# STM: SenTiMetalchemy
# Multi-modal approach to Sentiment Analysis

Team Member: Tom Liu, Kexin Gao, Lexie Wang, Yutong Li

# Main Task (D2 - D3)

- Task: Sentiment Intensity Classification
  - Input: **Text** (Uni-modal), Audio, Video (Multi-modal)
  - Output: Sentiment intensity score within [-3, 3]
- Dataset: CMU-MOSI Dataset (Zadeh et al., 2016)
  - a collection of **2199** opinion video clips annotated with **sentiment in the range [-3,3]**
  - Annotated with labels for subjectivity, sentiment intensity, per-frame and per-opinion annotated **visual features**, and per-milliseconds annotated **audio features**.
- Approach:
  - Uni-modal method
    - 3 models trained on only text modality
  - Multi-modal method
    - 1 fusion model trained on text, audio, video modalities

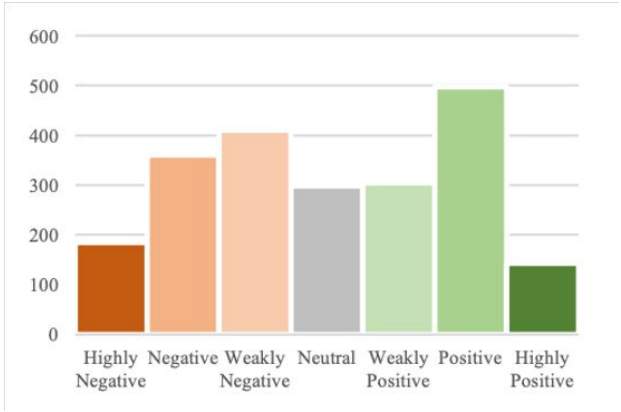



Figure: Distribution of sentiment over the entire dataset.

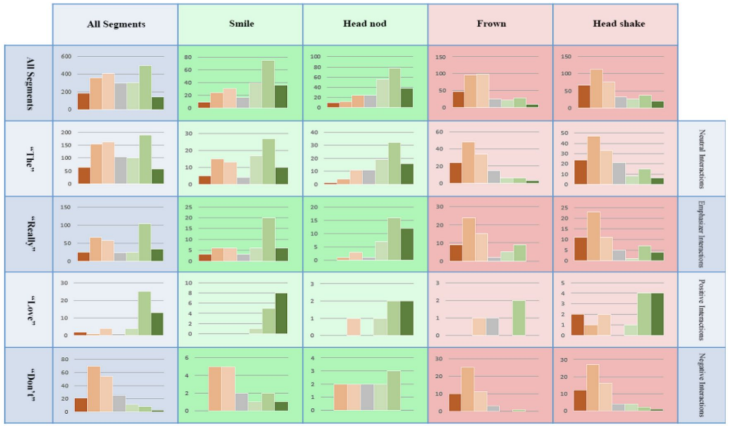
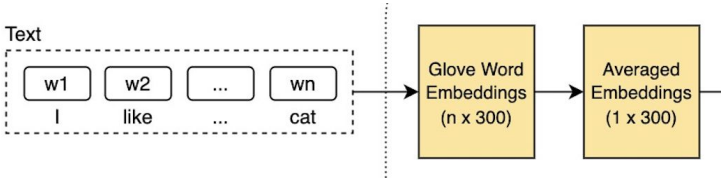



**Figure 3**: Sentiment intensity histograms for different spoken words and visual gestures. In each histogram y-axis is the frequency of co-occurrence and x-axis is sentiment intensity as in Figure 2.

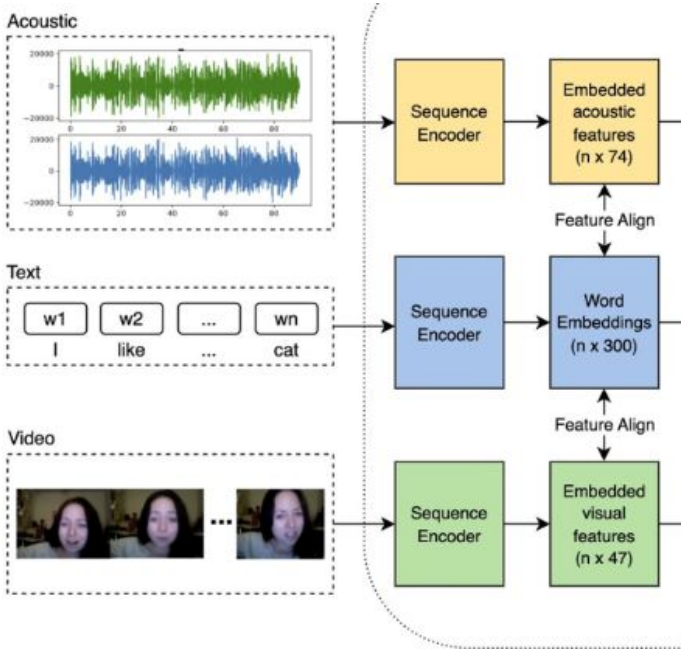
# System Overview

- Data Loading
  - CMU-Multimodal SDK - get the embeddings of the three modalities of each instance
- Data Processing
  - Uni-modal (text only)
    - Get the embeddings of each token of the sentence
    - Average into a sentence embedding vector
  - Multi-modal (text + audio +video)
    - SDK built-in feature alignment method
    - Align acoustic and visual features with textual features
- Dataset Split
  - dataset is split into train (58%), test (10%), and development (32%)
- Training
  - Train 4 models in total (details in the next part)
- Evaluation
  - Metrics: ACC-7, ACC-2, F1 score, MSE, R^2

Uni-modal processing



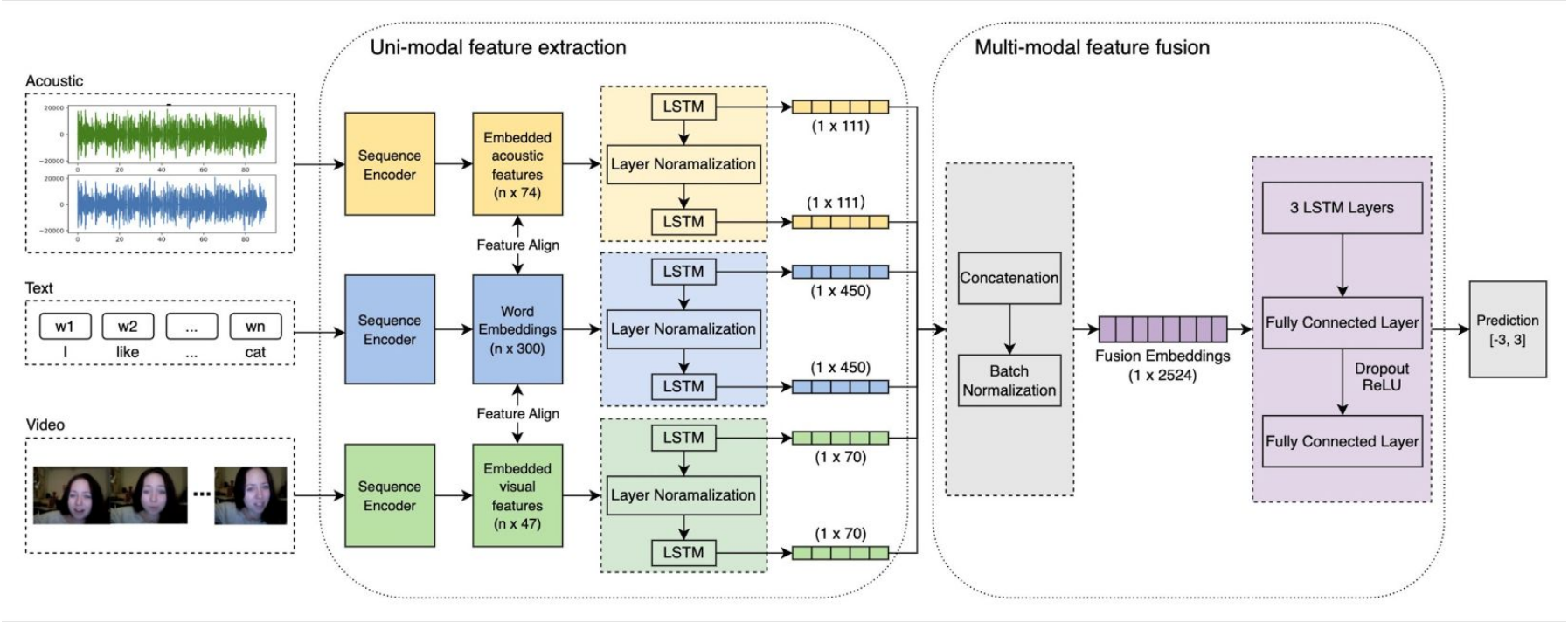

Multi-modal processing



# Uni-Modal Model

- Baseline - Support Vector Regression
  - Radial Basis Function kernel
- Neural Network
  - STM–NN: two **fully connected layers** with size of 512 and 256
  - STM–LSTM: **LSTM** layer followed by one **fully connected layer**
  - *STM–UniModalLSTM: Trained on each of the three modalities individually

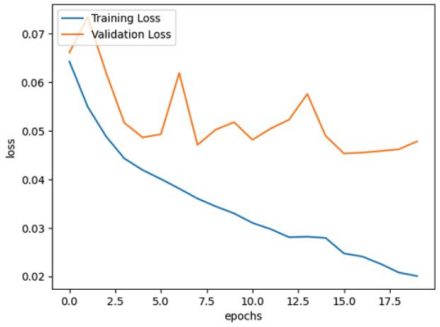
# Multi-Modal Model



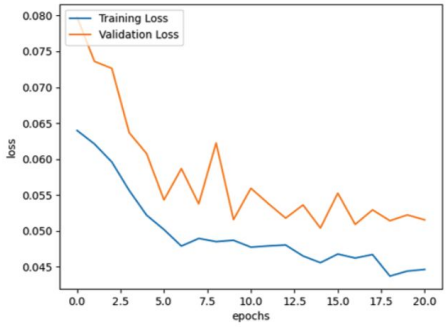
# Result (Main Task)

- STM-NN and STM-LSTM outperform the baseline model in ACC-2
- ACC-7 of the fusion model is significantly higher than the ACC-7 of uni-modal models.
- STM - UniModalLSTM reaches ACC-2 of 48% for audio modal, 52% for video, 71% for text.

ATM-NN




ATM-LSTM




| Method | Model | MAE ↓ | Corr ↑ | ACC-7 ↑ | ACC-2 ↑ | F1 ↑ |
|---|---|---|---|---|---|---|
| Uni-modal | STM-SVR | 1.60 | 0.39 | 20.52 | 70.74 | 72.88 |
| | STM-NN | **1.02** | **0.43** | 20.08 | 73.80 | 75.63 |
| | STM-LSTM | 1.09 | 0.37 | 21.40 | **74.24** | **75.69** |
| Multi-modal | STM-Fusion | 1.08 | 0.31 | **22.16** | 70.99 | 71.73 |
| | *UniMSE* | *0.69* | *0.81* | *48.68* | *86.90* | *86.42* |

Table 1: Results on CMU-MOSI. Contents in italic denote the current SOTA model (trained on larger dataset MOSEI). Contents in bold denote the best performance from our models.

# Issues and Successes

**Issues**

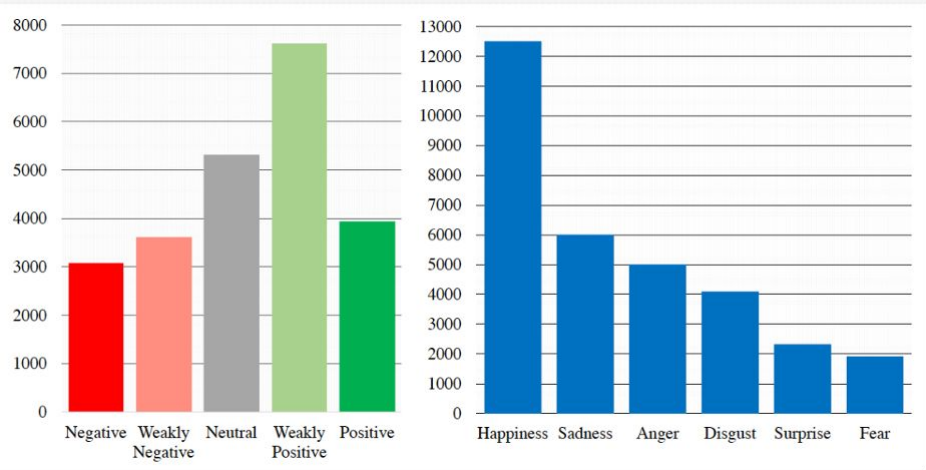
- STM–LSTM outperformed STM-Fusion in ACC-2 and F1
- Computational resources
- Simple concatenation late fusion
- CMU-MOSI dataset size is small

**Successes**

- Improvement in 7 class classification
- STM-Fusion performed better than STM - UniModalLSTM in visual and acoustic modalities

# Adaptation Task

| Dataset | # V | # S | Mod | Sent | Emo | TL (hh:mm:ss) |
|---|---|---|---|---|---|---|
| MOSEI | 23,500 | 1,000 | $\{l, v, a\}$ | ✓ | ✓ | 65:53:36 |
| CMU-MOSI [64] | 2,199 | 98 | $\{l, v, a\}$ | ✓ | ✗ | 02:36:17 |




- CMU- MOSEI (Bagher Zadeh et al., 2018)
  - More than **23,500** sentence videos
  - **1000** online YouTube speakers
  - Gender balanced
  - Annotated with **sentiment intensity** and **emotion**

# Next Step - Adaptation Task

- We may try the Attention mechanism
- Current output: Sentiment Score Only (i.e., [-3, 3])
- Next step: Sentiment Score + Emotion Label (2.5, “happy”)

# Reference

Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph (Bagher Zadeh et al., 2018)

MOSI: Multimodal Corpus of Sentiment Intensity and Subjectivity Analysis in Online Opinion Videos (Zadeh et al., 2016)