1 Introduction

We used a preprocessed subset of the IEMOCAP dataset, which provides text extracted from audio conversations along with speaker information. The dataset contains dialogs split into five sessions, with a total of 5531 utterances labeled with one of four emotions. To maintain class balance, we applied a stratified split across training, validation, and test sets. The following table presents the distribution of utterances across each emotion category:

Class	Train	Validation	Test
Angry	660	165	221
Happy	980	245	327
Neutral	1025	256	342
Sad	651	163	217

Table 1: Emotion distribution across the dataset.

2 Model Architecture

We implemented two types of models to capture both sentence-level and dialog-level emotional context. Both models employ Recurrent Neural Networks (RNNs) and attention mechanisms to capture the dependencies between words, sentences, and dialogues.

2.1 Sentence-Level Model

The sentence-level model processes individual sentences and classifies them into one of the four emotion categories. The model architecture is as follows:

- **Word-Level Encoding:** Each word in a sentence is embedded using a randomly initialized embedding layer of size 100. A Bidirectional LSTM with 128 hidden units processes the embedded words to capture both forward and backward dependencies, producing a sentence representation.
- **Classification Layer:** The final sentence representation is passed through a dense layer with a softmax activation function to classify the sentence into one of the four emotion categories: *angry*, *happy*, *neutral*, and *sad*.

2.2 Dialog-Level Model

The dialog-level model is designed to capture the interactions and dependencies across multiple sentences within a conversation. We experimented with two window sizes, 5 and 10, for context length. The architecture is described below:

- **Sentence Encoding:** Each sentence is first encoded using a Bidirectional LSTM, similar to the sentence-level model. This produces a sentence representation for each utterance in the dialog.
- **Dialog Encoding:** The sentence representations are passed through a GRU with 128 hidden units. This step captures the sequential dependencies across sentences in the dialog, allowing the model to consider both past and present information when making predictions.
- **Self-Attention Mechanism:** A self-attention layer is applied to focus on the most relevant utterances in the dialog. The attention mechanism allows the model to weigh the importance of each sentence based on its relevance to the overall emotion of the dialog.
- **Classification Layer:** The final attention-weighted dialog representation is passed through a dense layer with softmax activation to classify the dialog into one of the four emotion categories.

3 Implementation Details

We implemented the models using TensorFlow and Keras. Key aspects of the implementation include:

- **Embedding Layer:** We used an embedding size of 100 with a vocabulary size of 5000 words.
- **LSTM Layer:** A Bidirectional LSTM with 128 hidden units captures word-level information in both the sentence-level and dialog-level models.
- **GRU Layer:** For the dialog-level model, a GRU layer with 128 hidden units captures the relationships across multiple sentences.
- **Self-Attention Layer:** A self-attention mechanism is used in the dialog model to focus on relevant portions of the dialog, improving the classification by considering dialog context.

The dataset was split into 80% for training, 10% for validation, and 10% for testing. To address class imbalance, we computed class weights based on the training distribution.

4 Results and Observations

4.1 Sentence-Level Model Results

For the sentence-level model, we achieved an accuracy of 60%. Table below summarizes the class-wise precision, recall, and F1-scores:

Class	Precision	Recall	F1-Score
Angry	0.62	0.60	0.61
Happy	0.75	0.60	0.67
Neutral	0.55	0.61	0.57
Sad	0.53	0.62	0.57

Table 2: Sentence-level classification report.

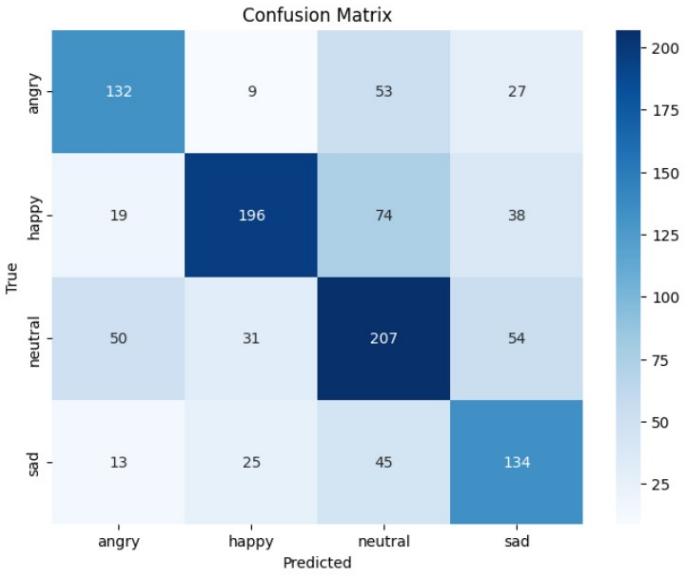

Figure 1: Sentence-level Confusion Matrix

Observation: The model performed best on the *happy* class, with a precision of 0.75, while the *neutral* class had the lowest performance with an F1-score of 0.57. The confusion matrix in Figure 1 shows misclassifications between *neutral* and *sad* as the most common errors.

4.2 Dialog-Level Model Results

We experimented with two sliding window sizes for the dialog-level model: 5 and 10.

Experiment 1: Window Size 5

- With a window size of 5, the model achieved an overall accuracy of 60%, closely matching the performance of the sentence-level model. However, the additional context provided by the dialog allowed for slight improvements, particularly in the angry class, where precision increased to 0.62. This suggests that even a smaller window of dialog context can help the model better differentiate between certain emotions, enhancing its ability to interpret emotional cues beyond isolated sentences.

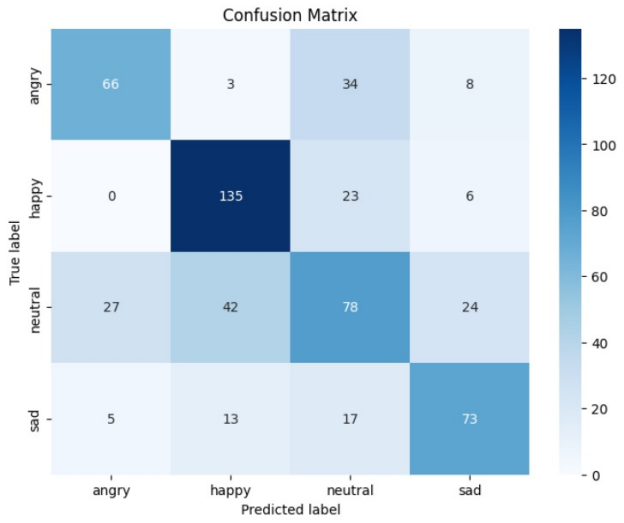

Figure 2: Dialog-level Confusion Matrix (Window Size 5)

Classification Report:

	precision	recall	f1-score	support
angry	0.67	0.59	0.63	111
happy	0.70	0.82	0.76	164
neutral	0.51	0.46	0.48	171
sad	0.66	0.68	0.67	108
accuracy			0.64	554
macro avg	0.64	0.64	0.63	554
weighted avg	0.63	0.64	0.63	554

Figure 3: Dialog-level Classification Report (Window Size 5)

Experiment 2: Window Size 10

- With a window size of 10, the model significantly improved its ability to capture the context of conversations, resulting in a notable boost in performance. The recall for both the angry and happy classes saw substantial increases, rising to 0.77 and 0.88, respectively. This indicates that with more dialog history, the model became better at correctly identifying these emotions, which often depend on surrounding context for accurate interpretation.
  Additionally, the larger window helped improve the F1-scores, reflecting the model's ability to balance both precision and recall. The overall accuracy rose to 70%, demonstrating that more context enables the model to better understand the flow of conversation and differentiate between emotions more effectively. While the neutral class remained challenging, the increased context did help reduce some confusion between emotions, further improving the model's robustness.

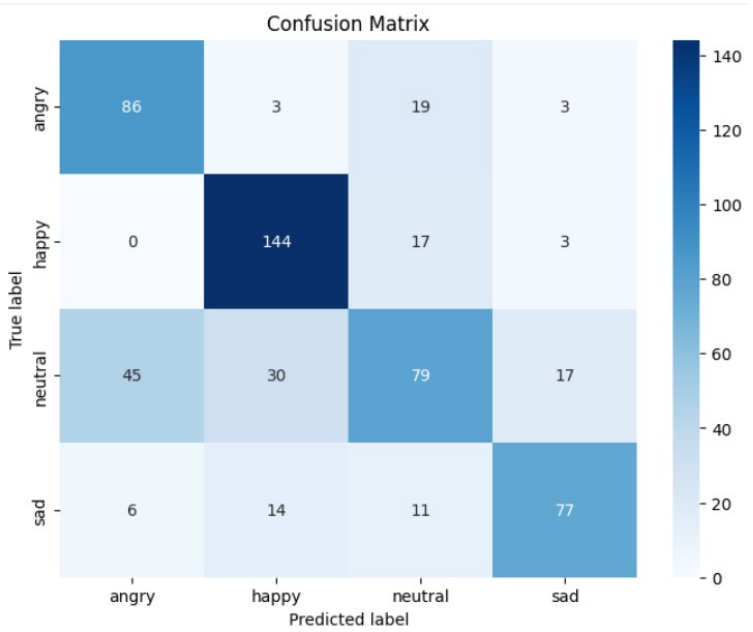

Figure 4: Dialog-level Confusion Matrix (Window Size 10)

Classification Report:

	precision	recall	f1-score	support
angry	0.63	0.77	0.69	111
happy	0.75	0.88	0.81	164
neutral	0.63	0.46	0.53	171
sad	0.77	0.71	0.74	108
accuracy			0.70	554
macro avg	0.69	0.71	0.69	554
weighted avg	0.69	0.70	0.69	554

Figure 5: Dialog-level Classification Report (Window Size 10)

5 Comparison with Paper Results

The original paper reported an accuracy of 69% using a self-attentive model for emotion recognition, leveraging both sentence- and dialog-level context. Our results, particularly with the dialog model using a window size of 10, demonstrate comparable performance, achieving an accuracy of 70%. Several key observations align with the paper's findings:

- The happy class consistently showed the highest accuracy and F1-scores, indicating that this emotion is easier to classify due to clearer distinguishing features in both sentence and dialog context.
- Despite increased context through a larger window size, the neutral class remained the most challenging to predict accurately, mirroring the paper's observation that neutral emotions tend to overlap with other classes, particularly sad and angry.
- Expanding the window size significantly improved the model's recall for both the angry and happy classes, suggesting that capturing broader dialog context helps identify emotional transitions and increases the model's ability to capture subtle shifts in emotional tone.

6 Conclusion

In this work, we implemented a self-attentive emotion recognition network, focusing on both sentence-level and dialog-level emotion classification. Using a subset of the IEMOCAP dataset, we explored various architectures and window sizes to capture the dynamics of emotional expression in conversations. Our results show that increasing dialog context with a larger window size leads to significant improvements, particularly in recall for more easily distinguishable emotions like angry and happy.

By applying a hierarchical structure involving LSTM, GRU, and self-attention mechanisms, our model effectively captures both word-level and sentence-level dependencies. The dialog-level model, especially with a larger window size of 10, demonstrated notable gains in performance compared to the sentence-level approach, highlighting the importance of dialog context in emotion recognition tasks.

Our experiments underscore the potential of self-attention mechanisms in enhancing emotion recognition by allowing the model to focus on the most contextually relevant parts of a conversation. The results provide a strong foundation for future work aimed at refining these techniques and further improving classification accuracy, particularly for more ambiguous emotions like neutral.