M-SENA: An Integrated Platform for Multimodal Sentiment Analysis

Huisheng Mao[1, 2]*, **Ziqi Yuan**[1, 2]*, **Hua Xu**[1, 2]†, **Wenmeng Yu**[1, 2], **Yihe Liu**[1, 3], **Kai Gao**[3]
[1]State Key Laboratory of Intelligent Technology and Systems,
Department of Computer Science and Technology, Tsinghua University
[2]Beijing National Research Center for Information Science and Technology(BNRist)
[3]School of Information Science and Engineering, Hebei University of Science and Technology
{mhs20,yzq21}@mails.tsinghua.edu.cn
xuhua@tsinghua.edu.cn

Abstract

M-SENA is an open-sourced platform for Multimodal Sentiment Analysis. It aims to facilitate advanced research by providing flexible toolkits, reliable benchmarks, and intuitive demonstrations. The platform features a fully modular video sentiment analysis framework consisting of data management, feature extraction, model training, and result analysis modules. In this paper, we first illustrate the overall architecture of the M-SENA platform and then introduce features of the core modules. Reliable baseline results of different modality features and MSA benchmarks are also reported. Moreover, we use model evaluation and analysis tools provided by M-SENA to present intermediate representation visualization, on-the-fly instance test, and generalization ability test results. The source code of the platform is publicly available at https://github.com/thuiar/M-SENA.

1 Introduction

Multimodal Sentiment Analysis (MSA) aims to judge the speaker's sentiment from video segments (Mihalcea, 2012; Soleymani et al., 2017; Guo et al., 2019). It has attracted increasing attention due to the booming of user-generated online content. Although impressive improvements have been witnessed in recent MSA researches (Tsai et al., 2019; Rahman et al., 2020; Yu et al., 2021), building an end-to-end video sentiment analysis system for real-world scenarios is still full of challenges.

The first challenge lies in effective acoustic and visual feature extraction. Most previous approaches (Zadeh et al., 2017a; Hazarika et al., 2020; Han et al., 2021a) are developed on the provided modality sequences from CMU-MultimodalSDK[1]. However, reproducing exact identical acoustic and visual feature extraction is almost impossible due to the the vague description of feature selection and backbone selection (both COVAREP[2] and Facet[3] can not be directly used in Python). Moreover, recent literature (Tsai et al., 2019; Gkoumas et al., 2021; Han et al., 2021b) observe that the text modality stands in the predominant position while acoustic and visual modalities have few contributions to the final sentiment classification. Such results further arouse the attention on effective feature extraction of acoustic and visual modalities.

With the awareness of the importance of acoustic and visual feature extraction, researchers attempt to develop models based on customized modality sequences instead of provided features (Dai et al., 2021; Hazarika et al., 2020). However, performance comparison with different modality features is unfair. Therefore, the demand for reliable comparison of modality features and fusion methods is increasingly urgent.

Another factor that limits the application of existing MSA models in real scenarios is the lack of comprehensive model evaluation and analysis approaches. Models obtained outstanding performance on the given test set might degrade in real-world scenarios due to the distribution discrepancy or random modality perturbations (Liang et al., 2019; Zhao et al., 2021; Yuan et al., 2021). Besides, effective model analysis is also crucial for researchers to explain the improvements and perform model refinement.

The **M**ultimodal **SEN**timent **A**nalysis platform (M-SENA) is developed to address the above challenges. For acoustic and visual features, the platform integrates Librosa (McFee et al., 2015), OpenSmile (Eyben et al., 2010), OpenFace (Baltrusaitis et al., 2018), MediaPipe (Lugaresi et al., 2019) and provides a highly customized feature extraction API in Python. With the modular MSA pipeline, fair comparison between different features

[*] These authors contributed equally to this work.

[†] Hua Xu is the corresponding author.

[1] Features provided by CMU

[2] https://github.com/covarep/covarep

[3] https://imotions.com

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
System Demonstrations, pages 204 - 213
May 22-27, 2022 ©2022 Association for Computational Linguistics

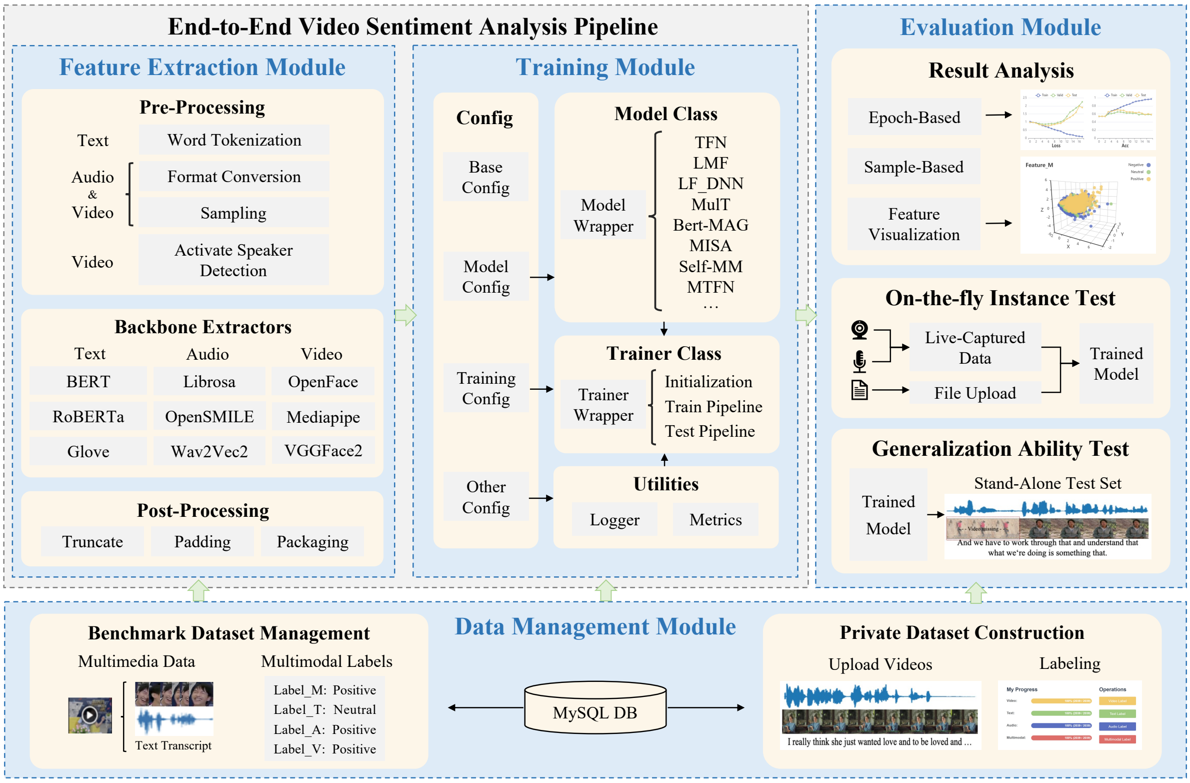

Figure 1: The overall framework of the M-SENA platform contains four main modules: data management module, feature extraction module, model training module and model evaluation module.

and MSA fusion models can be achieved. The results can be regarded as reliable baselines for future MSA research. Furthermore, the platform provides comprehensive model evaluation and analysis tools to reflect the model performance in real-world scenarios, including intermediate result visualization, on-the-fly instance demonstration, and generalization ability test. The contributions of this work are briefly summarized as follows:

1. By providing a highly customized feature extraction toolkit, the platform familiarizes researchers with the composition of modality features. Also, the platform bridges the gap between designing MSA models with provided, fixed modality features and building a real-world video sentiment analysis system.

2. The unified MSA pipeline guarantees fair comparison between different combinations of modality features and fusion models.

3. To help researchers evaluate and analyze MSA models, the platform provides tools such as intermediate result visualization, on-the-fly instance demonstration, and generalization ability test.

2 Platform Architecture

M-SENA platform features convenient data access, customized feature extraction, unified model training pipeline, and comprehensive model evaluation. It provides a graphical web interface as well as Python packages for researchers with all features above. The platform currently supports three popular MSA datasets across two languages, seven feature extraction backbones, and fourteen benchmark MSA models. Figure 1 illustrates the overall architecture of the M-SENA platform. In the remaining parts of this section, features of each module in Figure 1 will be described in detail.

2.1 Data Management Module

The data management module is designed to ease the access of multimedia data on servers. Besides providing existing benchmark datasets, the module also enables researchers to build and manage their own datasets.

Benchmark Datasets. M-SENA currently supports three benchmark MSA datasets, including CMU-MOSI (Zadeh et al., 2016), CMU-MOSEI (Zadeh et al., 2018b) in English, and CH-SIMS (Yu et al., 2020) in Chinese. Details of integrated datasets are shown in Appendix A. Users can filter

Acoustic Feature Sets	
ComParE_2016 (Schuller et al., 2016)	Static (HSFs)
eGeMAPS (Eyben et al., 2015)	Static (LLDs)
wav2vec2.0 (Baevski et al., 2020)	Learnable
Visual Feature Sets	
Facial Landmarks (Zadeh et al., 2017b)	Static
Eyes Gaze (Wood et al., 2015)	Static
Action Unit (Baltrušaitis et al., 2015)	Static
Textual Feature Sets	
GloVe6B (Pennington et al., 2014)	Static
BERT (Devlin et al., 2018)	Learnable
RoBerta (Liu et al., 2019)	Learnable

Table 1: Some of the supported features in M-SENA.

and view raw videos conveniently without downloading them to the local environment.

Building Private Datasets. The M-SENA platform also provides a graphical interface for researchers to construct their own datasets using uploaded videos. Following the literature (Yu et al., 2020), M-SENA supports unimodal sentiment labelling along with multimodal sentiment labelling. The constructed datasets can be directly used for model training and evaluation on the platform.

2.2 Feature Extraction Module

Emotion-bearing modality feature extraction is still an open challenge for MSA tasks. To facilitate effective modality feature extraction for MSA, M-SENA integrates seven most commonly used feature extraction tools and provides a unified Python API as well as a graphical interface. Part of the supported features for each modality are listed in Table 1 and described below:

Acoustic Modality. Various acoustic features have been proven effective for emotion recognition (El Ayadi et al., 2011; Akçay and Oğuz, 2020). Hand-crafted acoustic features can be divided into two classes, low level descriptors (LLDs), and high level statistics functions (HSFs). LLDs features, including prosodies, spectral domain features and others, are calculated on a frame-basis, while HSFs features are calculated on an entire utterance level. In addition to the hand-crafted features, M-SENA also provides pretrained acoustic model wav2vec2.0 (Baevski et al., 2020) as a learnable feature extractor. Researchers can also design and build their own customized acoustic features using the provided Librosa extractor.

Types	Scenarios		
	Films(TV)	Variety Show	Life(Vlog)
Easy	10 (en:4 ch:6)	8 (en:4 ch:4)	8 (en:4 ch:4)
Common	9 (en:4 ch:5)	11 (en:6 ch:5)	8 (en:4 ch:4)
Difficult	9 (en:4 ch:5)	9 (en:5 ch:4)	8 (en:4 ch:4)
Noise	9 (en:4 ch:5)	8 (en:4 ch:4)	7 (en:2 ch:5)
Missing	9 (en:4 ch:5)	9 (en:5 ch:4)	7 (en:3 ch:4)

Table 2: Statistics of the generalization ability test dataset, where "en" represents "English", "ch" represents "Chinese".

Visual Modality. In existing MSA research, facial Landmarks, eyes gaze, and facial action units are commonly used visual features. The M-SENA platform enables researchers to extract visual feature combinations flexibly using OpenFace and MediaPipe extractors.

Text Modality. Compared with acoustic and visual features, semantic text embeddings are much more mature with the rapid development of pretrained language models (Qiu et al., 2020). Following previous works (Zadeh et al., 2017a; Rahman et al., 2020; Lian et al., 2022), M-SENA supports GloVe6B (Pennington et al., 2014), pretrained BERT (Devlin et al., 2018), and pretrained RoBerta (Liu et al., 2019) as textual feature extractors.

All feature extractors above are available through both Python API and Graphical User Interface(GUI). Listing 1 shows a simple example of default acoustic feature extraction using Python API. The process is similar for other modalities. Advanced usage and detailed documentation is available at Github Wiki[4].

```
from MSA_FET import
    FeatureExtractionTool

# Extract Audio Feature for MOSI.
fet = FeatureExtractionTool("librosa")

feature = fet.run_dataset(
    dataset_dir='~/MOSI',
    out_file='output/feature.pkl'
)
```

Listing 1: An example of acoustic feature extraction on the MOSI dataset using MMSA.

2.3 Model Training Module

M-SENA provides a unified training module which currently integrates 14 MSA benchmarks, including tensor fusion methods, TFN (Zadeh et al., 2017a), LMF (Liu et al., 2018), modality factorization methods, MFM (Tsai et al., 2018), MISA (Hazarika et al., 2020), SELF-MM (Yu et al., 2021), word-level fusion methods, MulT (Tsai et al., 2019), BERT-MAG (Rahman et al., 2020),

[4] https://github.com/thuiar/MMSA-FET/wiki

Feature Combinations	TFN		GMFN		MISA		Bert-MAG	
	Acc-2 (%)	F1 (%)	Acc-2 (%)	F1 (%)	Acc-2 (%)	F1 (%)	Acc-2 (%)	F1 (%)
CMU-SDK†	78.02	78.09	76.98	77.06	82.96	82.98	83.41	83.47
[T1]-[A1]-[V1]	77.41	77.47	77.77	77.84	83.78	83.80	83.38	83.43
[T2]-[A1]-[V1]	70.40	70.51	71.40	71.54	75.22	75.68	-	-
[T3]-[A1]-[V1]	80.85	80.79	80.21	80.15	79.57	79.67	-	-
[T1]-[A2]-[V1]	76.80	76.82	78.02	78.03	83.72	83.72	82.96	83.04
[T1]-[A3]-[V1]	77.19	77.23	78.44	78.45	82.16	82.23	83.57	83.58
[T1]-[A1]-[V2]	77.38	77.48	78.81	78.71	83.2	83.14	82.13	82.20
[T1]-[A1]-[V3]	76.74	76.81	78.23	78.24	84.06	84.08	83.69	83.75

Table 3: Results for feature selection. For text, [T1] refers to BERT, [T2] refers to GloVe6B, [T3] refers to RoBerta. For acoustic, [A1] refers to eGeMAPS, [A2] refers to customized feature including 20-dim MFCC, 12-dim CQT, and f0, [A3] refers to wav2vec2.0. For visual, [V1] refers to action units, [V2] refers to landmarks, [V3] refers to both landmarks and action units. CMU-SDK† refers to modified CMU-SDK features with BERT for text.

multi-view learning methods: MFN (Zadeh et al., 2018a), GMFN (Zadeh et al., 2018b), and other MSA methods. Detailed introduction of the integrated baseline methods is provided in Appendix B. We will continue following advanced MSA benchmarks and put our best effort into providing reliable benchmark results for future MSA research.

2.4 Result Analysis Module

The proposed M-SENA platform provides comprehensive model evaluation tools including intermediate result visualization, on-the-fly instance test, and generalization ability test. A brief introduction of each component is given below, while a detailed demonstration is shown in Section 4.

Intermediate Result Visualization. The discrimination of multimodal representations is one of the crucial metrics for the evaluation of different fusion methods. The M-SENA platform records the final multimodal fusion results and illustrates them after decomposition with Principal Component Analysis (PCA). Training loss, binary accuracy, F1 score curves are also provided in M-SENA for detailed analysis.

Live Demo Module. In the hope of bridging the gap between MSA research and real-world video sentiment analysis scenarios, M-SENA provides a live demo module, which performs on-the-fly instance tests. Researchers can validate the effectiveness and robustness of the selected MSA model by uploading or live-feeding videos to the platform.

Generalization Ability Test. Compared to the provided test set of benchmark MSA datasets, real-world scenarios are often more complicated. Future MSA models need to be robust against modality noise as well as effective on the test set. Driven by the demand from real-world applications and observations, the M-SENA platform provides a generalization ability test dataset (consists of 68 Chinese and 61 English samples), simulating as many complicated and diverse real-world scenarios as possible. The statistics of the proposed dataset is shown in Table 2. In general, the dataset contains three scenarios and five instance types. Specifically, the three scenarios refers to films, variety shows, and user-uploaded vlogs, while the five instance types refer to easy samples, common samples, difficult samples, samples with modality noise, samples with modality missing. In addition, the dataset is balanced in terms of gender and scenario to avoid irrelevant factors. Examples of the generalization ability test dataset are shown in Appendix C.

3 Experiments on M-SENA

In this section, we report experiments conducted on the M-SENA platform. Comparison of different modality features are shown in Section 3.1, and comparison of different fusion models are shown in Section 3.2. All reported results are the mean performances of five different seeds.

3.1 Feature Selection Comparison

In the following experiments, we take BERT [T1], eGeMAPS (LLDs) [A1], and Action Unit [V1] as default modality features, and compare them with the other six feature sets. Specifically, we utilize GloVe6B [T2], RoBerta [T3] for text modality comparison; customized acoustic feature[A2](including 20 dimensional MFCC, 12 dimensional CQT, and 1 dimensional f0), wav2vec2.0 features [A3] for acoustic modality comparison; facial landmarks [V2], facial land-

Model	MOSI				MOSEI				SIMS			
	Acc-2	F1	MAE	Corr	Acc-2	F1	MAE	Corr	Acc-2	F1	MAE	Corr
LF_DNN	79.39	79.45	0.945	0.675	82.78	82.38	0.558	0.731	76.68	76.48	0.446	0.567
EF_LSTM	77.35	77.43	0.995	0.644	81.23	81.02	0.588	0.695	69.37	56.82	0.591	0.380
TFN	78.02	78.09	0.971	0.652	82.23	81.47	0.573	0.718	77.07	76.94	0.437	0.582
LMF	78.60	78.61	0.934	0.663	83.83	83.68	0.562	0.735	77.42	77.35	0.438	0.578
MFN	78.78	78.71	0.938	0.665	83.30	83.23	0.570	0.720	78.55	78.23	0.442	0.575
GMFN	76.98	77.06	0.986	0.642	83.48	83.23	0.575	0.713	78.77	78.21	0.445	0.578
MFM	78.63	78.63	0.958	0.649	83.49	83.29	0.581	0.721	75.06	75.58	0.477	0.525
MulT	80.21	80.22	0.912	0.695	84.63	84.52	0.559	0.733	78.56	79.66	0.453	0.564
MISA	82.96	82.98	0.761	0.772	84.79	84.73	0.548	0.759	76.54	76.59	0.447	0.563
BERT_MAG	83.41	83.47	0.761	0.776	84.87	84.85	0.539	0.764	74.44	71.75	0.492	0.399
MLF_DNN	-	-	-	-	-	-	-	-	80.44	80.28	0.396	0.665
MTFN	-	-	-	-	-	-	-	-	81.09	81.01	0.395	0.666
MLMF	-	-	-	-	-	-	-	-	79.34	79.07	0.409	0.639
Self_MM	84.30	84.31	0.720	0.793	84.06	84.12	0.531	0.766	80.04	80.44	0.425	0.595

Table 4: Experiment results for MSA benchmark comparison. All models utilize the Bert embedding and the provided acoustic and visual features in CMU-MultimodalSDK. Due to the requirement of unimodal labels, multi-task models, including MLF_DNN, MTFN, and MLMF, are tested on SIMS only.

marks and action units [V3] for visual modality comparison. Besides, we also report the model performances using the modality features provided in CMU-MultimodalSDK.

Table 3 shows the experiment results for feature selection. For Bert-MAG which is designed upon the Bert backbone, experiments are conducted only for Bert as text feature. It can be observed that, in most cases, using appropriate features instead of original features in CMU-MultimodalSDK helps to improve model performance. For textual modality, Roberta feature performs best for TFN and GMFN model, while Bert feature performs best for MISA model. For acoustic modality, wav2vec2.0 embeddings (without finetune) perform best for GMFN and Bert-MAG model. According to literature (Chen and Rudnicky, 2021; Pepino et al., 2021), finetuning wav2vec2.0 can further improve model performance which might provide more effective acoustic features for future MSA research. For Visual modality, the combination of facial landmarks and action units achieves the overall best result, revealing the effectiveness of both landmarks and action units for sentiment classification.

3.2 MSA Benchmark Comparison

Experiment results of benchmark MSA models are shown in Table 4. All models are improved using Bert as text embeddings while using original acoustic and visual features provided in CMU-MultimodalSDK. Besides recording reliable benchmark results, the M-SENA platform also provides researchers with a convenient approach to reproduce the benchmarks. Again, both GUI and Python API are available. We show an example of the proposed Python API in Listing 2. Detailed and Advanced usage is included in our documentation at Github[5]. We will continuously catch up on new MSA approaches and update their performances.

```python
from MMSA import MMSA_run

# Load Default Training Config.
config = get_config_regression(
    model_name='tfn',
    dataset_name='mosi'
)

# Using User Designed Hyper-parameter.
config['post_fusion_dim'] = 32

# Modality Feature Selection.
config['featurePath'] = 'feature.pkl'

# Start Model Training.
MMSA_run(
    model_name='tfn',
    dataset_name='mosi',
    config=config,
    seeds=[1111]
)
```

Listing 2: An example to train model with M-SENA.

4 Model Analysis Demonstration

This section demonstrates model analysis results using the M-SENA platform. Intermediate result analysis is presented in Section 4.1, on-the-fly instance analysis is shown in Section 4.2, and generalization ability analysis is illustrated in Section 4.3.

[5] https://github.com/thuiar/MMSA/wiki

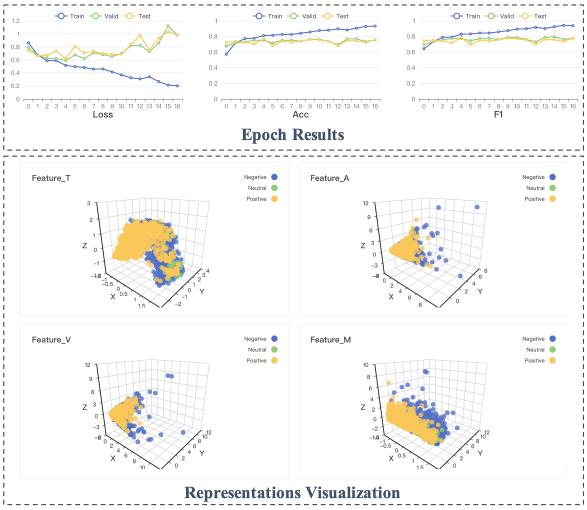

Figure 2: Intermediate Result Analysis for TFN model trained on MOSI dataset.

4.1 Intermediate Result Analysis

The intermediate result analysis submodule is designed to monitor and visualize the training process. Figure 2 shows an example of training TFN model on MOSI dataset. Epoch results of binary accuracy, f1-score and loss value are plotted. Moreover, the learned multimodal fusion representations are illustrated in an interactive 3D figure with the aim of helping users gain a better intuition about the multimodal feature representations and the fusion process. Unimodal representations of text, acoustic, and visual are also shown for models containing explicit unimodal representations.

4.2 On-the-fly Instance Analysis

M-SENA enables researchers to validate the proposed MSA approaches using uploaded or live-recorded instances. Figure 3 presents an example of the live demonstration. Besides model prediction results, the platform also provides feature visualization, including short-time Fourier transform (STFT) for acoustic modality and facial landmarks, eye gaze, head poses for visual modality. We will continuously update the demonstration to make it a even more intuitive and playable MSA model evaluation tool.

4.3 Generalization Ability Analysis

We utilized the model trained on MOSI dataset with [T1]-[A1]-[V3] modality features in Section 3.1 for generalization ability test. Experimental results are reported in Table 5. It can be concluded that all models present a performance gap between original test set and real-world scenarios, especially for the instances with noisy or missing modalities. Another observation is that the noisy instances are usually more challenging than modality missing for MSA models, revealing that noisy modality feature is worse than none at all. In the future, for the demand of real-world applications, MSA researchers may consider analyzing model robustness as well as performances on the test set, and design a more robust MSA model against random modality noise.

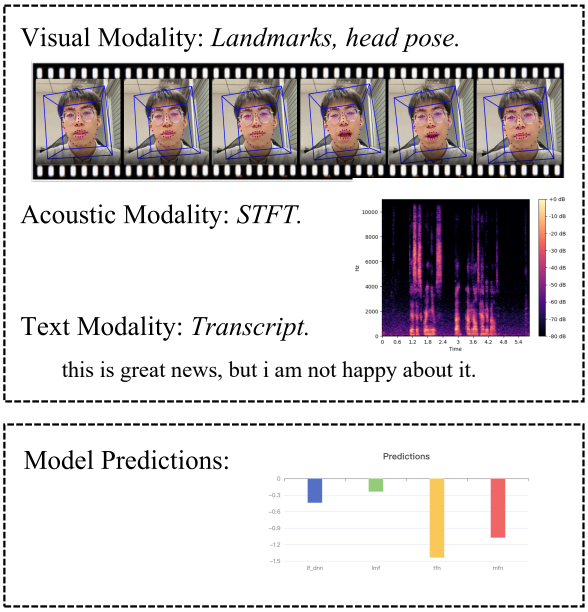
Visual Modality: *Landmarks, head pose.*
Acoustic Modality: *STFT.*
Hz
Time
Text Modality: *Transcript.*
this is great news, but i am not happy about it.
Model Predictions:
Predictions
lf_dnn
lmf
tfn
mfn

Figure 3: On-the-fly instance test example. The M-SENA platform also provides real-time modality feature visualization along with the model prediction results.

Types	TFN Acc-2 / F1	GMFN Acc-2 / F1	MISA Acc-2 / F1	Bert-MAG Acc-2 / F1
Easy	83.3 / 84.4	75.0 / 76.1	75.0 / 76.7	66.7 / 66.7
Common	71.4 / 74.5	85.7 / 82.3	71.4 / 75.8	78.6 / 78.6
Difficult	69.2 / 69.2	61.5 / 60.5	53.9 / 54.4	84.6 / 84.6
Noise	60.0 / 50.5	50.0 / 44.9	50.0 / 35.7	60.0 / 51.7
Missing	63.6 / 60.6	81.8 / 77.8	63.6 / 60.6	63.6 / 61.5
Avg	70.0 / 68.4	71.7 / 69.3	63.3 / 62.4	71.7 / 69.7

Table 5: Results for English generalization ability test. Binary accuracy and F1 scores are reported to show the effectiveness and robustness of the model.

5 Related Works

To the best of our knowledge, there are two widely used open-source repositories from CMU team[6] and SUTD team[7]. Both of them provide tools to load well-known MSA datasets and implement sev-

[6] https://github.com/A2Zadeh/CMU-MultimodalSDK

[7] https://github.com/declare-lab/multimodal-deep-learning

eral benchmarks methods. So far, their works have attracted considerable attention and facilitated the birth of new MSA models such as MulT (Tsai et al., 2019) and MMIM (Han et al., 2021b).

In this paper, we propose M-SENA, compared to previous works, the M-SENA platform is novel from the following aspects. For data management, previous work directly loads the extracted features, while the M-SENA platform focuses on intuitive raw video demonstration, and provides user with a convenient means for private dataset construction. For modality features, M-SENA platform first provides user-customized feature extraction toolkit and a transparent feature extraction process. Following the tutorial, Users can easily reproduce the feature extraction steps and develop their research on designed feature set. For model training, the M-SENA platform first utilizes a unified MSA framework and provide an easy-to-reproduce model training API integrating fourteen MSA benchmarks on three popular MSA dataset. For model evaluation, the M-SENA is the first MSA platform consisting of comprehensive evaluation means stressing model robustness for real-world scenarios, which aims to bridge the gap between MSA research and applications.

Conclusion

In this work, we introduce M-SENA, an integrated platform that contains step-by-step recipes for data management, feature extraction, model training, and model analysis for MSA researchers. The platform evaluates MSA model in an end-to-end manner and reports reliable benchmark results for future research. Moreover, we further investigate comprehensive model evaluation and analysis methods and provide a series of user-friendly visualization and demonstration tools including intermediate representation visualization, on-the-fly instance test, and generalization ability test. In the future, we will continuously catch up on advanced MSA research progress and update new benchmarks on the M-SENA platform.

Acknowledgement

This paper is funded by The National Natural Science Foundation of China (Grant No. 62173195) and Beijing Academy of Artificial Intelligence (BAAI). The authors thank the anonymous reviewers for their valuable suggestions.

References

Mehmet Berkehan Akçay and Kaya Oğuz. 2020. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication*, 116:56–76.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv preprint arXiv:2006.11477*.

Tadas Baltrušaitis, Marwa Mahmoud, and Peter Robinson. 2015. Cross-dataset learning and person-specific normalisation for automatic action unit detection. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 6, pages 1–6. IEEE.

Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 59–66. IEEE.

Erik Cambria, Devamanyu Hazarika, Soujanya Poria, Amir Hussain, and RBV Subramanyam. 2017. Benchmarking multimodal sentiment analysis. In *International Conference on Computational Linguistics and Intelligent Text Processing*, pages 166–179. Springer.

Li-Wei Chen and Alexander Rudnicky. 2021. Exploring wav2vec 2.0 fine-tuning for improved speech emotion recognition. *arXiv preprint arXiv:2110.06309*.

Wenliang Dai, Samuel Cahyawijaya, Zihan Liu, and Pascale Fung. 2021. Multimodal end-to-end sparse model for emotion recognition. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5305–5316.

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*.

Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray. 2011. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern recognition*, 44(3):572–587.

Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. 2015. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE transactions on affective computing*, 7(2):190–202.

Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462.

Dimitris Gkoumas, Qiuchi Li, Christina Lioma, Yijun Yu, and Dawei Song. 2021. What makes the difference? an empirical comparison of fusion strategies for multimodal language analysis. *Information Fusion*, 66:184–197.

Wenzhong Guo, Jianwen Wang, and Shiping Wang. 2019. Deep multimodal representation learning: A survey. *IEEE Access*, 7:63373–63394.

Wei Han, Hui Chen, Alexander Gelbukh, Amir Zadeh, Louis-philippe Morency, and Soujanya Poria. 2021a. Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, pages 6–15.

Wei Han, Hui Chen, and Soujanya Poria. 2021b. 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.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1122–1131.

Zheng Lian, Bin Liu, and Jianhua Tao. 2022. Smin: Semi-supervised multi-modal interaction network for conversational emotion recognition. *IEEE Transactions on Affective Computing*.

Paul Pu Liang, Zhun Liu, Yao-Hung Hubert Tsai, Qibin Zhao, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2019. Learning representations from imperfect time series data via tensor rank regularization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1569–1576.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. 2018. Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2247–2256.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. 2019. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*.

Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, volume 8, pages 18–25.

Rada Mihalcea. 2012. Multimodal sentiment analysis. In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, pages 1–1.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Leonardo Pepino, Pablo Riera, and Luciana Ferrer. 2021. Emotion recognition from speech using wav2vec 2.0 embeddings. *arXiv preprint arXiv:2104.03502*.

Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, pages 1–26.

Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2359–2369.

Björn Schuller, Stefan Steidl, Anton Batliner, Julia Hirschberg, Judee K Burgoon, Alice Baird, Aaron Elkins, Yue Zhang, Eduardo Coutinho, Keelan Evanini, et al. 2016. The interspeech 2016 computational paralinguistics challenge: Deception, sincerity & native language. In *17TH Annual Conference of the International Speech Communication Association (Interspeech 2016), Vols 1-5*, pages 2001–2005.

Mohammad Soleymani, David Garcia, Brendan Jou, Björn Schuller, Shih-Fu Chang, and Maja Pantic. 2017. A survey of multimodal sentiment analysis. *Image and Vision Computing*, 65:3–14.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2019, page 6558. NIH Public Access.

Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2018. Learning factorized multimodal representations. *arXiv preprint arXiv:1806.06176*.

Erroll Wood, Tadas Baltrusaitis, Xucong Zhang, Yusuke Sugano, Peter Robinson, and Andreas Bulling. 2015. Rendering of eyes for eye-shape registration and gaze estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3756–3764.

Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. 2020. CH-SIMS: A Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3718–3727, Online. Association for Computational Linguistics.

Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. 2021. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10790–10797.

Ziqi Yuan, Wei Li, Hua Xu, and Wenmeng Yu. 2021. Transformer-based feature reconstruction network for robust multimodal sentiment analysis. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4400–4407.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017a. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114.

Amir Zadeh, Yao Chong Lim, Tadas Baltrusaitis, and Louis-Philippe Morency. 2017b. Convolutional experts constrained local model for 3d facial landmark detection. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 2519–2528.

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018a. Memory fusion network for multi-view sequential learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6):82–88.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018b. 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.

Jinming Zhao, Ruichen Li, and Qin Jin. 2021. Missing modality imagination network for emotion recognition with uncertain missing modalities. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2608–2618.

A Integrated Datasets

CMU-MOSI. The MOSI (Zadeh et al., 2016) dataset is a widely-used dataset that consists of a collection of 2,199 video segments from 93 YouTube movie review videos.

CMU-MOSEI. The MOSEI (Zadeh et al., 2018b) dataset expands the MOSI dataset by enlarging the number of utterances and enriching the variety of samples, speakers, and topics. For both MOSI and MOSEI datasets, instances are annotated with a sentiment intensity score ranging from -3 to 3 (strongly negative to strongly positive).

CH-SIMS. The SIMS dataset (Yu et al., 2020) is a Chinese unimodal and multimodal sentiment analysis dataset. It contains 2,281 refined video segments in the wild with both multimodal and independent unimodal annotations of a sentiment intensity score ranging from -1 to 1 (negative to positive, the score interval is 0.2).

B Integrated Benchmarks

LF-DNN. The Late Fusion Deep Neural Network (Cambria et al., 2017) first extracts modality features separately and performs late fusion strategy for final predictions.

EF-LSTM. The Early Fusion Long-Short Term Memory (Cambria et al., 2017) is based on input-level feature fusion and conducts Long-Short Term Memory (LSTM) to learn multimodal representations.

TFN. The Tensor Fusion Network (TFN) (Zadeh et al., 2017a) calculates a multi-dimensional tensor (based on outer product) to capture uni-, bi-, and tri-modal interactions.

LMF. The Low-rank Multimodal Fusion (LMF) (Liu et al., 2018) is an improvement over TFN, where the low-rank multimodal tensors fusion technique is performed to improve efficiency.

MFN. The Memory Fusion Network (MFN) (Zadeh et al., 2018a) accounts for continuously modeling the view specific and cross-view interactions and summarizing them through time with a Multi-view Gated Memory.

Graph-MFN. The Graph Memory Fusion Network (Zadeh et al., 2018b) is an improvement of MFN, which can change the fusion structure dynamically to obtain the interaction between the modalities and improve the interpretability.

MulT. The Multimodal Transformer (MulT) (Tsai et al., 2019) extends multimodal transformer architecture with directional pairwise cross-modal attention which translates one modality to another using directional pairwise cross-attention.

BERT-MAG. The Multimodal Adaptation Gate for Bert (MAG-BERT) (Rahman et al., 2020) is an improvement over RAVEN on aligned data with applying multimodal adaptation gate at different layers of the BERT backbone.

MISA. The Modality-Invariant and -Specific Representations (Hazarika et al., 2020) is made up of a combination of losses including similarity loss, orthogonal loss, reconstruction loss and prediction loss to learn modality-invariant and modality-specific representation.

MFM. The Multimodal Factorization Model (Tsai et al., 2018) is a robust model, which can learn multimodal-discriminative and modality-specific generative factors, then reconstructs missing reconstruct missing modalities by adjusting for independent factors.

MLF_DNN. The Multi-Task Late Fusion Deep Neural Network (Yu et al., 2020) first extracts modality features separately and performs late fusion strategy for final predictions through unimodal labels training.

MTFN. The Multi-Task Tensor Fusion Network (Yu et al., 2020) calculates a multi-dimensional tensor (based on outer product) to capture uni-, bi-, and tri-modal interactions through unimodal labels training.

MLMF. The Multi-Task Low-rank Multimodal Fusion (Yu et al., 2020) is an improvement over MTFN, where low-rank multimodal tensors fusion technique is performed to improve efficiency through unimodal labels training.

Self_MM. The Self-Supervised Multi-Task Multimodal (Yu et al., 2021) design a label generation module based on the self-supervised learning strategy to acquire independent unimodal supervisions, which can balance the learning progress among different sub-tasks.

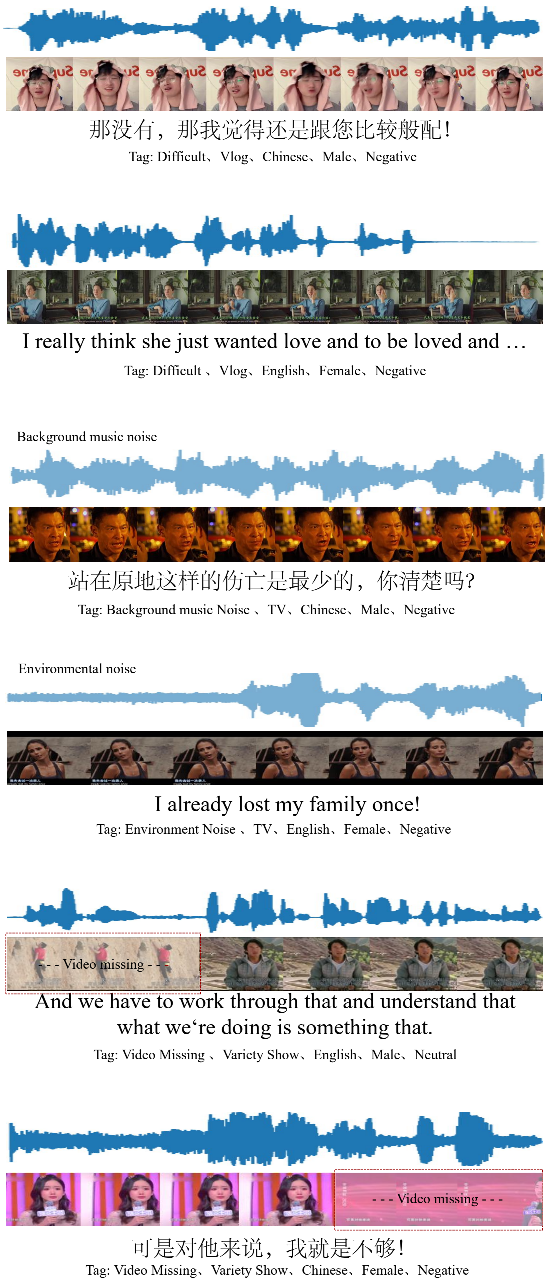
那没有，那我觉得还是跟您比较般配!
Tag: Difficult、Vlog、Chinese、Male、Negative
I really think she just wanted love and to be loved and …
Tag: Difficult 、Vlog、English、Female、Negative
Background music noise
站在原地这样的伤亡是最少的，你清楚吗?
Tag: Background music Noise 、TV、Chinese、Male、Negative
Environmental noise
I already lost my family once!
Tag: Environment Noise 、TV、English、Female、Negative
- - - Video missing - - -
And we have to work through that and understand that what we‘re doing is something that.
Tag: Video Missing 、Variety Show、English、Male、Neutral
- - - Video missing - - -
可是对他来说，我就是不够!
Tag: Video Missing、Variety Show、Chinese、Female、Negative

Figure 4: Examples of the constructed generalization ability test dataset.

C Generalization Ability Test Datasets

The examples of the proposed generalization ability test dataset are shown in Figure 4.