(TBD) ATM: Alchemist Transformers-based Multi-modal Sentiment Analysis Model (Deliverable 2)

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

Abstract

In this project, we plan to train a **A**lchemist **T**ransformers-based Multi-modal Sentiment Analysis **M**odel **(ATM)** on the Multimodal Corpus of Sentiment Intensity **(CMU-MOSI)** dataset. Starting from a monomodal statistic-based machine learning model as the baseline, we analyze the performance of the current state-of-art models and develop new or improved strategies for this task. Lastly, we attempt to perform an adaptation task on CMU Multimodal Opinion Sentiment and Emotion Intensity **(CMU-MOSEI)** dataset.

1 Introduction

CMU-MOSI dataset is a collection of 2199 opinion video clips (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, per-frame and per-opinion annotated visual features, and per-milliseconds annotated audio features.

CMU-MOSEI dataset is the largest dataset of 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. The dataset is gender balanced. All the sentences utterance are randomly chosen from various topics and monologue videos. The videos are transcribed and properly punctuated.

2 Task description

Approach: For our baseline approach, we will use Naive Bayes or SVM (Joachims, 2005) to build a sentiment classifier and only use text data.

In our baseline approach II, We plan to use the Transformer model(Vaswani et al., 2017), e.g. fine tune BERT (Devlin et al., 2018), for the sentiment analysis task on text data of CMU-MOSI dataset. Inspired by the multimodal analysis (Poria et al., 2017), we will also experiment with multimodal fusion methods to improve the performance further.

Comparison: After completing the training of our baseline model and multimodal model, we will 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 selection.

Improvement: As mentioned above, we will identify the possible strengths and weaknesses of our model by comparing the performances of our model to that of the the state-of-the-art models (Hu et al., 2022). We will further identify and analyze the possible sources of these strengths and weaknesses and make changes to different aspects of our model accordingly.

Adapation: 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 prediction task.

Evaluation: For the main task 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,

†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.

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. The goal of our baseline model is predicting sentiment based on sentences, so we only extract embeddings of text data. In the next step, we split the dataset into train, dev, and test sets. The seconds part contains basline models that takes vector-based representations as input and output predictions. In the last part, we evaluate the models and visualize the errors.

4 Approach

We built two sentiment classifiers trained on the text data of the CMU-MOSI dataset as our baseline using Support Vector Regression (SVR) and Fully Connected Neural Networks (FCNN).

Text feature vectors were 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 data was split into train (58%), test (10%), and development (32%) according to the GOLD metrics from CMU-MOSI.

We tuned 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 tuned hyperparameters of batch size, epochs, number of layers, layer size, and activation function for the Feed-forward Neural Networks using grid search.

5 Results

We set the optimal parameters for each of the models by the grid search approach introduced in the last part. For the SVR model, we choose the kernel of 'rbf,' C of 200, epsilon of 1, and gamma of 'auto.' For Neural Networks, we designed two hidden layers with the relu activation with neurons (512, 256). The output layer uses sigmoid activation. The training process has a batch size of 256.

While training the neural network model, the loss decreases quickly in the first few epochs as shown in Figure 1. However, it starts to overfit after the 10th epoch and the loss on the validation set reaches the minima at the 20th epoch.

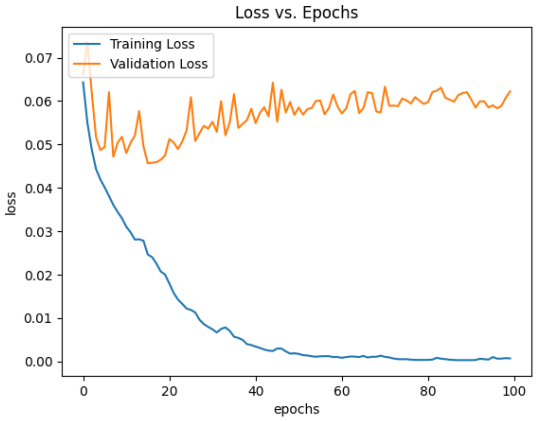

Figure 1: Neural Network Loss vs. Epochs

We follow previous works (Hu et al., 2022) and adopt mean absolute error (MAE), Pearson correlation (Corr), seven-class classification accuracy (ACC-7), binary classification accuracy (ACC-2) and F1 score computed for positive/negative and non-negative/negative classification as evaluation metrics.

Table 1 shows our results on the CMU-MOSI test set using ATM SVR and neural network methods. The same results from the current state-of-art multimodal model, UniMSE, are also reported (Hu et al., 2022).

Model	Metric	Score
ATM-SVR	MAE	1.596
ATM-NN	MAE	1.073
UniMSE	MAE	0.691
ATM-SVR	Corr	0.389
ATM-NN	Corr	0.429
UniMSE	Corr	0.809
ATM-SVR	ACC-7	20.52
ATM-NN	ACC-7	20.08
UniMSE	ACC-7	48.68
ATM-SVR	ACC-2	70.74
ATM-NN	ACC-2	73.80
UniMSE	ACC-2	86.90
ATM-SVR	F1	72.88
ATM-NN	F1	75.69
UniMSE	F1	86.42

Table 1: Results on CMU-MOSI

In general, the neural network model outper-

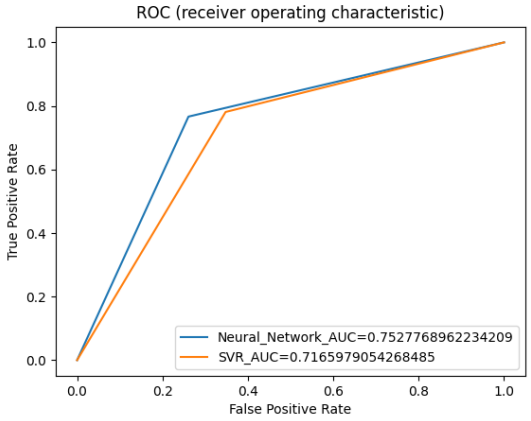

Figure 2: ROC

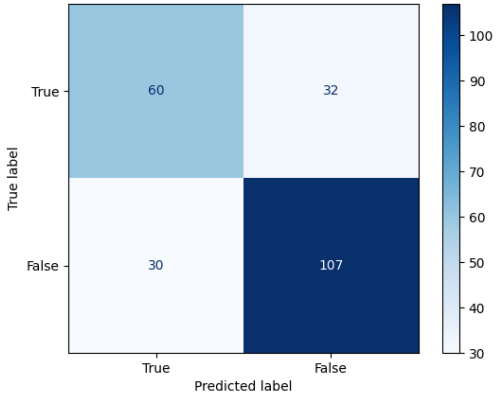

Figure 4: SVR Confusion Matrix

forms the SVR model. We conducted some further experiments to compare their results by results visualization. Figure 2 shows two curves of ROC (receiver operating characteristic) of two models. The neural network curve is closer to the top left corner of the plot, indicating the model performs better at classifying the data.

Figure 3 and figure 4 display the confusion matrix from the binary classification results. The neural network model has more true positive predictions than the SVR model. However, SVR has fewer false positive predicts and has more correct predictions in negative sentiment. Thus, in the task of distinguishing negative affect, this SVR model is still valuable compared to the neural network model.

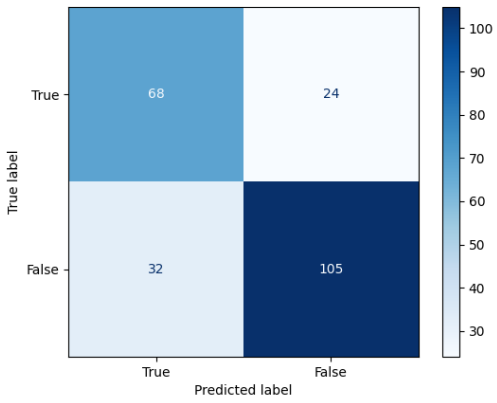

Figure 3: Neural Network Confusion Matrix

6 Discussion

To briefly recapitulate our approach, we developed three models to perform sentiment classification task on the CMU-MOSI dataset using SVR and FCNN. We averaged word embedding dimensions from the original shape $n \times 300$ to 1×300. However, one limitation of our current method with averaging word embedding dimensions 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.

Additionally, we employed the grid search method to tune our hyperparameters. This approach to hyperparameter tuning has many benefits, and the first benefit is that it allows us to exhaustively explore many different combinations of hyperparameters to find the best-performing one. The second benefit is that this method is easy to implement, as it requires a lot less work than manual tuning. The third benefit is that it is reproducible.

In terms of our choice of activation functions for the neural network model, we selected ReLU activation function for hidden layers because it helps mitigate the vanishing gradient problem, preventing gradients from decreasing to extremely small values as they get propagated through the two-layer network. It is also computationally efficient compared to other activation functions.

To begin the discussion of our results, it is important to highlight that both of our baseline models achieved an F1 score that is ≥ 70, which suggests that both models were able to capture the underlying sentimental patterns encoded in textual data.

The ATM-SVR model achieved an F1 score of 72.88, and our ATM-NN model achieved an F1 score of 75.69. The current state-of-the-art model, UniMSE, achieved a higher F1 score of 86.42. We are satisfied with the performance of our baseline models in terms of F1 score, especially considering the fact that UniMSE employs visual and acoustic cues in addition to textual cues and our baseline models rely on textual information only.

Similarly, both of our baseline models achieved an binary classification accuracy score that is ≥ 70. The SVR model has an ACC-2 score of 70.74 and the NN model has an ACC-2 score of 73.80. Interestingly, the accuracy scores achieved by the two models on the seven-class classification task are much lower. Both baseline models have a ACC-7 score that is approximately equal to 20, whereas the UniMSE model (ACC-7 $=$ 48.68) performs significantly better at the seven-class classification task. We have two interpretations of this interesting result. The first interpretation is that the seven-class classification task requires a classifier to be able to capture and extract subtle cues from granular data, and many of that granularity is encoded in visual and acoustic cues. A slight change in the intonation of a sentence could completely reverse the meaning of a sentence. Our ATM-SVR and ATM-NN models do not have access to these subtle cues, and, hence, it is reasonable that they are less meticulous at predicting the sub-class of a data instance. Another possible interpretation is that the development and tuning of our model is based on its performance on the binary classification task, and if we were to redesign the models based on their performance of the seven-class classification task, our baseline models could achieve much higher ACC-7 scores. Specifically, having more hidden layers could enable our ATM-NN model to have greater expressiveness and capacity, in that it will be able to understand more intricate and nuanced patterns.

Our ATM-SVR model achieved a MAE score of 1.596, and our ATM-NN model achieved a MAE score of 1.073. UniMSE outperforms our SVR and NN models in terms of MAE, achieving a score of 0.691. This suggests that the UniMSE model fits the data better, with less deviation between predicted and true values. We expect our future multi-modal models to achieve lower MAE scores, similar to or lower than that of UniMSE.

Furthermore, the UniMSE model (Corr = 0.809) outperforms our ATM-SVR (Corr = 0.389) and ATM-NN (Corr = 0.429) at Pearson correlation. Both of our baseline models achieved a correlation score that is in between weak and moderate correlation. Our hypothesis is that, in general, the UniMSE model yields more accurate predictions, with fewer false positive and false negatives, which could affect the correlation score to a significant extent.

Out next step is to develop models that take advantage of acoustic and visual cues in addition to texts. By doing so, we expect our future models to have better understandings of the context and meaning of data, which might lead to higher accuracy scores in both binary and multi-class classification tasks. In addition, we expect our multi-modal models to be better at handling ambiguities, which is a major drawback of models trained on text-only data. Visual and acoustic information are likely to help with disambiguating these situations.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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.

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.

Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. 2017. A review of affective computing: From unimodal analysis to multimodal fusion. *Information fusion*, 37:98–125.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

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.

A Appendix

PLACEHOLDER