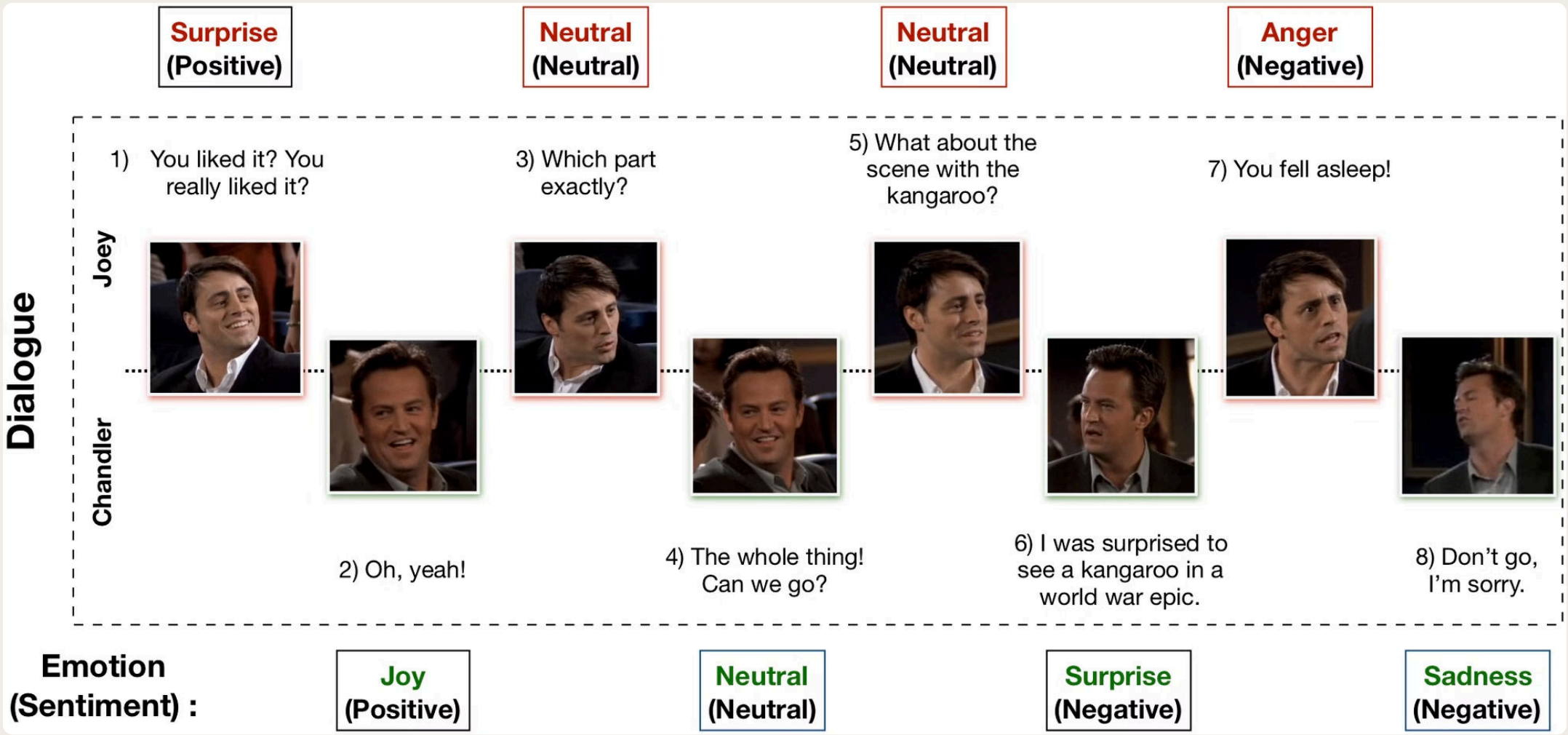
Dialogue Emotion Recognition using the MELD Dataset

The MELD dataset is a significant development in the field of multimodal deep learning. It provides a large-scale collection of conversational data with synchronized audio, text, and visual modalities, enabling researchers to explore novel techniques for integrating diverse information sources.

Example dialogue from MELD



Dataset Statistics for Emotion

Statistics	Train	Dev	Test
# of modality	{a,v,t}	{a,v,t}	{a,v,t}
# of unique words	10,643	2,384	4,361
Avg. utterance length	8.03	7.99	8.28
Max. utterance length	69	37	45
Avg. # of emotions per dialogue	3.30	3.35	3.24
# of dialogues	1039	114	280
# of utterances	9989	1109	2610
# of speakers	260	47	100
# of emotion shift	4003	427	1003
Avg. duration of an utterance	3.59s	3.59s	3.58s

	Train	Dev	Test
Anger	1109	153	345
Disgust	271	22	68
Fear	268	40	50
Joy	1743	163	402
Neutral	4710	470	1256
Sadness	683	111	208
Surprise	1205	150	281

Multimodal Deep Learning Models

Unimodal Models

Developing separate models for audio and text to understand the individual modalities.

1

2

Early Fusion

Concatenating the audio and text embeddings as input to a single neural network.

3

Late Fusion

Employing separate subnetworks for audio and text, then combining the outputs at a later stage.

Retrieving Audio and Text Embeddings

Audio Embeddings

For audio features, the openSMILE toolkit is employed to extract a set of 6,373 features. Due to the high dimensionality of the audio features, the researchers used L2-based feature selection with sparse estimators, such as SVMs, is used to derive a dense representation of the audio segments.

Text Embeddings

To extract textual features, each token is initialized with pre-trained 300-dimensional GloVe embeddings. These embeddings are subsequently processed using a 1D-CNN to generate 100-dimensional textual feature representations.

Multimodal Fusion

Early fusion: Combining the audio and text embeddings to form a unified representation that captures the rich interactions between the two modalities.

Methodology and Results: Emotion Recognition

1 Emotion Classification

Leveraging the MELD dataset to train models for recognizing emotional states expressed through audio, text, and their combination.

2 Multimodal Synergy

Demonstrating that the integration of audio and text features leads to significant improvements in emotion recognition accuracy compared to unimodal approaches.

3 Cross Validation

Cross validation was used to accurately define the metrics and be able to compare with the state of the art models

Models

1

Standard LSTM

The **Standard LSTM** model is structured to handle sequential data, leveraging LSTM units to capture temporal dependencies, with dropout applied for regularization and a final fully connected layer to produce the desired output.

2

Bidirectional LSTM

The **Bidirectional LSTM** model enhances the standard LSTM by using bidirectional processing, allowing it to capture dependencies in both forward and backward directions.

3

Bidirectional LSTM with attention

The **BiLSTM with Scaled Dot-Product Attention** model enhances the bidirectional LSTM architecture by adding an attention mechanism.

Methodology and Results: Hyperparameters, and training parameters

Input Dimension

Variable, based on the input feature size

Hidden Dimension

128

Output Dimension

Variable, based on the number of target classes

Number of layers

3 lstm layers were chosen for all experiments

Dropout rate

0.25 was chosen as the dropout rate after the output layer.

Learning Rate

10^(-3) was chosen as the learning rate for all models.

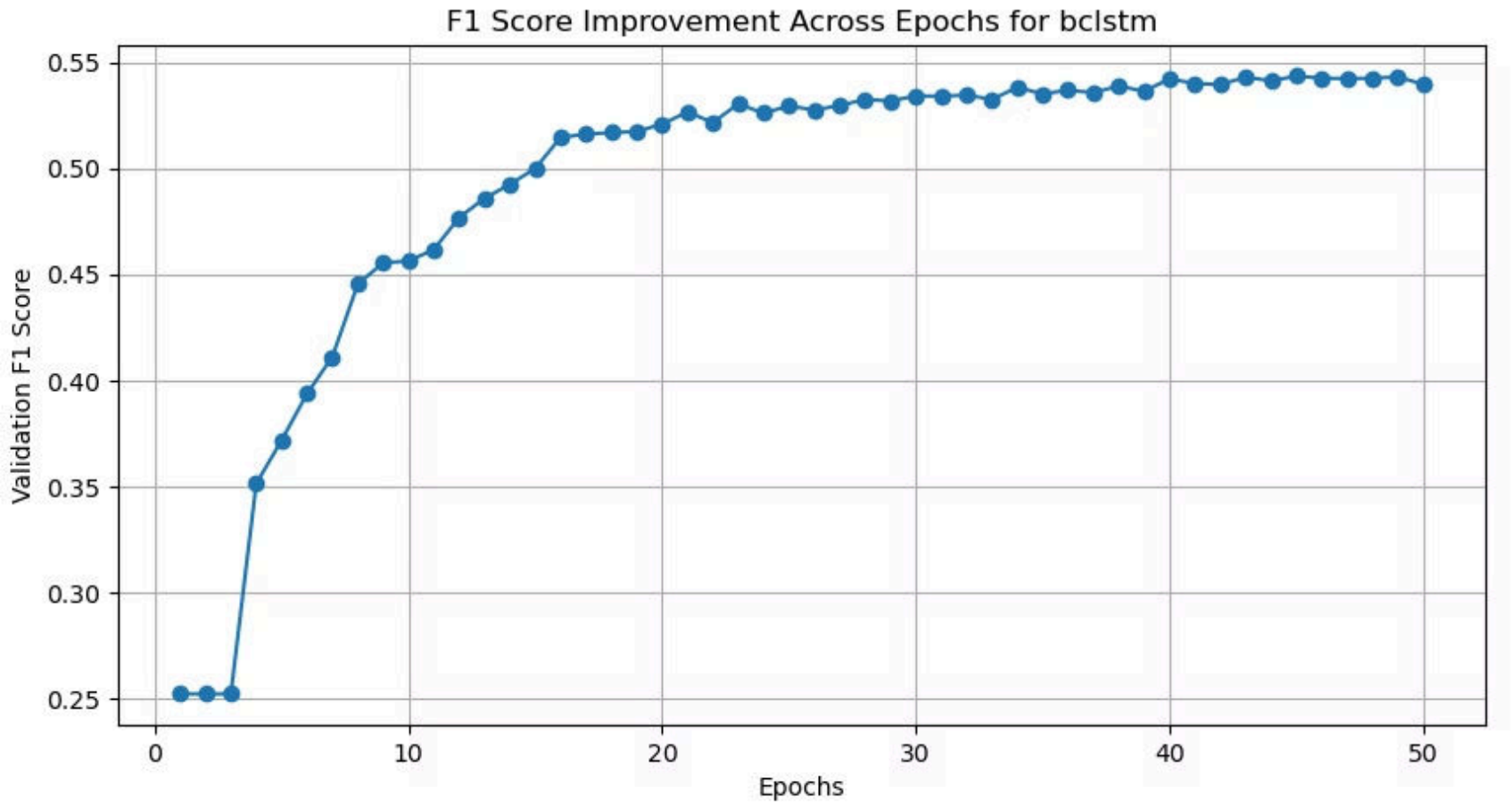
Epochs

For the results section, 50 epochs were chosen. Recommended: 100 epochs.

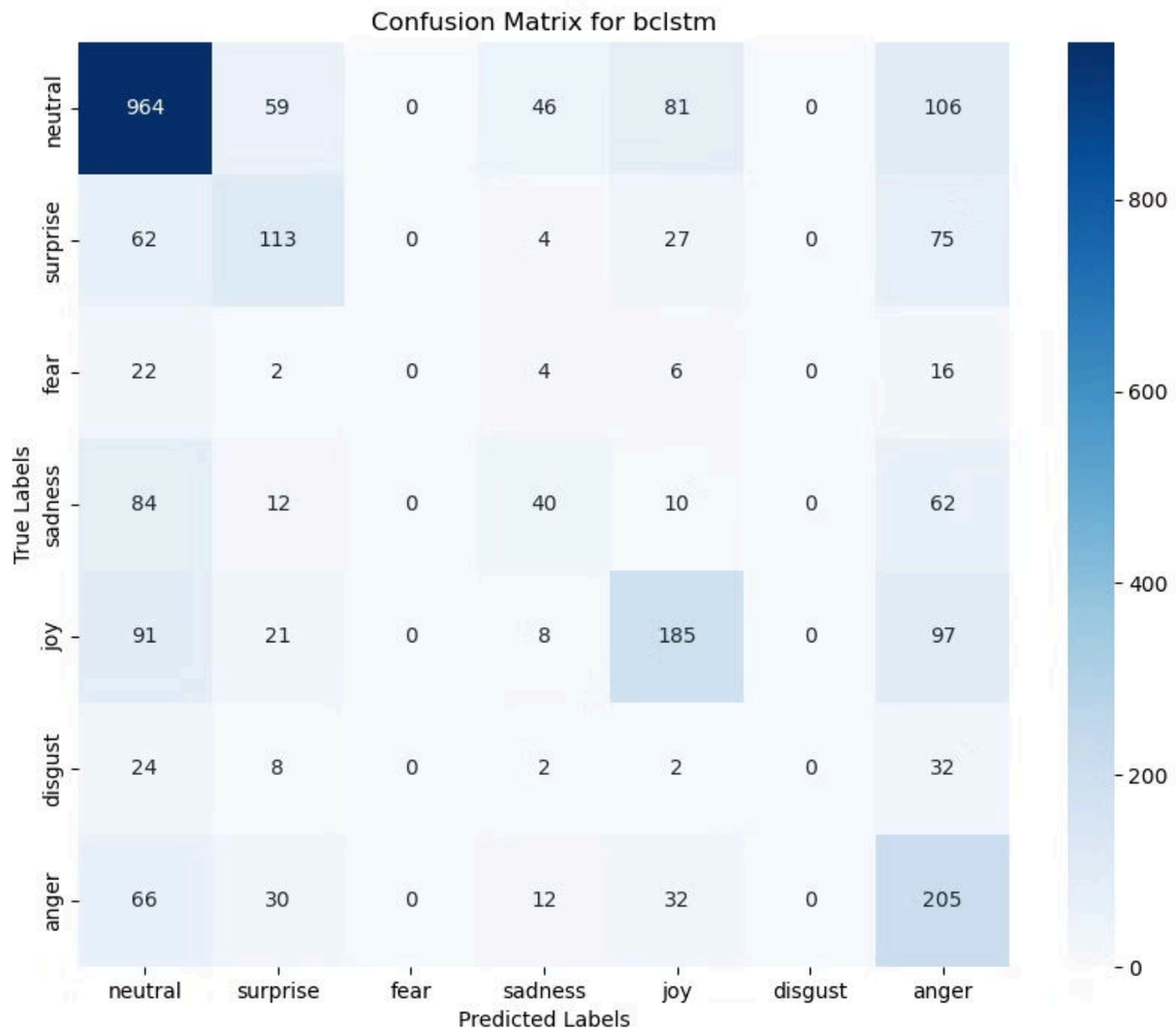
Patience

Early stopping will be called if the model's f1 score during training doesn't improve over **15** epochs. For unimodal models better results were observed with higher patience, around 20-25.

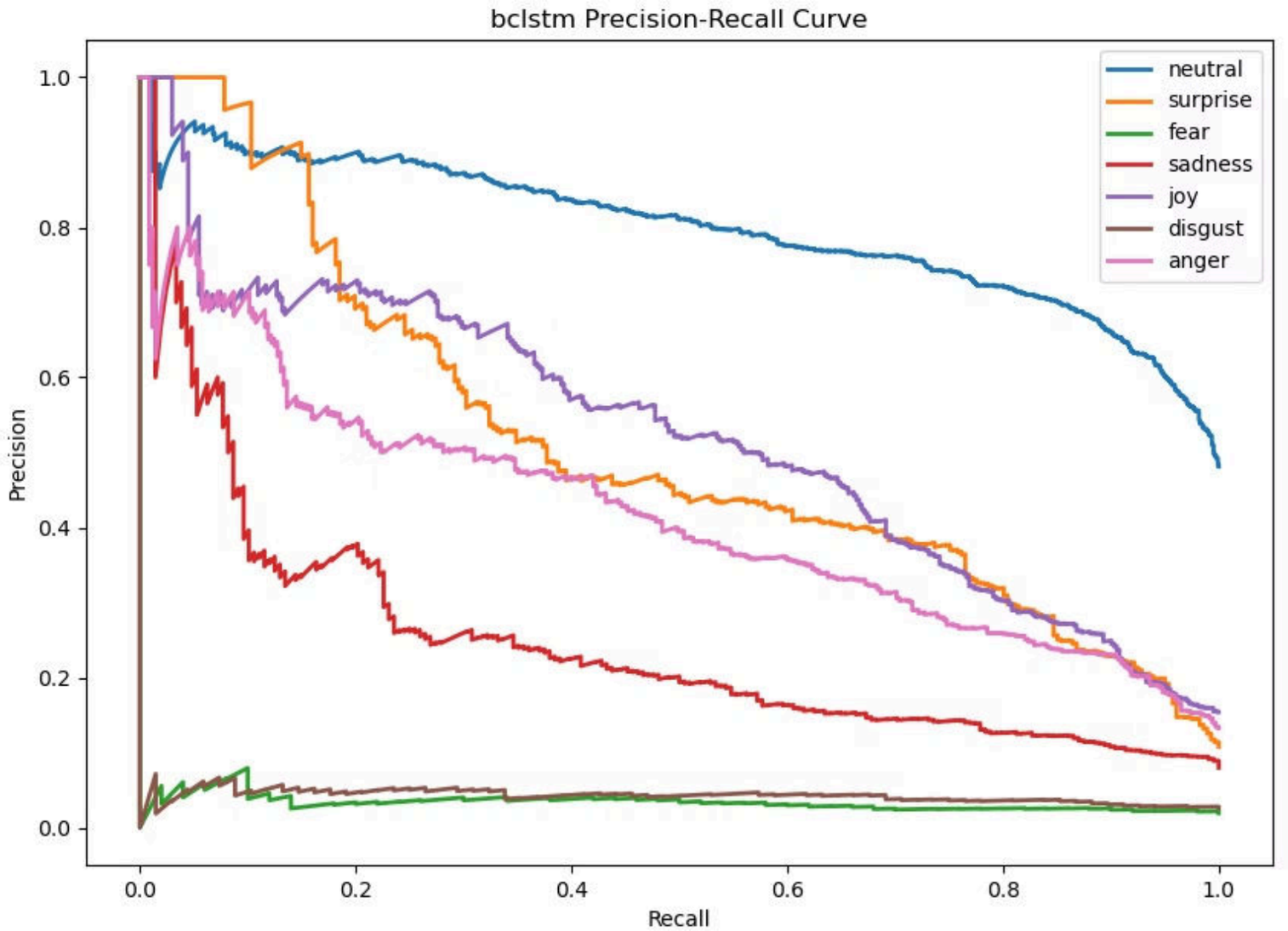
Methodology and Results:
Best model: bcLSTM



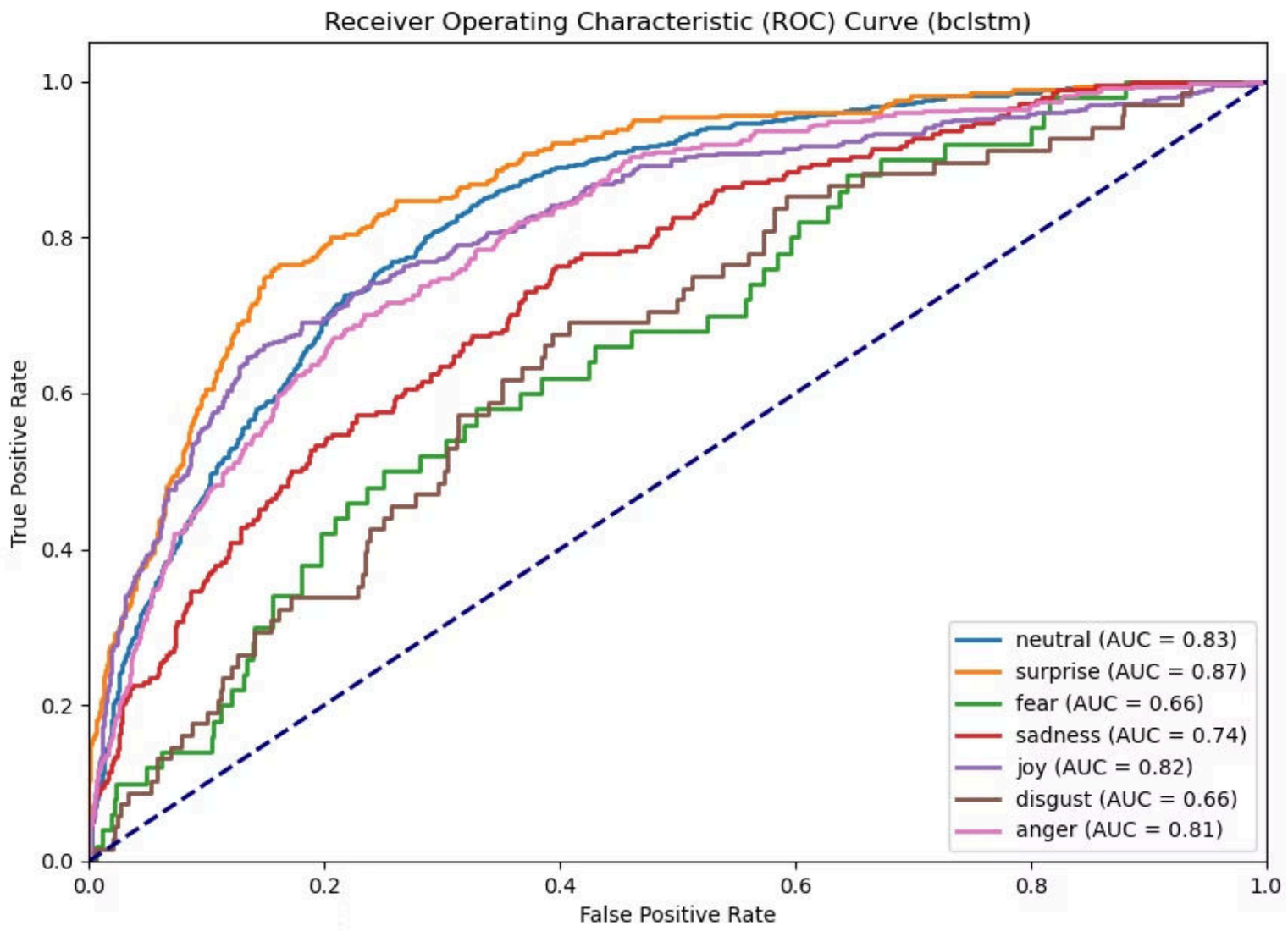
Methodology and Results: Best model: bcLSTM



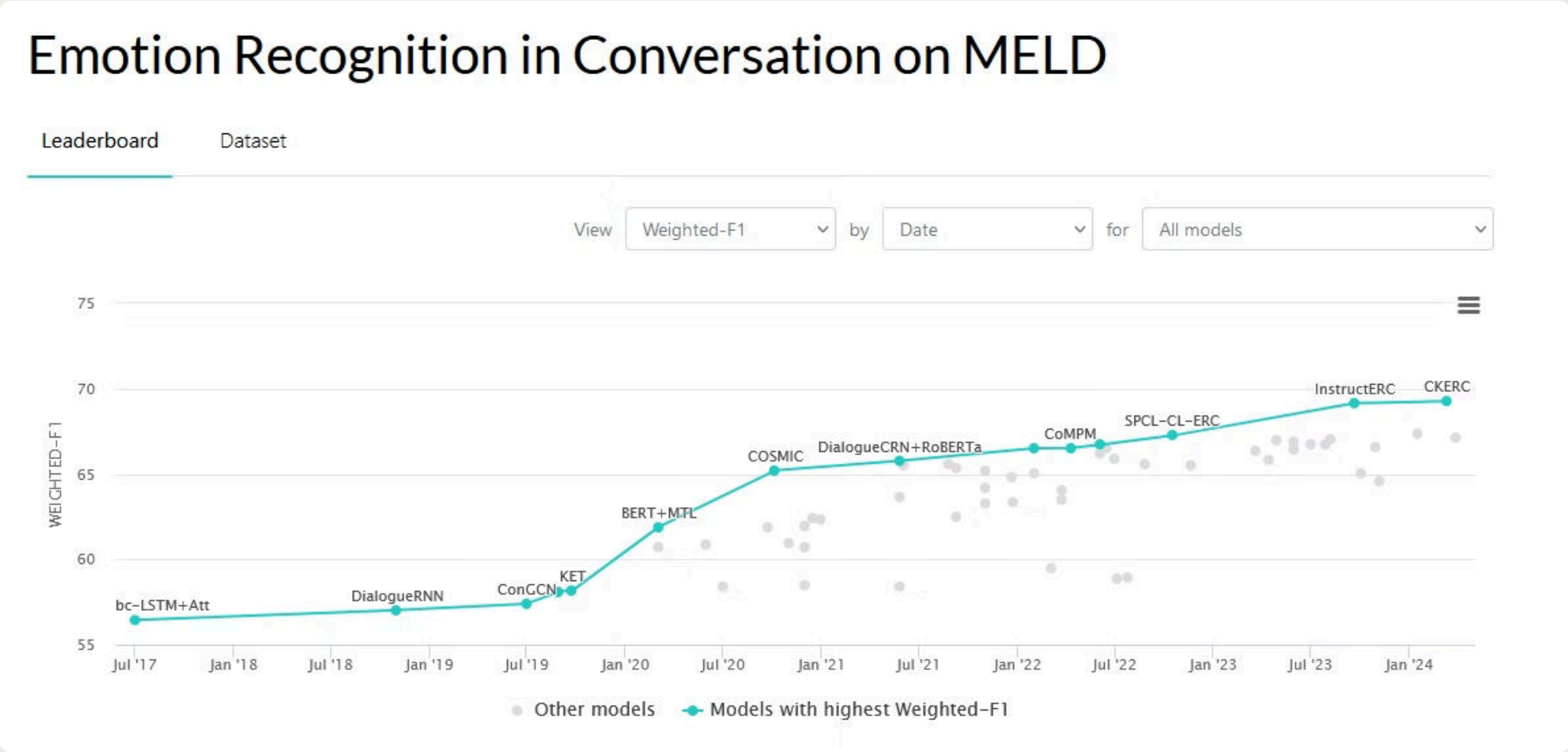
Methodology and Results:
Best model: bcLSTM



Methodology and Results:
Best model: bcLSTM



Comparison with State of the Art models



Conclusions and Demo