(TBD) ATM: Alchemist Transformers-based Multi-modal Sentiment Analysis Model (Deliverable 1)

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.

†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.

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,

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, ethical risks and future work of our study.

3 System Overview

PLACEHOLDER

4 Approach

PLACEHOLDER

5 Results

PLACEHOLDER

6 Discussion

PLACEHOLDER

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