-LSTM to give the model more context. The experiment was implemented with an early stopping with 8 as its patience because of LSTM's deep learning nature that may take many epochs to get to its best performance. The resulting model managed to go toe to toe with some of the best results that existing research proposes, and even beat them.

Table 3: Audio Model Comparison

Model	Accuracy
Proposed Audio Model	**60.6%**
MDREA - ARE [22]	54.6%
BiLSTM-64attDim [5]	60.00%
BiLSTM-128attDim [5]	60.8%
STSER – Audio [15]	53.35%

With both the best performing unimodal model chosen, test data was fed into it and the resulting prediction data are stacked and used to train the bimodal model which is just a logistic regression model. The resulting 5-fold validation of the model with its training data is 74.072% with a standard deviation of 1.040%. The model was then tested with the pre-separated data and yielded a result of 75.181%.

The resulting confusion is presented in Figure 10. The matrix shows that our model is best at predicting the neutral emotion with the highest number of true positive and is weakest when predicting both the sad and the angry emotions.

To dive even further into the result, a comparison of the recall, precision, and F1 scores is presented in Figure 11 below. Here, the angry class continues its trend of being the class with the highest recall, as the audio model has angry as its highest recall class, and the text model has angry class recall as the second highest. The neutral class having the lowest numbers out of all the four classes follows the research that states that the neutral class is located right in the center of activation-valence space, further complicating its discrimination from other classes [24].

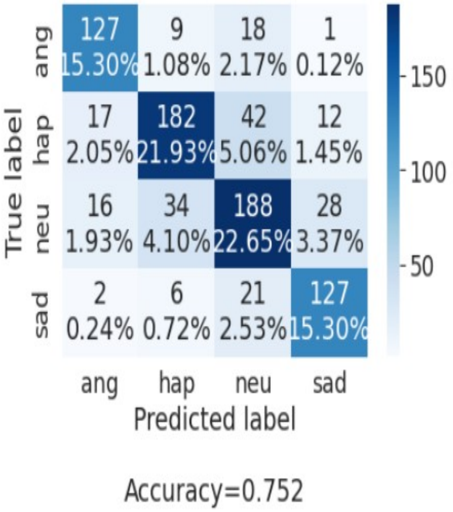


Figure 10: Bimodal Model Confusion Matrix

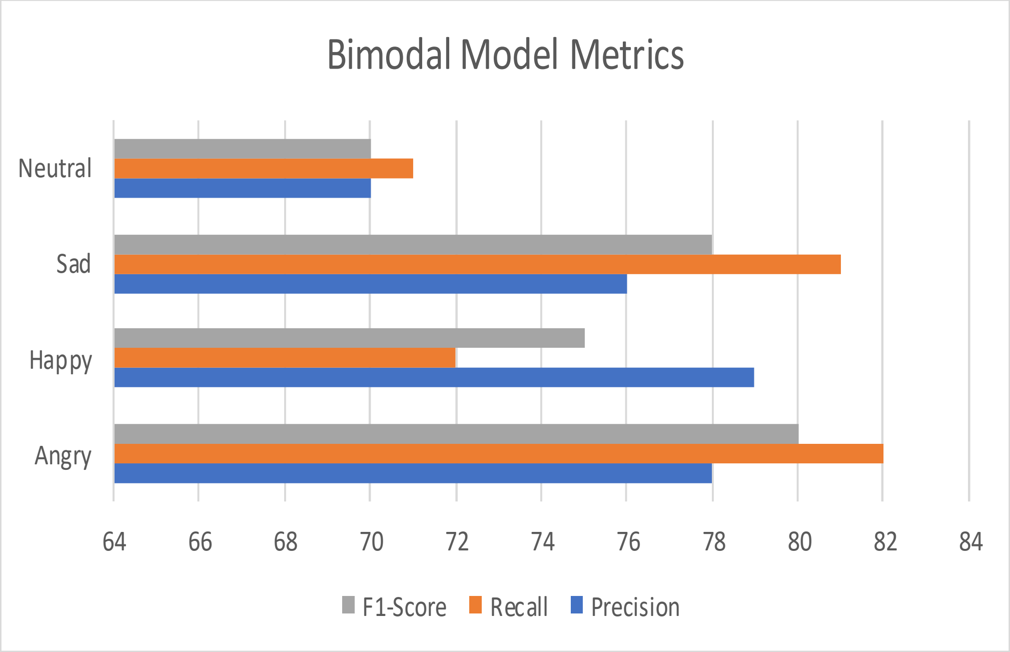


Figure 11: Bimodal Model Metric After Tested on Testing Dataset

Comparisons between the proposed model and its other bimodal counterpart are also presented in Table 4 below. Many components differentiate the experiments in the table below, such as data splitting, unimodal models used, and the feature set that is used by the audio model. Even though the individual unimodal models have lower accuracy our proposed model achieved an overall better score in terms of its bimodal performance. But the edge that the proposed model has when compared to other models is the stacking ensemble technique implemented, as it combines the result of both unimodal models and stacks it to make cohesive data, rather than just using whichever is higher.

Table 4: Bimodal Model Comparison

Model	Accuracy
Proposed Bimodal Model	**75.181%**
IEmoNet[5]	73.5%
STSER[15]	72.05%
MDREA [22]	71.8%

5. CONCLUSION

Albeit emotion recognition is not an easy task, improvement in this field will greatly improve and push human-computer interaction for the better. This research shows that just by changing up the ensemble method, we can see great improvement in the bimodal model. The model in this research managed to achieve an accuracy score of 75.181% which surpasses that of IEmoNet's 73.5%, STSER's 72.05%, and MDREA's 71.8%.

This research was done with limitation on device computing power, as this research only makes use of the Google Colab free feature. It is also done under the assumption that using the same dataset (i.e., IEMOCAP) to train both models will make the model more robust, rather than using a different dataset to train each modality.

As the field of emotion detection is still a very fresh topic, there are many possible ways to continue this research, such as the implementation of better unimodal models, for example, XLNet for the unimodal text. Not only that, but an implementation of a better ensemble technique may also show leaps and bound improvement in the model's accuracy as shown in this research.

Another thing to implement may be the third modality, the facial expression elicited during the moment of utterance. With it, the multimodal model will be more robust and be better at predicting the elicited emotion classes.

REFERENCES:

[1] N. Cummins, J. Epps, M. Breakspear, and R. Goecke, "An investigation of depressed speech detection: Features and normalization," Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, pp. 2997–3000, 2011.

[2] W. Budiharto and M. Meiliana, "Prediction and analysis of Indonesia Presidential election from Twitter using sentiment analysis," J. Big Data, vol. 5, no. 1, 2018, doi: 10.1186/s40537-018-0164-1.

[3] I. Lefter, L. J. M. Rothkrantz, D. A. Van Leeuwen, and P. Wiggers, "Automatic stress detection in emergency (telephone) calls," Int. J. Intell. Def. Support Syst., vol. 4, no. 2, p. 148, 2011, doi: 10.1504/IJIDSS.2011.039547.

[4] 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 Conf. Affect. Comput. Intell. Interact. ACII Asia 2018, 2018, doi: 10.1109/ACIIAsia.2018.8470378.

[5] V. Heusser, N. Freymuth, S. Constantin, and A. Waibel, "Bimodal speech emotion recognition using pre-trained language models," arXiv, 2019.

[6] Z. J., M. X., and C. L., "Speech emotion recognition using deep 1D & 2D CNN LSTM networks," Biomed. Signal Process. Control, vol. 47, pp. 312–323, 2019, [Online]. Available: http://www.embase.com/search/results?subaction=viewrecord&from=export&id=L2001092397%0Ahttp://dx.doi.org/10.1016/j.bspc.2018.08.035.

[7] C. Busso et al., "IEMOCAP: Interactive emotional dyadic motion capture database," Lang. Resour. Eval., vol. 42, no. 4, pp. 335–359, 2008, doi: 10.1007/s10579-008-9076-6.

[8] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018.

[9] A. Vaswani et al., "Attention is all you need," arXiv, 2017.

[10] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "Glue: A multi-task benchmark and analysis platform for natural language understanding," 7th Int. Conf. Learn. Represent. ICLR 2019, 2019.

[11] R. Matteo and V. Giorgio, "Ensemble methods: a review," Data Min. Mach. Learn. Astron. Appl., pp. 3–40, 2001.

[12] K. Liu, Y. Li, N. Xu, and P. Natarajan, "Learn to combine modalities in multimodal deep learning," arXiv, 2018.

[13] D. H. Wolpert, "Stacked generalization," Neural Networks, vol. 5, no. 2, pp. 241–259, Jan. 1992, doi: 10.1016/S0893-6080(05)80023-1.

[14] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge," Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, pp. 312–315, 2009, doi: 10.21437/interspeech.2009-103.

[15] M. Chen and X. Zhao, "A multi-scale fusion framework for bimodal speech emotion recognition," Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, vol. 2020-Octob, pp. 374–378, 2020, doi: 10.21437/Interspeech.2020-3156.

[16] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north American english," PLoS One, vol. 13, no. 5, 2018, doi: 10.1371/journal.pone.0196391.

[17] B. J. Abbaschian, D. Sierra-Sosa, and A. Elmaghraby, "Deep Learning Techniques for Speech Emotion Recognition, from Databases to Models," Sensors, vol. 21, no. 4, p. 1249, Feb. 2021, doi: 10.3390/s21041249.

[18] G. Sahu, "Multimodal Speech Emotion Recognition and Ambiguity Resolution," arXiv, 2019.

[19] Mustaqeem, M. Sajjad, and S. Kwon, "Clustering-Based Speech Emotion Recognition by Incorporating Learned Features and Deep BiLSTM," IEEE Access, vol. 8, pp. 79861–79875, 2020, doi: 10.1109/ACCESS.2020.2990405.

[20] B. McFee et al., "librosa: Audio and Music Signal Analysis in Python," Proc. 14th Python Sci. Conf., pp. 18–24, 2015, doi: 10.25080/majora-7b98e3ed-003.

[21] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," IEEE Trans. Signal Process., vol. 45, no. 11, pp. 2673–2681, 1997, doi: 10.1109/78.650093.

[22] S. Yoon, S. Byun, and K. Jung, "Multimodal Speech Emotion Recognition Using Audio and Text," 2018 IEEE Spok. Lang. Technol. Work. SLT 2018 - Proc., pp. 112–118, 2019, doi: 10.1109/SLT.2018.8639583.

[23] S. Tripathi, S. Tripathi, and H. Beigi, "Multi-Modal Emotion recognition on IEMOCAP Dataset using Deep Learning," 2018, [Online]. Available: http://arxiv.org/abs/1804.05788.

[24] M. Neumann and N. T. Vu, "Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2017-Augus, pp. 1263–1267, 2017