LLaVAC: Fine-tuning LLaVA as a Multimodal Sentiment Classifier

A Preprint

Thodsaporn Chay-intr[*1,2], **Yujun Chen**[†3], **Kobkrit Viriyayudhakorn**[‡1,2,4], **and Thanaruk Theeramunkong**[§1,3]
[1]Intelligent Informatics and Service Innovation Research Center, Thailand
[2]iApp Technology Co., Ltd., Thailand
[3]Panasonic Research and Development on Artificial Intelligence (AI), Japan
[4]Artificial Intelligence Entrepreneur Association of Thailand (AIEAT), Thailand
[5]Sirindhorn International Institute of Technology, Thammasat University, Thailand

Abstract

We introduce LLaVAC, a method for constructing a classifier for multimodal sentiment analysis. This classifier is capable of classifying both text and image modalities by performing fine-tuning on the Large Language and Vision Assistant (LLaVA). In this work, we design a prompt to consider unimodal and multimodal labels and fine-tune LLaVA for classifying multimodal sentiment labels by generating predicted labels. Our method outperforms baselines by up to 7.31% in accuracy and by 8.76% in weighted-F1 in the MVSA-Single dataset across three dataset processing procedures.

Keywords Multimodal large language model · Prompt tuning · Sentiment classification

1 Introduction

Multimodal Sentiment Analysis (MSA) refers to the process of detecting polarities or attitudes by considering multiple modalities, such as images, text, and speech. The polarities (label) in each modality is commonly 3-class classification (positive, negative, and neural) [Lopes et al., 2021].

Existing studies in MSA has focused on fusing multiple modalities by presenting complex approaches [Cheema et al., 2021, Jiang et al., 2020, Li et al., 2022] along with pre-trained models such as BERT [Devlin et al., 2019], RoBERTa [Liu et al., 2019], and CLIP [Radford et al., 2021], to improve sentiment classification. Meanwhile, Large Language Models (LLMs) have shown their effectiveness in various language processing tasks, such as text classification [Sun et al., 2023a]. However, their applications have primarily focused on the text domain [Naveed et al., 2023], which may not be entirely suitable for the multimodal domain, especially in the context of MSA.

To address this limitation, Multimodal Large Language Models (MLLMs) have been developed, broadening the scope and enhancing the versatility of processing to include multiple modalities [Sun et al., 2023b, Yang et al., 2023]. However, the exploration of MLLMs within MSA remains unexplored in a context similar to that of LLMs.

Given the above statement, in this work, we introduce LLaVAC, a method that involves fine-tuning a MLLM, particularly LLaVA, for constructing a multimodal sentiment classifier. This method aims to leverage the strengths of MLLMs in the processing and analyzing of multimodal data, including image and text. The goal is to improve MSA capabilities while simultaneously minimizing the reliance on complex and manual feature engineering.

Our contributions are summarized as follows:

[*] t.chayintr@gmail.com
[†] chingyokukun@gmail.com
[‡] kobkrit@aieat.or.th
[§] thanaruk@siit.tu.ac.th

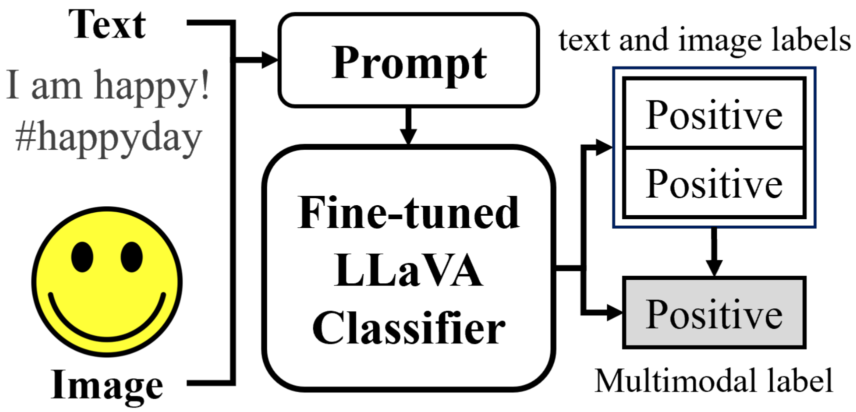

Figure 1: Our LLaVAC method that utilizes fine-tuned LLaVA classifier to predict image, text, and multimodal labels from a prompt containing image and text data.

- We propose LLaVAC, a method that constructs a classifier for multimodal sentiment analysis, incorporating prompt design and fine-tuning of LLaVA. Our code is publicly available at `https://github.com/tchayintr/llavac`.
- Experimental results demonstrate the effectiveness of our method by outperforming all baselines.

2 Background and Related Work

2.1 Multimodal Sentiment Analysis

MSA is an evolving research field focused on analyzing sentiments in various data types, such as images, videos, audio, and text [Cheema et al., 2021]. By Combining computer vision, natural language processing, and machine learning, MSA has shown success in areas such as social media, customer service, and product reviews [Gandhi et al., 2023].

Previous work has focused on fusing multiple modalities with pre-trained models to predict a multimodal label that encapsulates sentiment labels for each modality. For example, Cheema et al. [2021] concatenated image and text features initialized from CLIP [Radford et al., 2021] and RoBERTa [Liu et al., 2019], respectively. Wang et al. [2023] fused both features through Convolutional Neural Networks (CNNs) along with Convolutional Block Attention Module (CBAM) [Woo et al., 2018] where the image and text features were initialized from Residual Networks (ResNet) [He et al., 2015] and BERT [Devlin et al., 2019].

These approaches achieved state-of-the-art performance with their complicated methods, emphasizing the effectiveness of using pre-trained models such as CLIP, RoBERTa, and BERT. However, to the best of our knowledge, MLLMs have not been applied to initialize image and text features in MSA.

2.2 Multimodal Large Language Models as a Classifier

Applying LLMs has recently proven effective in various NLP tasks, which has inspired the development of MLLMs integrating images, videos, and audio [Naveed et al., 2023]. However, to the best of our knowledge, Sun et al. [2023a] is the only study that has incorporated LLMs, including RoBERTa and GPT-3, using prompts to build a classifier specifically for Sentiment Analysis, but not for MSA. Recognizing this gap, our work aims to utilize a MLLM, specifically LLaVA [Liu et al., 2023a,b], to develop a classifier tailored for MSA.

3 Methodology

We propose LLaVAC, a method to build a classifier for multimodal sentiment analysis, specifically for image and text data, to classify image, text, and multimodal labels.

Our method involves designing a prompt along with its corresponding response, and then fine-tuning LLaVA. This enables us to utilize its prior knowledge and adapt the LLaVA as a multimodal sentiment classifier to predict the multimodal label, as shown in Figure 1.

3.1 Prompt Design

Previous studies [Xu and Mao, 2017, Cheema et al., 2021, Zhang et al., 2023] utilized only the multimodal label without considering text and image labels. Therefore, we designed a prompt to manipulate the LLaVA fine-tuning considering both modalities when assigning image, text, and multimodal labels subsequently.

Figure 2 illustrates an example of our prompt and the corresponding response used to fine-tune LLaVA. This prompt, along with its corresponding response, contains image, text, and multimodal labels for each image-text pair, formatted to be compatible with LLaVA fine-tuning.[5]

3.2 Fine-tuned LLaVA Classifier

We simply use our fine-tuned LLaVA classifier in the zero-short classification scenario to inference a prompt similar to Figure 2 without its response to predict image, text and multimodal labels as output.

Prompt:
Consider the following Image and Text.
Image: <image>
Text: RT @babeshawnmendes: "that was really energetic"

Classify the sentiment labels (negative, neutral, positive) for the Image and Text labels, separately.
Finally, jointly analyze and classify the multimodal label for both Image and Text.
Provide a short answer with 3 labels for Text, Image, and Multimodal labels, respectively.

Response:
positive, positive, positive

Figure 2: Example prompt with its response, including image, text, and multimodal labels, where <image> denotes the LLaVA-compatible image data, used for fine-tuning the model.

4 Experiments

4.1 Dataset

In this study, we chose the MVSA-Single dataset[6] to evaluate our approach. This dataset comprises image-text pairs from Twitter. The MVSA dataset is divided into two subsets: MVSA-Single and MVSA-Multiple. In MVSA-Single, a single annotator assigns separate labels to both the image and the text, while in MVSA-Multiple, three annotators are used for involved in labeling each pair.

We used three different procedures to process the dataset. First, we divided the dataset into 10 splits, with each split comprising train, validation, and test sets, as described in Cheema et al. [2021] on fairness. Second, we randomly divide the dataset into train, validation and test sets in an 8:1:1 ratio, as performed by Xu and Mao [2017], Wang et al. [2023]. Lastly, we use the separation from Zhang et al. [2023].

4.2 Experimental Settings

We utilized LLaVA (v1.5-7b)[7] as our base model. We fine-tuned it using LLaVA hyperparameters with LoRA [Hu et al., 2021], as suggested in Liu et al. [2023a,b]. All relevant settings are accessible at `https://github.com/haotian-liu/LLaVA`.

[5] `https://github.com/haotian-liu/LLaVA/blob/main/docs/Finetune_Custom_Data.md`

[6] `https://mcrlab.net/research/mvsa-sentiment-analysis-on-multi-view-social-data`

[7] `https://huggingface.co/liuhaotian/llava-v1.5-7b`

Unlike previous works, such as Cheema et al. [2021], which removed hashtags and links to achieve the best scores on the test set, our approach retains these elements in the text samples during the fine-tuning and testing phases. As LLaVA does not require a validation set in the fine-tuning phase, we utilized only the train set for fine-tuning and the test set for the evaluation step. This differs from previous work Xu and Mao [2017], Cheema et al. [2021], Li et al. [2022], Wang et al. [2023] that employed a validation set. In addition, we use Apache Spark[8] to process the dataset, ensuring high performance, scalability, and adaptability for data handling.

4.3 Evaluation Metrics

We evaluated our approach on three dataset processing procedures using accuracy and weighted-F1 scores, as employed in Cheema et al. [2021]. Specifically, in the case of dividing the dataset into 10 splits, as outlined in Section 4.1, we computed the average scores in all splits to demonstrate overall performance.

4.4 Results

Tables 1, 2, and 3 illustrate the evaluation results comparing our model with the baseline models. According to these results, our model outperformed all baseline models in all evaluation metrics, with the exception of the model by Zhang et al. [2023] in terms of accuracy. This suggests the potential effectiveness of using MLLMs in fine-tuning for classification tasks.

Table 1: Comparison of our model's results with previous works using dataset splits similar to Cheema et al. [2021]. Both Acc and F_1 are averaged over the splits.

Models	**Acc**	**F_1**
MultiSentiNet [Xu and Mao, 2017]	63.27	59.12
FENet-BERT [Jiang et al., 2020]	69.02	67.30
Se-MLNN [Cheema et al., 2021]	75.33	73.76
Ours	**76.24**	**76.36**

Table 2: Comparison of our model's results with previous works using a random dataset split, as outlined by Wang et al. [2023].

Models	**Acc**	**F_1**
FENet-BERT [Jiang et al., 2020]	74.21	74.06
CMCN [Peng et al., 2022]	73.61	75.03
CLMLF [Li et al., 2022]	75.33	73.46
CBAM [Wang et al., 2023]	**77.11**	76.55
Ours	77.05	**76.76**

Table 3: Comparison of our model with previous work using the dataset split similar to Zhang et al. [2023].

Models	**Acc**	**F_1**
QMF [Zhang et al., 2023]	78.07	76.30
Ours	**85.36**	**85.06**

5 Conclusion

We presented LLaVAC, a method to construct a classifier for multimodal sentiment analysis by fine-tuning LLaVA to generate only image, text, and multimodal labels. We archived this by creating a prompt to manipulate the LLaVA in considering both image and text modalities, and generate only their labels. LLaVAC outperformed the baselines on the MVSA-Single dataset, achieving up to 7.31% and 8.76% in accuracy and weighted-F1, respectively.

References

Vasco Lopes, Antonio Gaspar, Luis A. Alexandre, and Joao Cordeiro. An automl-based approach to multimodal image sentiment analysis. In *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, July 2021. doi:10.1109/ijcnn52387.2021.9533552. URL `http://dx.doi.org/10.1109/IJCNN52387.2021.9533552`.

[8] `https://spark.apache.org`

Gullal S. Cheema, Sherzod Hakimov, Eric Müller-Budack, and Ralph Ewerth. A fair and comprehensive comparison of multimodal tweet sentiment analysis methods, 2021.

Tao Jiang, Jiahai Wang, Zhiyue Liu, and Yingbiao Ling. Fusion-extraction network for multimodal sentiment analysis. page 785–797, Berlin, Heidelberg, 2020. Springer-Verlag. ISBN 978-3-030-47435-5. doi:10.1007/978-3-030-47436-2_59. URL https://doi.org/10.1007/978-3-030-47436-2_59.

Zhen Li, Bing Xu, Conghui Zhu, and Tiejun Zhao. CLMLF:a contrastive learning and multi-layer fusion method for multimodal sentiment detection. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2282–2294, Seattle, United States, July 2022. Association for Computational Linguistics. doi:10.18653/v1/2022.findings-naacl.175. URL https://aclanthology.org/2022.findings-naacl.175.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi:10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. Text classification via large language models, 2023a.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models, 2023.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. Aligning large multimodal models with factually augmented rlhf, 2023b.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v(ision), 2023.

Ankita Gandhi, Kinjal Adhvaryu, Soujanya Poria, Erik Cambria, and Amir Hussain. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Information Fusion*, 91:424–444, 2023. ISSN 1566-2535. doi:https://doi.org/10.1016/j.inffus.2022.09.025. URL https://www.sciencedirect.com/science/article/pii/S1566253522001634.

Huiru Wang, Xiuhong Li, Zenyu Ren, Dan Yang, and chunming Ma. Exploring multimodal sentiment analysis via cbam attention and double-layer bilstm architecture, 2023.

Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023b.

Nan Xu and Wenji Mao. Multisentinet: A deep semantic network for multimodal sentiment analysis. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, page 2399–2402, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450349185. doi:10.1145/3132847.3133142. URL https://doi.org/10.1145/3132847.3133142.

Qingyang Zhang, Haitao Wu, Changqing Zhang, Qinghua Hu, Huazhu Fu, Joey Tianyi Zhou, and Xi Peng. Provable dynamic fusion for low-quality multimodal data, 2023.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.

Cheng Peng, Chunxia Zhang, Xiaojun Xue, Jiameng Gao, Hongjian Liang, and Zhengdong Niu. Cross-modal complementary network with hierarchical fusion for multimodal sentiment classification. *Tsinghua Science and Technology*, 2022. URL https://api.semanticscholar.org/CorpusID:245025098.