Sentimetalchemy (STM): A Multi-modal Approach to Sentiment Analysis (Deliverable 3)

Tongxi Liu†, Yutong Li†, Lexie Wang†, Kexin Gao†, Gina-Anne Levow, Haotian Zhu
Department of Linguistics
University of Washington
{ltxom, lyt826, lexwang, kexing66, levow, haz060}@uw.edu

1 Introduction

Sentiment analysis, a subfield of natural language processing (NLP), aims to automatically identify and categorize emotions, attitudes, and opinions conveyed in textual data. Traditionally, sentiment analysis has primarily relied on textual information to infer sentiments. However, with the rise of social media, image sharing, and multimedia content, the inclusion of visual and acoustic modalities has proven to be valuable in capturing a more holistic understanding of sentiment.

Multi-modal sentiment analysis, an emerging research area, integrates multiple modalities such as text, images, audio, and video to uncover the rich and nuanced aspects of human sentiment. By combining these diverse sources of information, multi-modal sentiment analysis techniques strive to achieve a more accurate and comprehensive representation of human emotions and opinions expressed in online platforms.

In this project, we discover different approach towards the task of sentiment analysis and emotion detection. We implement uni-modal classifiers using statistic-based machine learning model and deep learning based models. We also discover the way to fuse different modalities of text, video and audio to robust the classifier. We analyze the performance of the current state-of-art model UniMSE (Hu et al., 2022) and discover new strategies for this task. Lastly, we perform an adaptation task on a sentiment analysis and emotion detection on a larger dataset.

†Four alchemists equally contributed to this work. (TBD) Liu focuses on the methodology of chrysopoeia, the process of fitting raw material into gold. Wang controls the alloying process to fuse multimodal materials into one. Li creates panaceas to cure overfitting/underfitting. Gao devotes to making an elixir of life for the model to adapt to new tasks.

2 Task description

2.1 Dataset

As we plan to build sentiment classifiers based on a fusion of different modalities, we choose two multi-modal datasets for our tasks.

CMU-MOSI dataset is a relatively small collection of 2199 opinion video clips for sentiment analysis task (Zadeh et al., 2016). Each opinion video is annotated with sentiment in the range [-3,3]. The dataset is rigorously annotated with labels for subjectivity, sentiment intensity, visual features (annotated per-frame and per-opinion), and audio features (annotated per-milliseconds).

Another dataset, CMU-MOSEI, is the largest dataset for multimodal sentiment analysis and emotion recognition to date (Bagher Zadeh et al., 2018). The dataset contains more than 23,500 sentence utterance videos from more than 1000 online YouTube speakers. All the sentences utterance are randomly chosen from various topics and monologue videos. The videos are transcribed and properly punctuated, and the overall dataset is gender balanced.

2.2 Main Tasks (D2-D3)

We approach the sentiment analysis task from two ways: build uni-modal classifiers with only textual features, and fuse different modalities of text, video and audio to robust the models.

As our baseline approach, we build a Support Vector Regression (SVR) model (Joachims, 2005) based on the text data of CMU-MOSI dataset. As comparison for the uni-modal method, we also implement two deep learning based models - a vanilla fully-connected neural network model (NN) (Rumelhart et al., 1986), and a Long Short-Term Memory model (LSTM) (Hochreiter and Schmidhuber, 1997), using only the textual features. For further improvement, we also build a multi-modal classifier (Fusion), which includes features ex-

tracted from text, video and audio from CMU-MOSI.

We compare our models' performances to that of the state-of-the-art models that have achieved high performance on the CMU-MOSI dataset (Hu et al., 2022). We expect the comparison results to reveal the advantages and limitations of our model architecture, which would consequently guide us to potential improvements in data-preprocessing methods, architecture design, and parameter tuning.

2.3 Adaptation Tasks (D4)

We will adapt our pre-trained model to the CMU-MOSEI dataset, an upgraded version of MOSI, annotated with sentiment and emotion (the MOSI dataset only contains sentiment labels). We plan to finetune our model with a slice of MOSEI dataset and test the adaptation results on the new task on emotion classification.

2.4 Evaluation Metrics

For the main tasks on MOSI and the adaptation task on MOSEI, we follow the evaluation methods in previous works (Han et al., 2021; Hu et al., 2022), using mean absolute error (MAE), Pearson correlation (Corr), seven-class classification accuracy (ACC-7), binary classification accuracy (ACC-2) and F1 score as performance evaluation metrics. We will also analyze model limitation, ethical risks and future work of our study.

3 System Overview

Our system contains three main components. In the first part, we load the multimodal datasets using CMU-Multimodal SDK. Since the dataset contains text, video, and acoustic feature data with different frequencies, we align data from different modalities to a pivot modality, which is words. In the next step, we split the dataset into train, dev, and test sets. The second part contains baseline models and multimodal models that take vector-based representations as input and output predictions. In the last part, we evaluate the models and visualize the errors.

4 Approach

4.1 Uni-modal

We build three sentiment classifiers trained on the text data of the CMU-MOSI dataset as our baseline using Support Vector Regression (SVR), Fully Connected Neural Networks (NN), and LSTM network.

Text feature vectors are extracted from the CMU-MOSI dataset and aligned with each data point's labels. We reduce the dimension of the word embedding from $n \times 300$ to 1×300 by average. The output label was linearly transformed from [-3, 3] to [0, 1] to apply the sigmoid function in the output layer of the Neural Networks. The dataset is split into train (58%), test (10%), and development (32%) according to the GOLD metrics from CMU-MOSI.

We tune kernel ('linear,' 'poly,' 'rbf,' 'sigmoid'), kernel coefficient (gamma), epsilon, and squared l2 penalty (C) hyperparameters on the development split using grid search for the SVR model. The coefficient of determination of the prediction was calculated on the test set. We tund hyperparameters of batch size, epochs, number of layers, layer size,

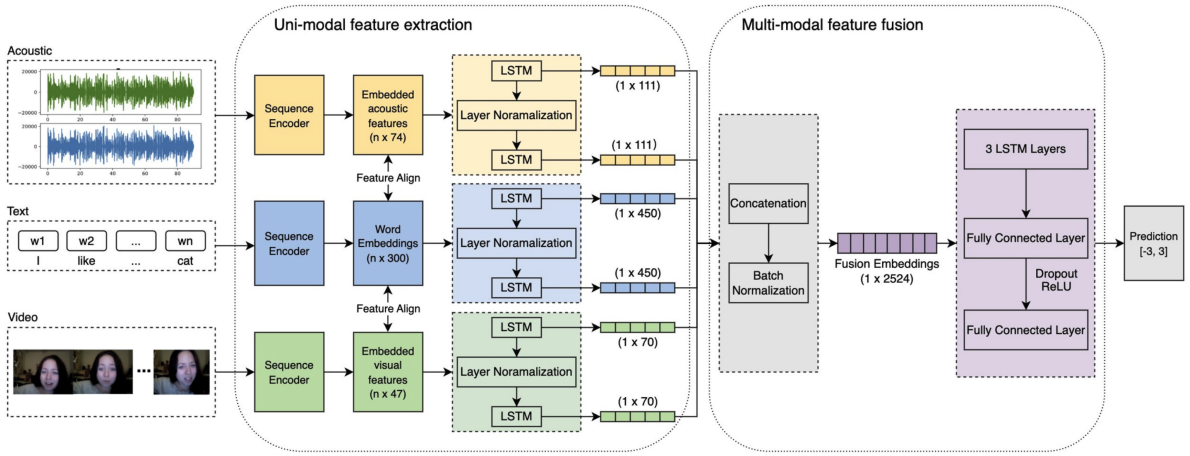

Figure 1: Overview of Fusion Model Architecture

and activation function for the Feed-forward Neural Networks using grid search.

4.2 Multi-modal

We train a Multi-modal classifier (STM-Fusion) to capture more context information than Uni-modal baseline models.

In addition to the text feature vectors, we also extracted visual and acoustic feature vectors from the CMU-MOSI dataset and aligned them with text data. The size of each text feature vector is $n \times 300$, where n is the length of the instance and it varies in the dataset. After applying the alignment function provided by the CMG-Multimodal SDK, the dimensions of visual and acoustic feature vectors became $n \times 47$ and $n \times 74$. We combine three types of feature vectors and then split the data into train, development, and test datasets.

Our model architecture is shown in (Figure 1). After the data preparation step, we pass text, visual and acoustic features to the unified feature extractor layers. Then, we perform late modality fusion and feed the result into three LSTMs and two fully connected layers.

Similarly to the approach mentioned in 4.1, we use the development set to tune hyperparameters including batch size, hidden layer size, dropout, weight decay, number of layers, and epochs.

5 Results

We set the optimal parameters for each model by grid searching. For the SVR model, we choose the Radial Basis Function kernel, C of 200, epsilon of 1, and automatic gamma value. For the fully-connected neural networks, we setup two hidden layers with size of 512 and 256 respectively, followed by ReLU activation function. Sigmoid function is applied to the output layer. For the LSTM model, we implement one LSTM layer (with ReLU activation and dropout of 0.5), followed by two fully connected layers (the former one uses ReLU and the output one uses Sigmoid).

Table 1 shows the performance of each model on the CMU-MOSI test set, compared with the current state-of-art multimodal model, UniMSE (Hu et al., 2022).

For the uni-modal method, deep learning models (i.e. fully-connected Neural Networks and the LSTM model) significantly outperform the baseline SVR model in ACC-2, whereas the SVR model gives better result on ACC-7. LSTM leads to higher ACC-7 compared to SVR and NN.

Comparing the confusion matrix from the binary classification results, the neural network model has more true positive predictions than the SVR model, whereas SVR has fewer false positive predicts and has more correct predictions in negative sentiment. This indicates that, in the task of distinguishing negative affect, the SVR model is still competitive compared to the neural network model.

By comparing the training and validation loss over epochs (Figure 2), it also shows that the LSTM architecture significantly improves the over-fitting issue that exists in the vanilla NN architecture. This indicates that STM-LSTM is more robust to make generalization to unseen data.

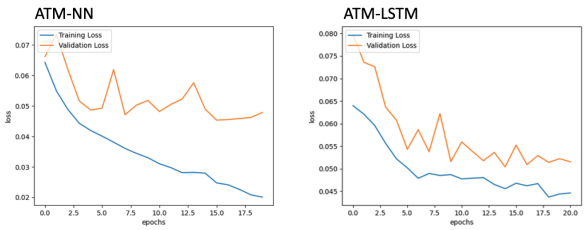

Figure 2: Loss vs. Epochs of STM-NN and STM-LSTM

Two other modalities (i.e. video and audio) are added in the multi-modal model Fusion. We expect the results outperform the uni-modal methods, as we assume two extra feature resource could lead to more accurate classification. The ACC-7 of the fusion model beats the one from LSTM, whereas its ACC-2 only slightly outperform the SVR baseline.

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.

Though there is still a gap between our results and the SOTA model of UniMSE, this is already a relative good performance given the fact that our models are trained on a smaller dataset.

6 Discussion

To briefly recapitulate our approach, we trained three uni-modal models (STM-SVR, STM-NN, STM-LSTM) based on texts and one multi-moda model (STM-Fusion) based on textual, acoustic, and visual data. We tested all models on a binary classification task and a seven-class classification task.

Amongst all three uni-modal models, the STM-LSTM model achieved the highest performance on both binary and seven-class classification tasks. STM-LSTM (ACC-2=74.24) performs only slightly better on the binary classification task compared to STM-SVR (ACC-2=70.74) and STM-NN (ACC-2=73.80). However, STM-LSTM (ACC-7=21.40) performs much better than the other two models on the seven-class classification task. The STM-LSTM achieved the highest F1 score (F1=75.69) compared to STM-NN (F1=75.63) and STM-SVR (F1=72.88). However, the STM-NN model (MAE=1.02, Corr=0.43) slightly outperforms the other two uni-modal models on MAE and Corr.

One big advantage of the STM-LSTM model is that it is obviously better at avoiding over-fitting than STM-NN. This is because the architecture of LSTM includes a mechanism called "memory cell" that controls the flow of information within the network by selectively remembering and forgetting information passed on from previous time-steps. The LSTM architecture also has a drop-out mechanism that prevents the model from relying too much on certain neurons. These two mechanisms enables the LSTM architecture to learn long-distance dependencies while avoiding over-fitting.

As mentioned previously, the STM-NN model has more true positive predictions than the STM-SVR model, while the STM-SVR model has fewer false positive predictions and more correct predictions in negative sentiment. This suggests that the STM-SVR model is advantageous at distinguishing negative affect compared to STM-NN.

Moreover, one limitation of our current method with averaging word embedding dimensions, used on uni-modal models, is that we could lose the meanings encoded in dimensions that were "averaged away." Another limitation is that, by averaging the word embeddings' dimensions, we are making an assumption that all words are equally important, whereas linguistically speaking, certain words may carry more sentiment information than other ones. Our method "dilutes" the significance of these words by performing the averaging on all embeddings.

Moving on to the STM-Fusion model, which is trained on textual, acoustic, and visual information. This model employs late fusion of features of different modalities. All three types of features pass through one LSTM layer, normalization, and another LSTM layer, and then all resulting embeddings (two for each modality) are concatenated and normalized to get fusion embeddings. Eventually, the fusion embeddings get passed through three LSTM layers and two fully connected layers, with drop-out and ReLU mechanisms included.

Unsurprisingly, the STM-Fusion model achieves the highest seven-class classification accuracy score (ACC-7=22.16), which is higher than any of the uni-modal models. This is also in line with our hypothesis from D2. There are several ways to interpret this result. First, the STM-Fusion model has access to multiple modalities, which provides richer and more diverse set of information. This could have enabled the model to learn more robust representation of the data. Second, many of the subtle differences in between sentiment and/or affect classes might be encoded in acoustic and visual data. For example, the same sentence could mean entirely different things when spoken with different tones. Also, the speaker's facial expressions encode sentiment cues. Thus, the STM-Fusion model is able to handle ambiguities and identify subtle differences. Third, having multiple modalities also helps the model with overcoming the weaknesses of one specific modality.

Even though the STM-Fusion is the best performing model of ours, it is important to note that the STM-Fusion model's accuracy scores are still far from those of the state-of-the-art UniMSE model (ACC-7=48.68, ACC-2=86.90). Our hypothesis is that this is due to the small size of our CMU-MOSI dataset. The UniMSE model is trained on the CMU-MOSEI dataset, which contains 65 hours of video data, whereas the CMU-MOSI dataset we trained our models on only includes 2 hours of video data.

One limitation of the multi-modal STM-Fusion

model is that it may require more computational resources compared to uni-modal models due to its complex architecture. Additionally, the current method of averaging word embedding dimensions may result in the loss of important information and dilution of the significance of certain words.

To address these limitations, future improvements could include adapting the models on larger datasets, such as the CMU-MOSEI dataset, to gain a better understanding of the context and meaning of the data, and potentially improve the models' accuracy. We believe that our STM-Fusion model will achieve a much higher score on both classification tasks after we adapt the model on larger dataset. Additionally, alternative methods of encoding word embeddings, such as weighted averaging, could be explored to better capture the significance of different words.

Our next step is to adapt our model on the CMU-MOSEI dataset, which contains large amounts of data. We believe that this will help improve the performance of our models, especially the STM-Fusion model, significantly. By doing so, we expect our models to gain better understandings of the context and meaning of data, which will lead to higher accuracy scores in both binary and multi-class classification tasks. Additionally, by adapting our model on the CMU-MOSEI dataset, our models will be able to perform emotion detection, since that information is included in CMU-MOSEI but is lacking in our current dataset.

7 Ethical considerations

PLACEHOLDER

8 Conclusion

PLACEHOLDER

References

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, Melbourne, Australia. Association for Computational Linguistics.

Wei Han, Hui Chen, and Soujanya Poria. 2021. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9180–9192, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9:1735–1780.

Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu, and Yongbin Li. 2022. UniMSE: Towards unified multimodal sentiment analysis and emotion recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7837–7851, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Thorsten Joachims. 2005. Text categorization with support vector machines: Learning with many relevant features. In *Machine Learning: ECML-98: 10th European Conference on Machine Learning Chemnitz, Germany, April 21–23, 1998 Proceedings*, pages 137–142. Springer.

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. 1986. Learning representations by back-propagating errors. *Nature*, 323:533–536.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. MOSI: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *CoRR*, abs/1606.06259.